Mary Mellon. “The Use of Cyrillic Metadata for Enhancing Discovery of Russian Digital Collection Items: A Case Study of the Bowman Gray World War I Postcards Digital Collection.” A Master's paper for the M.S. in I.S. degree. November, 2014. 35 pages. Advisor: Denise Anthony.

This paper examines the online discoverability of multilingual digital collections, focusing on the effectiveness of romanized and original script metadata for providing access to materials in non-roman script languages. Using the *World War I Postcards from the Bowman Gray Collection* digital collection at the University of North Carolina at Chapel Hill as a case study, the dynamics of Russian-language user access to postcards with and without Cyrillic description were compared with those of other major language user groups accessing the collection.

While limited on a dependence on Google’s system of determining user language, the results suggest that the nature of the Cyrillic metadata included in postcard records, limited to title, publisher, and other information transcribed from the resource in a bibliographic cataloging context, did not enhance the discoverability of the postcards. Moreover, every language group was at a distinct disadvantage compared to English-language users in terms of numbers of items discovered. In conclusion, I discuss various factors that may have affected these results, as well as implications for cultural heritage institutions with multilingual and multi-script collections.

Headings:

Digital libraries.

Transliteration.

THE USE OF CYRILLIC METADATA FOR ENHANCING DISCOVERY OF RUSSIAN DIGITAL COLLECTION ITEMS: A CASE STUDY OF THE BOWMAN GRAY WORLD WAR I POSTCARDS DIGITAL COLLECTION

by
Mary Mellon

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

November, 2014

Approved by:

_____

Advisor

Table of Contents

Introduction

Romanization, or the transliteration of a non-roman script into roman lettering, is a descriptive practice that is widely used in American cultural heritage institutions to describe materials. Its use is mandated in the two major bibliographic descriptive standards, AACR2 and RDA, which influence descriptive practices in archives and other contexts. According to the "Romanization Landscape" statement from the Policy and Standards Division of the Library of Congress (2011), romanization, or the transformation of a non-roman script to roman lettering, is used "primarily for LC staff and staff at other libraries without language expertise" performing functions in circulation, acquisitions, and other areas. In other words, it was not designed as a user-centric means of accessing materials.

It is therefore not surprising that catalogers of Russian, Chinese, Arabic, and other non-roman script materials have historically had difficulty in applying Anglo-centric rules and guidelines in a way that realistically reflects not only the content of the resources but the information-seeking strategies of potential researchers. For cultural heritage institutions, these difficulties can extend beyond the realm of the library OPAC into other areas influenced by bibliographic description, such as creating finding aids for archival collections and applying metadata to digital collection objects. For instance, the use of controlled terms such as name authorities and subject headings in contexts beyond

the OPAC perpetuate an aspect of description that does not accommodate non-roman script searching.

The Bowman Gray World War I Postcards digital collection at the University of North Carolina at Chapel Hill provides opportunity to look at how non-English language groups access a multilingual digital library collection that includes a non-roman script: Russian (Cyrillic). A collection first made available online in 2009, it is made up of 528 sets of postcards originating from seven European countries in addition to the United States, including 115 sets from Russia. As the fundamental content of the collection is images, it is not a text-rich collection, limiting the benefits of keyword searching and increasing the importance of cataloger-supplied description. Most of the metadata for the postcard sets are shared with bibliographic records for the sets in the library OPAC, with some enhancements for discovery and access, such as image tags from the Library of Congress Thesaurus for Graphic Materials (TGM) and user tagging and commenting functionalities.

All postcards in the World War I postcard collection have information transcribed from the item itself, such as title, caption, and publisher information, meaning that all foreign language postcards have some level of vernacular description as long as text appears on the postcard. In the case of Russian postcards, however, this information was originally included in the postcard metadata fields only in romanized form, as permitted under the AACR2 rules applied to their description.

In 2012, staff working in Slavic and East European Resources and Digital Projects divisions of the UNC Chapel Hill Libraries coordinated the addition of the original Russian Cyrillic script transcribed from the Russian postcard subset to their item

records in the digital collection. The motivations behind this project were to increase the accuracy of description and improve discovery of the postcards for users searching in Cyrillic Russian, rather than transliteration. Due to the image-oriented nature of the collection, however, it is by no means certain that the level of description in original Russian will have a significant impact on discovery and access.

By analyzing user demographic and website traffic data for the entire postcard collection, I address the following question: *Does inclusion of Cyrillic for basic bibliographic fields affect the discoverability of digital collection items for Russian-language users?* The intention behind this approach is to help determine whether including vernacular scripts adds value to description, or whether additional options need to be considered, like cataloger-supplied description and access points in multiple languages and scripts. In the literature review that follows, it should be clear that many language specialists consider vernacular script to be an essential part of bibliographic description, whether or not its inclusion is required by content standards. What remains less clear is whether meeting the minimum bibliographic description requirements will translate into discoverability of resources when more and more users are relying on keyword and full-text search to find sources online.

Literature review

Transliteration has existed as a written communication tool much longer than the professionalization of library and information services. People with no knowledge of a particular script still rely on phonetic representations in their own script to interpret and reproduce names of people, places, untranslatable concepts, and other terms as needed. In American cataloging, transliteration's primacy over vernacular scripts grew upon the

automation of cataloging, when existing systems could not handle the input or display of non-roman scripts. Catalogers of non-roman script material thus have had a long, uneasy relationship with transliteration, which on the one hand has been vital to providing a certain level of access to such resources, but on the other hand, is often far from sufficient in meeting users' needs in terms of discovery and access.

Even after technological improvements, including the release of Unicode-enabled MARC21 in 1999 and OCLC's 2005 conversion of the WorldCat database to handle Unicode character encoding, and revision of content standards to include provisions for including non-roman scripts, library and information professionals continue to mull improved access to such materials. The limitations of transliteration as an avenue for discovery and intellectual access of information resources in non-roman scripts has been extensively documented for a variety of scripts and languages, including Russian and other Cyrillic-script languages (Aissing 1995, Brewer 2009, Husic 2009), Chinese, Japanese, and Korean, or "CJK" scripts (Arsenault 2002, Kudo 2010, Park 2007), Arabic, Persian, and Hebrew, and others (Aliprand 1992, Molavi 2006, Lawson 2010). While each language has its own idiosyncrasies when it comes to script conversion, similar themes include lack of standardization or the existence of competing standards, loss of information through transliteration, the inability to return transliterated script to the original vernacular, and conflicting user expectations regarding searching for non-roman scripts.

Despite the advantages of transliteration for users who are unable to either intellectually or technologically access resource description or content in non-roman scripts, it still presents challenges for scholars knowledgeable in such scripts. The

information loss that occurs when transliterating scripts and its consequences have been extensively documented. Joan Aliprand provides a broad overview of the problem of "information distortion" (1992, p. 105) resulting from transliteration of non-roman scripts in the bibliographic context, using examples from various languages. For instance, she notes that representation distinctions between homophones in Chinese script can be lost when transliterated, resulting in homonyms that can confound users when searching for materials. Another major aspect of information loss is the inability to reconstruct original script accurately from transliteration, for instance to search for a publication cited in a research article. There is usually only one corresponding roman character or set of characters for a non-roman character, but there can be multiple options when trying to convert back to the original script, especially if diacritics are not used.

