Using OCLC WorldCat Language Indexes to Quantify Slavic and Eurasian Language Collections and Answer Complex Bibliographic Questions

GEOFF HUSIC
University of Kansas Libraries, Lawrence, Kansas, USA

ABSTRACT

The author presents strategies that any librarian or researcher with access to OCLC WorldCat may use to answer many kinds of questions concerning foreign-language materials in their or other libraries. These advanced searching strategies, some not obvious and some even hidden in the WorldCat interface, may also be used to help collection managers generate valuable information about their library's collection strengths in the respective languages, and by extension, make comparisons with collections at other institutions. Special attention is devoted to certain problems associated with Slavic and Eastern European languages.

KEYWORDS: OCLC WorldCat, MARC format, language codes, collection strengths by language, collection development, Slavic and East European languages and literatures

INTRODUCTION

Most academic librarians are familiar with OCLC WorldCat, the world's largest bibliographic database and utility, serving most US and Canadian academic libraries and almost 17,000 libraries in approximately 170 other countries and territories.[1] WorldCat exists in several different "flavors" which offer varying services: 1) a free, public version, Open WorldCat (http://www.worldcat.org), 2) a subscription-based version with additional searching functionality, via the OCLC FirstSearch interface (WorldCat), sometimes also serving in a similar interface as a library's local catalog (branded as WorldCat Local, and 3) a technical-services interface used by libraries primarily for cataloging records for import into local library catalogs, and which currently exists in the PC-client version called *Connexion*. The information presented in this article pertains primarily to the subscription-based WorldCat First-Search version, hence referred to as simply "WorldCat." While information similar to that shown here can be mined by using the Connexion client, the techniques used are somewhat different, and some of the export functionality described below is absent.

There could be a variety of reasons that a librarian might want to research detailed information about how various foreign languages are represented in their library's collections. Some reference questions I have fielded over the years include, for example: "How many Russian books do you have in your collection?" "How many journals in Polish?" How can I find books in Slovenian published in Italy?" or "How many books in both Russian and Tajik?" More elaborate information demands can also be prompted in preparing narratives for grants, such as the Title VI National Resource Center grant, which includes a "Strengths of the Library" section.

Address correspondence to Geoff Husic, MA, MS, Slavic & Near East Studies Librarian, University of Kansas Libraries, Room 519, 1425 Jayhawk Blvd., Lawrence, KS 66045-7544, USA. E-mail: husic@ku.edu

These kinds of question can be very difficult to answer relying solely on a library’s own online catalog, especially without expert knowledge of the underlying structure of catalog bibliographic records. This kind of information can be culled somewhat more efficiently by cataloging staff who have access to the local catalog server and have the ability to create specialized reports based on expert searches, but these individuals may not have the language expertise to understand the nature of some of these more complex linguistic questions.

SCOPE AND LIMITATIONS

In this article, I will present advanced searching strategies that any librarian or even sufficiently motivated patron with access to WorldCat can use to answer many of these kinds of questions. Some of these search options are not obvious, and some are even hidden in the WorldCat interface. I believe these strategies can also be used broadly in a manner of ways to generate valuable information about a library’s collection strengths in the respective languages. WorldCat was never intended to be a collection analysis tool, but rather a bibliographic-discovery one. Nevertheless, with a little ingenuity, much collection information can be extracted by employing the techniques I describe here. What these techniques will **not** be able to do is allow the searcher to quickly compare holdings across libraries or do in-depth subject analysis of entire collections. While comparison is possible, it will require performing the same searches institution by institution. Detailed subject analysis will require identifying specific subject areas and searching based on subject metadata. OCLC offers a separate product, called WorldShare Collection Evaluation, which will allow, with various degrees of success, cross-library comparison of collections and granular subject analysis based on pre-set subject and genre categories.[2] However, some of the caveats I mention below concerning MARC language codes in the underlying data will also apply to data extracted using WorldShare Collection Evaluation. Using these techniques to generate numeric counts of books in a particular language will be yielding information on the number of titles, not volume count, since books and, of course, serials, are often published in more than one volume. Determining actual volume count usually requires also determining how many items are attached to each catalog record, and this usually requires library-systems server access, and is outside of the scope of this article. While the techniques used here can be useful to gather information about any language represented in WorldCat database, I will also be focusing on some issues specific to the Slavic and other Eurasian languages, for example, problems related to the languages formerly referred to collectively as Serbo-Croatian, and confusion arising in labeling Slovenian versus Slovak. Since the functionalities described here are made possible based on how the various WorldCat indexes are configured, these strategies could change at any time if those configurations change.

LITERATURE REVIEW

There have been only a very small number of articles published that either discuss the searching functionality of WorldCat, even in general terms, or mention some of the deficiencies found in the WorldCat data. There does not appear to be any other literature that specifically addresses how to make the most of language-related data and the challenges associated with doing so. In 2011, my colleagues Amalia Monroe (Social Sciences and Collection Assessment Librarian, University of Kansas Libraries) and Lea Currie (Head of Collection Development, University of Kansas Libraries) published an article on the University of Kansas’ experience using the

WorldCat Collection Analysis tool, the predecessor of a newer version called WorldShare Collection Analysis.[3] They had two specific comments about how the tool dealt with language, 1) that WCA did not allow sorting by language, and 2) that several area-studies librarians, including myself, were dissatisfied with the limited number of languages for which the data sets could be filtered. The University of Kansas Libraries subsequently subscribed to the newer WorldShare tool and librarians had the opportunity to review the results in the spring of 2017. It did not make a great deal of sense limiting languages to a few dozen of the most numerically represented in the WorldCat Analysis tool. Materials on the lesser-known languages are also very important for academic libraries and linguistic researchers, and are often what makes a library's collection distinctive. The situation appears to have improved somewhat in the WorldShare product, although I have only had the opportunity to view a few sample reports. The selection of languages is no longer limited, however some cells in the language column of the report spreadsheets are populated with corrupted data such as call numbers, genre headings (e.g. "adult"), and others. These changes should improve the utility of the tool for assessing language and area-studies collections. Some of the strategies discussed below may help others fill in some of the gaps in the WorldShare Data Collection if collection managers have chosen to use that tool. Many of these tasks, with moderate added effort, can be accomplished using the institutional WorldCat subscription rather than the separate WorldShare product, at no additional cost.

