Data-Intensive Tools for Modeling and Visualizing Mass Reading

The 'Reading Chicago Reading' Project

Principal Investigators:
John Shanahan (English, DePaul University)
Robin Burke (Information Science, University of Colorado, Boulder)
Ana Lučić (Library, DePaul University)

Research and Technical Support:
Ke Feng (DePaul University)
Mihaela Stoica (DePaul University)
Nandhini Gulasingam (DePaul University)

This project was made possible in part by a 2018 award from the Catalyst Fund at LYRASIS

1. GOALS

Our principal aim was creation of an interactive visualization dashboard for data associated with the "One Book One Chicago" (OBOC) program of the Chicago Public Library (CPL). This was to be a prototyping path for an easy-to-use interactive tool to visualize circulation and other book data, to be used by librarians, academic researchers, city officials, and any interested members of the public. Our interactive dashboard can stand alone, but might ultimately serve as the centerpiece of a more elaborate open-access website about "reading life" in Chicago across multiple devices (print books, ebooks, playaways, etc.) that facilitates browsing by the public so that they might place themselves among the contemporary cultural pulse of their city.

We are producing outcomes that focus on helping public libraries better understand their roles as community anchors as well as the differential impact of targeted promotional events in the delivery of city-scale reading. A dashboard for visualizing data about book checkouts together with associated outreach and text measures illustrates how a book can be a node of multimodal exchanges of value to library administration. Our project, instead of looking only at data that can be gathered through traditional library systems such as circulation statistics, also archives the outreach work of the library and integrates data sources originating outside the library such as demographic data about the neighborhoods surrounding branches. Our five main sources of data are (1) demographic data, such as census data and other aggregated information about city residents and patrons, (2) associated social media data, especially from Twitter and Goodreads, (3) an archive of Chicago Public Library outreach events, (4) anonymized CPL book circulation transactions, and (5) full-text book content available through the HathiTrust Digital Library.

Four main research questions animated our work:

- how can wide-focus data sources be collected, integrated, and analyzed to make reliable quantitative predictions of circulation patterns both system-wide and branch-by-branch
- how can these data sources be collected, integrated, and analyzed to evaluate, and predict, the impact of different library promotional activities
- how can these data sources provide evidence of linkage of promotional choices on other books throughout the library's catalog
- can these kinds of models become the basis for data-driven decision making in library promotions and collection development?

2. PROCESS

(a.) what did you do?

The work from July 2018 to September 2019 comprised the following major activities:

1) In summer and fall 2018, Burke and Feng collected CPL, city of Chicago, and social media data into a database. From January 2019 to September 2019, they created and tested an interactive dashboard made with Bokeh, a python-based visualization tool.

2) From fall 2018 to fall 2019, Lucic and Shanahan worked on feature/text measures extraction and text measures for 76 works. Because the texts were in copyright, the HathiTrust Research Center (HTRC) analytics platform was used to extract several non-consumptive features, including different measures of reading difficulty, named entities, and sentiment data from 7 recent OBOC selections and 69 CPL-recommended works that did not receive additional promotion but were on the OBOC recommended reading lists at CPL branches during the promotional period.

3) In Summer 2019, Stoica in collaboration with Lucic and Gulasingam, developed maps for dashboard interactions of CPL OBOC events and circulation for selected OBOC seasons.

(b.) *what worked and what did not?*
Capturing the reach and differential impact of the promotion of mass reading events has proven to be difficult when reading data is embedded in an array of social media platforms and is not regularly archived. The Chicago Public Library has been an excellent partner to us in sharing the data about One Book One Chicago, and yet the format of the data we have received presented challenges to us. Burke created a project mySQL database and we added to it all CPL circulation data we were given, covering 2011-2017 for the OBOC choices but also a set of just over 300 other books made up of all the titles CPL recommended each OBOC season. Originally, we planned to use the data about all of these books in the dashboard, but that goal proved too ambitious. We were, however, able to create the interactive dashboard for our original OBOC set using the Bokeh interface. See below in "Accomplishments."

(c.) what modifications did you make from the original proposal, or would you recommend others make if they want to adapt your model?
Chicago Public Library and Chicago city data is key to our interactive dashboard and the analysis that we have done throughout the duration of our project. Given that our project deals with the circulation of books throughout a large heterogeneous library system (80 branches in the case of CPL), being able to understand the data and what it conveys, how it was captured, at what time, and by which system is an important element to take into consideration when analyzing circulation of items throughout a system of libraries (e.g. any library share consortium or a system of public libraries).

