East/West
Journal of Ukrainian Studies

The Ukrainian Kyrylytsia, Restored: An Automation Project for Adding the Cyrillic Fields to Ukrainian Records in OCLC WorldCat

Jenny Toves, Roman Tashlitskyy and Lana Soglasnova

Volume 8, Number 2, 2021

URI: https://id.erudit.org/iderudit/1083564ar
DOI: https://doi.org/10.21226/ewjus626

See table of contents

Publisher(s)

Canadian Institute of Ukrainian Studies University of Alberta

ISSN

2292-7956 (digital)

Explore this journal

Cite this article

Toves, J., Tashlitskyy, R. & Soglasnova, L. (2021). The Ukrainian Kyrylytsia, Restored: An Automation Project for Adding the Cyrillic Fields to Ukrainian Records in OCLC WorldCat. *East/West*, *8*(2), 307–320. https://doi.org/10.21226/ewjus626

Article abstract

This report from the field concerns a collaborative project which resulted in successfully adding the Cyrillic fields to about 30,000 Ukrainian bibliographic records in OCLC WorldCat, the world's largest online catalogue. Historically, the Ukrainian records in English-speaking libraries were only provided in transliteration according to the Library of Congress Romanization Table. However, the current standards also require the original script, such as the Ukrainian Kyrylytsia. While automating the Cyrillicization of Ukrainian legacy records is theoretically straightforward, in practice it faced more than one challenge, from poor quality of transliteration to the historical changes in Ukrainian orthography. The report presents the OCLC Ukrainian Cyrillicization project and discusses the steps in its implementation as an example of a successful collaboration in the areas of bibliographic automation, Ukrainian philology and culture, Slavic cataloguing, and linguistics.

The Ukrainian *Kyrylytsia*, Restored: An Automation Project for Adding the Cyrillic Fields to Ukrainian Records in OCLC WorldCat

Jenny Toves
OCLC

Roman Tashlitskyy
University of Toronto

Lana Soglasnova
University of Toronto

Abstract: This report from the field concerns a collaborative project which resulted in successfully adding the Cyrillic fields to about 30,000 Ukrainian bibliographic records in OCLC WorldCat, the world's largest online catalogue. Historically, the Ukrainian records in English-speaking libraries were only provided in transliteration according to the Library of Congress Romanization Table. However, the current standards also require the original script, such as the Ukrainian Kyrylytsia. While automating the Cyrillicization of Ukrainian legacy records is theoretically straightforward, in practice it faced more than one challenge, from poor quality of transliteration to the historical changes in Ukrainian orthography. The report presents the OCLC Ukrainian Cyrillicization project and discusses the steps in its implementation as an example of a successful collaboration in the areas of bibliographic automation, Ukrainian philology and culture, Slavic cataloguing, and linguistics.

Keywords: Ukrainian language, Cyrillic script, automatic de-transliteration, Library of Congress transliteration, cataloguing, OCLC WorldCat.

0. The Background and Introduction

In Western scholarship, it is standard practice to transliterate the Ukrainian (and, indeed, all non-Roman script) names and terms, including the bibliographical information. The Anglophone countries mostly adhere to the American Library Association—Library of Congress Romanization Table (ALA-LC RT) for the Ukrainian language. In English-language cataloguing, best practices require for bibliographic records to include both

Volume VIII, No. 2 (2021) DOI: https://doi.org/10.21226/ewjus626

transliteration and the original script (Non-Latin Script Materials Affinity Group; "PCC Guidelines"). However, the legacy records in library catalogues often lack the original script, thus falling short of current standards. Figure 1 below shows an example of a bibliographic field in transliteration only (OCLC record No. 1114559628; MARC Field 245 "Title and Statement of Responsibility").

