
Michael Neubert

ABSTRACT. Google is pursuing “mass digitization” of Russian books as part of its efforts that now involve more than a dozen partners, although so far there is no separate Russian-language search interface. How easy is it to find materials in books.google.com that are in Russian? What do users need to know to maximize their results?

KEYWORDS. Mass digitization, Russian books, searching, Cyrillic, Google Book Search

Google’s efforts at mass digitization of books with its network of library partners has received considerable attention in the press. However, this discussion has been generally at a high level and focused on English-language materials. This article examines aspects of the Google efforts to digitize Russian-language books in partner libraries and some of the difficulties in accessing these materials.

At the time of this writing, the URL http://books.google.ru/ does not offer a specifically Russianized search interface, rather just some pages of Russian language information for users and a pointer to the English-language search interface at http://books.google.ru. Notwithstanding the absence of a Russian search interface, books in
Russian are included in the digitization efforts and available in books.google.com. Google describes the language aspect of its “collection development policy” for digitization of books as follows: “Google Book Search now accepts books in all languages. Books in English, French, Italian, German, Spanish, Dutch, Portuguese, Chinese (simplified), Chinese (traditional), Japanese, Korean and Russian are integrated in the index.”

A paper given by Terence Emmons at the annual Crimea library conference in 2006 described the potential benefits to scholars of having well-developed access to a large body of books digitized by Google from its partner libraries. He gave positive examples of English-language subject keyword searching and searches to attempt to locate items containing specific text and posited similar benefits as Russian materials were more heavily included. (He also noted some problems with the system, most of which he associated with not enough scanning having yet taken place at that time.)

Google (on its own behalf) suggests through presentation of positive feedback that it has received that Google Book Search is a very useful tool for librarians. For example, Tom Bruno, Library Assistant at Harvard University is quoted as saying:

For the third time in the past week, I’ve been able to answer thorny reference questions using Google Book Search that I otherwise would have simply given up on. I tried Google Book Search and lo and behold—there was my mystery reference. That was more than enough to steer me to the proper volume on the shelves here at the library. Mission accomplished! Google Book Search and similar endeavors have given me a burst of optimism. Not only will scholarship be enhanced as a result of digitization, but it will become more relevant as well. The more that is scanned and made searchable, the more humanity as a whole will benefit.

Searching books.google.com for Russian-language books, many of which Google has decided are in the public domain and thus made available full-text, already generates results, although subject coverage is uneven since the effort to digitize millions of items in the partner libraries generally organized by subject will take years, and many materials remain to be done. How “findable” are these
materials? What is their “behavior” as items to be accessed in this online environment?

Google is secretive, so it is not easy to get accurate information about what they are doing or what they are trying to do. In addition, they impose a fair amount of secrecy on the partner libraries. Informal requests for information from contacts at partner libraries resulted in answers such as, “I know who has the information that would be useful to know but he (or she) can’t give that out to anyone.” There were no substantive responses to questions that I sent to the Google “librarian center” as to the number of items that had been scanned in the Russian language. They did state, “While different library partners are stronger in some areas than in others, we don’t have any language restrictions on what we digitize. It’s likely that the language breakdowns of each library’s holdings can be found out by giving the libraries a call—it’s doubtful that they keep that confidential.”

The number of partner libraries is steadily growing. Google now has thirteen partners, including three foreign partner libraries. The foreign partners at this time are all in countries (Spain, Germany) where the language issues are relatively simple compared to Russian. For librarians and users interested in Russian materials, the digitization initiatives at the University of Michigan and Stanford University are for the foreseeable future the most likely to yield useful results, as both have strong Slavic collections and Google is digitizing the entire general collections of these institutions. The University of Michigan’s work is particularly interesting because it is actively making available online the materials digitized by Google via its own Online Public Access Catalog (OPAC) (although it trails behind books.google.com, judging by what can be found in the one and not in the other). Thus materials are digitized by Google and added to books.google.com, and after some delay are added to the University of Michigan’s Mirlyn online catalog.