First, the key for possible future use by others will be their own access to various kinds of data. Our focus on Chicago has been fortuitous as the city is very proactive in making large data sets available via the City of Chicago data portal. We are not certain that other localities and library systems can be as forthcoming.

Second, given technological changes in integrated library systems and given differences in how libraries record and preserve their data the format of circulation data capture is very important and will have an influence on what kind of analysis can or cannot be done with the data in the future.

3. ACCOMPLISHMENTS

We created a prototype dashboard with several major features (more are in development). It was created with Bokeh interactive visualization https://bokeh.pydata.org/en/latest/index.html The Bokeh library is particularly suitable for creating interactive plots, dashboards, and data applications.

Our dashboard will help illuminate the following dimensions of data:

Circulation time series

The first interactive visualization, the time series, shows a general circulation trend for 7 OBOC selections. This type of visualization will be of interest to:

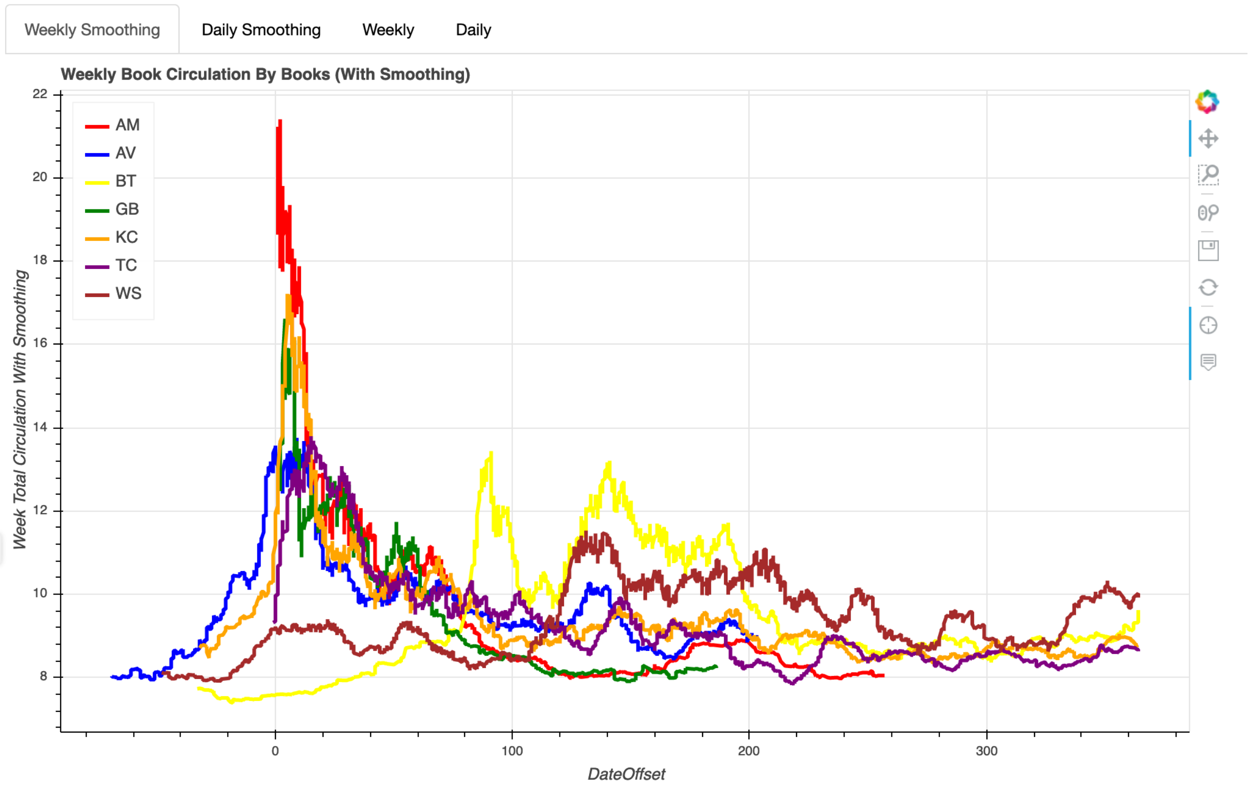
1) CPL staff as they work on preparing programming for future events
2) Librarians from community, public, school, academic libraries that plan promotional event around one book, movie, author
3) General public who are interested in the effect of media advertisements, promotional strategies, and live events throughout the metropolitan city on the circulation of *OBOC* selections, and identifying differences between individual titles across the city.