Brewer (2009) addresses how romanization acts as a major contributing cause to information literacy problems specifically in the realm of Slavic studies, although many of the problems he identifies exist in other language contexts. Based on his observations of student research as a Slavic Studies librarian, Brewer identifies a lack of understanding of the multiplicity of transliteration systems and search strategies to deal with the problem, especially in the light of increasing student reliance on full-text searching. Not only is text transcribed from the resource problematic, but indexing aids such as uniform titles, name authorities, and Library of Congress subject headings can cause confusion, especially since many older authorized forms are not transliterated according to the current Library of Congress standards. For example, "Tchaikovsky, Peter Ilich," the long-used Western variant of the famous Russian composer's name, is favored in the authority

form over “Chaikovskii, Petr Il’ich,” the proper transliteration according to the Library of Congress romanization standard for Russian.

Husic’s (2009) discussion of Russo-Serbian transliteration illuminates the challenge of representing archaic orthographies in a way that reflects how modern users might search for resources containing them, which is another problem that can extend beyond the realm of Slavic languages. Since this nineteenth-century Russo-Serbian script in question appears as a mishmash of Cyrillic orthographies, past attempts to romanize were drawn in conflicting directions in terms of applying existing romanization systems for Russian and Serbian, resulting in a lack of consistency and a greater intellectual burden on researchers. In cases such as these, it might even be desirable to include a transliteration into the modern version of the non-Roman script to aid in discovery and access, as it is not a given that potential users would be familiar enough with obsolete scripts to be able to search using them.

Transcription of information from resources has not been the sole focus of the romanization debate, as researchers have also explored the problematic aspects of providing subject and name access for non-roman script materials in a meaningful way for end users. While Library of Congress subject headings already are not always intuitive, the practice of using English equivalents and romanized forms for non-roman script materials risk rendering such headings useless for item discovery.

El-Sherbini and Chen (2011) address the issue of transliteration and subject access through a survey of library professionals and end users (faculty and student researchers), evaluating their subjects’ experiences when performing subject heading searches for non-roman script materials. As in the case of Brewer, the authors found a

gap in terms of library professionals' knowledge of the use of romanization in resource description and the awareness of students. Library staff were less likely to use vernacular scripts in subject heading searches than keyword searches and also tended to be more comfortable with English-equivalent headings and LCSH structures in general. End users reported a variety of problems related to romanization, including unfamiliarity with transliteration standards and inconsistent romanization on the part of catalogers, especially for East Asian languages, Arabic, and Hebrew. Not surprisingly, a majority of librarians and end users expressed the opinion that inclusion of subject headings in original script would be beneficial for research, although librarians expressed concern over the time and resources that would be involved in implementation of this goal. The idea of providing subject headings in non-roman scripts is an important example of a growing consideration of going beyond the text that appears on a resource when providing intellectual access to end users.

The issues of description of multilingual collections and representation of non-roman scripts have also gained the attention of the archival community, evidenced by a panel devoted to the subject at the 2014 Society of American Archivists annual meeting.[1] DACS provides guidelines for the identification of languages found in a collection but provides little guidance for leveraging multiple languages and/or scripts in describing collection content. Elements of archival description such as subject access and name authorities are still governed by bibliographic descriptive standards, meaning that non-roman scripts are also missing from these areas. In the absence of guidelines for display

[1] The panel session "Many Languages, One Archives: Creating Multilingual Finding Aids and Digital Collections" took place on 15 August 2014 at the SAA Annual Meeting, Washington, DC, August 13-16, 2014. Participants included Liz Phillips, Lisa Nguyen, John R. Nemmers, and Margarita Vargas-Betancourt (in absentia).

of multiple languages, institutions have turned to a variety of ways to reflect the language content of archives and special collections, from minimal foreign language inclusion (e.g. transcription of document and folder titles that appear in a collection), to full description in multiple languages (including translation of archivist-supplied notes at the collection level), to mirror resource guides and website interfaces for multiple languages.

The SAA panel presentations (2014) by Liz Phillips and Lisa Nguyen of the Hoover Institution Archives and John Nemmers and Margarita Vargas-Betancourt of the University of Florida highlighted the possibility of lingering technological and logistical issues of dealing with resource description in multiple languages and/or scripts. One technological challenge is updating existing data environments to properly display non-roman scripts, as the Online Archive of California had to do before being able to accommodate the Hoover's first finding aids with CJK scripts in 2010. As Nemmers pointed out, the structural standard EAD will not have the capacity to encode for multiple language elements until the release of its next version. Perhaps the biggest obstacle to implementing multilingual or multi-script finding aids on a large scale are is the necessity of specialized language knowledge and extra labor to provide original description and translation, especially when this might mean taking resources away from other projects (Nguyen 2014, Vargas-Betancourt 2014).

The SAA panelists performed preliminary evaluations of their resources and were able to identify some benefits from their multilingual approaches. Using geographic location and language data gathered through Google Analytics, both institutions showed an increase in users outside of English-speaking countries visiting their resources, which in the case of the Hoover Institution's CJK finding aids—the use of the Chinese Cultural

Revolution guide increased by 257% after incorporating vernacular script—points to tangible benefits in terms of discovery of non-roman script materials (Phillips 2014). These preliminary studies do not address the issue of describing low-text, image-oriented digital collections, but the use of Google Analytics as a nonintrusive means of collecting data on language and access to web resources has informed the methodology of the present study.

Overview of World War I Postcard Collection

The postcard digital collection is made up of approximately 528 postcard sets representing a variety of themes and original functions, including nationalistic and anti-enemy propaganda and domestic fundraising for the war efforts of the respective nations. For the most part, "sets" are composed of one or more postcards from a particular series and publisher, although some thematically related cards not connected through a publishing series were grouped together for description and cataloging purposes.

The country of publication of each postcard set is reflected by a subcollection code in the call number suffix, and the language of the titles and captions printed on the postcards (if any) roughly corresponds to the country of publication, with some exceptions.[2] Table 1 shows the composition of the postcard collection by country subcollection, showing the largest to be from the United Kingdom, followed by Russia, France, and Germany. In terms of language, therefore, English is the best represented through a combination of United Kingdom and United States subcollections, but foreign languages still make up more than 50% of the collection.

[2] For instance, the United States subcollection includes a handful of postcards published by émigré societies in Polish and German.

Table 1: Country subcollections in the World War I Postcard Collection

Country	Call number suffix	Number of postcard sets
United Kingdom	[n/a]	178
Russia	rur	115
France	fr	94
Germany	gw	77
Italy	it	32
USA	u.s.	24
Poland	pl	8
	TOTAL:	528

In both the library catalog and CONTENTdm, postcards are described first at the set level and then at the item level using AACR2 rules.[3] Each postcard set in CONTENTdm has a set-level record ("Object Description") that summarizes the contents of the set. Choosing an individual postcard from the set will also display an item-level record ("Description") field that describes that particular postcard, including a summary of the image and transcription of any text appearing on the postcard. The set level description corresponds to the MARC records for physical postcard sets in the library OPAC.