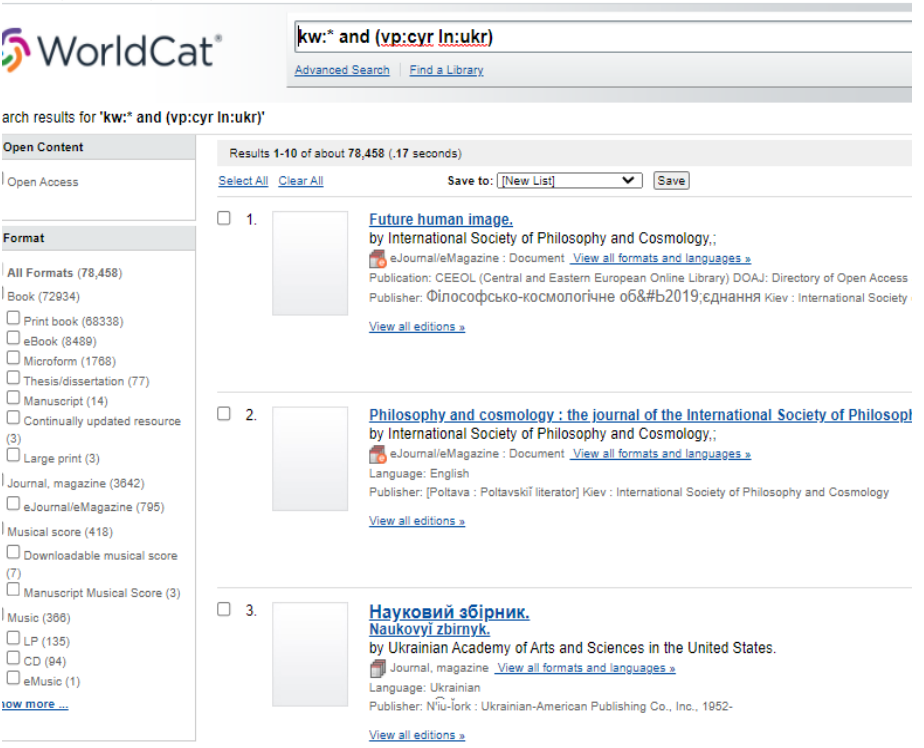
Figure 1. Bibliographic field "Title and statement of responsibility" in transliteration according to the Library of Congress Ukrainian Romanization Table ("Ukrainian [2011]").

245 10 Kruta arkhitektura: Superovi fakty dliā͡diteĭ - malykh i velykykh / Saĭmon Armstrong, pereklad z anhliĭs'koï Hanny Leliv.

OCLC WorldCat, "the world's largest catalog," contained about 2 billion items, according to its web site at the time of writing this report, in December 2020 (*WorldCat*). For the Ukrainian language, WorldCat contained 757,789 records for all types of materials: books, serials, maps, musical scores, visual materials, digital contents, etc., in many different languages of cataloging. Out of this total, only about 30,000 Ukrainian records with English as language of cataloguing contained the Cyrillic fields before our project was implemented. Precise extraction of such statistics, especially regarding the presence of the Cyrillic fields, was performed on OCLC raw data by the first author.[1] To estimate the number of materials in the Ukrainian language in WorldCat using its public interface may require some bibliographic prowess, namely, the use of OCLC Expert Search index labels ("Index Labels"). One may use the following search string in the search box in order to get all records with Cyrillic in Ukrainian: kw:* and (vp:cyr ln:ukr). As can be seen from the screenshot of the public interface of OCLC WorldCat provided in Figure 2, this combination of search terms yielded 78,457 records. Specifically, the search term kw:* retrieves records with at least one character (*) in the Keyword index (kw), which are also ('and') indexed in the Language index (ln) as being in the Ukrainian language (ukr) and in the Character set index (vp) as having Cyrillic characters present (cyr) ("Character Sets"). While these numbers may not be as accurate as the results obtained with raw data manipulation, one can still get a general idea about the scope of representation of Ukrainian materials in WorldCat, as well as monitor the changes.

[1] J. Toves. [Calculated from a research copy of the data from worldcat.org] [Unpublished raw data]. OCLC.

