Books as Expressions of Global Cultural Diversity

Data Mining for National Collection Analysis

Timothy J. Dickey

A number of bodies have been jointly interested in book publication data as measures of cultural diversity. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics is especially interested in global patterns in book publication as expressions of cultural diversity and heritage. Such data, however, are not widely collected by national publishing organizations and library statistics agencies. The increasingly global reach of the WorldCat database, on the other hand, makes it an obvious source for data mining. This paper presents results from an OCLC Research project that produced a rich data portrait of global book publishing, with emphasis on collection analysis by country. Researchers were able to compare the annual publishing for every country of the world (as reflected in WorldCat), the libraries that collect and import a country's works, the monographs their libraries import from other countries, and the proportion of publications in various official and native languages. The results provide a global overview of book publishing and a wealth of case studies in single countries' practices in book publishing and the preservation of their literary heritage. The present paper compares the book publishing and book collections in libraries in six countries around the world and demonstrates the power of data mining within this sphere.

Timothy J. Dickey (tdickey1@kent.edu) is an adjunct faculty member at Drexel University, Philadelphia; Kent State University, Kent, Ohio; and San Jose State University, San Jose, California.

Submitted November 11, 2010; tentatively accept for publication December 13, 2010, pending modest revision; revision submitted February 16, 2011; accepted for publication March 15, 2011.

This research was conducted when the author was a postdoctoral researcher at OCLC Research in Dublin, Ohio. Outcomes from the research project have been previously reported in an OCLC Research Webinar (Sept. 16, 2010), at the Library History Seminar XXII: Libraries in the History of Print Culture (Madison, Wisc., Sept. 11, 2010), at the XXIX annual Charleston Conference (Charleston, S.C., Nov. 5, 2009), and at the 4th International Conference on the Arts in Society (Venice, Italy, 28–31 July 2009, not presented in person), The author wishes gratefully to acknowledge the helpful comments and contributions of Lynn Silipigni Connaway, Jeremy Browning, Karen Smith-Yoshimura, Eric Childress, and the anonymous reviewers in the preparation of this paper.

> "The limits of my language mean the limits of my world."
> —Ludwig Wittgenstein, *Tractatus Logico-Philosophicus*, 5.6

Globally and nationally, book publishing represents a central kind of cultural heritage. A number of bodies have found themselves jointly interested in any statistics to measure book publication as a measure of cultural diversity. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics has been exploring library statistics for worldwide book consumption and helped found the European Expert Meeting on Book and Library Statistics.[1] These bodies, as well as the International Federation of Library Associations (IFLA), are especially interested in any global patterns in the publication world as expressions of cultural diversity and heritage. Such data, however, are not widely collected by any national publishing organizations or library statistics agencies. The increasingly global reach of library records in the WorldCat database, on the other hand, makes it an obvious source for data mining. OCLC's bibliographic database represents more than 200 million items, with 1.6 billion copies held by libraries worldwide.[2] It is well suited to serve as a global aggregate library collection, to be mined for data on national publication and library collection profiles.

This paper presents the fruits of an OCLC Research project that aimed to compare national publishing profiles and to determine whether WorldCat provides an adequately diverse bibliographic collection to allow comparison worldwide.[3] The project used the method of data mining to research specific aspects of the global literary arts, with emphasis on collection analysis by country and region. Researchers specifically attempted to profile the annual publishing for every country of the world (as reflected in the WorldCat database), the libraries that collect and import a country's works, the monographs their libraries import from other countries, and the proportion of publications in various official and native languages, as well as data on translated works. The results provide a global overview of the publishing arts and a wealth of case studies in single countries' practices in both literary publishing and the preservation of their literary heritage.

This paper compares case studies of book publishing and library book collections from six countries: Bolivia, Chile, Germany, Poland, South Africa, and Thailand. As a set of test cases, the six were chosen to highlight non-English works and non-English cataloging, and to reflect a mix of continents across the world, of development levels, and of OCLC member libraries' contributions to the database. All data presented are limited by librarians' contributions to WorldCat. However, the six case studies amply demonstrate the strength of the data-mining methodology within this sphere and the richness of the data possible to mine for historical profiles of national bibliographies.

Literature Review

The UNESCO Institute for Statistics has been exploring library statistics for worldwide book consumption, actively promoting the 2005 Convention on the Protection and Promotion of the Diversity of Cultural Expressions.[4] Specifically, one tenet of the expected results is "Linguistic diversity [being] promoted through publishing and translation."[5] A number of bodies—UNESCO, IFLA, the International Publishers' Association (IPA), and the International Booksellers' Federation (IBF)—met at the European Expert Meeting on Book and Library Statistics in November 2008 to explore potential pilot projects to collect national bibliographic statistics; some of the individuals involved include Michael Heaney, executive secretary of Oxford University Library Services and chair of the ILFA Committee for Statistics and Evaluation; Simon Ellis, head of Cultural Expressions and Creative Industries, UNESCO Institute for Statistics (Montreal); and Mauro Rossi, UNESCO Division of Cultural Expressions (Paris).[6] Similar meetings occurred at the request of IFLA in June and September of 2008, resulting in a consultant's report on book statistics, though funding for further steps at that time was not available; contact with OCLC Research offered one option for further research.[7]

One specialized expression of the interest these bodies have is tracking any available statistics on global book publishing in the Index Translationum, an "international bibliography of translations" published by the UNESCO Division of Cultural Expressions in print since 1932 and in digital form since 1979.[8] This resource tracks the translations of a culture's books into other languages, and the online database contains more than 2 million records. However, this excellent tool is somewhat limited in scope because it is dependent on data donated annually by participating national libraries and uses a proprietary data format—it has, for instance, no ISBN field for linking its data to other bibliographic tools. In addition, neither book publishing organizations, nor national libraries as a whole, nor library statistics agencies have been collecting data on book publishing on any global scale.[9]

For countries around the world, some literature addresses national bibliographies. Recent examples include articles relating to the national bibliographies of Canada, Korea, a collection of Eastern European nations, and Mauritius.[10] Since 2004, IFLA has published the journal *International Cataloging and Bibliographic Control*; this publication has presented much valuable research, especially in the area of standards for international cooperation in cataloging. However, a noticeable gap remains in global bibliographic control and in comparative studies.

OCLC Research had recently developed one prototype service, the OCLC WorldMap, for visualization of global library data.[11] OCLC Researchers Lynn Connaway, Timothy Dickey, and Jeremy Browning, with OCLC librarian Lawrence Olszewski, developed the prototype service to mine, collect, and compare library data from both WorldCat and standard print reference sources and to compare bibliographic information of different countries. The WorldMap graphically portrays and compares library and bibliographic data, including titles published in a country, holdings worldwide of those titles, libraries in a country and their type, national expenditures on those libraries, and concentrations of archives, museums, and other cultural heritage institutions.[12] UNESCO interest in this particular prototype and in OCLC Research experience in the method of data mining led to the research project reported here.

Research Method

The method selected was data mining. OCLC's Office of Research has invested significant effort in the area of data mining.[13] Data mining is the computer-aided analysis of databases and other large datasets (such as records of

website hits, logs of transactions within an automated system such as a library catalog or online retailer, or electronic stores of demographic data) to expose new information derived from the aggregate of the data. The technique first appeared as a tool for business intelligence, only later to be adopted by libraries; the success of Google and Amazon has taught the library field that greater value exists within bibliographic data as well. Libraries have made huge investments in creating and maintaining rich, structured information describing the resources in their collections. These data already embody considerable value by supporting basic local access and inventory control. They also represent potential value in terms of knowing more about the characteristics of library collections. OCLC Research sees data mining as an effort to get increased return on libraries' investment in this bibliographic data.

Specifically, research projects have demonstrated the value of the WorldCat database as an "aggregate collection" of bibliographic data.[14] As a global-scale dataset of potential value, it can "not only provide librarians data for decision-making for collection and service development, but also provide users with enhanced discovery and access methods."[15] The WorldCat database is an increasingly global and increasingly comprehensive source of bibliographic data and remains strongest in its data on books. WorldCat contains more than 200 million records, with more than 1.6 billion holdings of those resources; approximately 54 percent are non-English catalog records, illustrating the increasingly global reach of the "aggregate collection."[16] Its member libraries are located in more than 100 countries, and the data go beyond those countries to include works from countries that are collected in other OCLC member libraries. Perhaps most importantly for this project, WorldCat contains publications in more than 470 languages; language was a central part of the definition UNESCO and OCLC considered while developing the OCLC Research project "Books as an Expression of Cultural Diversity." WorldCat, although somewhat limited by its Anglo-centric roots, offers a globally aggregated source of bibliographic data to examine the question at hand.

The basic objective of the project, then, is to mine WorldCat's overwhelmingly monographic records for data on global book publication and collection patterns. Researchers parsed the monographic data by country of publication, year of publication, and language use as a measure of cultural diversity. An axiomatic concept on the importance of language exists in cognitive anthropology that (in the oft-quoted thoughts of Benjamin Lee Whorf) "language shapes the way we think, and determines what we can think about."[17] The so-called Sapir-Whorf Hypothesis offers the concept that language functions not only as a framework for communication between speakers, but also as a framework for our basic comprehension of the world; it has been tempered somewhat in regards to linguistic relativity starting in the 1980s, but remains an influential interpretation.[18] The language or languages spoken by a culture help determine that culture's perception of the world and its expression of itself within that world; thus language data remain important to tracking books as expressions of any culture.