The “dual access” to the same materials via two systems (Mirlyn and books.google.com) makes it possible to get a somewhat better sense of what has been accomplished than when using books.google.com alone. According to information provided on the University of Michigan Web site, the Michigan Library has about 183,000 Russian titles, both monographs and serials. A recent article in the Detroit News revealed that Google is digitizing as many
as 30,000 volumes in a week at the University of Michigan and that at the present rate the project to digitize much of the library’s seven million volumes would be complete in five years. Since mass digitization assumes taking advantage of maximum efficiency in pulling items from the shelves for digitization, it appears that books in Russian are digitized and made available as the Google effort moves through the shelves, which are likely arranged by classification. To put it differently, books in English and in Russian in a particular classification range are digitized and made available at the same time.

Further details can be gleaned by using the University of Michigan online catalog. If a user goes to Mirlyn (http://mirlyn.lib.umich.edu/) and uses the Command Language search option and then enters \textit{wln=rus and wrd=``Michigan Digitization Project''}, he or she will get (as of April 16, 2007) 17,898 hits—signifying that something approaching ten percent of the approximately 183,000 Russian titles at University of Michigan have been digitized and made available (either in full-text or with “snippets”). Since there is a lag, presumably even more University of Michigan materials are available in books.google.com (more than 18,000 volumes).

The Google presentation of books seems to focus primarily on the 	extit{physical volume} rather than on a particular 	extit{title}. In other words, a multi-volume work will have separate “records” (the information found from the “About This Book” link is as close as one gets to a bibliographic record) for each physical volume and no “parent” record for the title as a whole. In some cases, it can be difficult to find individual volumes of a multi-volume work or to discern if the problem is that Google digitized some but not all volumes of a particular work. Figure 1 shows an example. Here we see the record for Volume Three of an 1843 multi-volume work. Further searching in Google does not turn up any other volumes, so (apparently) only the one volume was digitized.

This situation can be contrasted with the University of Michigan’s approach as it presents these same files from its own OPAC, which provides access from full Machine Readable Cataloging (MARC) records and for multi-volume works offers separate links to individual volumes from a holdings record. Figure 2 shows a page of links for individual volumes of a title in Mirlyn. This record deals with a complex situation in which twelve volumes are contained in
four physical volumes, according to the bibliographic record (although the digitized materials and holdings record seem to indicate that there are only three physical volumes). The holdings record also makes clear that volumes 9–12 were not digitized (or not yet available, at any rate). This is in contrast to a search result in books.google.com, where a user cannot see clearly whether particular volumes are available and must wonder if he or she has performed the correct search to retrieve the particular volumes.

Figure 1 also presents another aspect of the books.google.com search interface for Russian books that users should understand. The volume record ("About This Book") presents the title and author in Latin characters in transliteration rather than in Cyrillic (or in both Cyrillic and transliteration). As a result, it is possible to search for title and author keywords in transliteration. The search that retrieved the title shown in Figure 1 was done searching for the term slovar. By contrast, a search for словарь would retrieve all those items where that word appears anywhere in the full text. Note that generally diacritics that comply with Library of Congress transliteration are included, but in this case (Figure 1) they were not. In fact, there is not even representation there of the soft sign; this presumably is a quality-control issue rather than a policy decision. (With this sort of "black box" system it is difficult to know for sure.) Where diacritics are present in a books.google.com transliterated title, it is necessary to search using them in order to retrieve that term (except for instances where they were left out, of course). Unfortunately, it is not particularly easy for most users to generate these diacritic characters—a simple apostrophe, for example, is not the correct match for the symbol used in transliterating the soft sign. How are users to generate the needed ligatures for a search on this title shown in Figure 3? Given these complications with the transliterated titles and authors, it seems advisable to search in Cyrillic. (While the titles do not display in Cyrillic, the ranking of search results suggests that searches in Cyrillic in books.google.com system search on an index of title information in Cyrillic that gives items with those title keywords a higher rank in search results.)

Searching in Cyrillic is complicated by the way books.google.com deals with the changes in orthography that occurred in 1919. Since typical Russian users do not have readily available keyboards to generate pre-reform characters such as yat’, users should search for such characters with the modern equivalent characters (е in the case of the letter yat’). Characters that remained the same after the reform but for which usage changed should be searched using the pre-reform spelling. To take one example, a user would search русскаго rather than русскаго in order to capture items with that word that were published prior to the orthography reform. To capture items with that word published both before and after, it would be necessary to search both ways. This suggests that users need
a reasonable understanding of Russian orthography to get good results.\textsuperscript{13}

Test attempts to identify phrases in Russian in items and then to find that text via searching were often disappointing, although not enough searches were performed to provide a reliable statistical sample. The impression is that the Optical Character Recognition (OCR) of the Cyrillic text is not especially reliable. Searching for bound phrases, in other words, may not mean that the particular phrase is not in a book that has been digitized, but rather that the OCR has failed to allow the user to find all instances of that phrase.