I have only identified one article, other that my own article on Romani records in WorldCat (referenced in the discussion or Romani below), which specifically addresses, although somewhat obliquely, problems associated with language information encoded in WorldCat records. In their article on next-generation library catalogs, Susan C. Wynne and Martha J. Hanscom, touch on the topic of the impact of human error in MARC encoding, especially regarding encoding for language. This deficiency is revealed in their interviews with catalogers. They note: "In describing a language code issue, an interviewee who wishes to remain anonymous wrote, 'The code mismatch is caused more by human error than anything else' "[4]

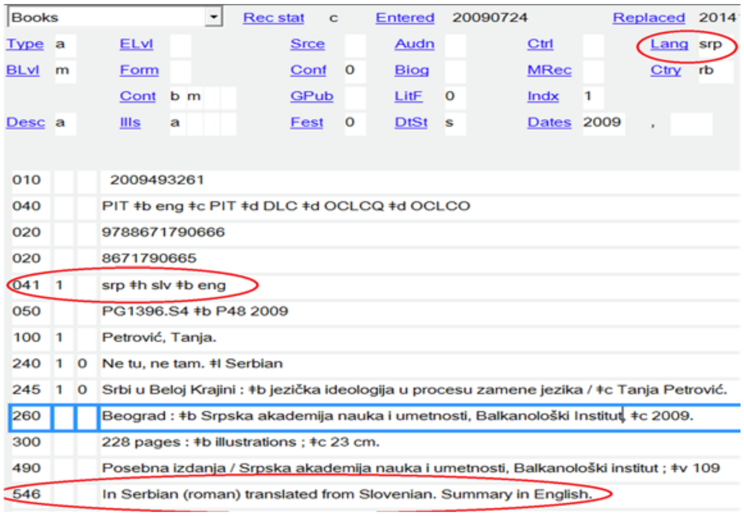
MARC FORMAT AND LANGUAGE CODES

It is beyond the scope of this article to go into elaborate detail about the MARC bibliographic format, but I will simply summarize that this is the specific computer data format in which bibliographic records are cataloged and stored in bibliographic utilities such as WorldCat and library catalogs. The raw MARC data is manipulated by local library systems to present a more user-friendly display in the online catalog. As many librarians, even non-catalogers, will have at least had the experience of viewing an OCLC MARC record, I hope I will be forgiven for glossing over this topic. I will be focusing here primarily on the WorldCat interface, which is more user friendly for non-catalogers than raw OCLC MARC data. However, I will need to briefly explain a few MARC-format data elements before continuing to the topic at hand, how to maximize the utility of the MARC language codes to tease out language-related collection information.

The language codes, which are used to code language-specific information in the MARC records, consist of three letters. In many cases these are based on the first three letters of the English name of the language, e.g. **eng** (English), **ger** (German), **rus** (Russian), etc., while some are based on the vernacular name of the language, e.g. **srp** (Serbian, from *srpski*) and **hrv** (Croatian, from *hrvatski*). Several languages share the first letters of the language name in

English, for example, Slovenian and Slovak, so it was necessary to create different strings to distinguish these codes. Although the lists of language codes are freely available online and need not be memorized, confusion with language codes has created an environment rife for human error in coding MARC records.[5] This will be discussed in more detail after presenting the search strategies.

Information concerning the language of the text of a work may be found in three distinct areas of the MARC record, and is represented by these three-letter language codes: 1) In the required fixed 008 field language code, which is labeled in the Connexion interface with the more user-friendly **Lang**, 2) the 041 field (languages associated with the text, and required if applicable), and 3) sometimes in a 546 field language note (required if applicable, but nevertheless frequently absent).[6] All three are represented in the example below (Figure 1).



Books				Rec stat	c	Entered	20090724			Replaced	2014
Type	a	ELvl		Srce		Audn		Ctrl		Lang	srp
BLvl	m	Form		Conf	0	Biog		MRec		Ctry	rb
		Cont	b m	GPub		LitF	0	Indx	1		
Desc	a	Ills	a	Fest	0	DtSt	s	Dates	2009 ,		

010			2009493261
040			PIT ‡b eng ‡c PIT ‡d DLC ‡d OCLCQ ‡d OCLCO
020			9788671790666
020			8671790665
041	1		srp ‡h slv ‡b eng
050			PG1396.S4 ‡b P48 2009
100	1		Petrović, Tanja.
240	1	0	Ne tu, ne tam. ‡l Serbian
245	1	0	Srbi u Beloj Krajini : ‡b jezička ideologija u procesu zamene jezika / ‡c Tanja Petrović.
260			Beograd : ‡b Srpska akademija nauka i umetnosti, Balkanološki Institut, ‡c 2009.
300			228 pages : ‡b illustrations ; ‡c 23 cm.
490			Posebna izdanja / Srpska akademija nauka i umetnosti, Balkanološki institut ; ‡v 109
546			In Serbian (roman) translated from Slovenian. Summary in English.

