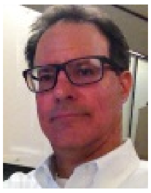
How information about library collections represents a treasure trove for research in the humanities and social sciences



WorldCat, an aggregate database of library catalogues worldwide, was primarily set up to aid libraries in carrying out their work in areas such as cataloguing or resource sharing. But the information it carries about much of the world's accumulated published output is also a a unique source of information for answering a wide range of questions about world literature and other forms of creative expression. ***Brian Lavoie*** *offers an insight into the types of questions WorldCat data can provide answers to, and how research of this kind also amplifies the value and impact of library collections.*

What is the most popular work by an Irish author? Where could one go to find out?

The answer to the first question is *Gulliver's Travels*. The answer to the second is WorldCat.

WorldCat is an aggregate database of library catalogues, encompassing the collections of many thousands of libraries worldwide. It contains more than 425 million bibliographic records, each describing a distinct publication, and more than 2.6 billion library holdings, each indicating a particular library holding a particular publication in their collection. The publications represented in WorldCat include materials of all descriptions – books, serials, maps, recordings, etc. – and from all time periods, from ancient times to the present. WorldCat is produced and maintained by OCLC, a non-profit global library cooperative that provides shared technology services, original research, and community programmes to libraries worldwide. A public interface to WorldCat is available online.

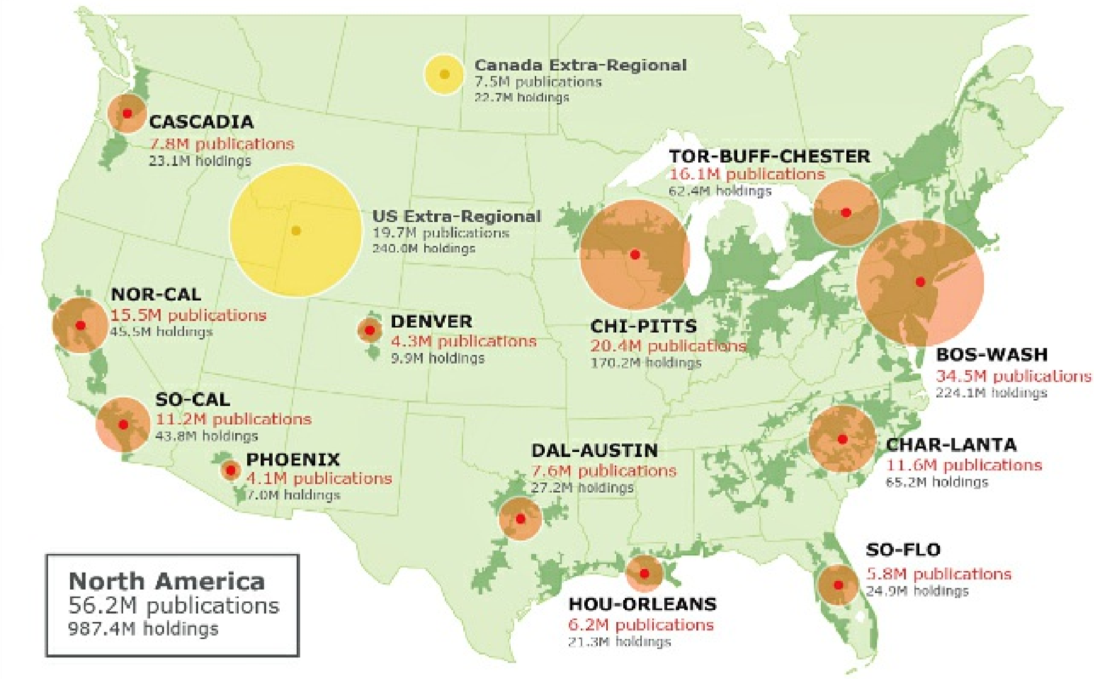
WorldCat's primary function is to aid libraries in carrying out their work in areas such as cataloguing or resource sharing. But it is also a treasure trove for research in the humanities and social sciences. WorldCat contains bibliographic information about much of the world's accumulated published output – especially monographs – and is therefore a unique source of information for answering a wide swathe of questions about world literature and other forms of creative expression. The answers are derived from data that libraries create to describe the materials in their collections, which is then aggregated in WorldCat. Using this data, interesting discoveries can be made not by looking in a book, but by knowing information *about* the book. And not just one book, but potentially tens of millions of books and other materials!

