Towards a Chicago place name dataset:

From back-of-the-book index to a labeled dataset

Ana Lucic and John Shanahan

University of Illinois and DePaul University

May 31, 2020

**Introduction**

Reading Chicago Reading[[1]](#footnote-0) is a grant-supported digital humanities project that takes as its object the “One Book One Chicago” (OBOC) program[[2]](#footnote-1) of the Chicago Public Library. Since fall 2001, One Book One Chicago has fostered community through reading and discussion. On its “Big Read” website, the Library of Congress includes information about One Book programs around the United States,[[3]](#footnote-2) and the American Library Association (ALA) also provides materials with which a library can build its own One Book program and, in this way, bring members of their communities together in a conversation.[[4]](#footnote-3) While community reading programs are not a new phenomenon and exist in various formats and sizes, the One Book One Chicago program is notable because of its size (the Chicago Public Library has 81 local branches) as well as its history (the program has been in existence for nearly 20 years). Although relatively common, book clubs and community-based reading programs have not been subjects of sustained study with quantitative methods of data science.

The following research questions have been guiding the Reading Chicago Reading project so far: can we predict the future circulation of a book using a predictive model based on circulation, demographics, and text characteristics? How did different neighborhoods in a diverse but also segregated city respond to particular book choices? Have certain books been more popular than others around the city as measured by branch-level circulation, and can these changes in checkout totals be correlated with CPL outreach work? A related question to this is the focus of this paper: the association of place name with sentiment in Chicago-themed books, namely, what are the trends that emerge from the spatial analysis that adds sentiment to geographic location?

Exploration of these questions, and our attempt to find a solution for some of them, has made us reflect on innovative services that libraries can offer. We will discuss this possibility in the last section of this paper.

## **Chicago as a place name**

Thus far, the Reading Chicago Reading project has focused its analysis on seven recent OBOC book selections and their respective “seasons” of public outreach programming:

* Fall of 2011: Saul Bellow’s *The Adventures of Augie March*
* Spring of 2012: Yiyun Li’s *Gold Boy, Emerald Girl*;
* Fall of 2012:Markus Zusak’s *The Book Thief*
* 2013 – 2014:Isabel Wilkerson’s *The Warmth of Other Suns*
* 2014 – 2015: Michael Chabon’s *The Amazing Adventures of Kavalier and Clay*
* 2015 – 2016: Thomas Dyja’s *The Third Coast*
* 2016 – 2017: Barbara Kingsolver’s *Animal Vegetable Miracle: A Year of Food Life*

All of the listed works above, spanning categories of fiction and non-fiction, are still in copyright. Of the seven works, three were categorized as Chicago-themed works because they take place in the Chicago area in whole or in substantial part: Saul Bellow’s *The Adventures of Augie March*, Isabel Wilkerson’s *The Warmth of Other Suns*, and Thomas Dyja’s *The Third Coast.*

As part of ongoing work of the “Reading Chicago Reading” project, we used the secure data portal of the HathiTrust Research Consortium to access and pre-process the in-copyright novels in our set. The HathiTrust research portal permits the extraction of non-consumptive features of the works included in the digital library, even those that are still under copyright. Non-consumptive features do not violate copyright restrictions as they do not allow the regular reading (“consumption”) or digital reconstruction of the full work in question. An example of a non-consumptive feature is the part of speech information extracted either in aggregate or without the connection to its source word. The locations in the text are another example of a non-consumptive feature as long as we do not aim to extract locations with the surrounding context: while the extraction of location from a work under copyright will not violate copyright law, the extraction of location with its surrounding context will do so. Similarly, the sentiment of a sentence also falls under the category of a “non-consumptive” feature as long as we do not extract the entire sentence and its sentiment score. Thus it was possible to utilize the HathiTrust research portal to access and also extract both the locations as well as sentiment of individual sentences from copyrighted works. As later paragraphs will reveal however, we also needed to verify a number of these extractions, which was done in a manual way: by checking the extracted references against the actual text of the work.