*Figure 1: OCLC Connexion MARC record[7]

The 008 field can accommodate only one language code, and is intended to reflect the primary language of the text.[8] The 041 field can accommodate codes for any number of languages also represented in the text, and has several secondary value tags that can indicate whether the associated language is substantive, supplementary (such as summaries), or the original language of a translated work. If present, the 041 field will also repeat the language code found in the 008 field, i.e. the primary language.[9] The 546 field language note is usually present when some further clarification about the text is useful, such as a language appearing in an atypical script, for example, Bosnian in Arabic script, Romanian in Church Slavic script, in a non-standard dialect, or there is some other situation that cannot be explained with the codes alone. The nature of the script of a text is not always obvious from the catalog record, as catalog records are based on the transliterated form of non-Latin script languages, with vernacular script information being optional. The transliteration scheme used for Serbian in Cyrillic, for example, is often difficult to distinguish from Croatian especially for the non-specialist, so the 546 language note note is helpful for those who can only read one script or the other to select books in their preferred script.

Before the creation of three distinct language codes for Bosnian, Croatian, and Serbian (BCS) in 2009[10] to replace the earlier designation “Serbo-Croatian,” it was common practice among Slavic catalogers to create a language note to specify the script of the text: “Serbo-

Croatian (Cyrillic)," which was in most cases Serbian, Montenegrin, or Bosnian in Cyrillic, and "Serbo-Croatian (roman)" which could be either Serbian in roman script[11], Bosnian in roman script, or Croatian.[12] There were only two language codes available before 2009 to cover these three languages, **scc** and **scr** respectively, and this has presented a number of new problems down the line that I will discuss further below. In current practice however, because of the availability of a distinct code for each of the three languages, the 546 language notes for these languages is most typically found in association with Serbian works, which are printed in Latin script (546 Serbian ǂb Latin) almost as often as in Cyrillic (546 Serbian ǂb Cyrillic).[13] Fortunately, BCS is the only group of modern Slavic languages that have such pronounced script complications, although similar complexities also occur in Romanian and several of the Central Asian languages that have had turbulent script histories.[14] While these 546 language notes are very useful to the catalog user, they are less reliable for data retrieval than 008 and 041 fields mentioned above, as they are, in many cases, not standardized.

OCLC WORLDCAT (VIA FIRSTSEARCH)

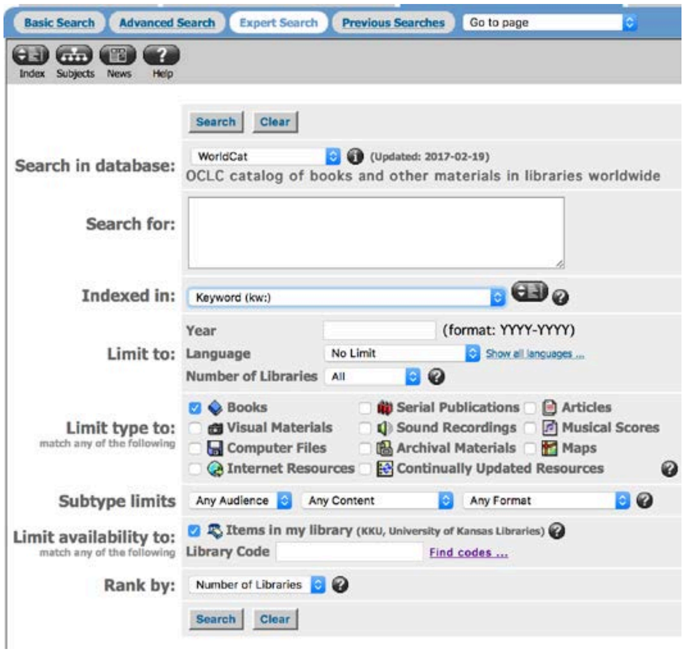
The OCLC WorldCat database user interface is quite user friendly and offers three different search screens targeted at different users, basic, advanced, and expert, each adding progressively more options. While the basic and advanced search screens are perfectly adequate for most users, we will focus on the expert screen, which is the only search that can incorporate the strategies shown here. There is not a great deal of direction available on this search page to guide the user, and it assumes that the user will be aware of the basics of Boolean searching. There are however several external documents that go into much greater detail about how the indexes of WorldCat, and the other OCLC versions such as Connexion, can be searched.[15] While this OCLC documentation is very helpful, it is also very technical and combines information about the various WorldCat products besides the FirstSearch interface. What makes the documentation especially unwieldy is that it presupposes an expert knowledge of the MARC format. The audience for this technical documentation is therefore the cataloging specialist and not the casual user of the WorldCat search interfaces such as FirstSearch. I believe even a reference specialist without cataloging experience will likely have difficulty parsing this information on WorldCat indexes and its more sophisticated search capabilities.

We will be using the Expert Search screen for all our searches. This screen is the one highlighted in the white tab at the top of the screen in Figure 2 below. Before formulating our search strategies, it is helpful to understand the distinction between an "index search" and a "search limit," both of which are possible options on this screen. A search limit must be used with at least one index label entered in the main search window. The search limits options can be seen in the "Limit type to" section in Figure 2. Index label searches, on the other hand, can be entered independently. While the language of a work can be chosen as an option in the "Limit to" section, it will also require at least one additional search term. This is not evident to the user from on the interface and has limitations explained below. I recommend manually entering all index terms as shown in the examples.

Searches can be entered in the "Search for" box by pulling down the desired index search in the "Index section." The search index labels can also be entered manually in the "Search for" box and combined using the usual Boolean operators. This latter is the strategy we will use.

For the purposes of this demonstration, I will say I am interested in printed monographs in my library's collections. I will therefore check the "Books" box, and I will enter my libraries

OCLC code, KKU (University of Kansas), to limit results to our library's holdings.[16] Others may wish to do a study of serials, maps, or others formats in the same manner. Some of the "Subtype limits," such as fiction, may also be of interest, to distinguish, say, strengths of literature in various languages, or to compare fiction versus non-fiction holdings.