WorldCat provides us with the opportunity to engage in a "distant reading" of published materials. In a print-centric world, scholars were usually limited to "close reading": studying the contents of individual books, articles, and other materials. With the advent of powerful computing resources and digitised information, new possibilities have arisen for computational analysis of mass aggregations of digitised texts. Led by scholars such as Franco Moretti (who coined the term), distant reading brings the tools of data-intensive analysis to the humanities and other disciplines. While distant reading usually refers to mass analysis of the contents of books and other publications, we can also think of it as mass analysis of information *about* a collection of materials. WorldCat is the largest aggregation of information about the world's publications, and as such, it provides the opportunity for a distant reading of society's collective creative output, or some subset of that output.

For example, OCLC recently produced a study that explored the Irish presence in the published record, including materials published in Ireland, by Irish authors, and/or about Ireland. Using library holdings as a rough measure of popularity, we identified *Gulliver's Travels* as the most popular work by an Irish author, and Jonathan Swift as the most popular Irish author. We were able to determine that Oscar Wilde is the most translated Irish author. Using bibliographic data, we traced the patterns by which various Irish authors slipped in and out of fashion, or in the case of some, achieved enduring popularity. For example, the writer Bram Stoker enjoyed new heights of popularity (measured in terms of total number of publications) in the late 20th century, likely driven by a string of big-budget films based on his masterwork *Dracula*. And of course, the works of authors such as James Joyce and George Bernard Shaw continue to enjoy sustained interest.

Studying Ireland’s contributions to the published record demonstrates how library data can introduce a degree of precision to otherwise indistinct concepts. For example, what is the size of the Irish presence in the published record? Using the library data in WorldCat, we can say it consists of 892,814 distinct works. While that number is subject to caveats associated with WorldCat’s coverage, it nevertheless provides a reasonable approximation of a distinct corpus of materials within the published record. (We have also done similar analyses for Scotland and New Zealand, as well as a limited study of Poland.)

Library data can also be used in conjunction with socio-economic frameworks. OCLC produced an analysis of the distribution of library collections based on “mega-regions” work led by urban studies scholar Richard Florida. A mega-region is a regional conglomeration of multiple urban centres and their hinterlands, connected through networks of infrastructure, economic interdependencies, and cultural ties. For example, the Boston-Washington corridor on the US east coast is one such mega-region. Using WorldCat bibliographic and holdings data, we determined the size and characteristics of the “collective collection” formed by aggregating all library holdings in each of the 12 US and Canadian mega-regions identified by Florida and his colleagues. While this data illuminates many issues of special interest to libraries, it can also be used to correlate regional library collecting activity with a wide range of economic and social metrics.



Data represents print book holdings in WorldCat as of January 2018. Circles are scaled to reflect the number of distict print book publications in each mega-region.

Lavoie, unpublished data, OCLC Research, 2018

Figure 1: the North American mega-regional collective print book collections.

We also used WorldCat data to study the diversity of languages present in US public library collections. Bibliographic and holdings data helped us to establish a list of materials held by public libraries in each state, as well as determine the language of the contents of each publication. In this way, we produced a ranking of the most popular languages in the public library collections in each state. In all states, English – unsurprisingly – was the most popular language, and in most states, Spanish was the second most popular. Excluding English and Spanish, however, revealed a much more diverse tapestry of languages spread across the 50 US states. We explored the correlation of these findings with the most commonly spoken languages in each state, connecting public libraries’ collecting activity to the demographic features of the communities they serve.

