Library Top 500

Anna Preus and Aashna Sheth

2024-04-10

Table of contents

Data Essay

Introduction

This dataset contains information on the top 500 novels most widely held in libraries, according to OCLC, a global library organization with over 16,000 member libraries in over 100 countries. The dataset includes information on authors' biographies, library holdings, and online engagement for each novel, as well as the full text for all works that are not currently under copyright (190 novels).

```
//| echo: false

/*Inputs.table(search, data)*/

Inputs.table(search, {
  layout: "fixed",
  rows: 50,
  sort: "top_500_rank",
  reverse: false,
  format: {
    /*RecreationVisits: x => d3.format('.2s')(x),*/
    pub_year: x => d3.timeFormat(x),
    author_birth: x => d3.timeFormat(x),
    author_death: x => d3.timeFormat(x),
    gr_num_ratings_rank: x => html`<div style='background:${color(x)}'>${d3.format('.2s')(x)
  }
})
```

Download Data

This dataset is based on a list of the Top 500 Novels compiled by OCLC from information in their online database WorldCat, the largest database of library records. The first section of the list was published online with great fanfare as the Library Top 100 in 2019, accompanied by the claim that for novels, "literary greatness can be measured by how many libraries have a copy on their shelves."

We wondered about the implications of this claim and more broadly about what it means to base ideas of "literary greatness" on the number of libraries that hold a particular work. How do historical biases in systems of literary production, preservation, and circulation figure into these kinds of claims? And how do we even define what counts as a novel?

To contextualize the initial list and dig into its claims about literary greatness, we collected information on each novel from a number of other databases, including Wikipedia, Goodreads, Project Gutenberg, the Virtual International Authority File (VIAF), and Classify (a now-shuttered OCLC tool), which we have compiled here.

The dataset was created by Anna Preus and Aashna Sheth, who are also the authors of this data essay.

HISTORY

To start, what is a novel? Today, novels are so prevalent that the term is often applied to a much wider range of books than it actually describes. "Novel" is an umbrella term that applies to works of longform fiction in a range of genres: romance, sci-fi, historical fiction, horror, detective fiction, westerns, etc. The term generally does not apply– although these distinctions can sometimes be fuzzy–to short fiction, poetry, plays, biographies, or other non-fiction forms. The term novel was first used to describe a "long fictional prose narrative" in the 1600s, and the form increased in popularity across the eighteenth and nineteenth centuries. Interestingly, OCLC's list of top 500 novels extends much further back than this. The oldest work on the list is *The Tale of Genji*, a classic work of Japanese literature written over 1,000 years ago.

A key issue in literary studies is which books from the past we continue to read in the present, and which books from the present we will continue to read in the future. The vast majority of novels fall out of circulation shortly after they are published, quickly becoming part of what Margaret Cohen has called "the great unread." When teachers assign texts in their classes and when literary scholars conduct research, they're making choices about which texts continue to be valuable and important for people to read and study.

The term "canon" in literary studies refers to texts that continue to be considered important over time. Ankhi Mukherjee defines the canon as "a set of texts whose value and readability have borne the test of time" and notes that it " involves not merely a work's admission into an elite club, but its induction into ongoing critical dialogue and contestations of literary value." Canonical works continue to be read, taught, and discussed, and in popular terminology they're often considered "classics." These are works you might read in a high school English class: F. Scott Fitzgerald's _The Great Gatsby, _for example, or Harper Lee's *To Kill a Mockingbird.*

One of the things that defines a classic is the fact that it stays in print for a long period of time. When a print book is published, it is issued in an edition with a specific number of physical copies. If the book is profitable, it may be re-issued in different editions over many years. If it becomes a classic, it is likely to be issued in dozens or hundreds of editions even long after the author's death.

In addition to classics, libraries are invested in making popular new novels available to readers. The Top 500 list includes many recently published best-sellers, including books in the *Harry Potter*, *Twilight*, and *Hunger Games* series. Some of these popular books go on to become classics–Charles Dickens's *Great Expectations*, which was wildly successful in its day, is a good example. Many popular texts, though, do not become canonical.

By focusing on the books that librarians around the world have chosen to continue to make available to readers, OCLC was able to create a list of widely read novels that includes both classic works and more recent, popular works that may not have received the same levels of literary acclaim. This represents a unique opportunity to look at….

We wondered, though, how did OCLC's data compare to other potential indicators of popularity or canonicity? And, for that matter, how was the list actually constructed?

WHERE DID THE DATA COME FROM? WHO COLLECTED IT?

The initial list of Top 500 novels was collected by a team at OCLC, the non-profit organization that manages WorldCat. It was compiled based on analysis of data in WorldCat, which consists of catalog records created by librarians around the world, and entered into WorldCat.

Building on this list, we compiled data from a number of other databases, including Project Gutenberg, VIAF, and Wikipedia–a process that is described in greater detail below.