Within this method of data mining, OCLC researchers set specific data limits. To filter only textual materials, the MARC leader field "Type of Record" must carry the value T06 = a (for books), and leader field must be 07 = a or m (indicating a record for either a monograph component part or a complete monograph); this procedure excluded serials, theses, and dissertations, but the data mining otherwise cast a deliberately wide net. Publication dates in the catalog records (also taken from the fixed field elements where possible) had to be numeric and less than or equal to 2010; this filtered out works coded with publication dates of, for instance, "19xx," which could not be folded reliably into the rest of the data. Dates as early as 1000 AD were included, but these records tended to be outliers in the data for any particular country.

Technical staff extracted all bibliographic records that passed these filter into databases, which were subject to internal analysis and are available for further work by outside researchers. Each bibliographic record was counted in Functional Requirements for Bibliographic Records (FRBR) terms, so manifestations were counted rather than works because of the assumption that any new edition (be it of Shakespeare or the Bhagavad-Gita) is a fresh cultural artifact and pertinent to the cultural collection. Library holdings were counted for each manifestation—both worldwide holdings of the nation's cultural heritage and, where possible, the balance between worldwide holdings and holdings within the country of publication, a measure of how foreign libraries collect and value that cultural heritage. Researchers collected all language data possible from the fixed-field data elements, as well as from the 041 (language code) field, which contains not only three-letter codes for the language of the item being cataloged, but also subsidiary codes for the original language or languages if translation of the original has occurred.[19] The language data thus was able to represent how both "official" and "indigenous" languages are represented in a country's bibliographic heritage, and also how multiple linguistic content appears in the national collection.

This paper covers a sample of results, primarily case studies in the data from six countries (Bolivia, Chile, Germany, Poland, South Africa, and Thailand) in an effort to answer the following questions:

- How do these national bibliographic collections (as reflected in WorldCat) compare to one another?
- Does WorldCat provide an adequately diverse *global*

bibliographic collection to allow such comparisons worldwide? In other words, are the national bibliographies, as reflected in WorldCat, distinct enough from one another, and distinct in culturally and historically justifiable ways, to support the WorldCat data as globally reliable?

The six countries (Bolivia, Chile, Germany, Poland, South Africa, and Thailand) addressed in this paper deliberately highlight non-English works and non-English cataloging. They include some mix of continents across the world, as well as a mix of development levels, and a mix of OCLC member libraries' contribution to the database. South Africa has eleven official languages; the other countries have single dominant languages and a handful of minority tongues. Data have been extracted from WorldCat (at the time of writing) for all non-U.S. countries, totaling more than 66 million bibliographic records and 450 million holdings. The findings from the six profiles tend to support the integrity of the worldwide dataset mined from the data in the WorldCat database in the distinctiveness of the profile they give of each national bibliography.

Limitations

All of the data discussed below are subject to the caveat that a national bibliographic profile is being constructed as reflected in WorldCat. This means that it is subject to what libraries—and specifically libraries that participate in WorldCat—*have* collected and cataloged. For some countries in the present comparison, such as Thailand and Bolivia, very few libraries in each country have been OCLC members, and thus the data will tend to reflect more of what other Anglo-American libraries and other major national libraries participating in WorldCat have collected of their publications.

In addition, several issues of Anglo-American cataloging practice and other cataloging standards as translated into MARC 21 affect the following profiles. For example, different cataloging standards may have different (or changing) concepts of what will be coded as a book. Even more difficult for the purpose of this research is the definition of a country. The German data below, for example, reflect generations of different catalogers' judgments of several centuries of shifting boundaries that have finally coalesced into what we in 2011 call Germany. The national bibliographic profiles of the current Balkan nations may be completely corrupted by the shifting assignments of MARC country codes in the region. Furthermore, cataloging practice in the area of date of publication may vary, especially with reprintings of prominent historical works. Finally, none of the historical profiles below should be taken as an assertion of historical causality on national book publication and collection.

Findings and Discussion

Book Publishing Profiles: Basic Publication Data

The increasingly global and comprehensive nature of the WorldCat database should yield ample and relatively trustworthy representations of the six national bibliographic profiles being considered. In addition, the six basic historical publication profiles offer remarkably distinct and lucid images of each country's evolving publishing history and how books and library collections tend to reflect the history of each.

For the six case study countries, table 1 shows the scope of the basic dataset, with the ongoing caveat that these are the records represented in WorldCat. This gives an idea of the richness of the data possible to mine in WorldCat, country by country. German publications, not unsurprisingly, are the richest subset of these data; the records from Germany include, pertinently, the catalogs of the Deutsche Nationalbibliothek (DNB), the Bayerische Staatsbibliothek, and Hessische Bibliotheksinformationssystem (HeBIS), the consortium of the largest university libraries in the Federal Republic.[20] In addition, this country has spearheaded national efforts at bibliographic control under the leadership of the DNB.[21] South Africa has a relatively high rate of holdings for its national publications, perhaps because of the participation of the University of Cape Town and the University of South Africa (Unisa) as cataloging partners in WorldCat.[22] Of the six countries in this study, Bolivia has the fewest records in WorldCat; one can presume that their national bibliography is larger than 58,000 titles, but the lesser participation of Bolivian libraries in WorldCat to date somewhat weakens the publishing profile presented here.

Table 2 provides a comparative sample of several other countries' datasets within WorldCat. It appears that for many countries (depending somewhat, but not entirely, on libraries' contributions), the data mining will be able to profile a large set of publications throughout their history. Even in the case of a problematic country like Ukraine (a former Soviet republic), enough catalogers participating in WorldCat even during the Cold War have paid attention to specifying that Soviet publications from Kiev should receive the country code for the Ukrainian Soviet Socialist Republic (unr) rather than the blanket code for the Soviet Union (ur). Thus Ukraine's bibliographic heritage can remain separable from the Soviet Union in the bibliographic data. In the case of the national bibliography as seen in WorldCat for China, the data was mined in August of 2009 and will certainly offer a much more complete picture once bibliographic data from the National Library of China are fully integrated into the database.[23] The example of China highlights the importance of national libraries' contribution to the aggregate bibliographic data available for data mining in projects

Table 1. The Six Case Studies

Country	OCLC Member Libraries	Publications	Holdings	Translations
Bolivia	1	58,838	302,309	780
Chile	387	265,948	900,330	7,351
Germany	378	12,843,605	70,341,725	381,141
Poland	27	1,225,446	3,108,677	174,738
South Africa	1,082	299,574	2,014,327	10,063
Thailand	39	181,003	463,104	5,700

Table 2. A Sample of Other National Bibliographies

Country	OCLC Member Libraries	Publications	Holdings	Translations
China	1,160	3,208,114	8,160,062	238,315
Finland	71	905,252	1,476,734	122,150
France	253	4,948,620	29,127,570	211,341
India	37	1,121,311	6,636,530	52,107
Russia	45	1,883,418	8,151,097	94,319
Ukraine	3	236,166	710,372	9,061