Figure 2. OCLC WorldCat "Advanced Search" options to retrieve Ukrainian language records.[2]



The Cyrillic fields can be added to legacy records retroactively in order to improve the quality of library catalogues and access to library materials. Such retroactive enhancement, also known as Cyrillicization, or de-transliteration, can be performed automatically through programmatic matching of Roman characters to their Cyrillic correspondences in accordance with the Library of Congress Romanization Table. Jacobs and others reported a retroactive enhancement of 13,099 records for Russian language materials at the Queens Borough Public Library with a de-transliteration program named "Cyril," which was specially written in Perl programming language. Using an updated version of Cyril, the freeware MarcDeTrans freely available online, in 2011-13 the Slavic Section of the University of Toronto Libraries also implemented a pilot retroactive

2 https://www.worldcat.org/search?q=kw%3A*+and+(vp%3Acyr+ln%3Aukr). Accessed 25 Jun. 2021.

conversion for a few thousand Russian and Ukrainian records in the library catalogue (Summers). While for the Russian-language records that project delivered acceptable results with selective, manual quality control, for the records in the Ukrainian language the quality could not be maintained due to a high number of transliteration errors (a more detailed discussion of transliteration issues for the Ukrainian language follows below).

On a global scale, OCLC Research has recently launched a project "Kirillitsa v WorldCat" (Toves et al.). The first step in the project was to Cyrillicize, in collaboration with UCLA Metadata and Cataloging, the Russian language records, with an outcome of "about 958,000 records in the Russian language in WorldCat, representing 3,7 M holdings" retroactively enhanced with Cyrillic fields (Toves et al.). In March 2020, a call to "help OCLC with their Ukrainian Cyrillic auto-processing" was issued on the SlavLibs and SlavCats, two mailing lists for Slavic librarians and Slavic cataloguers (Fletcher). Over a dozen Slavic librarians responded as volunteers to participate in the project as reviewers. In addition, the three authors worked as a team on specific strategies to create a good set of records for Cyrillicization. Over the summer of 2020, several sample sets of enhanced records were sent to reviewers to assess the quality of Cyrillicization and identify the errors. As a final result, over the Labour Day weekend (4-7 Sept. 2020), the Cyrillic fields were added in a batch process to 28,333 Ukrainian records in OCLC WorldCat, which nearly doubled the number of Ukrainian records with Cyrillic fields to the total of 58,928. This report describes how the team did it and what remains to be done.

1. Criteria for Cyrillicization for Ukrainian Records in WorldCat

The records to be Cyrillicized had to meet quality criteria. These were of several kinds.

One set of criteria was based on bibliographic standards. To begin with, only the full-level records for physical materials created according to two current standards, Resource Description and Access (RDA) and Anglo-American Cataloguing Rules, Version 2a (AACR2a), were included. Also, the only language of cataloguing considered was English. Numerically, that left out 66,717 non-AACR2 and non-RDA records, as well as 370,269 records for digital assets, and 131,436 records in languages of cataloguing other than English, also excluding most vendor records (usually at less than full level of cataloguing).

The Library of Congress Romanization Table for the Ukrainian language is based on the Ukrainian orthography according to the orthographic reform adopted in 1933 by the government of the Soviet Ukraine (Maznichenko et al.; "Ukrainian [2011]"). For that reason, the chronological range for this first

stage of the project was limited to publications after 1933, which excluded about 18,000 records. OCLC's WorldCat contains a significant number of records for publications produced outside of Ukraine, especially by the Ukrainian diaspora in Canada, the United States, Germany, and other countries. However, most émigré and diaspora communities did not adopt the 1933 orthography until its post-Communist revisions, which represented an additional challenge for the project. Our working solution at the first stage of the process was to limit the geographical scope of the project to publications from Ukraine. The date and place of publication of Ukrainian bibliographic records are examples of bibliographic criteria determined by sociolinguistic and cultural considerations, requiring the knowledge of Ukrainian cultural history, especially the history of codification of the Ukrainian language. Importantly, Ukraine as a country of publication covers some publication places, such as Lviv, with a complex political history. Our working solution (in particular, regarding the letter "ґ") was to take into account such places of publication; it is discussed below in Section 3.1.