While the item-level records follow the same pattern as the set-level, elements specific to the individual postcards, such as captions and notes, reside only in the digital collection. Another example is the "Subject (tgm)" field, populated with keyword tags taken from the Library of Congress TGM. These tags were meant to improve searching and browsing options within the digital collection by expressing concepts and visual

[3] Cataloging and creation of the digital collection occurred before the release of RDA.

elements not encompassed by controlled subject headings. Table 2 demonstrates how fields in the CONTENTdm records correspond to the MARC records for Russian postcards.

Table 2: Matching metadata fields from the set-level and item-level CONTENTdm records to MARC fields in the library OPAC record

Field name	Set-level metadata fields to MARC	Item-level metadata fields to MARC
Rating	CONTENTdm only	n/a
Title	245: Transliteration or English (if cataloger-supplied); linked 880 field in Cyrillic	880 linked field (linked to 505) in Cyrillic: treated as entry in table of contents
Alternative Title	n/a	505: treated as entry in table of contents
Description	520	CONTENTdm only
Publisher	260 |b	(shared with set record)
Date	260 |c	(shared with set record)
Extent	300 |a	n/a
Size	300 |c	(shared with set record)
Note	500	CONTENTdm only
Subject (tgm)	n/a	CONTENTdm only
Subject topical	650	650
Subject name	600	600
Call Number	099	CONTENTdm only
OCLC Number	001	(shared with set record)

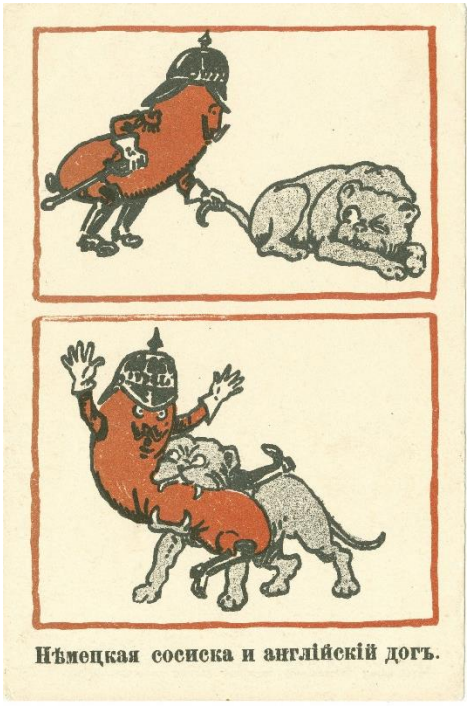
While all Russian postcards include romanized text, the Cyrillic fields were only added to catalog records and the CONTENTdm records after the digital collection had been published to improve access for users who, regardless of native language, search for sources in Russian Cyrillic. OCLC provides macros to aid in the Cyrillicization of transliterated fields in MARC records, but the Cyrillic text on individual postcards had to be transcribed manually for the CONTENTdm item-level records. This was

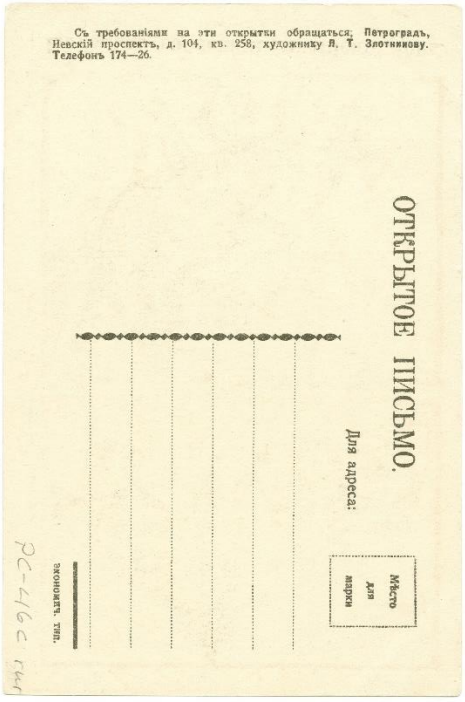
accomplished through the collaboration of graduate students and professional staff working in Slavic and East European Resources and Digital Projects divisions of the UNC Chapel Hill Libraries during the majority of the 2012 calendar year.

While the addition of Cyrillic was certainly desirable in terms of accurately identifying a resource (the RDA guiding principle of transcribing information "exactly as it appears on the source"), including added-value description in Russian was not considered, given limited time, resources, and language expertise that could be devoted to the project. As a result, the usefulness of the updated records for discovery by Russian-language users may depend in large part on how much text appears on the postcards and how well that text maps to major subjects or themes associated with World War I. The following two examples demonstrate how a text-poor postcard might be more difficult to discover through Russian-language searching than a text-rich postcard.

In Example 1, a cartoon postcard featuring a caricature of German emperor Wilhelm II as a sausage, the English description and subject headings are much more meaningful in conveying the nature of the postcard in text than the Russian caption alone (translates as "German sausage and English dog"). Since the image must be visually interpreted to understand its connection to World War I, it is unlikely that a text-based keyword search would return this item without the existence of the English-language, description, subject headings, and subject tags. The most meaningful information in Russian, however, appears to be the name of the artist, taken from the back of the postcard. Otherwise, there is no Russian text to contextualize the postcard as being related to World War I.

Example 1: Russian postcard with minimal Russian text. Image from *World War I Postcards from the Bowman Gray Collection* [digital collection], Rare Book Collection, Wilson Library, UNC-Chapel Hill. http://dc.lib.unc.edu/cdm/compoundobject/collection/graypc/id/9864/show/9838/





Title	Немецкая сосиска и английский дог.
Description	A colored drawing in two panels. The first panel shows a sausage dressed as William II pulling the tail of an English bulldog. In the second panel the bulldog is biting the sausage.
Alternative title	Nemetskaia sosiska i angliiskii dog.
Publisher	n/a
Date	1914-1918
Size	9 x 15 cm
Note	On back of card: Открытое письмо. С требованиями на эти открытки обращаться: Петроград, Невский проспект, д. 104, кв. 258, художнику Л.Т. Злотникову. Телефон 174-26. Экономич. тип.[4]
Subject (tgm)	Sausages; Dogs; Caricatures; Helmets
Subject topical	World War, 1914-1918--Russia.; Postcards--Russia.; World War, 1914-1918--Caricatures and cartoons.; Propaganda--Russian.; World War, 1914-1918--Propaganda.
Subject name	William II, German Emperor, 1859-1941.

[4] Field contains instructions for requesting postcards, including address and name of artist, and the name of the printing company.

Example 2 is a more text-rich postcard depicting Empress Alexandra (Aleksandra Feodorovna) and two of her daughters, Ol'ga and Tatiana, with wounded officers at a hospital outside of Petrograd. The caption transcribed from the postcard identifies the subjects of the photograph, thus providing useful keywords for searching in Russian. The postcard record is also a good example of how proper nouns in the original text can supply alternatives to Westernized or transliterated names that appear in Library of Congress subject headings, for instance providing "Императрица Александра Федоровна" ("*Imperatritsa Aleksandra Feodorovna*") as an alternative to the LCSH version "Alexandra, Empress…" which even English-speaking Slavic specialists might not search for when looking for primary sources.