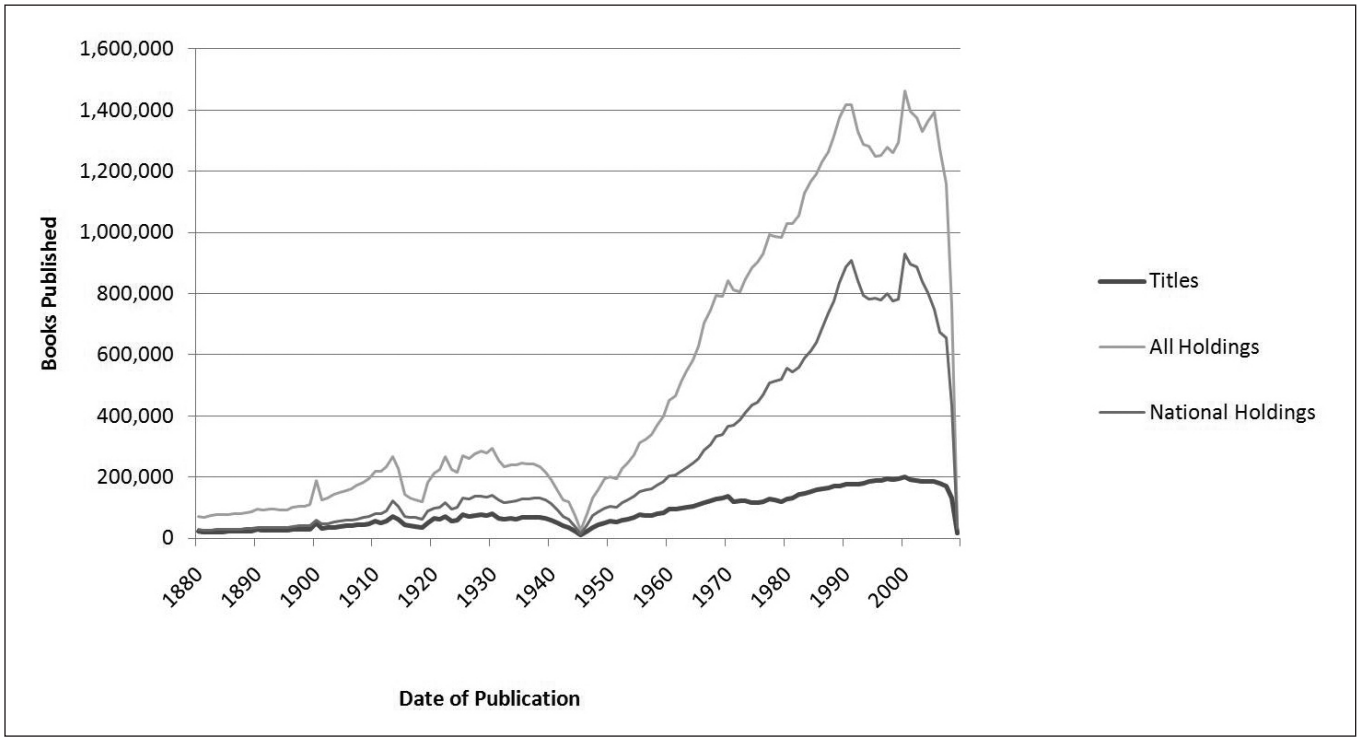

Figure 1. Book Publication in Germany 1880–2010, as Reflected in WorldCat

of this type: whereas acquisition of a country's publications by American research libraries can offer the beginnings of a robust dataset, the bibliographic universe represented in WorldCat, Google, or any such data source is highly fortified by including the contents of national libraries.

A graphical representation of the data *within* a subset can at times be even more revealing. The robust data on book publication and library collection from Germany in the twentieth and twenty-first centuries (figure 1) reflect the historical dips in book publication from 1914 to 1919 and the complete collapse of the industry in 1945 and 1946. (No necessary historical correlation is being asserted, but rather the overlap of historical events and changes in the publication profile as shown in the data mining.) After German unification in the 1990s, German book publication itself may not have waned, but the merging of libraries and university systems within the now-unified nation may have led to fewer copies being held (thus being represented graphically in a parallel dip in both national and international holdings of German books). From a peak of more than 900,000 indigenous library copies of German publications in 1991 (the bulk in these data of almost 1.5 million copies worldwide), by the middle of the 1990s, the German libraries' holdings of their own publications had dipped to around 780,000. The trough in "All Holdings" worldwide in figure 1 corresponds to the apparent culling of duplicate copies in German libraries.

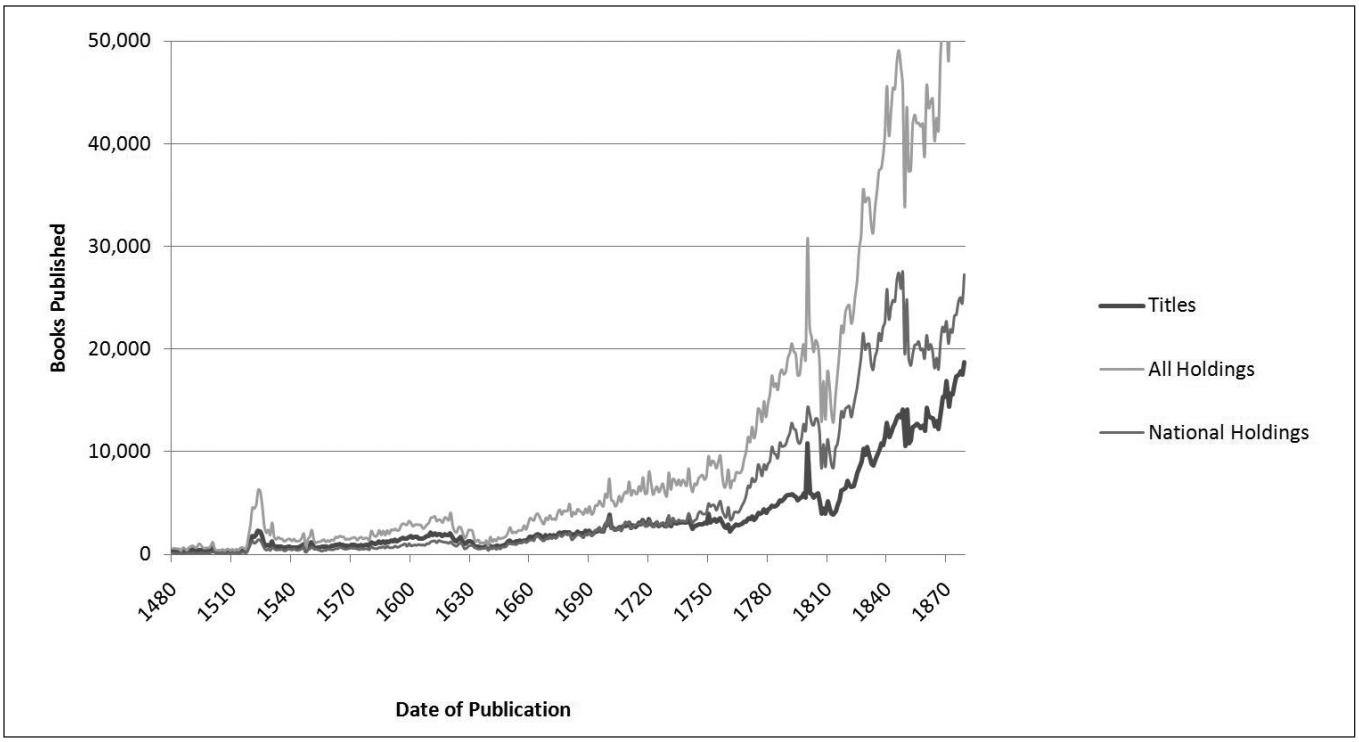

Figure 2. Book Publication 1480–1879 in Germany, as Reflected in WorldCat

The robust German dataset within WorldCat similarly documents some interesting overlaps with events earlier in the history of the German-speaking lands. In this earlier dataset (see figure 2), one can note several changes in recorded book publication that correspond to important historical developments. These include dips in both book printing and holdings during the Napoleonic Wars (1800s to the early 1810s); publications peaked at more than 4,500 in 1808 (with nearly 17,000 holdings) but would not return to similar numbers until 1815. During the revolutionary period and into the 1840s, total publications surged from around 9,000 books to 14,093 in the year 1848 (with a corresponding growth in libraries holding copies), immediately dropping nearly a third to 10,550 in the year after the tumultuous the 1848 revolutions. The WorldCat data even present a smaller but palpable falloff in book publishing activity much earlier, during the Thirty Years' War: a publication rate of almost 2,000 books a year from 1617 to 1620 falls to only 974 in 1627, to an abysmal 517 in 1639, and no real recovery until well after the war's end in 1648.

German data also show publication rates that may correspond to positive historical effects on the national literary scene. During the reign of Bismarck, the rates of publication within German-speaking lands surge from 14,404 in 1871 (Bismarck's ascension to Imperial Chancellor) to 28,440 in 1890 (the year of his resignation). The German publication data even experience an early peak during the Protestant Reformation that began in 1517. Only 327 publications survive from 1517, but that leaps immediately to 777 in 1518 and crests at 2,341 in 1523. This peak also could reflect the historical importance of Reformation materials, including the explosion of warring theological pamphlets, which libraries would tend to heavily collect and preserve.[24]

The unique nature of the German data also is revealed in comparison to the historical data from France (figure 3). The French data also contain a drop in publication rates in the revolutionary year of 1848, from 7,262 publications in 1847 to 5,815 in 1849. Even more pertinently, in 1871, the year the Prussians occupied Paris, the publication record crashes from 9,269 in the previous year to fewer than 6,000. The years 1789 and 1790, around the time of the French Revolution, see the most dramatic spike in both French national print publication and, perhaps tellingly, in libraries' collection and preservation of materials, visible in figure 3 as a quadrupling of printed publications collected by libraries to more than 20,000 titles.