Determining the criteria based on transliteration quality was a significant part of this project. Practising cataloguers are aware of some common errors in transliteration of Ukrainian records, and the apprehensions of two cataloguers among the authors were confirmed and expanded upon reviewing the first set of samples, also supported by the input from fellow reviewers. Table 1 summarizes the most common errors in Ukrainian transliteration and the resulting errors in Cyrillicization. Not only are these errors common, but they also occur with very high frequency.

The low quality of transliteration constituted one of the main challenges of the project. Cyrillicization relies on transliteration accuracy in order to generate a correct string of Cyrillic characters. The strategy is to apply automatic Cyrillicization only to "good" records which are free of errors, particularly with respect to transliteration. As the examples in the table demonstrate, if a wrong Roman character is used in error, it will also be Cyrillicized as the wrong Cyrillic character, or even as garbage. For example, there are two common errors in transliterating the Ukrainian "ї": instead of keeping the character as "ї," different characters are used: namely, "ĭ" (i with the diacritic "breve"), or "ǐ" (i with the diacritic "caron"). When encountering these errors, the Cyrillicization program will produce either the wrong character "й" for "ĭ" (according to the Ukrainian LC RT), or the garbage sequence и&#x0220E for "ǐ" (because "ǐ" is absent from the Ukrainian LC RT). Many errors listed in Table 1 are likely due to language interference with the Russian Romanization Table (for example, there is no character "ї" in Russian, only "й"). Overall, the high number of transliteration errors in the Ukrainian records led to a high number of errors in Cyrillicization in the initial set.

Table 1. Common errors in Ukrainian transliteration (ALA-LC Romanization Table).

Ukrainian character in Cyrillic	Library of Congress Romanization (transliteration)	Transliteration errors	Note
г	h	g	Wrong character
		g Cyrillicizes as ***ґ***	
є	i͡e (with a non-spacing ligature)	e, ie	Wrong character, missing diacritic
		e Cyrillicizes as **е** **ie** Cyrillicizes as **іе**	
ж	źh (with a non-spacing ligature)	zh	Missing ligature
		zh Cyrillicizes as **зг**	
и	y	i	Wrong character
		i Cyrillicizes as **і**	
ї	ï (i-diaeresis)	i, ĭ, ǐ (i-breve, i-caron)	Wrong characters
		ĭ Cyrillicizes as **й** **ǐ** Cyrillicizes as garbage и&#x0220E	
й	ĭ (i-breve)	ǐ (i-caron)	Wrong diacritic
		ǐ Cyrillicizes as garbage и&#x0220E	
ц	t͡s (with a non-spacing ligature)	ts	Missing ligature
		ts Cyrillicizes as **тс**	
ь	' (letter prime)	‘ (apostrophe)	Wrong character
ю	i͡u with a non-spacing ligature)	iu	Missing ligature
		iu Cyrillicizes as iy	
я	i͡a with a non-spacing ligature)	ia	Missing ligature
		ia Cyrillicizes as ia	

A working list of errors, such as Table 1 shows, was compiled. Our goal was to come up with the strategies to identify the errors in order to create an acceptable pool of records for Cyrillicization, usually by excluding the

ones that contain errors, or, in some cases, by fixing the error if possible. It is important to note that those empirical, data-driven strategies are not spelling correction algorithms, nor natural language processing modules. We were looking for uncommon patterns that are likely to signal an error. The main strategies are discussed in Section 2 below. Section 3 presents a discussion of selected transliteration errors, namely, the letter "ґ" vs "г"; the letter "ж"; and the soft sign "ь" vs the apostrophe '. Overall, about 19,000 records were excluded for which the patterns identified as suspicious could not be corrected.