WHY WAS THE DATA COLLECTED? HOW IS THE DATA USED?

OCLC's goal in producing the Top 500 list seems to be to encourage library patronage and reading. The website for the list includes a "Librarians Kit" with a variety of publicity materials–from printable bookmarks to Instagram tiles-that can help bring attention to books in the Top 500 list within libraries' collections.

`Need to link to local image file or to image URL ![alt_text](images/image1.png "image_tooltip")`

Our goal was to collect additional data to understand better how the list was constructed, and to contextualize and nuance its claims about literary greatness.

WHAT'S IN THE DATA? WHAT MAKES A BOOK A "TOP NOVEL"?

The Top 500 list represents a massive data extraction and analysis effort on the part of OCLC. While they do not provide detailed information on how the list was compiled, they do offer a brief explanation of the process that went into creating the list on their FAQ page (written in the context of the top 100, but also applies to the top 500):

> Materials in libraries are described and tracked in WorldCat in two ways. Any specific work of literature, music, art, history, etc., has an associated **catalog record**. This describes the item in a general sense. Every copy of the same book, for example, shares the same record. WorldCat also tracks library **holdings**, which indicate that a specific library has (or holds) at least one copy of that item.
>
> The Library 100 is based on the total number of holdings for a specific novel across all libraries that have registered that information in WorldCat. When a library tells OCLC, “We have a copy of that book available,” that counts as a holding, and in the case of The Library 100, counts as +1 toward its ranking on the list.

This process initially sounds straightforward: to create the Top 500 list, the OCLC team presumably searched the title of a work, counted the number of libraries that held each title, and published the first 500. But it turns out it’s actually much more complicated than that. In WorldCat, records are stored by edition, meaning that each edition has its own catalog record, and an individual title like, say, Miguel de Cervantes’s *Don Quixote*, may have been released in hundreds or thousands of editions since its initial publication. In the U.S., the most common format for digital library records is MARC, which stands for MAchine Readable Cataloging. Librarians with specialized training create detailed MARC records for each edition of a book, and librarians at public libraries, academic libraries, cultural heritage institutions, and private collections tag individual copies of the book that their library holds to that record.

This means that when developing the list, the OCLC team actually had to find all the editions of a specific title and sum the number of libraries that hold that edition across all editions. **Thus the top 500 list is not only a representation of how many libraries carry the work, but a representation of how many times a book has been re-edited and re-issued; the more editions a book has, the more records are created.** Often, there are duplicate records for individual editions, which may affect the overall count of copies tallied by the OCLC. And when a work is translated into different languages, all the editions of all the translations are also recorded in WorldCat, which also figures into the count of total holdings for each novel.

HOW WAS THE DATA COLLECTED? WHAT ADDITIONAL DATA WAS ADDED?

We wanted to contextualize the Library Top 500 list by compiling additional information on each novel from a range of other sources. We focused on gathering three main categories of information: information that could help us understand what types of works–and whose works–were included on the list, data that could potentially provide alternate measures of popularity or canonicity, and the full text of each novel that was in the public domain. We collected information from the following sources:

WorldCat: we used the now-shuttered OCLC tool Classify to gather data from WorldCat on the top 100 most widely held editions of each of the 500 novels on the list. We recorded total library holdings for these top 100 editions. We consider number of editions as a potential alternate measure of canonicity, although it is necessarily affected by the amount of time that has passed since the initial publication of the novel.

VIAF: The Virtual International Authority File is an OCLC-run database that contains structured records–called "name authority files"–for individual authors and creators. We used VIAF to gather information on authors whose novels were included on the list, including their birth and death dates, nationalities, genders, and occupations.

```
Need to link to local image file or to image URL ![alt_text](images/image2.png
"image_tooltip")
```

Wikipedia: we used Wikipedia, the popular, free, volunteer-authored encyclopedia, to identify the year of first publication for each novel on the list.

Goodreads: Goodreads, which is owned by Amazon, is the largest social networking site related to books, with over 150 million members. Itt allows users to rate, review, and discuss a huge range of texts. We drew on data from Goodreads as a potential alternate indicator of texts' popularity, collecting total number of reviews, total number of ratings, and average overall rating for each novel on the list.

Project Gutenberg: We used Project Gutenberg to access the full-text of all novels on the list that are currently in the public domain, or in other words, out of copyright. We chose Project Gutenberg because their eBooks are edited by volunteers, whereas many larger content repositories, like Internet Archive and HathiTrust, only make available machine-generated transcriptions of historical texts, which tend to be less accurate.

Our work creating this dataset not only builds on the work of the OCLC team who compiled the Top 500 list, but on the labor of the thousands of librarians who created records held in WorldCat and VIAF, of the volunteers who transcribed texts for Project Gutenberg and wrote articles for Wikipedia, and of the social media users who reviewed and rated books on Goodreads.