Alternatively, one can compare the twentieth-century data from Germany (figure 1) to the twentieth-century data from neighboring Poland (figure 4). The Polish data, as might be expected, do show a slump during their occupation by Nazi and Soviet forces, as well as a surprising dip in general worldwide holdings in the late 1970s, followed by a surge

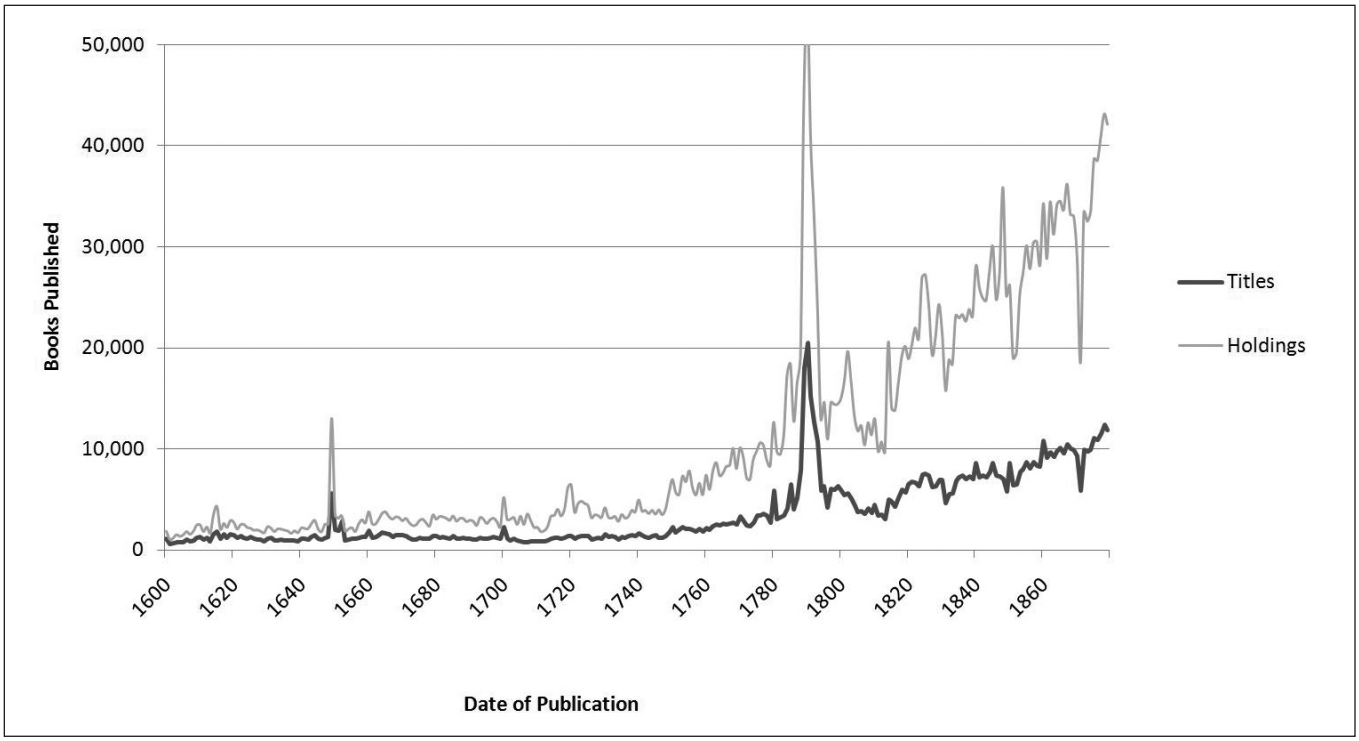

Figure 3. Book Publication in France 1600–1870, as Reflected in WorldCat

in publishing after the fall of Communism, a period that also corresponds to a very volatile period in the Polish publishing industry as the vibrant underground publishing community from the 1980s suddenly experienced the collapse of state censorship as well as turbulent economic conditions.[25] Also, one can note that the national holdings for Poland—holdings of Polish publications in Polish libraries—are in every year consistently lower than the total number of titles, an indicator of fewer data from libraries within Poland.

For a non-European country like South Africa, on the other hand, a comparable slump in publication activity during World War II would be less expected, and indeed it is less evident in the data. One major turning point in the South African data profile (see figure 5) instead follows the establishment of a national legal library deposit system in 1959.[26] South African publications, as reflected in the WorldCat data, exploded from 1,695 publications in 1959 to more than 9,500 in the year 1990. However, international pressure in the 1970s and 80s against Apartheid could be reflected in a tension between flatter growth in South African books within library collections outside of South Africa: over the same span of years, South African books discovered by this research in South African libraries increased more than tenfold, from 3,104 South African library holdings in 1959 to 41,234 in 1990, while non-South African library holdings of the same publications only increased from 7,000 to 20,000. The numbers are even more striking in proportion to the overall surge in publication rate. In 1959, each book published in South Africa was collected on average by at least four other libraries worldwide; by 1990, the proportion was fewer than two libraries per publication.

Book Publishing Profiles: Language Data

Reliable data on language of publication (see appendix A) also emerged from most of the sample national datasets and in most cases reflected appropriate and expected differences between them. Although the dominant language of publication in books from Bolivia was expectedly Spanish, note the presence in WorldCat of books published in indigenous languages, such as Aymara, Quechua, and Guarani. The data from Chile can be treated with even greater trust knowing that the country enjoys a national shared library network.[27] English-language publishing is apparently stronger in Chile than in Bolivia, despite the lesser dependence of the Chilean data on libraries external to Chile and perhaps can be explained by the dominant strain of minority English speakers in Chile.

For the larger European countries, conversely, the book publishing seems to be more Eurocentric, with less emphasis on a single dominant national language, greater emphasis on publishing in other European languages, and