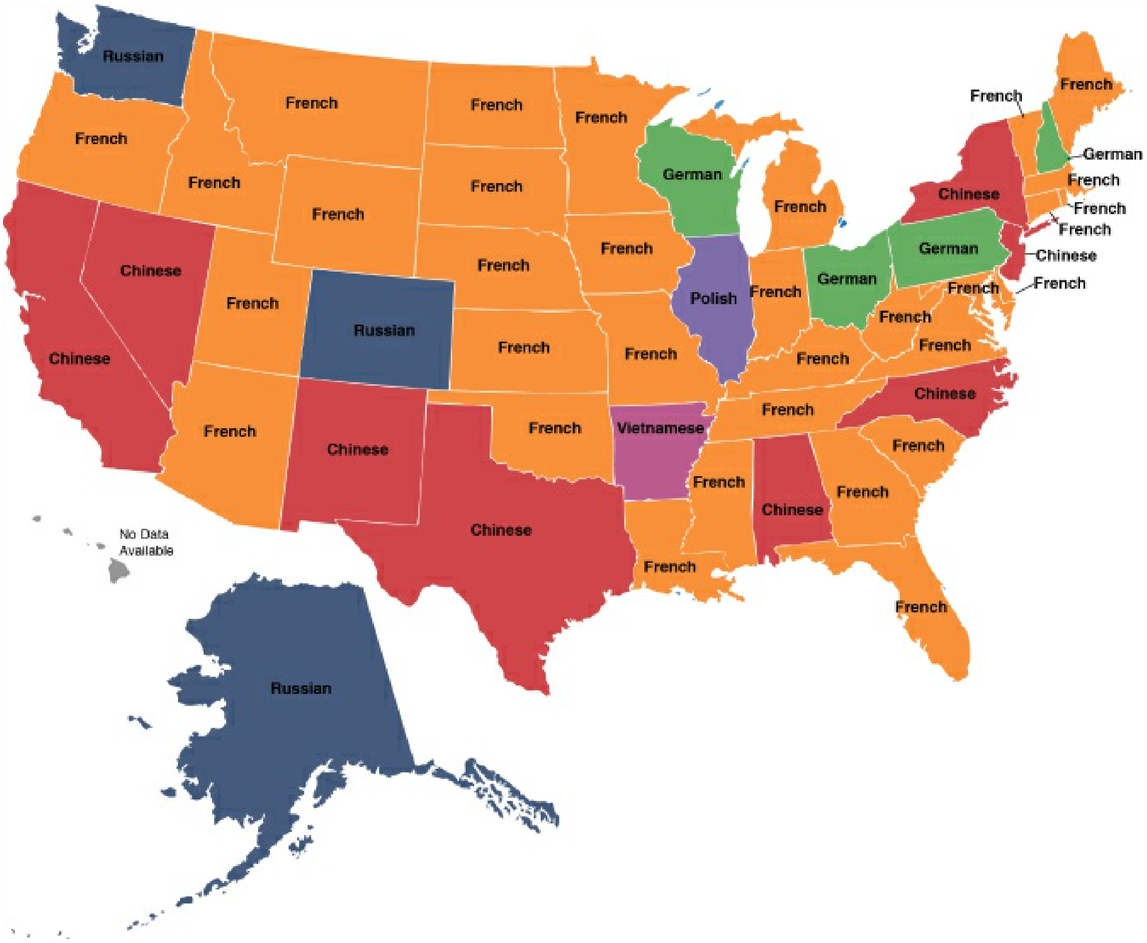


Figure 2: Most common language other than English and Spanish found in public library collections, by US state.

These studies provide a flavour of the types of analyses that can be done with library data. A database of WorldCat's proportions allows us to engage in new forms of "big data" analysis in literary and socio-economic research. WorldCat's coverage and comprehensiveness, though not perfect, is sufficient to build a big picture view of world literature and other interesting aggregations of material, which can then be explored at scale through computational analysis.

Research of this kind also amplifies the value and impact of library collections. Libraries create data to aid in the management and long-term stewardship of their collections, but at the same time they produce a spillover benefit in terms of the data's potential for revealing valuable information about society's published output. When aggregated in a database such as WorldCat, library data represents a detailed description of the world's published materials, as they appear in library collections. In this sense, libraries' investment in creating data about their collections not only supports care and use of the published record, but also broadens our understanding of the published record and its relationship to topics of interest in humanities and social science scholarship.

As libraries continue to fulfil their mission of connecting scholars to information resources, they will find new incentives to also connect scholars to their catalogues, not just as a means of discovering the treasures in their collections, but also as a window into patterns and trends in the evolution of creative expression.

The author is grateful to his colleague Lorcan Dempsey for valuable comments on an earlier version of this post.

Please note this post was updated on 20 September 2018 to feature a corrected version of Figure 2.

Note: This article gives the views of the author, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the author

***Brian Lavoie** is a Senior Research Scientist with OCLC. His research interests include analysis of library data, patterns and trends in the evolving scholarly record, and organisational and economic aspects of libraries and library capacities. His ORCID iD is: 0000-0002-7173-8753.*

Date originally posted: 2018-09-19

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/09/19/how-information-about-library-collections-represents-a-treasure-trove-for-research-in-the-humanities-and-social-sciences/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/