This paper arises from the finding that the three OBOC books that are set largely in or are about Chicago circulated differently than the OBOC books that are not, such as Marcus Zusak’s *The Book Thief*, Yiyun Li’s *Gold Boy*, and Michael Chabon’s *The Amazing Adventures of Kavalier and Clay*. Since we discovered that some CPL branches had higher circulation for “Chicago” books than others, we wanted to determine (1) which place names were featured in the three books, and (2) quantify and then examine the sentiment associated with these places.  Although recognizing a well-defined place in the text is no longer a difficult task thanks to the development of named entity recognizers such as the Stanford Named Entity Recognizer[[5]](#footnote-4) OpenNLP,[[6]](#footnote-5) spaCy[[7]](#footnote-6), and NLTK,[[8]](#footnote-7) recognizing whether a place name is a reference to a Chicago location is a harder task. If Chicago is the setting or one of the main topics of the book then we can assume that a number of locations mentioned will be Chicago places. However, if information about the topicality or locality of the book is not known in advance or if the plot in the book moves from location to location, then the task of verifying through automated methods whether a place name is a Chicago location is much harder.

With the help of LinkedGeoData[[9]](#footnote-8) we were able to obtain all the Chicago place names that were identified by volunteers through the OpenStreetMap project[[10]](#footnote-9) and download a listing that included Chicago buildings, theaters, restaurants, streets, and other prominent places. While this is very useful, we realized that we were missing historical place names. At the same time, the way that the place names are represented in a text will likely not always correspond to the way a place name is represented in the dictionary or a knowledge graph; for example, a sentence might simply note “that building” or “her home” instead of the named entity of the previous sentence. Moreover, there were many examples of generic place names: how many cities in the United States have a State Street, a Madison Street, or a 1st Avenue, and the like? A further hindrance was determining the type of place names we wanted to identify and collect from the text: it soon became obvious that for the purposes of visualizing a place name on the map, general references to Chicago went beyond the scope of the maps we wanted to create. We became more interested in tracking references to *specific* Chicago place names that included buildings (historical and present), named areas of the city, monuments, streets, theatres, restaurants, and the like. Given that our dataset comprised just three books, we were able to manually sift through the automatically identified place names and indicate whether they were indeed a Chicago place name or not. Prior to this we established the sentiment of each sentence in the three books using the Stanford Sentiment Analyzer.[[11]](#footnote-10) Our heuristic was that a sentiment score for the entire sentence would be assigned based on specific place(s) mentioned in the sentence. This assumption may not always be true, but our manual inspection of sentences and the sentiment assigned to them established that this method was fairly accurate. It should be mentioned that while we did examine some examples, we did not conduct an analysis of the accuracy of the sentiment scores assigned to the corpus.

Figure 1 indicates the result of our effort to integrate place names with the sentiment of the sentence.