Figure 4 gives some indication of problems with the Optical Character Recognition (OCR) for Russian. A search on \textit{пушкинъ} in the full text of \textit{A. S. Pushkin: trudy i dni} (remembering that \textit{Pushkin} in pre-orthographic reform Russian has a hard sign at the end) should demonstrate how well the google.books.com search engine works. It
quickly becomes clear that many occurrences of the search term are missed. (As a separate matter, one is reminded that with an inflected language it is important to search variants, such as пушкина. Searching that term was even more problematic than searching on пушкин alone, with books.google.com reporting the first occurrence on page 16; in fact it appears on more than half the pages with text that precede page 16.)

The above examples are for scanned text where the resulting images appear to a casual user to be of reasonable quality; thus the inability of the software to pick out variations on Pushkin is surprising. Referring back to Figure 3, clearly there is no OCR program that will be able to produce useful text from images of this quality. (It appears that a curvature correction program went awry in this case.) While that page is unusual in the severity of the problem, within that same book five of the first ten pages are so poorly reproduced that at least half of each page is unreadable by eye, much less recoverable by an OCR program.

A user might be hopeful that the partner libraries will provide better access. For example, the University of Michigan generally offers full access to the full text produced via OCR, in contrast to books.google.com, which hides this text from users. In addition, the PDFs made available by Michigan are not just images of the pages, as are the PDFs of items that are made available via books.google.com. Unfortunately at present for Russian materials, there is no full text at
all available for Russian-language items in Mirlyn’s presentation, and
the Mirlyn PDFs of Russian books are just images of the pages (as in
books.google.com). The “Search Within This Text” function does
not work for Russian in Mirlyn.

In the paper cited at the beginning of this article, Terence
Emmons identified several useful activities that he foresaw for
Russian scholars using books.google.com: finding quotes or passages
within works; locating copies of particular book titles; and keyword
searching that would identify books where those keywords were
present, with the ranking favoring books where the keyword(s)
appeared more often in the item. Given the problematic aspects of
searching books.google.com at the present time, it is difficult to have
much faith in the search results one gets for the first two of these
activities. Keyword searching to find books on a particular subject
seems likely to generate a relatively reliable result in the sense that
higher-ranked materials should contain more instances of a keyword
or keywords (and thus have relevance to that subject) but it also
seems as though many instances of a particular search term may be
missed. While it is still early and the amount of Russian content is
relatively low, it is not clear that the utility will improve unless
systematic changes are made to how the work is done overall. Will it
be possible in the future to improve the results of the OCR, for
example?14

For now, the results do not speak of a serious (and reliable)
research tool but rather a hit-or-miss sampling device that may
find what you are looking for or may not. (It may not find it
even though what you are looking for was digitized by Google and
its partners.) Given that it is unlikely that such an effort will be
repeated, it is disappointing at this point that the results seem so
poor.

NOTES

1. See Jill E. Grogg and Beth Ashmore, “Google Book Search Libraries and
Their Digital Copies,” Searcher 14, no. 7 (April 2007), http://www.infotoday.com/

2. This article was completed in early May 2007.


6. Google uses the 1923 cut-off for publications from the United States fairly consistently to determine whether an item is in the public domain; it is not clear what date is being used for materials published in Russia, other than that books published before 1900 all appear to be available full-text. Starting with about 1905, the Russian items are treated as not being in the public domain and only “snippets” are made available.

7. E-mail message received April 11, 2007, from the Google Library Center. Note that while the Library of Congress has partnered with Google for some book scanning activities, this has been compartmentalized within the Library of Congress such that I do not know the details of that effort.


9. While generally the Google efforts are described for these and some other libraries as including “all” the collections, it seems that Google focuses on materials that are readily digitized and thus does not digitize materials that are brittle.

10. The URL for Mirlyn, the University of Michigan libraries’ OPAC, is http://mirlyn.lib.umich.edu/


13. This approach, which is mechanical, can be contrasted with the Western bibliographic tradition, which has been to represent pre-reform orthography with separate transliterated elements, or with the Russian (and Soviet) approach, which has been to update the orthography to modern practices completely.