The screen capture below shows a snapshot of this visualization including 7 of the recent OBOC book titles and their circulation prior to and after their selection by CPL for the OBOC program. Interaction features include:

- The ability to choose which books are shown

- The ability to select the sample frequency (day or week) and whether smoothing is applied

- The ability to zoom in to particular date ranges

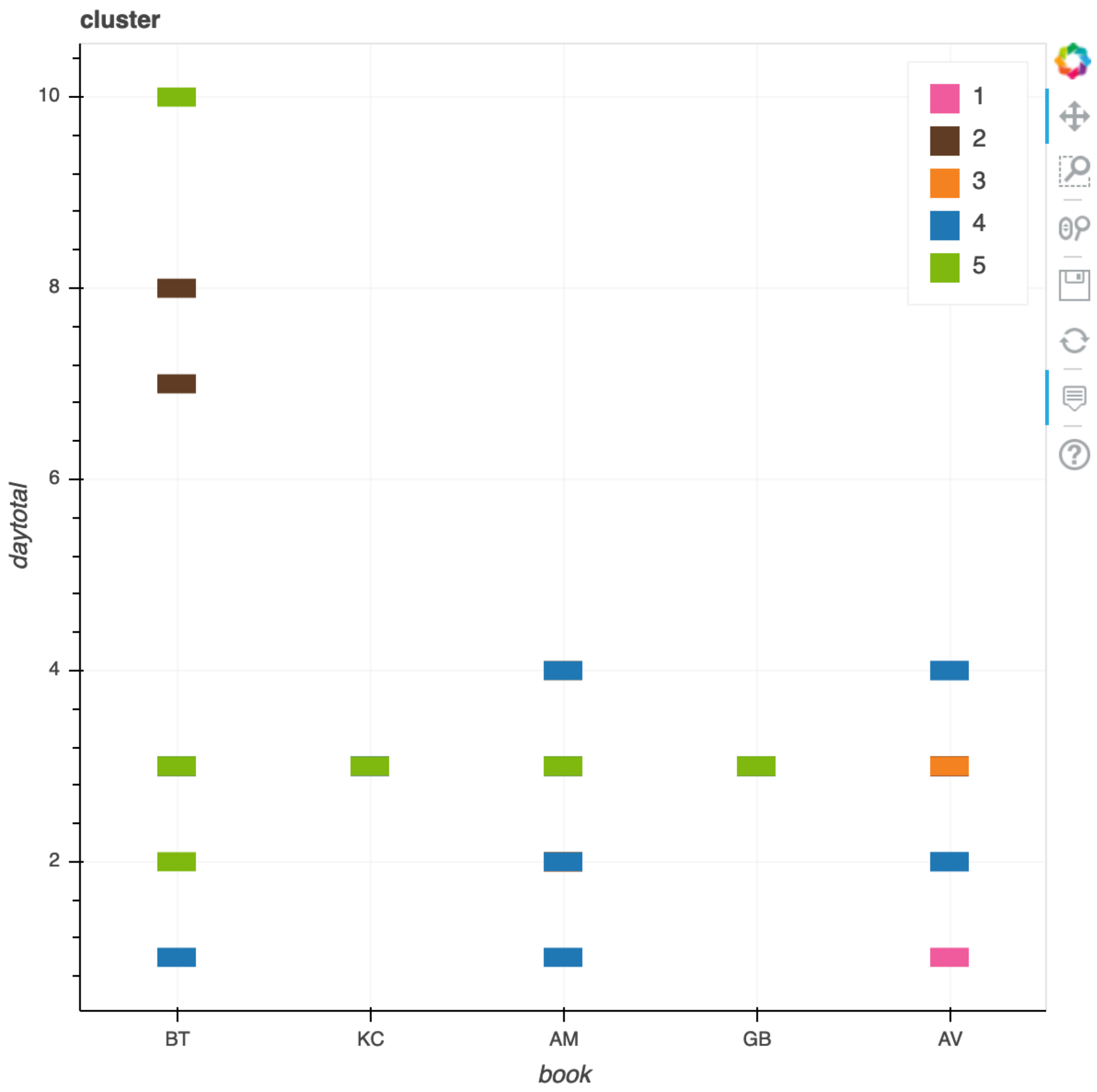


Circulation and demographics

The second interactive feature displays circulation per branch with demographic characteristics of the neighborhood associated with the branches, as represented in population clusters. This type of visualization highlights the popularity of OBOC books with demographic characteristics of the neighborhood, and will be of interest to public librarians as it highlights differences in the reception of the work and can inform planning of future promotional events. In addition to the audiences mentioned in the previous sub-section, results of this analysis will also be of interest to sociologists, urban studies researchers, historians, and archivists who study, for instance, contemporary reading practices.