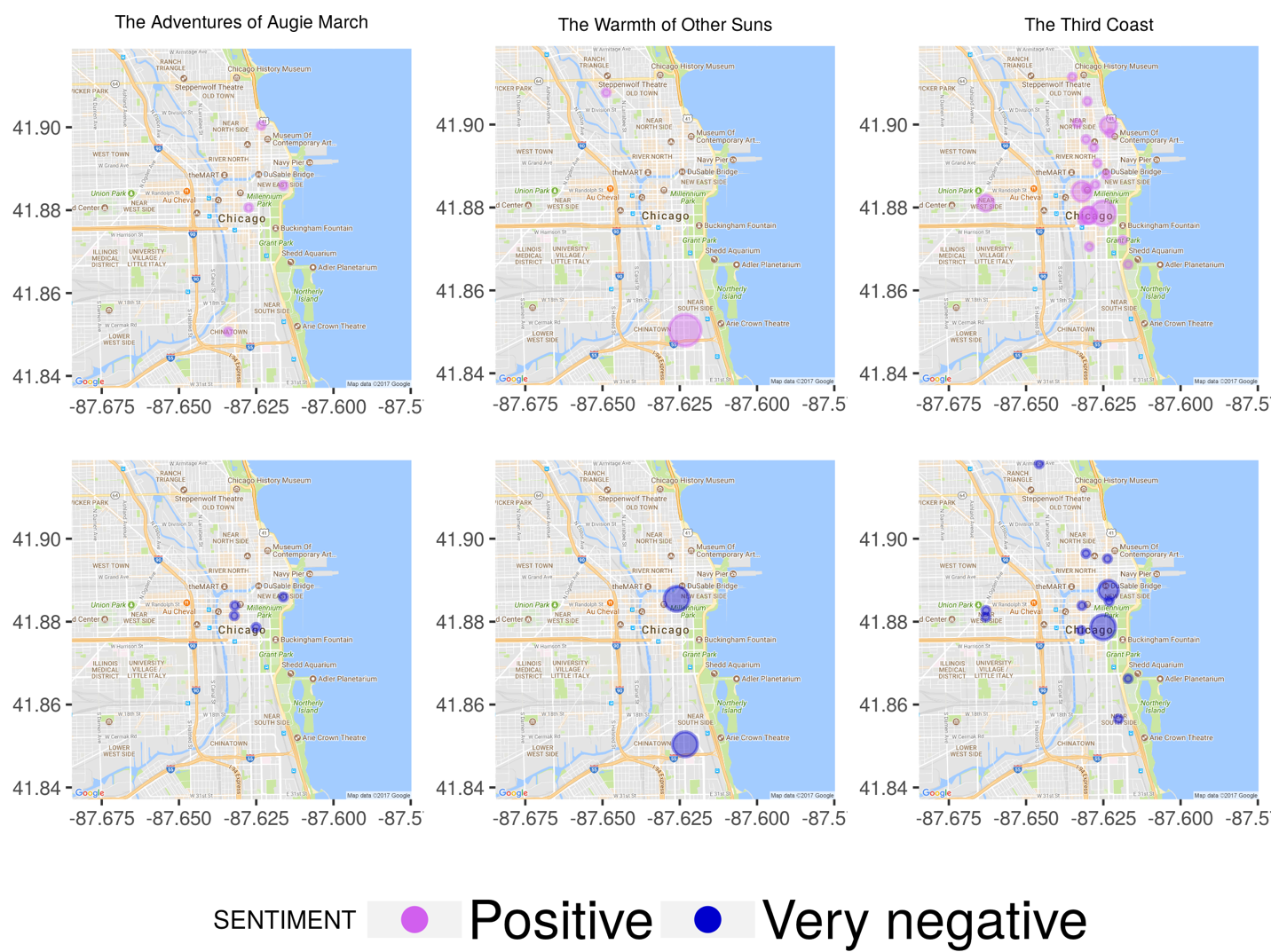


Fig. 1 – Visualizing place names associated with positive versus very negative sentiment extracted from three OBOC books.

Particularly telling here is *The* *Third Coast* quadrant which shows a concentration of positively-associated Chicago place names in the northern parts of the city along the shore of Lake Michigan. Negative sentiment appears to be more concentrated in the central part of Chicago and also in the southern parts of the city.

The place names extracted from our three Chicago-setting OBOC books allowed us to focus on particular areas of the city such as Hyde Park, which is mentioned in each of them. Larger circles correspond to a greater number of sentences that mention Hyde Park and are associated with a negative sentiment in both *The Adventures of Augie March* and *The Warmth of Other Suns*. Judging by Fig. 2, on the other hand, *The Third Coast* features sentences in which Hyde Park is mentioned in both positive and negative contexts.