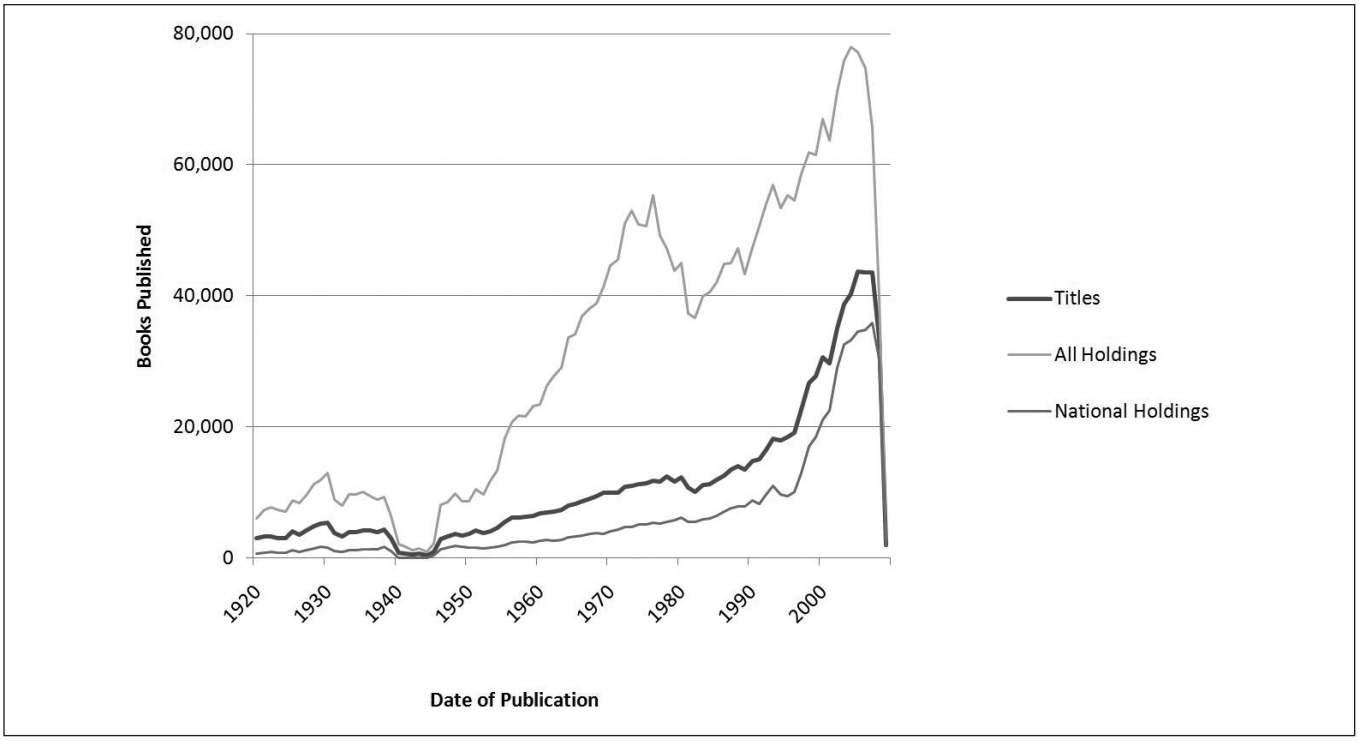

Figure 4. Book Publication in Poland 1920–2010, as Reflected in WorldCat

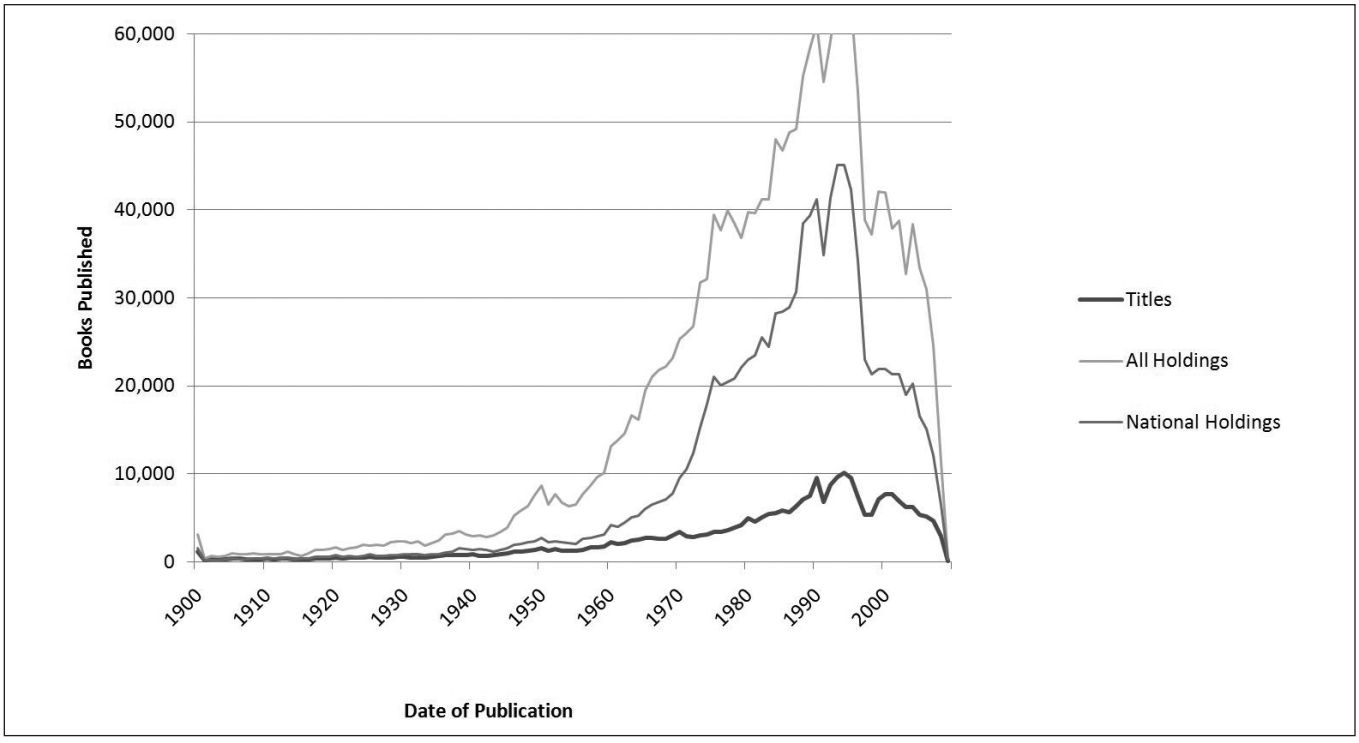

Figure 5. Book Publication in South Africa 1900–2010, as Reflected in WorldCat

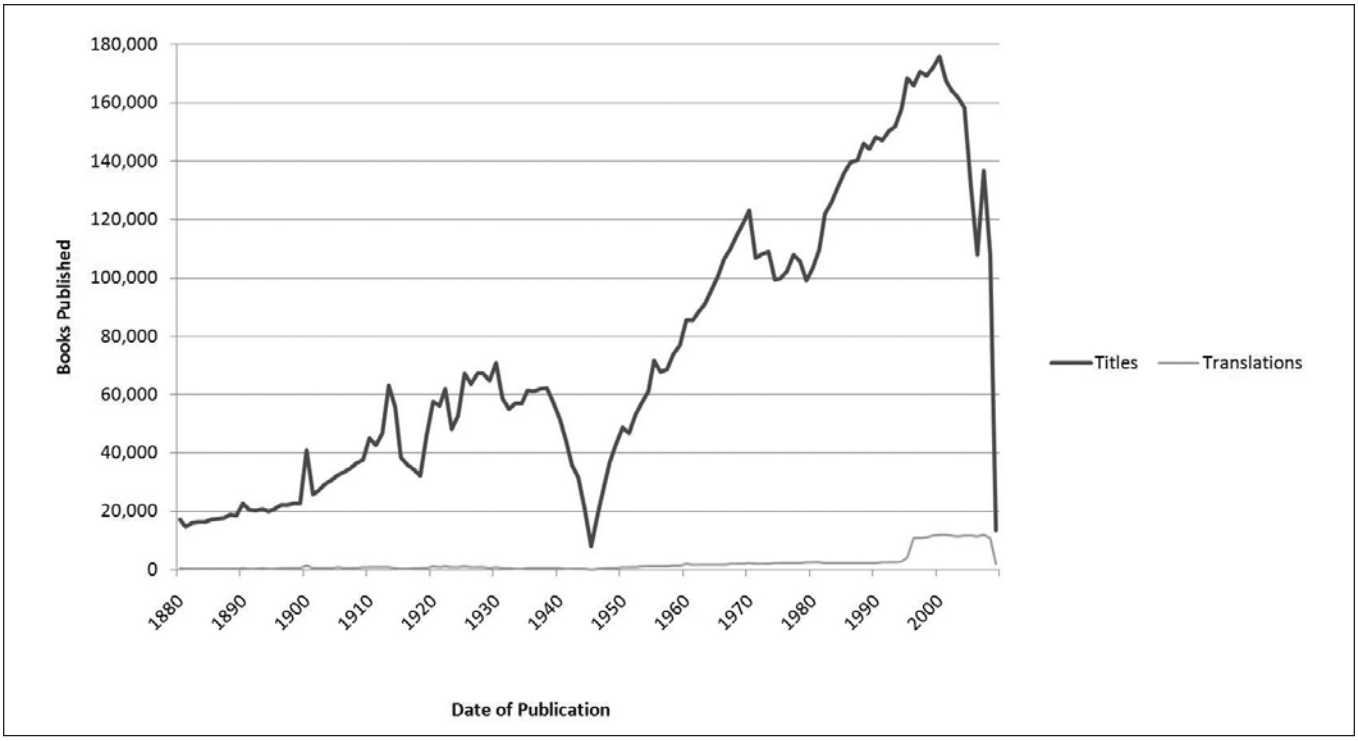

Figure 6. German-Language Publishing in Germany 1880–2010, as Reflected in WorldCat

fewer appearances of minority but native languages such as Danish, Frisian, and Sorbian in the German publications, and Lithuanian in those from Poland. On the other hand, these data include a strong presence of books published in Latin in the data from Germany and Poland—a concrete reflection of Western libraries' function as historical memory institutions. Not shown in appendix A but definitely present in the Polish and German data are works in Greek, Hebrew, and historical languages such as Low German and Middle High and Old High German, again reinforcing the function of the library collection in preserving these aspects of the historical culture. The data from South African book publishing, on the other hand, appropriately reflect the more complex linguistic heritage of that country—two dominant languages of the ex-colonial powers (and likely founders of much of the South African publishing industry) with a very healthy dose of both translations into, and works original to, a variety of indigenous languages.

The language data also can be parsed out by year across the historical span of a dataset. German-language publications in Germany, for instance (see figure 6), display interesting nodes around a spike in German-language publications in 1913 (leaping to more than 60,000 publications in the national language) as the country was gearing up for what would become the First World War, and a surprising dip through the 1970s. Compare that to the very different graphical shape of Germany's Latin-language publications (figure 7). Germany's Latin-language publishing first spiked in 1517 (with 245 Latin-language publications collected and preserved by libraries at the outset of the Reformation and appearing in WorldCat), and remained generally strong (around 1,000 titles per year published in Latin) up to the nineteenth century. At the start of the twentieth century, however, Latin-language publishing in Germany fell off to numbers around 800 titles per year; in the 1930s, during the Nazi era, this fell even more dramatically, and 1968 (the year the Second Vatican Council liberalized many Roman Catholic practices) ushered in another decline in Latin-language publication. (The spike in the year 1972 could be an outlier data point within the trajectory of the data; a random sample of Latin-language publications from Germany in 1972 revealed no explanation for such an increase in Latin-language publication in this year.)

Book Publishing Profiles: Translation Data

Data mining also was able to reveal patterns in book translations (see appendix B). Such data, allowing a view of the interactions between languages used by a culture, is among the most valuable to UNESCO as it seeks to measure cultural statistics. Once again, the data for Poland and Germany reflect a more Eurocentric vision, with translations from a