Example 2: Russian postcard with extensive Russian text. *World War I Postcards from the Bowman Gray Collection* [digital collection], Rare Book Collection, Wilson Library, UNC-Chapel Hill. http://dc.lib.unc.edu/cdm/compoundobject/collection/graypc/id/9922

(Example 2, continued)

Title	[Императрица Александра Федоровна и Великие княжны Ольга Николаевна и Татьяна Николаевна среди персонала Царскосельског о Дворцового Лазарета]
Alternative title	[Imperatritsa Aleksandra Feodorovna i Velikie kniazhny Ol'ga Nikolaevna i Tatiana Nikolaevna sredi personala TSarskosel'skogo Dvortsovogo Lazareta]
Description	A photograph of Alexandra, consort of Nicholas II, with her daughters Olga and Tatiana posing with the staff of the Tsarskoye Selo Palace military hospital.
Publisher	Т-во Р. Голике и А. Вильборг
Date	1914-1918
Size	9 x 15 cm
Note	On front of card: Со Всемилостивейшего соизволения издание газеты "Вечернее Время", Б.А. Суворина. С фот. худ. П.И. Волкова. On back of card: Открытое письмо. Carte postale. Всемирный почтовый союз. Россия. С соизволения Государыни Императрицы Александры Федоровны чистая прибыль от продажи этого издания пойдет на усиление средств лазаретов Царскосельского района, состоящего под Особым Покровительством Ее Величества. Перепечатка воспрещается. Т-во Р. Голике и А. Вильборг. Петроград. Звенигородская 11.
Subject (tgm)	Empresses; Nobility; Nurses; Medical personnel; Physicians; Military hospitals; Photographs; Group portraits; Uniforms
Subject topical	World War, 1914-1918--Russia.;
Subject name	Alexandra, Empress, consort of Nicholas II, Emperor of Russia, 1872-1918.; Olga Nikolaevna, Grand Duchess, daughter of Nicholas II, Emperor of Russia, 1895-1918.; Tatiana Nikolaevna, Grand Duchess, daughter of Nicholas II, Emperor of Russia, 1897-1918.; Romanov, House of.; TSarskosel'skii dvorets.

Of course, “text-rich” is a relative term—these postcards contain nowhere near the keyword-searching potential of born-digital or digitized, full-text documents processed with optical character recognition (OCR) technologies. The examples above thus demonstrate a reasonable concern that the effort of adding Cyrillic fields might not significantly aid discovery across the board, and that added-value description in

additional languages and or/scripts might be desirable for any institution wishing to linguistically or geographically diversify its digital collection audiences.

Methodology

Source of data

Data was gathered through the Google Analytics account for the CONTENTdm collections of the UNC-Chapel Hill Libraries. Data from categories relevant to the study were gathered for the 23 months from September 2012 through July 2014. This time period represents the longest uninterrupted and consistent tracking of data for the World War I postcard collection through Google Analytics. Data were gathered in monthly increments to ensure a complete view and avoid the automatic data sampling employed by Google Analytics over longer spans.

Definitions

(Google Analytics data terms appear in italics)

"Discovery" = *organic traffic* by *new users* to item-level *pages*

- *Organic traffic*: Traffic from links in organic (unpaid) search results. Examples of organic traffic sources: Google, Bing, Yahoo, Yandex, etc.
- *New users*- First-time visitors to website. This assumes that returning visitors are already aware of the WWI postcard collection and are no longer "discovering" it, regardless of traffic medium.

"Postcard" = *page* URL that provides full view of a postcard set and the individual postcard(s) it contains.

- Excludes collection landing pages, browsing pages, search pages and lists of search results.
- Operationalized, any URL containing the string "graypc/id/"
- "[language] postcard" (e.g., "Russian postcard") = postcard belonging to a particular language set as identified by its call number suffix (see Table 1).
- "vernacular postcard" = postcard belonging to language set that matches a particular language group. For example, a German postcard is a vernacular postcard for German-language users.

"Language group"=group of *users* that Google Analytics identifies with a particular *language*.

- For the purposes of this study, languages further distinguished by a country code were considered the same language. That is, "ru" and "ru-ru" were both counted as "Russian," and the various designations for English ("en" for English, "en-gb" for British English, "en-us" for American English, etc.) were also grouped together as one language.

Google Analytics tracks three categories of site traffic: organic, referrals, and direct/none. While referrals are an important aspect of how users discover and access digital collections, there is no obvious connection between the addition of Cyrillic to the Russian postcard subset and increased referrals, since there is no way of detecting how the entities that created the external links discovered the postcards--whether they reached the collection through a search engine or external link, what language they used, and so on. In other words, referrals cannot be used as a reliable surrogate for whether the inclusion of Cyrillic is effective for the discoverability of these postcards. Organic search

appears to be far better suited to uncovering the relationship between inclusion of Cyrillic Russian metadata and the discoverability of Russian postcards, since it is the indexed page content, including Cyrillic fields, that users are searching to reach the collection. Relying on the number of sessions by new users for this analysis, rather than the total number of sessions, helps limit the amount that repeat users skew the discoverability data (once a user returns to an item, they are no longer "discovering" it).

Selection of data

Inconsistencies in item URL formation and website tracking before and after a mid-2012 CONTENTdm version upgrade meant that a comparison of access data to Russian postcards before and after the Cyrillic was added would be unreliable to impossible. As an alternative, the effectiveness of the Cyrillic description as a means of discovery for Russian users was measured through two main data views: the percentage of postcards discovered by Russian-language users that are Russian postcards, and the percentage of postcards discovered by English-language users that are Russian postcards.

Since all postcards in the collection, regardless of place of publication or content language, have description in English, the discovery rate of Russian postcards by English users will be used to represent the discovery rate that would be expected if all postcards were equally discoverable, regardless of user or description language/script. In analyzing the above data, my expectations are that A) if the level of Cyrillic description IS sufficient for postcard discovery by Russian-language users, such users will discover a disproportionately high number of Russian postcards when compared to the English-language users, or B) if the level of Cyrillic description IS NOT sufficient for postcard

discovery by Russian-language users, the discovery rate will resemble that of the English-language users' discovery rate of Russian postcards.

This approach assumes that a user searching in a particular language will disproportionately land on items with description in that language. In order to test this assumption, the data from Russian users will be compared with the same statistics for the language groups corresponding to the other regional postcard subcollections: France, Germany, Italy, and Poland (see Table 1). This comparison will also serve to contextualize the difference (if any) between Russian and English user discovery of Russian postcards. In analyzing this data, my expectations are that A) if the level of Cyrillic description IS sufficient for postcard discovery by Russian-language users, the ratio of Russian postcard discovery rate by Russian users TO the Russian postcard discovery rate by English users will be similar or greater to that of the ratios for corresponding language groups (i.e. the ratio of German postcard discovery rate by German users TO the German postcard discovery rate by English users), or B) if the level of Cyrillic description IS NOT sufficient for postcard discovery by Russian-language users, the above-stated ratio will be significantly less than that of the ratios for corresponding language groups.

Limitations

While the ideal way to assess the effectiveness of Cyrillic metadata for discovery by Russian language users, as previously noted, would be a comparison of access data to Russian postcards before and after the Cyrillic was added, the necessary data from before Cyrillic was added is unavailable. In addition, there is no real control for varying cultural

or professional interest in postcards as an information resource and/or specific topics and genres, besides the fact that similar themes and genres are repeated across sets.