Fig. 2 – Visualization of the sentences that feature “Hyde Park” and their sentiment from three books

These results prompt us to continue with this line of research and to procure a larger dataset that contains Chicago place names extracted from other works of literature. This would allow us to focus on specific places such as “Wrigley Field” or the once-famous but no longer existing “Mecca” apartment building (which stood at the intersection of 34th and State Street on the South Side and was immortalized in a 1968 poetry collection by Gwendolyn Brooks). With a robust place name set, we could analyze the context in which these place names were mentioned in other literature, in contemporary or historical newspapers (*Chicago Tribune*, *Chicago Sun-Times, Chicago Defender*), or in library and archival materials. One such contextual element would be the sentiment associated with the place name.

Our interest in creating a dataset of Chicago place names extracted from literature led us to *The Chicago of Fiction,* a vast annotated bibliography by James A. Kaser. Published in 2011, this work contains entries on more than 1,200 works published between 1852 and 1980 that feature Chicago. Kaser’s book contains several indexes that can serve as sources of labeled data or instances in which Chicago locations are mentioned. Although we are still determining how many of the titles included in the annotated bibliography already exist in digital format or are accessible through the HathiTrust digital library, it is likely that a subset of the total can be accessed electronically. Even if the books do not exist in electronic format presently, it is still possible to use the index as a source of labeled data for Chicago place names. We anticipate that such a dataset would be of interest to researchers in Urban Studies, Literature, History, and Geography. A sufficiently large number of sentences featuring Chicago place names would enable us to proceed in the direction of a Chicago place name recognizer that can “learn” Chicago context or examine how much context Is sufficient to establish whether a “Madison Street” in a text happens to be located in Chicago or elsewhere.

### **How do libraries innovate? From print index to labeled data**

### Over the last decade, libraries have taken on additional services related to the development and preservation of digital scholarship projects. Librarians frequently assist faculty and students with the development of digital humanities or digital scholarship projects. They point patrons to portals where they can find data and help with licensing. Librarians also procure datasets and some perform data cleaning and pre-processing tasks, and yet it is not so common for librarians to participate in the *creation* of a dataset. A relatively recent initiative, however, Collections as Data,[[12]](#footnote-11) directly tackles the issue of treating research, library, and cultural heritage collections as data and providing access to them. This ongoing initiative aims to create 12 projects that can serve as a model to other libraries for making collections accessible as data.

The data that undergird the mechanisms of library workings – circulation records for physical and digital objects, metadata records, and the like – are not commonly available as datasets open to machine learning tasks. If they were, not only could libraries refer others to the already created and annotated physical and digital objects, but they could also participate in creating objects that are local to their settings. Creation and curation of such datasets could in turn help establish new relationships between area libraries and local communities. One can imagine a new “data challenge,” for instance, in which libraries assemble a community by building a dataset relevant to that community. Such an effort would need to be preceded by an assessment of the data needs and interest of that particular community. In the case of a Chicago place name dataset challenge, efforts could revolve around local communities adding sentences to the dataset from literary sources. The second step might involve organizing a challenge of building a place name recognizer model based on the sentences gathered.

One can also imagine turning metadata records into curated datasets that are shared with local communities and with teachers and university lecturers for use in the classroom. Once a dataset is built, scenarios can be invented for using it. This kind of work invites conversations with faculty members about their needs and about potential datasets that would be of particular interest. Creation of datasets based on unique materials at their disposal will enrich the palette of services already offered by libraries.

One of the main goals of the Reading Chicago Reading project was the creation of a model that can predict the circulation of a One Book One Chicago selection given parameters such as prior circulation for the book, text characteristics, and locality of the work. We are not aware of other predictive models that integrate circulation records with text features extracted from the books in this way. Given that circulation records are not commonly integrated with other data sources when they are analyzed, linking different data sources with circulation records is another challenging opportunity that this paper envisions.

Ultimately, libraries can play a dynamic role in both managing and creating data and *datasets* that can be shared with the members of local communities. Using back-of-the-book indexes as a source of labeled place name data is a tool that we have begun to prototype but which requires further exploration. While organizing a data challenge takes a lot of effort, a data challenge can be an effective way of reaching out to one’s local community and identifying their data needs.