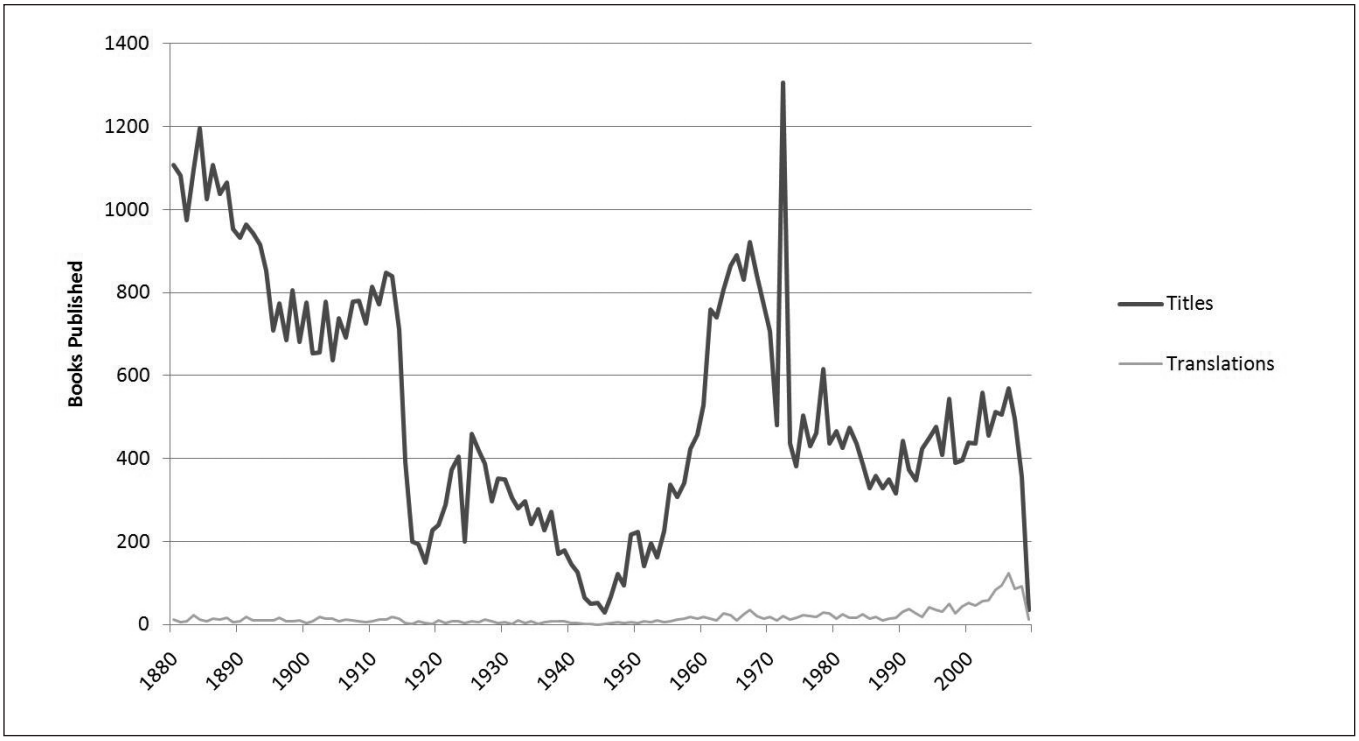

Figure 7. Latin-Language Publishing in Germany 1880–2010, as Reflected in WorldCat

variety of major European languages into German and Polish responsible for the majority of the translations; although in these data as well, translations into English from German and Polish also figure prominently. Latin and Greek translations also are represented in the WorldCat data from these two countries. The data on these translations were, however, somewhat less prominent. Greek to German translations were the tenth most prevalent, at 6,178 books; in Poland, Latin to Polish translations (2,687 books) were somewhat less common than Polish to German, and Greek to Polish lagged behind, although still recording 1,072 translations. The data on translations for Bolivia, Chile, and Thailand give a similar picture of both the predominant languages spoken in each country (Spanish and Thai), as well as the interaction of English and other languages with these dominant languages.

The data for the more culturally diverse country of South Africa, on the other hand, presents a much more varied tapestry of translation data, including translations between English, German, Afrikaans, and a number of tribal languages. Even more interesting historical patterns are revealed by plotting these data on a similar historical graph (figure 8). Whereas book publications from South Africa (figure 5) were more avidly collected after the 1959 deposit library system, the dominant type of translated works during the early decades of the Republic (1961 until the mid-1980s) are translations between English and Afrikaans, varying between around 25 and 150 per year. After the Zulu and Xhosa languages were decriminalized in 1991, however, translations from them enjoy a surge to more than 50 translations per year.

A different set of shifting cultural interactions may be viewed in the historical patterns of translated works from Poland (figure 9). The most overt pattern is the large increase in translations from English, French, and German after the fall of Communism; English to Polish translations, particularly, rocketed from fewer than 400 in 1989 to more than 6,000, literally off the chart, after 2000. But from the 1950s to the 1980s, the most important pair of languages tended to be Russian and Polish. In a lesser-populated dataset, such as the publication and collection profile from Thailand (figure 10), the spiky nature of the graph suggests that data mining in this instance is nearing the functional end of its reliability. The Thai data on translations also are the only instance in which translations into English lead the statistics (as reflected in WorldCat).

Future Research

These findings suggest several avenues for further research. First, data mining that compares copies of a country's national bibliography in foreign libraries to the holdings of