Search in database: WorldCat (Updated: 2017-02-19)
OCLC catalog of books and other materials in libraries worldwide

Search for:

Indexed in: Keyword (kw:)

Limit to: Year (format: YYYY-YYYY)
Language: No Limit — Show all languages ...
Number of Libraries: All

Limit type to: (match any of the following)
☑ Books □ Serial Publications □ Articles
□ Visual Materials □ Sound Recordings □ Musical Scores
□ Computer Files □ Archival Materials □ Maps
□ Internet Resources □ Continually Updated Resources

Subtype limits: Any Audience | Any Content | Any Format

Limit availability to: (match any of the following)
☑ Items in my library (KKU, University of Kansas Libraries)
Library Code — Find codes ...

Rank by: Number of Libraries

***FIGURE 2: WORLDCAT EXPERT SEARCH SCREEN**

When pulling down index options in the "Indexed in" section, the user will find a label that appears as **Language phrase (ln=)**.

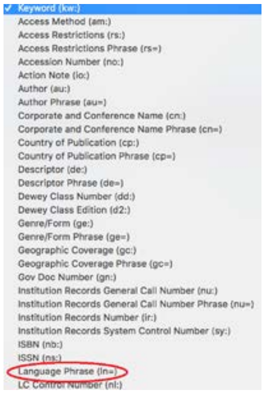

✓ Keyword (kw:)
Access Method (am:)
Access Restrictions (rs:)
Access Restrictions Phrase (rs=)
Accession Number (no:)
Action Note (io:)
Author (au:)
Author Phrase (au=)
Corporate and Conference Name (cn:)
Corporate and Conference Name Phrase (cn=)
Country of Publication (cp:)
Country of Publication Phrase (cp=)
Descriptor (de:)
Descriptor Phrase (de=)
Dewey Class Number (dd:)
Dewey Class Edition (d2:)
Genre/Form (ge:)
Genre/Form Phrase (ge=)
Geographic Coverage (gc:)
Geographic Coverage Phrase (gc=)
Gov Doc Number (gn:)
Institution Records General Call Number (nu:)
Institution Records General Call Number Phrase (nu=)
Institution Records Number (ir:)
Institution Records System Control Number (sy:)
ISBN (nb:)
ISSN (ns:)
Language Phrase (ln=)
LC Control Number (nl:)

***FIGURE 3: "INDEXED IN" PULL-DOWN MENU**

This is the only language index which is evident in the WorldCat expert search interface, but this is not the best index for our purposes. There is another, arguably much more powerful, index that we can use to achieve the goals outlined here much more efficiently. The existence of this label is documented only in the external documentation I noted above. The difference between the two labels is significant. Above I mentioned the MARC format fields 008 (primary language) and

041 (all associated languages). The WorldCat label **ln=** will search for language codes in both the mandatory 008 field and most subfields of the 041 field (if present), including the language of supplemental text in the work, such as abstracts, summaries, added tables of contents, which occur in the subfield ǂb. It will not, however, search for the original language of a translated work, which is encoded in the 041ǂh field.[17] However, there is a second "hidden" language index, **la=**, which is included in the OCLC documentation but does not display in the pull-down menu of the WorldCat "Indexed in" section. The search label **la=** will search **only** the 008 field, i.e. the primary language of the work. As many foreign language materials have summaries, tables of contents, or abstracts in multiple languages but with primary text in only one, this ability to limit one's search to the primary language is much more useful when trying to generate information on one's collection strengths based on language.[18]

When searching for multilingual items, say, books in both Russian and Tajik, then it is safer to use the more general search index **ln=** for both languages, since the user will not necessarily know which is coded in the MARC record as the primary language. Choosing the wrong language for the exclusionary **la=** search may lead to a failed result. In addition, any search combining two or more **la=** searches will return no results, as only one primary language can be coded in the MARC record 008 field. Table 1 summarizes how the two language indexes can be used and combined.

TABLE 1: SUMMARY OF LANGUAGE SEARCHES

ln=ukr	Will find items in Ukrainian in primary text or supplementary text such as summaries, abstracts, added tables of contents, etc. Will **not** search original language of a translation (unless original is included with the translation).
la=ukr	Will find items primarily in Ukrainian, but not items where Ukrainian is only the supplementary language.
la=ukr and **ln**=rus	Will find items primarily in Ukrainian, with supplementary text in Russian. Any number of additional languages can be added as long as using **ln=** with each.
la=ukr and **la**=rus	This search will **not** return results, since only one **la=** is possible per search.

The **la=** search strategy can be used to determine numerical strengths of the languages represented in the library's collections. These language searches can be combined with any number of the other WorldCat indexes to answer specific questions or even discover problems within the data. Two of the most useful indexes, in my experience, have been the index **cp:** (country of publication) and **pl:** (place of publication). These can be very helpful in identifying émigré publications in specific countries. The code **cp:** will search the two-letter MARC country code entered in the **Ctry** field in Connexion (see *Figure 1*).[19] On the other hand, the information searched by the index label **pl:** is not standardized. This will search the text in the 260/264 fields (place, publisher, date of publication).[20] In most cases, the text will include the city of publication solely, which is entered as it appears on the book and will not necessarily be consistent from book to book. For example, most Slovenian-language books from Italy are published in Trieste, which usually appears on the books and thus in the 260/264 field as "Trst," the Slovenian form of the city name rather than the Italian "Trieste." Some care must be taken when searching by city to accommodate these idiosyncrasies.

The WorldCat index **nt:** may also be useful for gathering information about a language or dialect that cannot be otherwise gathered by language codes. Some languages that have literary traditions in dialects other than the current official standard language will sometimes have such information added in the language note. The index **nt:** will search the catalog record note fields as a keyword. These notes are not standardized, so results will be based on what information the original cataloger has chosen to include. Please see the final example in "Additional Sample Searches" below.