Final note: We aim to make available a curated list of sentences that mention Chicago place names in the three OBOC selections that feature Chicago. We will invite the public and scholars to add sentences extracted from other literature. The ultimate goal is the creation of a labeled training dataset for creation of a Chicago place name recognizer.

References

American Library Association. n.d. “One Book One Community.” Programming & Exhibitions (website). Accessed May 31, 2020. <http://www.ala.org/tools/programming/onebook>.

Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python*. Sebastopol, CA: O’Reilly Media Inc.

Chicago Public Library. n.d. “One Book One Chicago.” Accessed May 31, 2020. <https://www.chipublib.org/one-book-one-chicago/>

“Collections as Data: Part to Whole.” n.d. Accessed May 31, 2020. <https://collectionsasdata.github.io/part2whole/>.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling.” In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* 363-370. <https://www.aclweb.org/anthology/P05-1045/>.

HathiTrust Digital Library. n.d. Accessed May 31, 2020. <https://www.hathitrust.org/>.

Kaser, A. James. 2011. *The Chicago of Fiction: A Resource Guide*. Lanham: Scarecrow Press.

Library of Congress. “Local/Community Resources.” n.d. Read.gov. Accessed May 31, 2020. <http://read.gov/resources/>.

LinkedGeoData. “About / News.” n.d. Accessed May 31, 2020. <http://linkedgeodata.org/About>.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. “[The Stanford CoreNLP Natural Language Processing Toolkit.](http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf)” In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60. <https://www.aclweb.org/anthology/P14-5010/>.

OpenStreetMap. n.d. Accessed May 31, 2020. <https://www.openstreetmap.org/>

Reading Chicago Reading. “About Reading Chicago Reading.” n.d. Accessed May 31, 2020. <https://dh.depaul.press/reading-chicago/about/>.

**Editor Comments for second draft:**

This is almost ready to go, after a few points in the comments have been addressed. The most important is Mark’s comment about the conclusion. It still seems to end abruptly. What if you combined the last two paragraphs with some additional glue, as follows?

Ultimately, libraries can play a dynamic role in both managing and creating data and datasets that can be shared with the members of local communities. Reading Chicago Reading continues to make steps in this direction. We are prototyping back-of-the-book indexes as a source of labeled place name data, and we aim to make available a curated list of sentences that mention Chicago place names in the three OBOC selections that feature Chicago. We will invite the public and scholars to add sentences extracted from other literature, perhaps issuing a data challenge to build local energy. The end goal, the creation of a labeled training dataset for creation of a Chicago place name recognizer, will, we hope, enable new avenues of research that we haven’t even foreseen.

Thank you for addressing our questions about working with in-copyright novels -- I think readers will find that very informative.

1. Reading Chicago Reading project (<https://dh.depaul.press/reading-chicago/>) gratefully acknowledges the support of the National Endowment for the Humanities Office of Digital Humanities, HathiTrust, and Lyrasis. [↑](#footnote-ref-0)
2. See <https://www.chipublib.org/one-book-one-chicago/>. [↑](#footnote-ref-1)
3. See <http://read.gov/resources/>. [↑](#footnote-ref-2)
4. See <http://www.ala.org/tools/programming/onebook>. [↑](#footnote-ref-3)
5. See <https://nlp.stanford.edu/software/CRF-NER.html>. [↑](#footnote-ref-4)
6. See <https://opennlp.apache.org/>. [↑](#footnote-ref-5)
7. See <https://spacy.io/>. [↑](#footnote-ref-6)
8. See <https://www.nltk.org/book/ch07.html>. [↑](#footnote-ref-7)
9. See <http://linkedgeodata.org/About>. [↑](#footnote-ref-8)
10. See [https://www.openstreetmap.org/](https://www.openstreetmap.org/#map=5/38.007/-95.844). [↑](#footnote-ref-9)
11. See <https://nlp.stanford.edu/sentiment/>. [↑](#footnote-ref-10)
12. See <https://collectionsasdata.github.io/part2whole/>. [↑](#footnote-ref-11)