Much of the validity of the study also depends on the accuracy of user language identification in Google Analytics. Since language is determined by a user's browser settings, it is possible that a user's native or browser language is correctly identified, while the user's search is being performed in a different language. Since this is the only source for interpreting language (external search terms are not available), it has to be assumed that this means of identification is more or less accurate. An evaluation of the resulting data may be able to shed light on the accuracy and effectiveness of the Google Analytics language category for granular analyses of collection use.

In summary, the above methodology will not be able to determine what impact, if any, the addition of Cyrillic made on postcard discovery, but at best indicate whether the level of description provided for all postcards served as effectively in allowing Russian-language users to discover Russian postcards.

Results

General characteristics of postcard traffic

As demonstrated in Figure 1, postcard discovery was dominated by far by English language new users. English language users had over twice as many views (n=1075) as other languages combined (n=474).

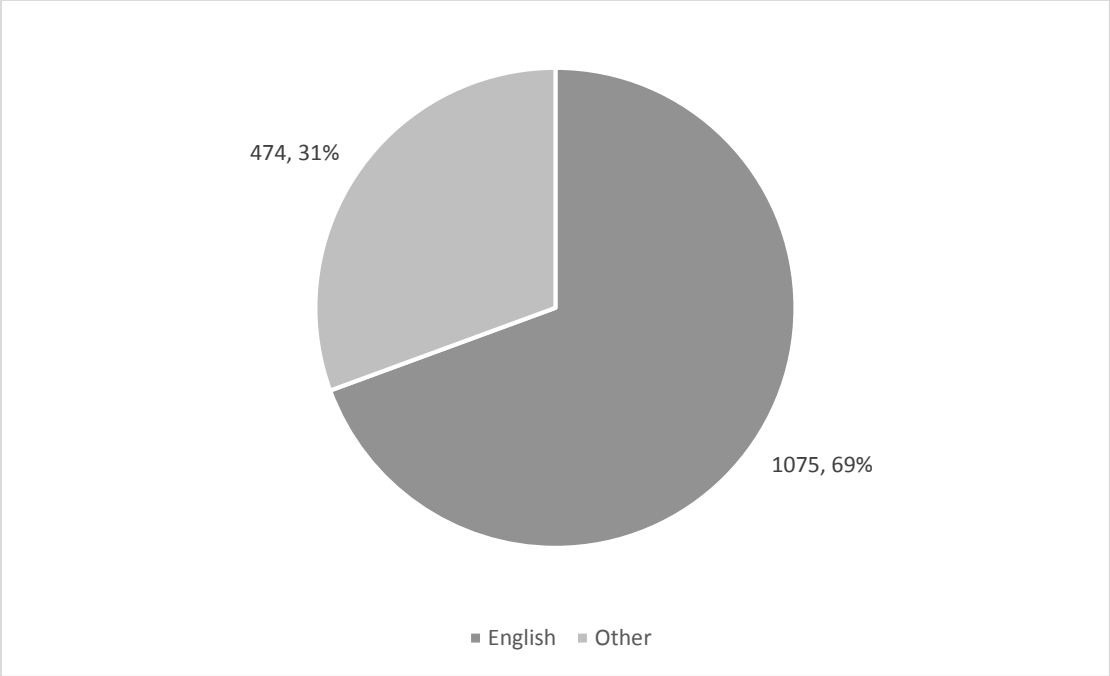


Figure 1: Postcards discovered by English-language users and other users

Figure 2 (below) includes the numbers of postcards discovered by the top 9 non-English language groups in terms of accessing postcards in the collection. Besides Dutch, the only other language groups to surpass 30 postcard discoveries are those that correspond with the postcard subcollections (French, German, Italian, Polish, and Russian).

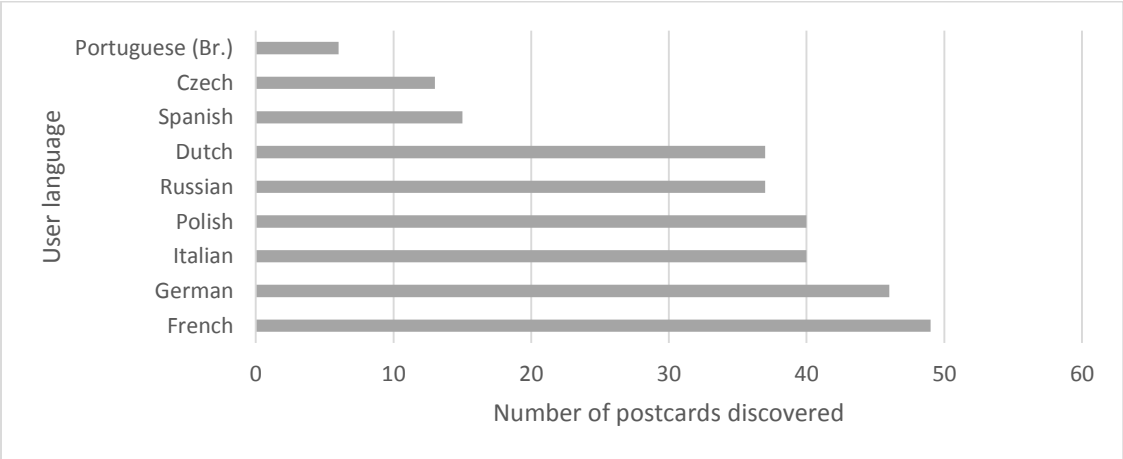


Figure 2: Postcards discovered by user language (top 9 groups)

Characteristics of Russian user traffic

Russian language users were the largest language group (n=106) after English in terms of total postcard traffic, but they were only fifth (n=37) in terms of postcard discovery. Fewer Russian-language new users discovered Russian postcards via an organic medium than by referrals: 35% (n=37) of all Russian new user traffic was organic, whereas 46% (n=49) came from referrals to the website. The remaining 19% (n=20) was direct or undetermined traffic.