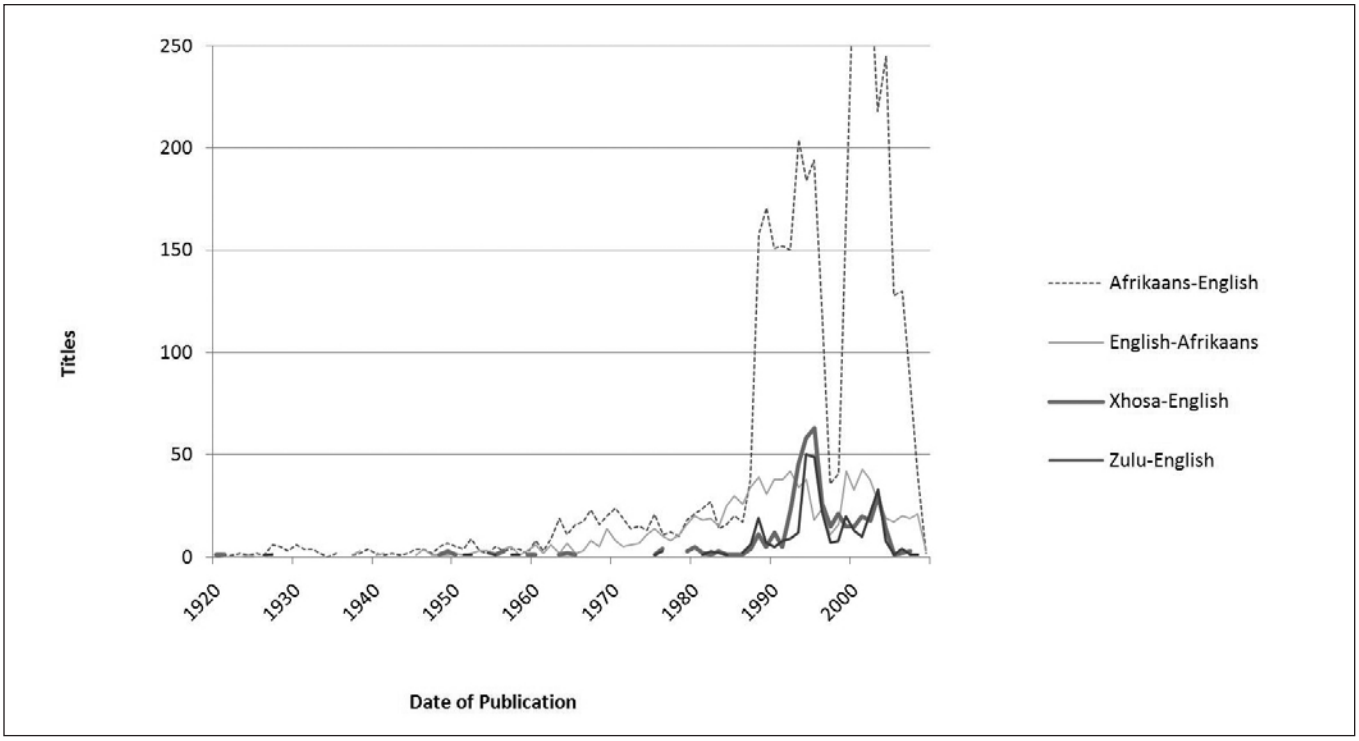

Figure 8. Translated Work from South Africa 1920–2010, as Reflected in WorldCat

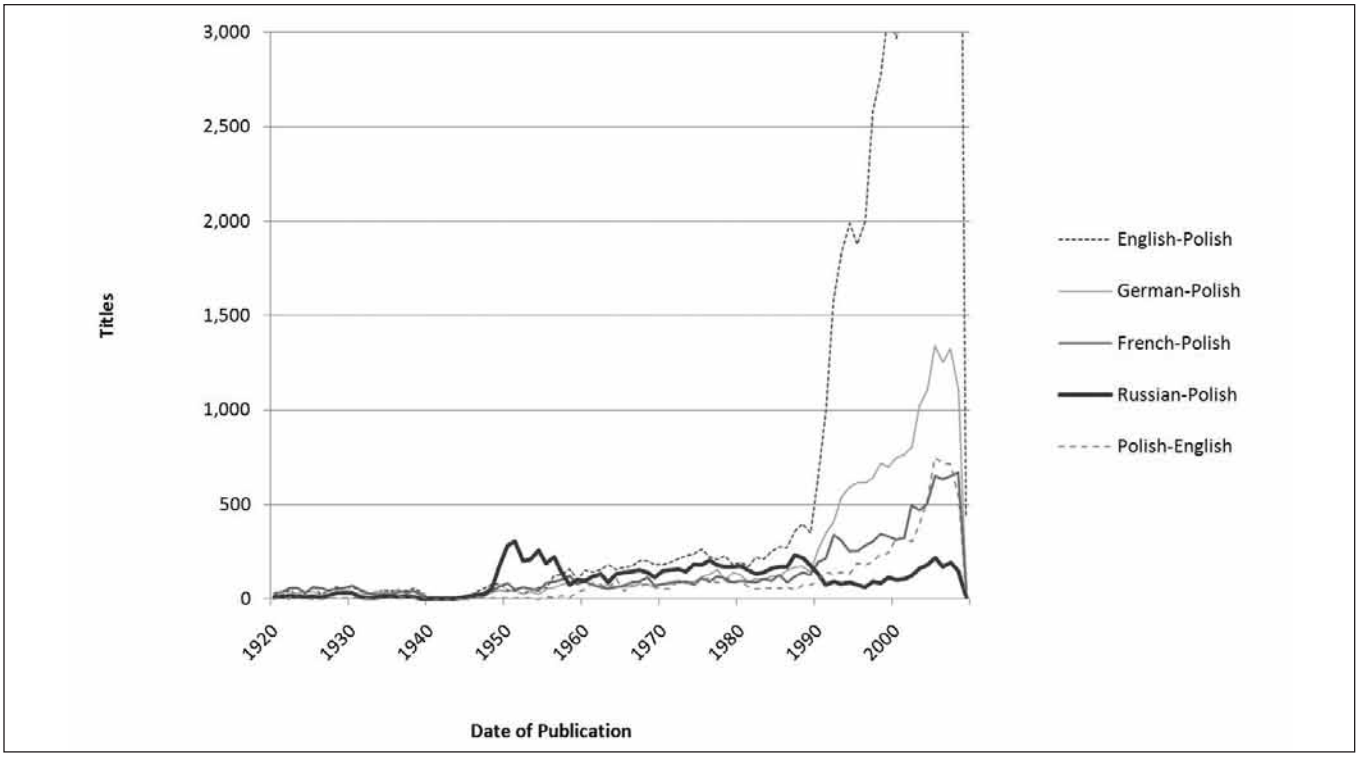

Figure 9. Translated Work from Poland 1920–2010, as Reflected in WorldCat

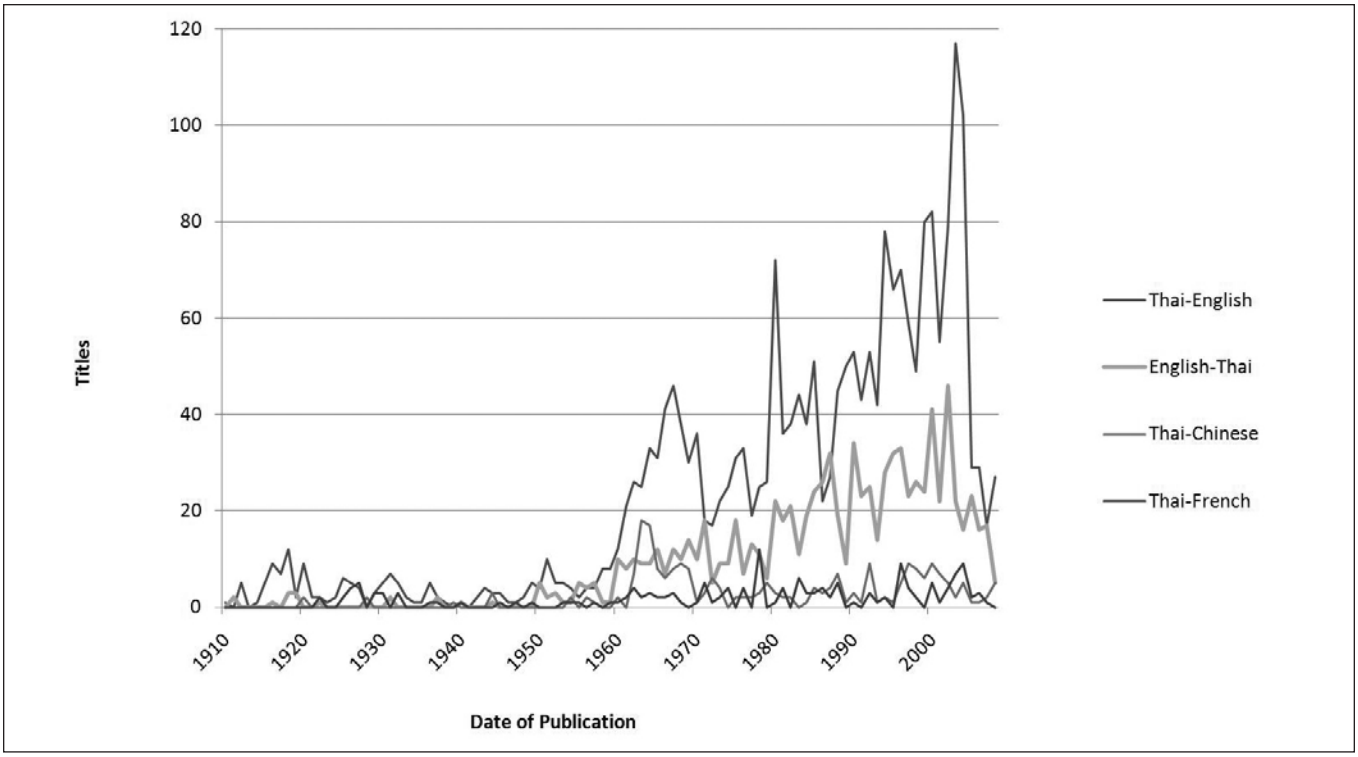

Figure 10. Translated Work from Thailand 1910–2010, as Reflected in WorldCat

libraries within the country should provide a different take on cultural production. How often do the books published as a reflection of the country's culture travel to other countries and end up in the collections of other libraries? What language materials and translations are being collected by others? Second, the potential exists to conduct longitudinal studies on books as an expression of cultural diversity: rechecking these global data every few years (as more libraries join WorldCat) to seek better data and trends of globalization in book collecting. Such longitudinal comparison also could offer UNESCO valuable evidence on progress toward its goals of preserving global linguistic diversity. Finally, as each country's national bibliographies are added more widely into the aggregate collection, the type of national publication profiles explored here can empower, via a data source, specialists to interpret historical trends in a country's literatures or publication history.

Conclusions

The research presented above began as an attempt to compare national publication profiles for countries around the world. Researchers used the techniques of data mining within the aggregate bibliographic database, WorldCat, to extract data for collection analysis, with emphasis on differentiation by country and region. The project successfully extracted data from each country (as their publication record is reflected in WorldCat) and parsed it according to publication patterns, publication languages, and data on translated works in those publication patterns. To demonstrate the validity of this approach, data from six countries were compared in some detail in this paper. Despite the limitations of data mining in a database with roots in Anglo-American traditions, the six countries' case studies appear robust and are sharply delineated from one another.

The data on book publication and library collections from each country, and the evidence of different patterns in language use and translations, offers different portraits of the literary arts in each country. Specifically, these patterns differ in ways that correlate to each country's historical trajectory. Different experiences during the Second World War (as well as earlier conflicts) correspond to different patterns in book production; political movements as different as German nationalism and the end of Apartheid in South Africa produce different patterns of linguistic publication; even global religious upheavals, such as the Protestant Reformation and the Second Vatican Council, may affect the data on book publication. In the countries whose publishing footprint in the WorldCat database is smaller, the data on book translations are weakest, yet even they offer strong indications of the interactions between languages

one might expect. In the case of a more complete national bibliography, as is likely represented by the data mined for Germany, even events deep in the country's history coincide with changes in the mined publication data and the presence of dead languages. This testifies to the importance of library collections as custodians of historical culture. These six case studies present a brief glimpse into the richness of information about library collections that can be mined from the catalog data already available.