Successful WorldCat searches will return both the number of records, which meet the search criteria, as well as the records themselves. In the following example, I am interested in books published primarily in the Albanian language in the five countries with the predominance of speakers, i.e. Albania, Kosovo, Macedonia, Montenegro, and Italy, since the year 2000. I am also interested in how many of each the University of Kansas Libraries owns. In addition to the search strings below, I have limited the year from 2000 to present (**2000-**). At the time of writing I obtained the following results:

Country	**Search**	**Held in WorldCat**	**Held in KU**
Albania	la=alb and cp:aa	11677	170
Kosovo Serbia	**la=alb and cp:kv** **la=alb and cp:yu**	1439 205 =1644	21 11 =32
Macedonia	**la=alb and cp:xn**	903	19
Italy	**la=alb and cp:it**	109	0

The alert reader may notice that there is a complication with searching for country code for Kosovo (**kv**), and s/he will be correct. Until April 2007, Kosovo was considered as part of "rump" Yugoslavia, then consisting only of Serbia and Montenegro (country code **yu**). From February 2007-May 2008, Kosovo was part of Serbia (country code **rb**). For an accurate count in this example, I have searched Serbia as well, and add the results, even if potentially a few of these might be books published in Serbia proper, rather than Kosovo.

In my introduction, I mentioned a real-life reference query about Slovenian-language books published in Italy, among others. Below are examples of how these queries can be answered by formulating the following Boolean searches in the WorldCat expert search box shown in *Figure 2.* The quote marks in the examples are not entered. Please note that all searches involving non-Latin script languages should be formed using the official American Library Association/Library of Congress Romanization schemes[21]. Transliteration based on these schemes must be used, because many pre-2005 OCLC records lack the optional Slavic Cyrillic fields. Non-Slavic-language Cyrillic scripts, e.g. Tajik, and many other non-Latin scripts became available in WorldCat only much more recently, in July 2016. As mentioned above, you may limit the search to your library's own holdings if you wish, search another library, or all WorldCat libraries' holdings, depending on one's intent. Date ranges can be specified as well to narrow chronological coverage.

Please note that while it is possible to search the labels **la=** and **ln=** using the English name of the language instead of the three-letter language code, i.e. one can search **la=Tajik** or **la=tgk**, I do not recommend this. Many language names have competing forms in English, e.g. Uyghur, Uygur, Uighur, or Kirghiz, Kyrgyz, etc., and choosing the wrong form will result in a failed search. In at least one case, the searching-label name is incorrect- Romani, when spelled out, i.e. **la=Romani**, does not return any results. The older name "Romany" must be used, even though the language-code list correctly has the newer accepted spelling "Romani."

COMPARING RESULTS ACROSS LIBRARIES

The index **li:** (holding library), can be combined with any number of additional holding library codes to show the overlap of items owned, and can be combined with any of the other indexes. For example, searching **li:kku and li:dlc** will show all overlap in holdings between University of Kansas Libraries and the Library of Congress. **li:kku and li:dlc and la=rus** will likewise show all items primarily in Russian owned in common.

ADDITIONAL SAMPLE SEARCHES

The following examples will illustrate the search strategies shown above. "Books" will be checked in the Expert Search window, except for Polish journal example:

- Slovenian books published in Italy: "**la=slv and cp:it"**
- Russian books published in Warsaw. This will be most successful if we use two searches and combine the results, since the city name may appear in the book in the Polish form or the Russian: "**la=rus and pl:Varshava"** and the additional search "**la=rus and Warszawa**"
- Number of Polish journals: "**la=pol and cp:pl"** ("Serial Publications" rather than "Books" will be checked in the Expert Search window)
- Books in both Russian and Tajik: "**ln=rus and ln=tgk**"
- Books in Korean published in Kazakhstan: "**la=kor and cp:kz**"
- Russian books with keyword "upyr'" (i.e. vampire) anywhere in the record: "**la=rus and kw:upyr*"** The asterisk is the wildcard and should be used since the noun may have case endings. While this search can also be done in the Advanced Search, limiting by language, as noted above, the Advanced Search will search the **ln=** index (including supplementary text) rather than the **la=** index (primary language only).
- Books in Polish that Library of Congress and University of Kansas own in common: **la=pol and li:dlc and li:kku**

- Fictional works in Czech: "**la=cze and mt:fic**"
- Books in the Croatian Kajkavian dialect (where this has been noted in the language note: "**la=hrv and nt:Kajkav***" I have truncated this search with the asterisk because the dialect name was found in notes in a variety of ways, e.g. Kajkavian, Kajkav, Kajkavščina, etc.[22]

REPURPOSING AND MIGRATING SEARCH RESULTS

When the user has retrieved a successful set of records there are several options available for saving and/or repurposing the data, depending on the size of the data set that results. For example, searching University of Kansas Libraries for the books in primarily Russian, returns a result of 109,376 as of May 2017. Since the WorldCat interface only allows users to mark ten records per page until a maximum of 100 is reached per download, it is not practical exporting very large sets of results. However record sets of a few hundred items can be exported fairly quickly. Even if exporting large sets in not an option, the number of books matching the criteria is nevertheless useful in itself for quantifying collection strengths by language, especially when comparing holdings with those of other libraries.

There are several options available for exporting results. Export filters for the citation-software products Endnote and RefWorks can be accessed by choosing the Export button after selecting one or more records. I personally prefer using the free citation tool Zotero to import selected records directly into a Zotero library of my choosing.[23] Zotero exists in a client version as well as a browser plugin available for most modern browsers. It seems to be currently most up-to-date in the Firefox version. Once imported into Endnote or Zotero, the records can be repurposed for lists of items by language, subject bibliographies, etc. With several extra steps and other software tools, data can also be converted to spreadsheet format. Those interested in this option are welcome to contact the author for details, as it is beyond the scope of this article.