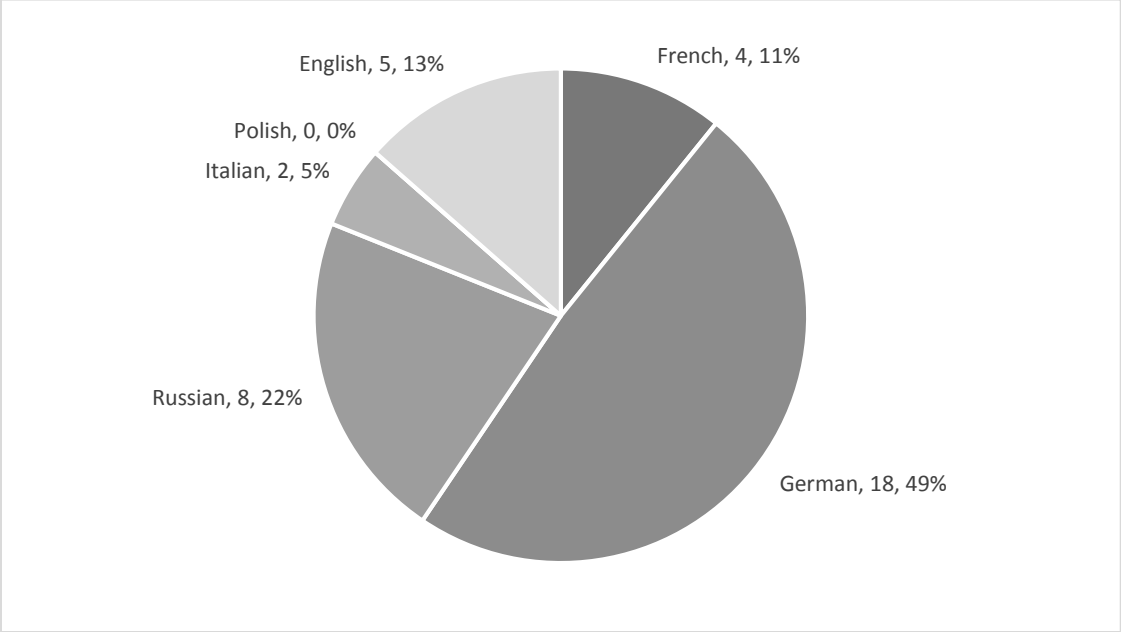


Figure 3: Postcard types discovered by Russian-language users

A breakdown of the different postcard types discovered by Russian users is shown in Figure 3 (above). The number of Russian postcards discovered (n=8) was in the

minority compared to the other language groups combined (n=29). Nearly half (n=18) of all the postcards discovered by Russian users were from German postcard sets.

Discovery of Russian postcards: English vs. Russian language users

In a comparison featuring vastly differing sample sizes, Russian postcard discovery (22%) by Russian users slightly exceeded that of English users (19%) in terms of the percentage of all postcards viewed by each language group that were from Russian postcard sets (see Figure 4).

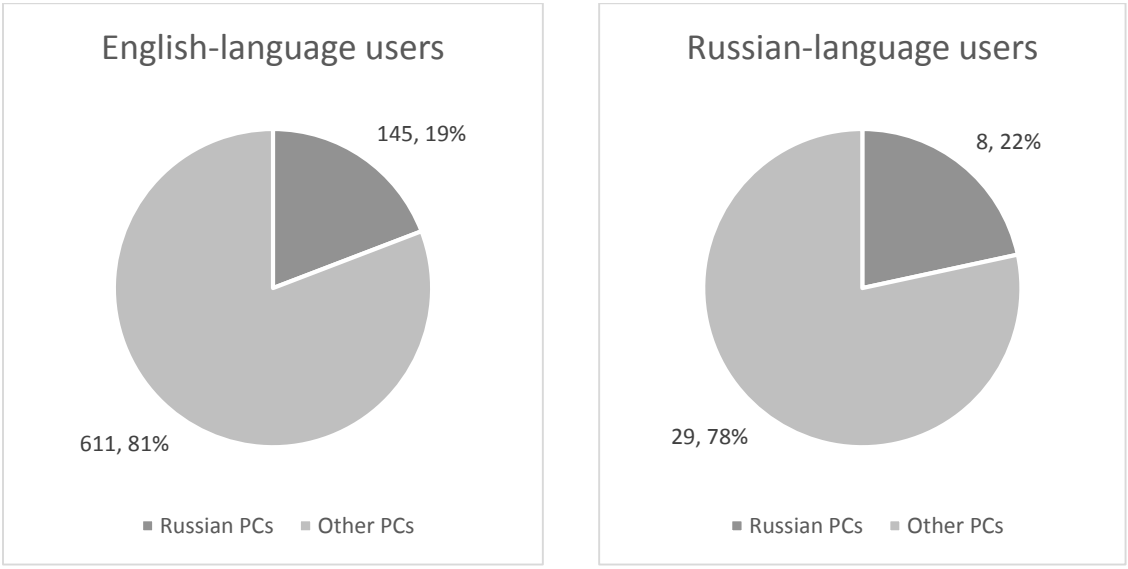


Figure 4: Discovery of Russian postcards as a percentage of all postcards discovered by English-language users (left) and Russian-language users (right)

Comparisons of vernacular postcard discovery among language groups

The number of postcards discovered through organic traffic over the 23 month period was very similar among the four language groups examined, ranging in number from 37 to 47. Figure 5 shows a breakdown of the percentages of all postcard types discovered by

each language group, while Figure 6 compares the percentage of vernacular postcards (i.e. postcards matching the language of a particular user group) and foreign language postcards viewed by each group.

According to these data views, the highest rates of vernacular postcards discovered were among German (35%) and French (31%) language groups. Russian, Polish, and especially Italian vernacular postcard discoveries happened at a much lower rate (22%, 18%, and 10%, respectively).

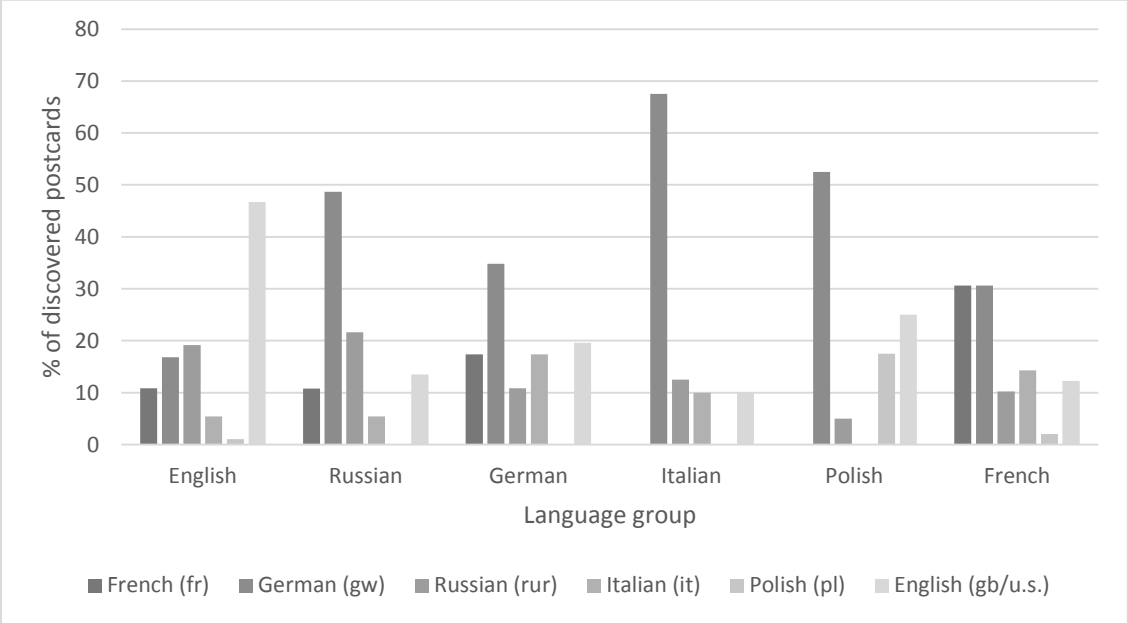


Figure 5: Postcard types discovered by language group as percentage of all postcards discovered by each language group