References and Notes

1. See, for example, UNESCO, Culture, Cultural Diversity, http://portal.unesco.org/culture/en/ev.php-URL_ID=34321&URL_DO=DO_TOPIC&URL_SECTION=201.html (accessed Nov. 8, 2010). The European Expert Meeting took place on Nov. 7, 2008, as a collaborative effort of publishers, librarians, and booksellers, and as part of Amsterdam's work to become labeled as World Book Capital City.
2. OCLC, WorldCat Facts and Statistics, www.oclc.org/us/en/worldcat/statistics/default.htm (accessed Nov. 8, 2010).
3. OCLC Research, Books as an Expression of Cultural Diversity, www.oclc.org/research/activities/globalbooks/default.htm (accessed Nov. 8, 2010).
4. UNESCO, *Convention on the Protection and Promotion of the Diversity of Cultural Expressions: Paris 20 October 2005* (London: Stationer's Office, 2007), www.unesco.org/new/en/unesco/themes/2005-convention/the-convention/convention-text (accessed Nov. 8, 2010).
5. UNESCO, Programme—Section of Creative Industries for Development, Major Programme 2010–2011 (Project 35 C/5) IV, Culture, http://portal.unesco.org/culture/en/ev.php-URL_ID=35864&URL_DO=DO_TOPIC&URL_SECTION=201.html, (accessed Nov. 8, 2010).
6. Michael Heaney, e-mail to the author, Dec. 4, 2008.
7. IFLA Statistics and Evaluation Section, "Minutes of the First Standing Committee Meeting" (Milan, Italy, Aug. 22, 2009), www.ifla.org/files/statistics-and-evaluation/minutes/august-2009.pdf, (accessed Jan. 22, 2011).
8. UNESCO, *Index Translationum: World Bibliography of Translation* (Paris: UNESCO, 1932–), http://portal.unesco.org/culture/en/ev.php-URL_ID=7810&URL_DO=DO_TOPIC&URL_SECTION=201.html (accessed Nov. 8, 2010).
9. Heaney, e-mail to the author.
10. Liz McKeen, "Canadiana: The National Bibliography for Canada, in the Digital Age," *International Cataloging & Bibliographic Control* 38, no. 2 (Apr. 2009): 19–22; Yeon-Kyoung Chung, "National Bibliographies Past, Present, and Future: The Korean Experience," presentation, World Library and Information Congress: 72nd IFLA General Conference and Council, August 2024, 2006, Seoul, Korea, http://archive.ifla.org/IV/ifla72/papers/109-Chung-en.pdf (accessed April 14, 2011): 1–16; Daniel M. Pennell, "The Fate of Book Chambers National Bibliographies in Belarus, Ukraine, and Moldova since 1991," *Slavic & East European Information Resources* 11, no. 1 (Jan. 2010): 10–20; Anna Katuna Chelidze and Janet Zmroczek, "National Bibliography of Georgia: Past, Present, and Future," *Slavic & East European Information Resources* 11, no. 1 (Jan. 2010): 41–45; Heghine Hakobyan, "National Bibliography in Armenia," *Slavic & East European Information Resources* 11, no. 1 (Jan. 2010): 46–53; Ibrhaim Ramjuan, "National Bibliographic Control in Mauritius: Issues and Challenges," *Information Development* 25, no. 4 (Nov. 2009): 296–303.
11. OCLC Research, WorldMap, www.oclc.org/research/activities/worldmap/default.htm (accessed Nov. 8, 2010). See also OCLC, Global Library Statistics, www.oclc.org/global-librarystats/default.htm (accessed Jan. 22, 2011).
12. A full list of the reference sources consulted for the data in the OCLC WorldMap may be found at OCLC, Global Library Statistics, www.oclc.org/globallibrarystats/sources.htm (accessed Jan. 22, 2011).
13. See OCLC Research, Data Mining Research Area, www.oclc.org/research/activities/past/orprojects/mining/default.htm (accessed Nov. 8, 2010), for a partial bibliography of studies.
14. Brian Lavoie, Lynn Silipigni Connaway, and Edward T. O'Neill, "Mapping WorldCat's Digital Landscape," *Library Resources & Technical Services* 51, no. 2 (2007): 106–15.
15. Lynn Silipigni Connaway and Timothy J. Dickey, "Beyond Data Mining: Delivering the Next Generation of Service from Library Data," part of the panel presentation "Transforming Data into Services: Delivering the Next Generation of User-Oriented Collections and Services," *Proceedings of the American Society for Information Science & Technology* 45, no. 1 (2008): 1062.
16. OCLC, WorldCat Facts and Statistics; Jay Jordan, "OCLC Update Breakfast," presentation, American Library Association Annual Conference, Washington D.C., June 27, 2010, http://vidego.multicastmedia.com/player.php?p=ma17k7dn (accessed Nov. 8, 2010).
17. Benjamin Lee Whorf, "Thinking in Primitive Communities," in *Language, Thought, and Reality*, ed. John B. Carroll, 65–87 (Cambridge, Mass.: MIT Pr., 1964).
18. For a summary of the basic tenets, see Harry Hoijer, "The Sapir-Whorf Hypothesis," in *Language in Culture: Conference on the Interrelations of Language and Other Aspects of Culture*, ed. Harry Hoijer, 92–105 (Chicago: Univ. of Chicago Pr., 1954); for later challenges, see George Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* (Chicago: Univ. of Chicago Pr., 1987); and John J. Gumperz and Steven C. Levinson, eds., *Rethinking Linguistic Relativity* (Cambridge, England: Cambridge Univ. Pr., 1996).
19. A special algorithm was developed to interpret the various ways multiple linguistic context has been encoded in that field's subfield $a and $h, which are both repeatable and which underwent a change in MARC coding instructions in 1981; this necessitated also mining the record creation date from each record.
20. OCLC, Brief History of OCLC Activities with National Libraries outside the U.S., www.oclc.org/us/en/worldcat/catalog/national/timeline/default.htm (accessed Nov. 8, 2010).
21. Claudia Fabian, Bibliographic Control in Germany, *Bollettino AB* 44, no. 1 (Mar. 2004): 18.
22. Ina Fourie and Marlene Burger, "Bibliographic Control in South Africa," *International Cataloging & Bibliographic*

Control 37, no.1 (Jan. 2008): 3–8.
23. For a timeline of national libraries' participation in WorldCat, to date involving forty-one countries worldwide, see OCLC, Brief History of OCLC Activities with National Libraries Outside the U.S., www.oclc.org/us/en/worldcat/catalog/national/timeline/default.htm (accessed Nov. 8, 2010).
24. For a recent historical view of this era of post-Reformation printing history, see Andrew Pettegree, *The Book in the Renaissance* (New Haven, Conn.: Yale University Press, 2010).
25. Izabella Tomljanovich, "Recent Publishing Trends and Developments in Poland," *Slavic & East European Information Resources* 1, no. 1 (2000): 83–96.
26. Peter Lor, "From a Trickle to a Torrent: Bibliographic Control of Books in South Africa, 1796 to 1996," *Mousaion* 23, no. 1 (2005): 19–38.
27. Elizabeth N. Steinhagen, "Leaders of Bibliographic Control: The Chilean Experience," *Cataloging & Classification Quarterly* 31, no. 1 (2000): 3–14.

Appendix A. Language Data from the Six Countries

Country	Publications	Language	Titles	Translations	Language Distribution (%)
Bolivia	58,838	Spanish (Official)*	56,326	529	95.73
		Quechua (Official)	202	46	0.34
		Aymara (Official)	228	28	0.39
		Guarani	28	10	0.05
		Other Amerindian	269	64	0.46
		(English)	887	69	1.51
Chile	265,948	Spanish (Official)	248,935	6,505	93.60
		English	10,965	533	4.12
		German	772	81	0.29
		Mapudungun	127	33	0.05
		(French)	935	55	0.35
Germany	12,843,605	German	10,558,895	317,448	82.21
		Danish	1,712	320	0.01
		Frisian	401	51	0.00
		Sorbian	83	22	0.00
		(English)	580,149	29,073	4.52
		(Latin)	378,698	8,041	2.95
		(French)	92,516	6,919	0.72
Poland	1,225,446	Polish	917,229	150,061	74.85
		German	85,304	5,041	6.96
		Ukrainian	2,105	282	0.17
		Lithuanian	327	81	0.03
		Belarusian	691	92	0.06
		(English)	63,137	10,916	5.15
		(Latin)	23,877	882	1.95
		(Russian)	9,927	912	0.81
South Africa	299,574	English	193,504	2,082	64.59
		Afrikaans	77,449	5,145	25.85
		Zulu	2,649	487	0.88
		Xhosa	2,364	553	0.79
		Sesotho	1,368	298	0.46
		Sepedi	1,344	177	0.45
		Setswana	1,290	154	0.43
		Xitsonga	729	132	0.24
		Venda	670	115	0.22
		Ndebele	167	31	0.06
Thailand	181,003	Thai	129,461	3,806	71.52
		English	44,233	1,514	24.44
		Chinese	1,224	65	0.68
		(French)	991	91	0.55
		(Pali)	920	47	0.51

* Languages whose names are in parentheses are not technically native to the country.

Appendix B. Translation Data from the Six Countries

Country	Publications	Translations	Original	Translation	Number	Language Distribution (%)
Bolivia	58,838	780	English	Spanish	211	27.05
			French	Spanish	61	7.82
			Spanish	English	58	7.44
			German	Spanish	50	6.41
			Aymara	Spanish	32	4.10
Chile	265,948	7,351	English	Spanish	2,571	34.97
			French	Spanish	1,365	18.57
			German	Spanish	575	7.82
			Spanish	English	442	6.01
			Italian	Spanish	410	5.58
Germany	12,843,605	381,141	English	German	149,318	39.18
			French	German	41,076	10.78
			Russian	German	14,254	3.74
			German	English	19,861	5.21
			Latin	German	12,811	3.36
			Italian	German	12,103	3.18
			Swedish	German	12,006	3.15
Poland	1,225,446	174,738	English	Polish	72,638	41.57
			German	Polish	20,626	11.80
			French	Polish	13,778	7.88
			Russian	Polish	9,828	5.62
			Polish	English	8,673	4.96
			Italian	Polish	4,885	2.80
South Africa	299,574	10,063	English	Afrikaans	4,077	40.51
			Afrikaans	English	1,008	10.02
			English	Xhosa	453	4.50
			English	Zulu	357	3.55
			Dutch	Afrikaans	292	2.90
			English	Sesotho	219	2.18
			German	Afrikaans	209	2.08
			English	Tswana	204	2.03
			German	English	165	1.64
			English	Sepedi	109	1.08
			English	Venda	80	0.79
			Zulu	English	54	0.54
			Afrikaans	Zulu	40	0.40
			Xhosa	English	31	0.31
Thailand	181,003	5,700	Thai	English	2,322	40.74
			English	Thai	983	17.25
			Chinese	Thai	237	4.16
			Pali	Thai	178	3.12
			French	Thai	165	2.89