PROBLEM AREAS

Now that I have demonstrated how to search by primary language in WorldCat with the Expert Search, I need to note some areas in which the user will need to exercise caution. These caveats pertain primarily to incorrect language codes in the OCLC MARC records. Trying to accurately quantify library holding for these languages will require some workarounds.

In their attempt to accommodate Croatian and Serbian national libraries after the breakup of Yugoslavia and the subsequent nationalist desire to rebrand Serbian, Croatian, and Bosnian as distinct languages, OCLC made several errors when bulk reassigning the former two codes that covered Serbo-Croatian in WorldCat. I believe this error was based on a fundamental misunderstanding of the language and script issues involved. In brief, what ultimately occurred was that records that were previously coded as **scc** (Serbo-Croatian in Cyrillic) were reassigned with the code **srp** (Serbian). Records with the code **scr** (Serbo-Croatian in Latin script), were reassigned with the code **hrv** (Croatian), regardless of whether these were actually texts in Croatian or Serbian in Latin script.[24] It could certainly not have been the intent of the Serbian National Library to have thousands of Serbian-language books to be suddenly relabeled as Croatian. I do not wish to relitigate the missteps that were made when these bulk changes were implemented. OCLC was making an honest effort to deal with a complex situation and was attempting to accommodate all stakeholders. However, in light of these bulk changes, when trying to quantify books in these languages, I would suggest considering Bosnian/Serbian/Croatian/Montenegrin as a linguistic whole and adding the results together. If it is necessary to segregate the results into the three languages, then combining the language search (**la=hrv**) with the country of publication Serbia (**cp:rb**) should lead to a fairly accurate result in separating out Serbian books in Latin script that were mislabeled as Croatian. Bosnian is a much more complicated issue. It is difficult to reliably segregate Bosnian from Croatian or Serbian, and as a result, language coding on records for works published in Bosnia tend to be unpredictable. Please see the relevant section in the *Slavic Cataloging Manual* for more information on Bosnian and BCS in general.[25] In a related, interesting sociolinguistic twist, Montenegrin is now considered a distinct language according to the 2007 Constitution of Montenegro. However, neither OCLC nor the Library of Congress yet recognize it, and there is no separate language code available for Montenegrin as of the time of writing.[26] For now, it continues to be labeled as **srp** (Serbian).

In addition to above example of mistaken data based on bulk changes, there are a small number of language codes that are frequently entered incorrectly by catalogers due to confusion or lack of knowledge of the languages involved. One notorious example, to which I have devoted an entire article, is the problem of confusion between Romanian and Romani.[27] I will just briefly summarize that in a large number of OCLC records for Romanian-language materials, the incorrect language code **rom** is frequently entered. However, **rom** is, in fact, the language code for Romani, an unrelated language. The source of the confusion is obvious. Most catalogers are more likely to come across books in Romanian than in Romani. Many people are not familiar with the older English spelling for Romanian, which was “Rumanian.” The proper language code **rum** was established based on this older spelling. Several years ago, I undertook a project to correct approximately 1600 of these erroneous codes in OCLC, and I occasionally check for new ones, however more mistaken codes appear frequently as new records are added.

An error that is more likely to be encountered by Slavic-language catalogers is confusion between Slovenian (language code **slv**) and Slovak (language code **slo**). In preparation for this article, I created a WorldCat query using the strategies above to confirm this significant problem in the database. A search for Slovak language (**la=slo)** and place of publication Slovenia (**cp:xv**) returned over 1700 results. This cannot be correct. While it is possible there has been a modest number of books in Slovak published in Slovenia, a result of 1700 can only indicate error in language coding, as I indeed verified by scanning through the records. The opposite problem, books from Slovakia misidentified as in Slovenian (**la=slv and cp:xo**), was smaller, only 515. While I have corrected the language codes in OCLC master records for books that University of Kansas Libraries own, I do not personally have the time to correct all the others in WorldCat.[28] One final language, which at first glance might seem problematic, turns out to be somewhat less so. During Soviet times, the official language of Moldova was called “Moldovan,” i.e. “limba moldovenească.” In 2013, the independent Moldova changed the name of its official language to Romanian, as the prior distinction was wholly artificial. However even during Soviet times, the language code **rum** used for both Romanian and Moldovan. Therefore, any search for **rum** will retrieve items in this language from either Romania or Moldova. The only complication for a user is that Romanian language books from Moldova prior to independence from the Soviet Union will be in the Cyrillic alphabet, rather than the Latin script now used. There is no easy way to separate out Cyrillic versus Latin-script Romanian in WorldCat. Some WorldCat records for Romanian in Cyrillic added after 2005 will have additional optional Cyrillic fields, and some older records have been manually updated with Cyrillic information by catalogers in the course of cataloging books based on these records. Searching for script information in a language note is not reliable since these notes are often lacking. Searching by script in which a language is printed is possible in some cases, however a script search will only identify those in which the vernacular information has been added, not how many books were actuallyß printed in that script. See “Searching by script” below how to identify these records with caveats. When viewing record by record, however, anyone with knowledge of Romanian will be able to recognize transliterated information in, for example, the title of the book, since the transliteration used for Cyrillic Romanian looks very different from Romanian in its conventional script (e.g. Romanian uses the letter “c” for the velar consonant in native Romanian words and not the letter ‘k”, which is how this common letter is transliterated from the Cyrillic letter “к.”