In the visualization below, the five demographic clusters are represented by colors. Branches belonging to each cluster are represented as rectangular glyphs. The different book events are shown on the x axis and the branches are located based on daily average circulation on the y axis. Interactive aspects of the visualization include:

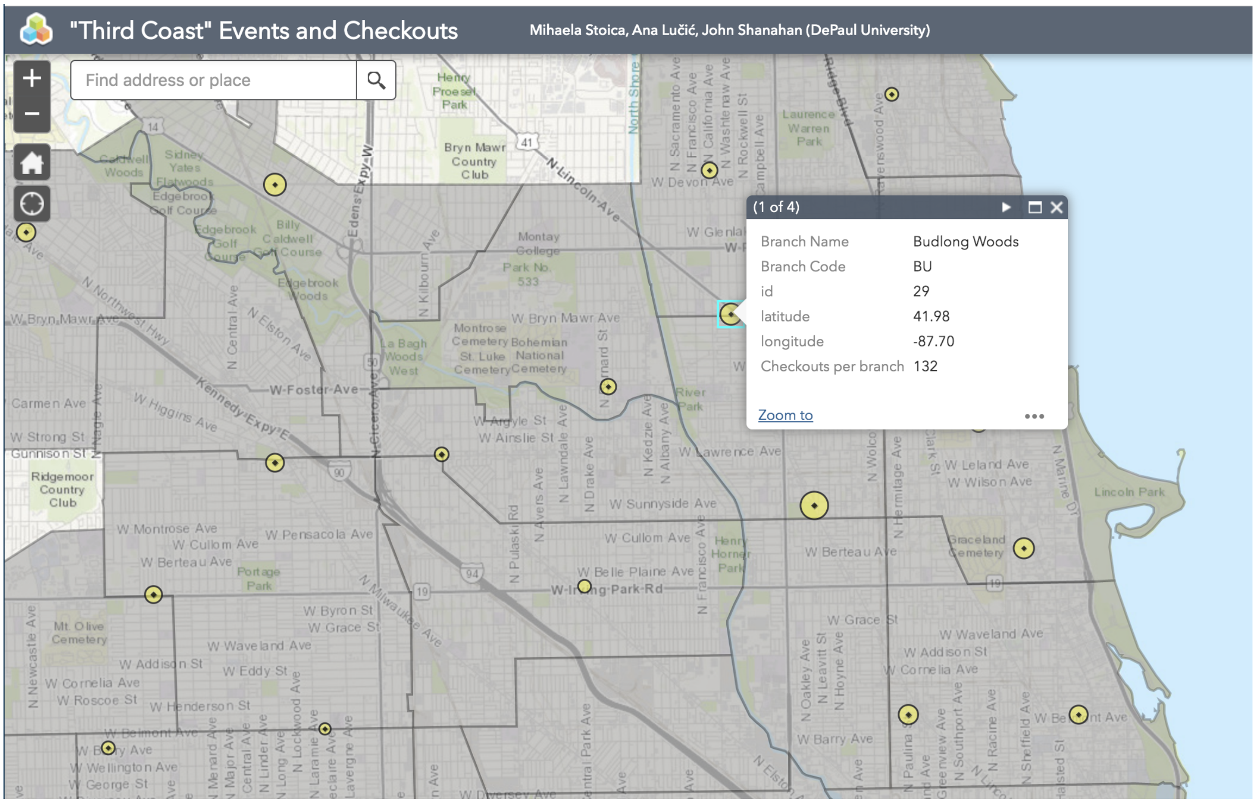
- Choice which books to include in x axis
- Choice of circulation statistic and range for the y axis
- Choice of textual branch code or graphic representation



Circulation and events distribution

The third interactive visualization below showcases a season's checkout totals per library branch and, at the same time, the number of OBOC-associated events that same season organized by CPL that helped promote the book throughout the city. This visualization can reveal the impact of promotional efforts. (Our next step is adding social media totals into the events.)

Using the interactive interface created with ArcGIS software, a librarian or other researcher can examine how a particular (OBOC-) selected work circulated throughout a system of 80 Chicago Public Libraries during the promotional period which ranges from four months (e.g. Bellow's *The Adventures of Augie March*) to nine months (e.g. Dyja's *The Third Coast*).