ACKNOWLEDGING BIAS [METADATA ANALYSIS?]

The Top 500 List provides an unprecedented opportunity to consider what works libraries around the world have on their shelves. This, in turn, serves as an important indicator of books' continued relevance to readers. As OCLC points out, "libraries offer access to trendy and popular books. But, they don't keep them on the shelf if they're not repeatedly requested by their communities over the years." Looking at what novels are held by the most libraries provides a sense of which works librarians consider important to continue to make available to readers.

At the same time, there are distinct biases in what gets kept on library shelves, and in which library shelves are being considered. The libraries that OCLC works with are disproportionately located in Europe and North America and OCLC uses cataloging systems developed in English-language contexts. The list is distinctly skewed toward works by White, European and American men, as is English literary history, but it would be impossible[difficult?] to tease apart this historical bias from potential compounding biases in WorldCat's underlying data, the data collection process, or library cataloging systems more broadly.

OCLC acknowledges broad biases in the list in a section titled, "Why isn't the list more diverse?"writing, "The list emphasizes many books that we tend to think of as 'classics,' because those are the novels most often translated, retold in different editions, taught and widely distributed in library collections. Because of this, the list tends to reflect more dominant cultural views." Although OCLC acknowledges this general bias towards works considered classics, more specific forms of bias aren't made apparent in the list itself, which only includes the title and author for each novel.

```
Need to link to local image file or to image URL ![alt_text](images/image3.png
"image_tooltip")
```

Each of these works and authors, of course, deserves careful consideration in their own right, but by compiling additional information into dataset, we can begin to see some of these biases as well as trends across the list more clearly.

[Metadata analysis prompts?

Over 85% of the novels on the top 500 list were originally published in English

Over 80% were written by authors from the U.S. or Great Britain

Over 70% were written by men]

We collected this data because we were interested in contextualizing the top 500 list. The top 500 list itself was compiled because… It is used….. We hope that our dataset will be used like…

- Reinforces what books are read (people pay more attention to the books that are labeled as "top 500"), could overshadow other important books
- OCLC put out this list to encourage publicity / reading. Wasn't made for academics to use!

WHAT'S IN THE DATA? WHAT MAKES A BOOK A "TOP NOVEL"? [Metadata Analysis?]

By building on OCLC's initial list and adding a range of information from other databases, we wanted to create a dataset that would offer opportunities to connect metadata analysis and full-text analysis in relation to a historically significant corpus of texts. To this end, we've included a brief exploration of the data below, as well as suggested activities, and some ideas for future areas of inquiry.

For starters, let's look into some author metadata. Who is represented in the most popular works and what may this reveal about literary canonization?

1. Author Gender
2. Which author names are most represented
3. Author Map
4. Number of books in each time period (buckets for 1800-1900, 1900-2000, etc).
5. Rank vs number of editions
6. Author languages/language of books?

Activity 1

```
&lt;**Activity** about looking into what "unnamed gender" may mean and discussion about pow
```

From our initial metadata analysis, we find that canonized works are those that are (1) written by white men, (2) primarily between 1900-2000, (3) primarily of x genre, and (4) have at least x editions.

Let's take a look at how OCLC books stack up against GoodReads rankings.

Activity about whether num_ratings or average_rating would be better to use. What factors affect them?>

We decide to use num_ratings as a fair comparison against OCLC num_libraries. Do the number of copies really match the number of people that check it out and read it, at least according to goodreads?

1. Look at top 5 goodreads num_ratings vs top 5 oclc
2. Stock chart of books that moved up and down

Activity: Which books had the most movement? Which had the least? Are there any trends?>

```
*ebook vs on shelf, translations into english or other languages
```

Now that we've explored questions of representation and popularity, let's dive into some full text analysis.

THIS IS VERY TBD

CONCLUSION

Text text.

- Goodreads data is skewed towards younger users (average goodreads demographic), and people who actually rate (also younger people)
- Relating goodreads num ratings is similar to how many libraries… how many libraries is inadvertently trying to measure, how many people read it ?
  - "How many libraries have the book on their shelves"
  - But there is a gap between publishing proliferation (canonization) and true popularity (people reading it).
  - https://www.tckpublishing.com/the-literary-canon/

QUESTIONS:

- Fighting words
- Genres
- Activities in laid with data essay? What do we wanna have as exercise and what in metadata analysis?

Possible activities:

1. Let's plot gender information, we see this 3rd category, what do you think it represents?
   1. Refer back to data, look up the history of a couple authors tagged with this? Make a point about how labels are important to provide context and can also be diminishing/obscuring/normative
2. Goodreads:
   2. What factors affect popularity on goodreads? Why do you think the top5 books are so different?
   3. Who liked vs number of ratings? What's the difference? What's fair to compare with OCLC?
   4. Which books had the largest change in ranking? What do you think this means?
   5. Incorporating voyant

Cut from history section:

Libraries keep records of the works they hold in their collections, and they create and organize these records in accordance with particular cataloging systems. Historically, these catalogs have been drawers of physical cards containing information about each book.