SEARCHING BY SCRIPT

There is a search index within WorldCat that will allow a user to search for catalog records based on the presence of non-Latin script information in the record. As I stressed above in the Moldavian example, this will only be useful in those cases where catalogers have added these non-Latin fields. Non-Latin script information is not required by the cataloging rules, and the primary indexing is based on the Latin-script information in the cataloging record. Some libraries chose to not use them at all. I mention this index here as it may be of potential benefit in the context of languages with multiple scripts. The index label **vp:** can be searched with the standardized abbreviation for the name of the script, e.g. **vp:Cyrl** (for Cyrillic).[29]

CONCLUSION

OCLC WorldCat data, while not perfect, as illustrated with some of the caveats above, provides a wealth of data that can be mined to help library collection managers quantify their Slavic and Eurasian language collections. I have demonstrated that by using a variety of search parameters, managers can generate valuable information regarding the relative strengths of the languages represented in their own, as well as other libraries’ collections. In addition, information can be gathered to answer very specific bibliographic questions that may be impossible to gather merely using a library’s own online catalog, at least without the assistance of a systems specialist. In those areas where these strategies do detect errors in the data, the searcher will have several options. If important, these errors can be reported to the library’s systems managers to correct in the local catalog data. If the library’s cataloging department is willing to correct OCLC master records, these corrections will benefit all other users of the WorldCat database. However, unless the library has a WorldCat Local installation, this latter step will not in itself correct the library’s local data automatically. I have also found OCLC very receptive to suggestions for correcting problems, and bringing more complex problems to their attention will generally be met with a positive response. In many cases, however, I feel that language-specialist catalogers in academic institutions will have a better understanding of the problems than will OCLC. Language specialists should be encouraged to correct language problems found in WorldCat data if their workflows and time permit.

APPENDIX 1: Slavic, East European, and Central Asian Language Codes

Albanian (**alb**), Armenian (**arm**), Belarusian (**bel**), Bulgarian (**bul**), Church Slavic (**chu**), Croatian (**hrv**), Czech (**cze**), Estonian (**est**), Georgian (**geo**), Hungarian, Kazakh, Kyrgyz (**kir**), Latvian, Lithuanian, Macedonian (**mac**), Polish (**pol**), *Romani (**rom)**, *Romanian (**rum**), Russian, Serbian (**srp**), *Slovak (**slo**), *Slovenian (**slv**), Tajik (**tgk**), Turkish (**tur**), Turkmen **tuk**), Ukrainian (**ukr**), Uzbek (**uzb**).
Languages for which there is frequent confusion or error are marked with *
For other languages see: "MARC Code list for languages": (https://www.loc.gov/marc/languages/langhome.html)

APPENDIX 2: East European, Caucasus, and Central Asia Country Codes

Albania (**aa**), ‡Armenia Republic (**ai**), ‡Azerbaijan (**aj**), ‡Belarus (**bw**), *Bosnia and Herzegovina (**bn**), Bulgaria (**bu**), Croatia (**ci**), †Czech Republic (**xr**), †Czechoslovakia (**cs**), ‡Estonia (**er**), ‡Georgia Republic (**gs**), Hungary (**hu**), ‡Kazakhstan (**kz**), *Kosovo Republic (**kv**), ‡Kyrgyzstan (**kg**), ‡Latvia (**lv**), ‡Lithuania (**li**), *Macedonia Republic, (**xn**), ‡Moldova (**mv**), *Montenegro (**mo**), Romania (**rm**), ‡Russia Federation (**ru**), Serbia (**rb**), *Serbia and Montenegro (**yu**), †Slovakia (**xo**), *Slovenia (**xv**), Ukraine (**un**), ‡Tajikistan (**ta**), ‡Turkmenistan (**tk**), *Yugolsavia (**yu**), ‡Uzbekistan (**uz**).

For other countries see: http://www.loc.gov/marc/countries/countries_name.html

The country codes below have undergone changes in the late 20th century due to political changes:

†Former Czechoslovakia:

Czech Republic (Coded **cs** for Czechoslovakia before May 1993)
Czechoslovakia (Coded **cs** for Czechoslovakia before May 1993)
Slovakia (Coded **cs** for Czechoslovakia before May 1993)

*****Former Yugoslavia:

Croatia (Coded **yu** for Yugoslavia before Oct. 1992)
Kosovo Republic (Coded **rb** for Serbia from February 2007-May 2008. From 1992-April 2007, coded **yu** for Serbia and Montenegro
Macedonia Republic (Coded **yu** for Yugoslavia before Oct. 1992)
Montenegro (Coded **yu** for Serbia and Montenegro from 1992-April 2007)
Serbia (Coded **yu** for Serbia and Montenegro from 1992-April 2007)
Slovenia (Coded **yu** for Yugoslavia before Oct. 1992)

APPENDIX 3: Summary of other useful WorldCat search indexes

Country of publication	**cp:**	**cp:un** Searches for works published in Ukraine	Searches the standardized Country of Publication code.
Geographic coverage	**gc:**	**gc:Crimea** Searches for the geographic topic “Crimea” in any part of a subject heading.	This will search for geographic aspects of the topic found anywhere in any of the subject headings. Library of Congress subject headings will be standardized based on the English name of place names.
Keyword	**kw:**	**kw:bosansk*** Searches for the truncated string “bosansk” anywhere in the record	This is the most general possible search available.
Holding library	**li:**	**li:dlc** Searches items held by Library of Congress **li:dlc and li:kku** This will show works owned in common by Library of Congress and University of Kansas	More than one library can be searched to show overlap. See Note 16 for holding-library codes. More than one can be searched at the same time.
Place of publication (usually)	**pl:**	**pl:Sverdlovsk*** Searches for “Sverdlovsk” in MARC 260/264 fields	Place of publication will typically be the city where the book is published. Best to truncate since places often appear with preposition and case endings. These are not standardized and are entered as they appear in the work.

Material type	**mt:**	**mt:fic** Searches for works of fiction	Useful for separating out fiction from non-fiction.
Notes	**nt:**	**nt:Cyrillic** Searches for string “Cyrillic” in notes	Will search several kinds of notes including language notes. May be helpful in distinguishing script in languages with multiple scripts.
Subject	**su:**	**su:Romanies** Searches for subject term “Romanies” **su:“Russian language”** Searches for phrase “Russian language” in subject headings	Will search any subject fields in the record. Queries of more than one word will require quotes as in the second example.
Year of publication	**yr:**	**yr:2013** **yr:1900-2000** **yr:2000-**	Will search the year of publication. A range of years can also be specified.