"Third Coast" Events and Checkouts

Mihaela Stoica, Ana Lučić, John Shanahan (DePaul University)

(1 of 4)

Branch Name	Budlong Woods
Branch Code	BU
id	29
latitude	41.98
longitude	-87.70
Checkouts per branch	132

4. LESSONS LEARNED

Throughout the year, CPL staff and branch librarians schedule hundreds of events across the city and increasingly use social media to draw attention to their activities. The library's "One Book One Chicago" programming is dynamic and multifaceted in this way too. The very qualities of its innovation and public appeal, however -- and what makes it such an exciting object of study for scholars and librarians -- also makes it difficult to capture and analyze. Nonetheless, our work has produced a number of insights and lessons of use to other researchers.

Most important for researchers to understand is how (1) the length of CPL promotional seasons has changed, and (2) the kinds of CPL community outreach have changed.

(1) Beginning in fall 2013, CPL changed OBOC from two book choices a year (i.e. fall and winter/spring) centered on branch-level “book club”-style discussion, to one book at the center of a longer “season” with a broad theme (e.g. Music, or Food) running September to May. This new form of OBOC programming also expanded outward from branches to “contiguous” programs of related interest sponsored by non-CPL organizations. This is important for a number of reasons. Previously, a “One Book” season as a data object was easily demarcated: the chosen book’s checkouts for a few months, system-wide. But each season since 2013 has become more complex as additional forms of social media outreach and contiguous programming have been added. In the past, each CPL OBOC season from the program’s launch in fall 2001 to spring 2012 was short in total span and largely limited to physical events like book discussions at branches on the model of book clubs; for the past five years -- the years of our dashboard -- an OBOC season is a hybrid physical/digital data event over time and across the city.

(2) The change to one book per season, beginning with Isabel Wilkerson’s *The Warmth of Other Suns* in 2013-14, reflects CPL’s new community outreach slogan “the book is just the beginning.” This means that the book selected becomes just one element in a more full-spectrum set of events centered on a larger “theme” with some events happening outside the walls of library branches. This is important for our work, and is reflected in the dashboard in event counts. Because of the changes that included more social media and more multi-modal outreach events (e.g. walking tours, maker lab sessions, Tweets, gardening tutorials, etc. depending on the book) it is even more challenging for librarians to determine what works in improving participation.

As a matter of data analytics, the change in length of season represents some difficulty as some of the earlier seasons in our data covering 2011-2017 are shorter and more concentrated and some are longer and, by virtue of this fact, have more total promotional events. Additionally, as our recent analysis has shown, the promotional strategy of the Chicago Public Library varies as well. Sometimes, an author appearance during the promotional period takes place, sometimes it does not; sometimes the author happens to be a Chicagoan (e.g. Greg Kot, a Chicago *Tribune* critic and author of the 2017-18 selection *I’ll Take You There*) and sometimes not. All of these factors have an influence on marketing decisions that typically show a great degree of variance from season to season. In our project analysis, for each recent season we counted and plotted all official CPL OBOC events, per branch, and mapped them with checkout statistics for the same book during the season. We wanted to know if there is a relationship between the number of OBOC events and checkouts. Our initial findings are that some branches (only) show a bump in checkouts after an event, and in most of those cases increases are not distinctly “punctual” (i.e. more checkouts in the week after a branch event).

The distribution for three OBOC Chicago-themed books demonstrate different pattern of events throughout the city. Promotion of the 2015/16 OBOC selection *The Third Coast* is particularly interesting, as Chicago Public Library made a concerted effort to organize at least one outreach event at every branch in the CPL system. A detailed analysis of this season's programming is the subject of another paper (in progress).

Even the announcement of the OBOC selections shows variation. Typically, the announcement is made by the Chicago Public Library and the Mayor's office by web post, radio, and social media. However, in the case of *The Book Thief* by Markus Zusak (2014-15 OBOC season), the news that this book was to be the next OBOC selection was leaked by a newspaper several months before the official announcement and the book rose in popularity (measured by CPL checkouts) even before the official launch date. This presented challenges to our dashboard visualization: for instance, how to present the official launch date for this book in the time series.