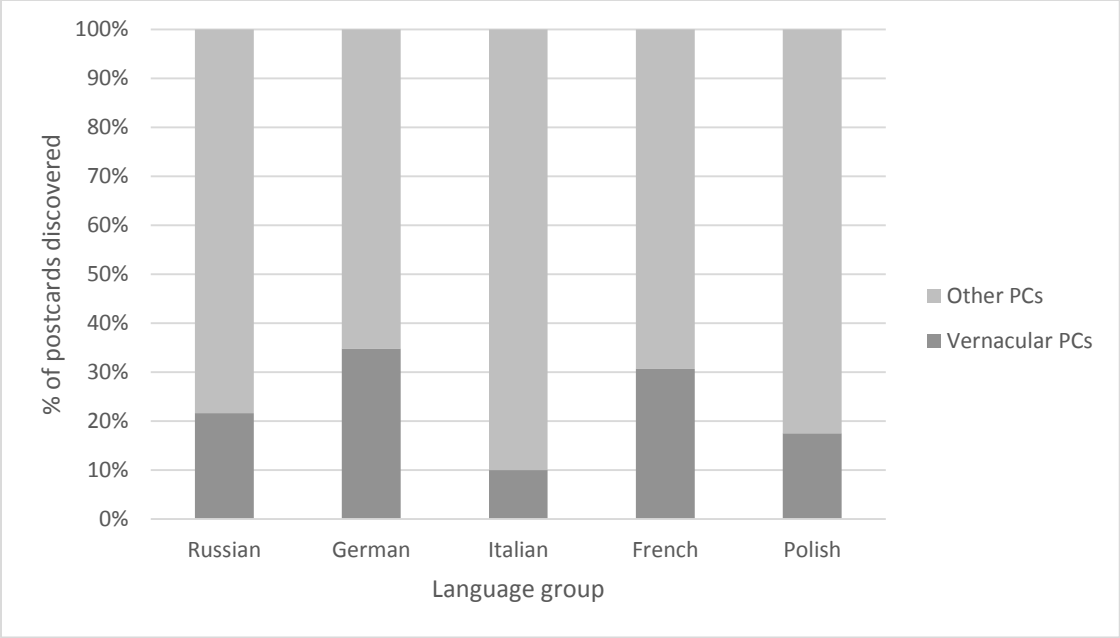


Figure 6: Percentage of postcards discovered by each group that were in the vernacular language of that group

Surprisingly, the top postcard type by percentage discovered across all non-English language groups was German postcards, including nearly half (48%) of postcards discovered by Russian users. Figure 7 compares the data from Figure 6 with the discovery rate of the same language postcard sets by English-language users, the latter data representing expectations of discovery if language were not a significant factor.

As previously noted, the rate of Russian postcard discovery by Russian users is slightly higher than that by English users, but the difference in the two percentage values is the lowest among all the comparisons between the various language groups and English users. When viewed as ratios (Table 3), the ratio of between Polish to English discovery rates of Polish postcards is greatest, followed by French (2.8:1) and German (2.1:1), while the ratio of Russian discovery rates of Russian postcards to English discovery rates of Russian postcards is closest to 1:1.

Table 3: Ratio of non-English user discovery rates of vernacular postcards to corresponding English user discovery rates

Language group	Ratio
Russian	1.1:1
Italian	1.8:1
German	2.1:1
French	2.8:1
Polish	16.5:1

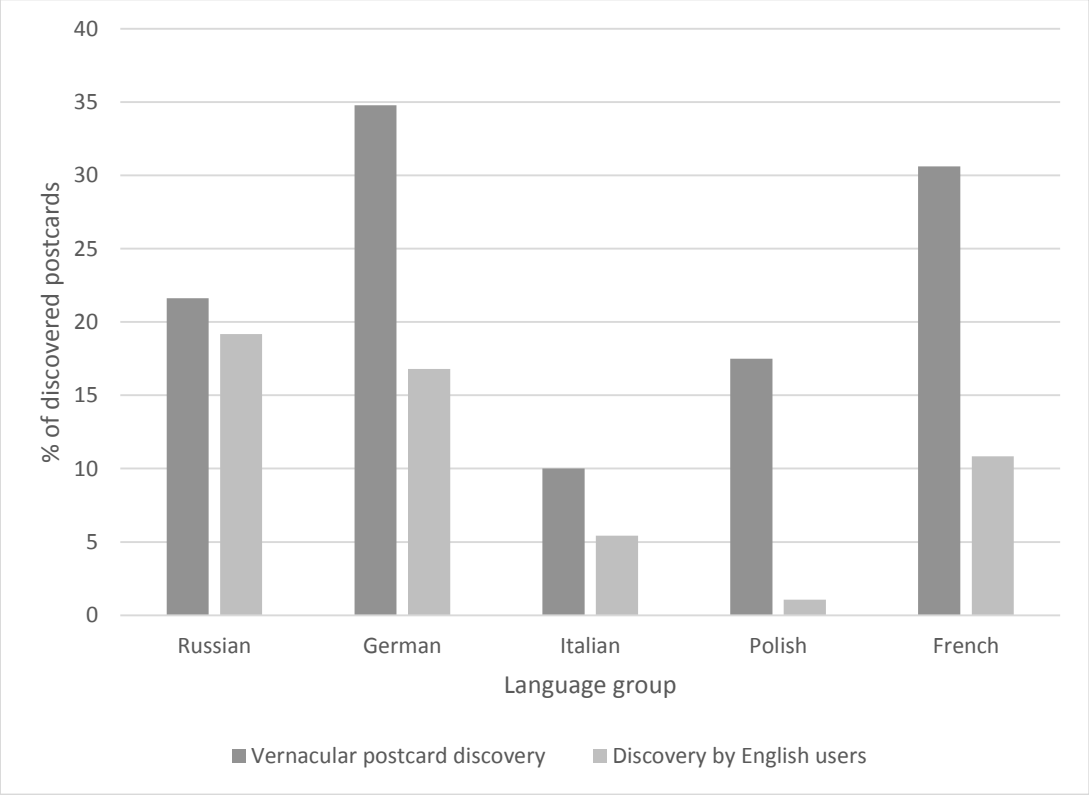


Figure 7: Vernacular postcard discovery rate by language group vs. discovery rate by English-language users

Discussion

As expected, English language users represented an overwhelming majority of users to discover postcards in the Bowman Gray World War I Postcards digital collection. Not only do all postcards have description in English, including subject keywords and physical description, but the website and interface languages are also English, which could impact a user's decision to follow a search results link to the collection. In terms of total postcard discovery, Russian users do not appear to be at a significant disadvantage relative to other non-English user groups. No single language group came close to the numbers of discoveries by English-language users, with the top language groups clustering around 40 discoveries over 23 months, compared to the 756 discoveries over the same time by English-language users.

When the data are broken down by the different types of postcards discovered across language groups, many of my assumptions regarding the relationship between user language and the language of postcards discovered are called into question, particularly that a non-English language user group would discover a predominantly large percentage of vernacular postcards. In the case of Russian-language users, the percentage of postcards they discovered that were Russian (22%) was slightly higher than the percentage of postcards discovered by English-language users that were Russian (19%), which, under the original assumptions, would mean that the level of Cyrillic description included in the postcards was not effective for postcard discovery by Russian users. Likewise, the nearly 1:1 ratio of the discovery rate of Russian postcards by Russian-language users to that of English-language users, which was much lower than the

corresponding ratios for other language group, suggests that Russian users are at a disadvantage relative to other groups when it comes to vernacular postcard discovery.