NOTES AND REFERENCES

[*] Image permissions: The screenshots included in this article are ©2017 OCLC Online Computer Library Center, Inc. and are used with OCLC's permission. WorldCat, FirstSearch, and Connexion are registered trademarks/service marks of OCLC.

[1] “Worldcat”: http://www.oclc.org/en/about.html (accessed 4/5/2017)
[2] “WorldShare Collection Evaluation”: http://www.oclc.org/en/collection-evaluation.html (accessed 4/11/2017)
[3] Monroe-Gulick, Amalia, and Lea Currie. “Using the WorldCat Collection Analysis Tool: Experiences from the University of Kansas Libraries.” *Collection Management* 36, no. 4 (October 1, 2011): 203–16. doi:10.1080/01462679.2011.604907.

[4] Wynne, Susan C., and Martha J. Hanscom. “The Effect of Next-Generation Catalogs on Catalogers and Cataloging Functions in Academic Libraries.” *Cataloging & Classification Quarterly* 49, no. 3 (February 28, 2011): 188. doi:10.1080/01639374.2011.559899.

[5] “MARC Code List for Languages”: http://www.loc.gov/marc/languages/ (accessed 04/19/2017)

[6] The information entered into the 546 note may be either formally tagged or can use free text, depending on the complexity of the information. In any case, it can generally not be used to reliably extract information through expert searches. For more detail see http://sites.google.com/site/seesscm/languages-in-multiple-scripts.

[7] The encoding shows that this work is a text in Serbian (041 ǂa srp), translated from the original Slovenian (041 ǂh slv), and includes an English summary (041 ǂb eng). The 546 language note indicates that the Serbian text in this case is in roman, i.e. Latin script. Translation information is also often repeated in the 546 field for the convenience of the catalog user as it will display on the catalog record.

[8] In items that are bilingual, the code will generally be assigned to prefer the language of the place of publication, or, if not clear, the first title in order of presentation in the piece.

[9] A fourth location in which MARC language codes can occur is in the subfield ǂb of the 040 field. This language code has nothing to do with the language of the text, per se, but rather indicates the language of description used by the cataloging agency, so it is not pertinent to this discussion. In almost all cases, for North American and British libraries this value will be **eng** (English).

[10] “Changes to MARC Code Lists for Languages”: “http://www.loc.gov/marc/languages/languagechg.html (accessed 04/18/2017)

[11] Prior to the current RDA cataloging code, the term “roman script” was used to describe languages using the Latin alphabet. The preferred designation under RDA is now “Latin script.”

[12] Language information will often be found, especially in older MARC records, in the general 500 notes.

[13] The character ǂ, always followed by another letter or number, is used to separate data elements occurring within one MARC field.

[14] Husic, Geoff. “Orthographic Reforms in the Former Soviet Union.” *Slavic & East European Information Resources* 2, no. 2 (February 28, 2002): 38. doi:10.1300/J167v02n02_07.

[15] “Searching WorldCat Indexes Online Help - Searching WorldCat Indexes.” http://www.oclc.org/support/help/SearchingWorldCatIndexes/Default.htm (accessed 04/12/2017) and “Searching WorldCat Indexes Online Help- Language” http://www.oclc.org/support/help/SearchingWorldCatIndexes/Default.htm#04_Indexes/Language.htm%3FTocPath%3DIndexes%7C_____41 (accessed 04/12/2017).

[16] The library code, if unknown, can be found using by clicking “Find codes.” University of Kansas Libraries’ Special Collections holdings have a separate code from the general libraries. Other libraries may also have more than one OCLC symbol for libraries throughout their systems. A list is also available here: http://www.loc.gov/marc/organizations/orgshome.html

[17] According to the OCLC documentation, it will also search the 377 field (associated language), however I do not believe I have ever seen this field in a MARC record.

[18] http://www.oclc.org/support/help/SearchingWorldCatIndexes/Default.htm#04_Indexes/Language.htm#kanchor445 (accessed 04/12/2017).

[19] The complete list of MARC country codes can be found at: http://www.loc.gov/marc/countries/countries_name.html#i

[20] The 264 field has recently replaced the 260 field for this function. The 260 will be still be found in many pre-2016 MARC records.

[21] "ALA-LC Romanization Tables": http://www.loc.gov/catdir/cpso/roman.html (accessed 4/20/2017)

[22] An example of a language note from which this keyword was extracted is: "In Croatian (Kajkavian dialect) and Latin."

[23] Zotero website: https://www.zotero.org (accessed 04/28/2017)

[24] It is very easy to tease out examples of these errors by searching WorldCat for books in, for example, Croatian language, published in Beograd (i.e. Belgrade).

[25] "Slavic cataloging Manual: Bosnian, Croatian, Serbian (BCS)": https://sites.google.com/site/seesscm/bosnian-croatian-and-serbian-bcs (accessed 04/29/2014).

[26] *Slavic Cataloging Manual* http://sites.google.com/site/seesscm/montenegrin-language (accessed 04/21/2017)

[27] Husic, Geoff. "Tracking the History of Romani Publications: Challenges Presented by Flawed Data." *Slavic & East European Information Resources* 13, no. 4 (December 1, 2012): 230–34. doi:10.1080/15228886.2012.730849.

[28] A bulk change of these codes may be desirable. The number of errors after a bulk change will be, at least, much less than in the current situation.

[29] "066 Script Codes": https://www.oclc.org/content/dam/support/connexion/documentation/client/international/066scriptcodes.pdf