This range of OBOC book genres, marketing efforts, readability measures, and season durations is not easy to capture and represent in a way that does justice to all of the factors. Although the interactive time series dashboard shows some of the commonalities and the common "bump" in circulation for all of the most recent OBOC selections, each season is unique and each line of the time series is different from another. Our future iterations of the dashboard will seek to capture this variance and to find ways to represent the richness and the variability of the factors that characterize each season. The challenge will be to find ways to represent this richness and variability and in that way capture the "season" in as many dimensions as possible.

The data that we received from the Chicago Public Library was reconstructed by us after the fact; the data, in other words, was not produced and stored with our work in mind. Most likely, this type of use of the circulation data was not anticipated. Most documentation of outreach events exists only as paper files (to which we were given access several times) and email by CPL staff (to which we did not have access). We have not been able to find examples of studies that, like ours, demonstrates the merger of circulation data with other types of data, for example, demographic and social media. Also, how books circulate throughout a system of libraries has not been a very well studied phenomena. This study, in a way, pioneers novel ways of studying circulation throughout *a system of libraries* while simultaneously merging circulation data with other data sources (demographic characteristics and the number of events for the book in different locations).

Finally, attention to future archiving of cultural programming is important. The record of several elements of the CPL OBOC programming, especially before 2014, are piecemeal. It is important that library systems preserve records of their programming in the interest of future historians, sociologists, and data scientists. CPL has been a wonderful partner, but prior to the inception of the “Reading Chicago Reading” project little attention was given to systematic archiving. That said, the public outreach programming records since late 2014, and circulation transaction data since 2011, have been much more thorough and accessible.

5. NEXT STEPS

The dashboard is, by its nature, an evolving platform -- additional data sources and interactive elements can and will be added over time. We will sustain our Lyrasis work by incorporating additional “One Book One Chicago” seasons and additional data types (e.g. social media). From maps in progress we can make animated maps that visualize branch-level circulation of books over time, as well as the network of book movement (i.e. place of checkout and return) across branches in the system.

We expect that our dashboard will be useful for predictions of other (i.e. non-”OBOC”) books, however to be able to do this requires information about the desired book, such as a “reading difficulty” score (we have calculated them for OBOC books in our set) and prior circulation in CPL to be at hand. We hope that by determining some generalizable numerical ranges for these measures we will be able to use the model with sufficient accuracy even in the absence of individual book data.

One major way we can expand our accomplishments is to join our findings about some books to additional data sets from the City of Chicago data portal. What are the correlations between, for instance, higher or lower branch checkouts for particular kinds of books and other demographic indicators such as voting participation and use of public transportation?

Future work will aim to highlight the uniqueness of each season by indicating the readability measure for the book, the targeted audience, subject topics for the book, how it circulated throughout the system of Chicago Public libraries, the number of events per branch/location that were organized. Such interactive visualization would allow users to examine differences between the seasons but also to study their unique characteristics.

We are especially interested in moving outward from our limited set of books (i.e. recent OBOC choices and associated “recommended titles”) into wider swathes of CPL circulation records. A new feature of the interactive dashboard is envisioned that can highlight time series circulation for the OBOC pick against recommended books that were not the “One Book” selection. Such visualization can help delineate the impact that special promotion has on the circulation of the book.

We believe that the true payoff of this project will be when it can draw upon more systemic CPL circulation access. With a larger set of circulation data, enhanced analysis is possible and more fine-grained results given: for example, with broad CPL holdings and checkout data, a librarian or other researcher can be truly data-driven in decision-making about book choices and events to sponsor: what books actually push other book checkouts; what events are most conducive to increased readership and library foot traffic, and helpful in augmenting participation in person and via social media?

We have several forms of outreach planned as well: our first audience is the Chicago Public Library and other public library systems. We presented examples of the prototype dashboard visualizations to CPL staff in the fall of 2018 and plan to do so again in fall 2019. It is hoped that we might also schedule an open dashboard "trial" session for users, both public librarians and any interested others, at the DePaul Library if not at CPL's Harold Washington (main) Library branch. We also expect to present work at ALA, LITA, and other academic venues.

We are presently in conversation with CPL about increased data access (especially in light of the 20-year anniversary of the OBOC program in fall 2021). We are in discussion of plans to archive a wide spectrum of data about OBOC for the benefit of future investigators.

6. FURTHER READING AND PROJECT MATERIALS

During the Lyrasis grant period we have produced, in addition to the prototype dashboard, several related technical papers, presentations, data sets, and code. See the project website link: https://dh.depaul.press/reading-chicago/. The site will also host the dashboard itself when the system reaches its first public release, subject to review by CPL staff.