»»> gd2md-html alert: inline image link here (to images/image4.png). Store image on your image server and adjust path/filename/extension if necessary. (Back to top)(Next alert)»»>

`Need to link to local image file or to image URL ![alt_text](images/image4.png "image_tooltip")`

»»> gd2md-html alert: inline image link here (to images/image5.png). Store image on your image server and adjust path/filename/extension if necessary. (Back to top)(Next alert)»»>

`Need to link to local image file or to image URL ![alt_text](images/image5.png "image_tooltip")`

When data storage on computers became powerful and popular in the mid-to-late 1900s, these cards were re-written to meet the MAchine Readable Cataloging (MARC) format. This format made it easier to digitize and standardize physical cards. Once a book's MARC record was online, libraries could print and distribute the same card.

Later on, as libraries gained personal computers, the physical cards themselves became obsolete and patrons could search an online database for book information. The contents of the LOC's MARC record database are the foundation of modern library browsing systems. The contents of the MARC record help populate a webpage about each book. Nowadays, more information is added to these webpages like reviews, star-ratings, and even some digitized pages.

Although each library or library system usually has its own browsing website, there are key central entities that contain information from many libraries. The current largest global database of library information is called WorldCat, which is maintained by OCLC and allows for searching across 2.7 billion records from libraries in over 100 countries.

Explore the Data

```
//| echo: false
//| output: false
library_data = d3.csv("https://raw.githubusercontent.com/melaniewalsh/responsible-datasets-in

use_data = d3.csv("https://raw.githubusercontent.com/melaniewalsh/responsible-datasets-in-co
```

```
//| echo: false
//| output: false



filtered = library_data.filter(function(penguin) {
  return bill_length_min < penguin.bill_length_mm &&
         islands.includes(penguin.island);
})
```

```
//| echo: false
color = d3
  .scaleLinear()
  .domain([0, 100, 300])
  .range(["#cafcc2", "#fce7c2", "#eb9494"])
```

Library Top 500

```
//| echo: false
viewof search = Inputs.search(library_data, {
  placeholder: "Search"
})
```

```
//| echo: false

/*Inputs.table(search, data)*/

Inputs.table(search, {
  layout: "fixed",
  rows: 50,
  sort: "top_500_rank",
  reverse: false,
  format: {
    /*RecreationVisits: x => d3.format('.2s')(x),*/
    pub_year: x => d3.timeFormat(x),
    author_birth: x => d3.timeFormat(x),
    author_death: x => d3.timeFormat(x),
    gr_num_ratings_rank: x => html`<div style='background:${color(x)}'>${d3.format('.2s')(x)
  }
})
```

Download Data

Exercises

R

Python

Discussion & Activities

Activity 1

It is inevitable that the devices that the National Park Service uses to count visits to the parks — like induction loop counters installed on the road — will break. But they will also get *fixed* at different rates, in different locations, as we could see in the case of Crater Lake National Park (where a counter was fixed quickly) and Carlsbad Caverns National Park (where a broken counter from 2019 still has not been fixed).

There are many reasons for these disparities, but some of the big ones might be geography and resources. The more remote a park, the harder it is to get a repair team to it. The less-resourced a park, the lower the likelihood they have on-site repair teams, or are prioritized by the repair teams that can be dispatched.

With this in mind, look at the locations of the following parks. Suppose that each one has an outage in their induction loop counter: which ones would you expect to be fixed first, and why? Research the parks, and rank them on a scale of 1 to 5 (1 being highest, and 5 being lowest) of which would be fixed quickest.

Park	Priority (1-5)	Reason
Acadia NP		
Lassen Volcanic NP		
Saguaro NP		
Yosemite NP		
Mammoth Cave NP		

Activity 2

The National Park Service sometimes fills in missing data with hard numbers or approximates data by applying special mathematical formulas. This is necessary work, but it is also under-explained work.

To see this in action, go to the NPS page that documents park reports and down the “Visitor Use Counting Procedures” PDF for three different parks.

How are the procedures for these three parks similar or different? What kind of effect do you think this has on the resulting data? What do you think is the best of documenting this information and communicating it to users of the data?

Activity 3

In 2014 and 2015, Kobuk Valley National Park reported that there were zero visitors to the park.

Use publicly available internet data - Twitter posts, Flickr photos, etc - to try and find evidence of people visiting the park (there is existing evidence!).

Based on your findings, how do you think, differently, if at all, about Kobuk Valley’s decision to record zero visits and about alternative methods for counting visits?