Such a disadvantage might be explained by the minimal Cyrillic description compared to English and romanized description for Russian postcards, especially if users are searching for keywords and subject terms in Russian that do not correspond with the Russian text that appears in title, caption, or publisher fields. For instance, a Russian-language researcher might be more likely to search for "Первая мировая война" [*Pervaia mirovaia voina*] vs. "World War I" or "World War, 1914-1918"; "Вильгельм II" [*Vil'gel'm II*] vs. "Wilhelm II," and so on, but such terms are not guaranteed to show up in the Cyrillic text of a postcard, even if they are reflected in the image content. Another interesting aspect is that the formal terms for "postcard" in Russian—"открытое письмо" [*otkrytoe pis'mo*] and "почтовая карточка" [*pochtovaia kartochka*], were used on postcard backs during the World War I era and therefore appear in the postcard metadata instead of the more common term used today: "открытка" [*otkrytka*]. This means that Russian users searching for postcards by genre, using the more common, colloquial term, might not locate items in the collection, depending on search engine functionality.

The fact that the highest percentage of postcard type across non-English users, including Russian users, was for German postcards, however, cannot easily be explained with regard to the language of users. The simplest explanation would be that user language as identified by Google does not strongly correlate to the language of search employed by users. For instance, Russian users that land on German postcards are searching in German, even though their browser settings indicate their primary language

to be Russian. Another explanation would be that topical interest is skewing the results more than expected, with the subjects of the German postcards holding greater interest in an international context. British and American postcards were the most popular for English-language users, at about 47% of postcards discovered.

Besides the possible language barriers to discovery and access, there is the possibility that all non-English-language users, including Russian language users, simply have less interest in the Bowman Gray collection, due to the existence of equivalent or superior alternatives in terms of digital postcard collections in other languages. For instance, Europeana, which acts as a "union digital library" of sorts, connects users with the digital collections of content partners while allowing users to browse in a variety of interface languages. Of the 49 postcards visits by Russian-language users that resulted from referrals, 59% of overall discoveries and 90% since January 1, 2014 occurred through referrals from Europeana 1914-1918, a special World War I-themed portal created due to the increased interest in the topic during the war's one hundredth anniversary.

Conclusion

All in all, it is impossible to draw definitive conclusions from the postcard discovery analysis, given the uncertainty over whether user language settings accurately reflect the search language of choice for a particular user and, perhaps, the differences in sample size between English-language user statistics and other language groups. It does appear, however, that the treatment of Russian according to bibliographic standards that were designed to aid in catalog searching is not well-suited to discovery through less structured environments like web searches. Indeed, the minimal foreign language description in an

Anglo-centric bibliographic description framework does not seem particularly effective for discovery for any foreign language group, regardless of whether it uses roman script or not: although non-English postcard types were well-represented within the digital collection, nearly 70% of all postcard discoveries were by English-language users.

Given the likelihood that the "Russian-language" users that discovered postcards may not even have been searching in Russian, there is reason to be concerned that Russian, and by extension, other non-roman script languages are even more problematic in context of repurposing bibliographic description for describing objects in digital library collections, especially given the known shortcomings of transliteration as described in the literature. Due to the aforementioned difficulties in data collection and analysis, more research would need to be done, preferably with "before-and-after" data, to determine the impact of script on digital resource discovery. More research is also needed to determine the web search habits of English-language scholars searching for materials in non-roman scripts to better inform information professionals about whether romanization serves this narrowly-defined audience, and if so, how.

For many institutions, an international audience is not a priority for designing resource guides and collection interfaces—achieving a global reach might not be feasible, given competing demands on limited resources or priorities to serve the surrounding community of scholars and the general public. In this case, alternatives to enhanced description such as involvement in digital library consortia and increased outreach might be more effective and sustainable than enhanced description. The inclusion of postcard surrogate records in Europeana 1914-1918 has already shown some benefits in terms of referring traffic to the postcard collection website.

Bibliography

Aissing, Alena L. 1995. “Cyrillic transliteration and its uses.” *College & Research Libraries* 56 (3): 207-19.

Aliprand, Joan. 1992. "Nonroman Scripts in the Bibliographic Environment." *Information Technology and Libraries* 11 (2): 105-19.

Arsenault, Clément. 2002. "Pinyin Romanization for OPAC retrieval—is everyone being served?" *Information Technology & Libraries* 21 (2): 45-50.

Brewer, Michael M. 2009. “Romanization of Cyrillic Script: Core Competencies and Basic Research Strategies for Slavic Students, Scholars, and Educators.” *Slavic and East European Information Resources* 10 (2-3): 244-256.

El-Sherbini, Magda and Sherab Chen. 2011. "An Assessment of the Need to Provide Non-Roman Subject Access to the Library Online Catalog." *Cataloging & Classification Quarterly* 49 (6): 457-483. doi: 10.1080/01639374.2011.603108.

Husic, Geoff. 2009. "Russo-Serbian Orthography: Cataloging Conundrum and a Proposed Solution." *Slavic & East European Information Resources* 10 (1): 45-60.

Kudo, Yoko. 2010. "A Study of Romanization Practice for Japanese Language Titles in OCLC WorldCat Records." *Cataloging & Classification Quarterly* 48 (4): 279-302.

Lawson, David R. 2010. "An Assessment of Arabic Transliteration Systems," *Technical Services Quarterly* 27 (2): 164-177.

Molavi, Fereshteh. 2006. "Main Issues in Cataloging Persian Language Materials in North America," *Cataloging & Classification Quarterly* 43 (2): 77-82.

Nemmers, John R. 2014. "Implementing Multilingual Access to Archival and Digital Resources.” Presentation at the Society of American Archivists annual meeting, Washington, DC, August 13-16, Accessed 5 November 2014. http://files.archivists.org/conference/dc2014/materials/307-Nemmers.pdf .

Nguyen, Lisa. 2014. "Many Languages, One Archives: Creating Multilingual Finding Aids and Digital Collections" panel presentation, Society of American Archivists annual meeting, Washington, DC, August 13-16.

Park, Jung-ran. 2007. "Cross-lingual Name and Subject Access: Mechanisms and Challenges." *Library Resources & Technical Services* 51 (3): 180-9.

Phillips, Liz. 2014. "Many Languages, One Archives: Creating Multilingual Finding Aids and Digital Collections" panel presentation, Society of American Archivists annual meeting, Washington, DC, August 13-16.

Policy & Standards Division, Library of Congress. 2011. "Romanization Landscape." http://www.loc.gov/catdir/cpso/romlandscape_Oct2011.html. Accessed 20 July 2014.

Society of American Archivists. 2013. *Describing Archives: A Content Standard*. 2nd edition. Chicago, IL: Society of American Archivists. Accessed 5 November 2014. http://files.archivists.org/pubs/DACS2E-2013.pdf

Vargas-Betancourt, Margarita. 2014. "Creating Multilingual Finding Aids: Cuban Collections at the University of Florida." Presentation at the Society of American Archivists annual meeting, Washington, DC, August 13-16, Accessed 5 November 2014. http://ufdc.ufl.edu/IR00004308/00001.

World War I Postcards from the Bowman Gray Collection [digital collection]. Rare Book Collection, Wilson Library, University of North Carolina at Chapel Hill. http://www2.lib.unc.edu/dc/graypc/.