2. Strategies for Error Identification and Correction

2.1 Generally, records that contained characters absent from the Library of Congress Romanization Table would be skipped. That strategy excluded such transliteration errors as "ǐ" for "ї" or for "ï**.**"

2.2 The Vowel Combinations Table

A table listing vowel combinations in the data was compiled in order to identify the more common vowel combinations. This allowed identification of the less common patterns likely signaling error. For example, words ending in the vowel combinations "oǐ" and "oi" were identified as errors and excluded from Cyrillicization, since they usually were a common transliteration error of the Ukrainian genitive singular feminine adjectival ending "*-oï*." As well, a particularly common error is the missing ligatures. In cataloguing, the Romanization table requires ligatures over the digraphs representing the "soft" vowels "є," "ю," "я," in order to distinguish them from combinations of simple letters. For example, "є" is transliterated as i͡e, and "я" is transliterated as i͡a. When transliteration omits a ligature in error, this will inevitably lead to Cyrillization errors, as demonstrated in Table 1 above: for example, "ie" will be Cyrillized as "ie" (instead of "є"), and "ia" as "ia," not "я." The vowel combinations without ligatures were examined, and, when identified as wrong, excluded from Cyrillicization (many were in fact words from Russian parallel fields, e.g., the word "vossoedinenie"). For some words, corrections were made, for example, 'komediia' was changed to 'komediiâ*,*' or 'derzhȁvotvorennia' to 'derzhȁvotvorenniâ*,*' based on the analysis of the list of two thousand most common words, as described in the next section.

2.3 The List of 2000 Common Words

Another strategy was to compile a list of 2,000 most common words occurring in Ukrainian bibliographic descriptions. Our hope was that the list would allow for some limited cleanup of transliteration to increase the number of records that get Cyrillic text. The fields for titles, statement of responsibility, edition, publication information (including place and publisher), series, the general note, and the contents note were examined (MARC fields codes '245', '246', '250', '260', '264', '362', '490', '500', '505'). Most of the fields that were transliterated were "Title and Statement of Responsibility" (MARC Code 245) and the publication information (MARC fields codes 260 and 264). The lists provided ample examples of transliteration errors.

All the words in the lists were manually examined by the second author, with some assistance from the third author, and the correct forms were identified. For many words, no variants were present, for example, for the most common word та=*ta* 'and'. In many cases, as expected, in addition to the correct transliteration, variant transliterations with errors were found. Usually there was more than one way to make an error. An absolute anti-champion in that category turned out to be the word 'української = ukraïns'koï' 'Ukrainian' (the genitive singular case, feminine gender). In Cyrillic, the word 'української' contains the soft sign "ь," commonly mis-transliterated as an apostrophe or even omitted, and two occurrences of the character "ї," commonly mis-transliterated in at least three different ways. Courtesy of various combinations of these common errors, the transliteration 'ukraïns'koï' was attested with errors in twenty-five (!) different ways.

For cases where the variants stood for a single correct form of the word (e.g., 'ukraïns'koï'*)* all transliteration errors were corrected before Cyrillicization was applied. Finally, in some cases two different forms were both possible, usually corresponding to different grammatical forms of the word, e.g., tradytsiï 'traditions' in the nominative plural case and tradytsiĭ 'traditions' in the genitive plural case. In some cases, where the less common variant only occurred a few times, the correct spelling could be confirmed through manual examination, e.g., 'shkil' ('school', in the genitive plural case), 'Shkil'' (can be a relatively common surname); or 'im.' (abbreviation of 'imeni') vs. 'ïm' (can be a verb 'eat' in the first person singular, present tense, or a pronoun 'they' in the dative case). However, in other cases both variants occurred with high frequency, e.g., 'tradytsiï' 'traditions' in the nominative plural case and 'tradytsiĭ' 'traditions' in the genitive plural. In absence of reliable computational linguistic routines to automatically confirm the correct spelling, such records were skipped.

2.4 Personal Names and Proper Names

Personal and proper names in the Ukrainian language constitute a distinct category of nouns challenging for automatic analysis due to some grammatic and stylistic peculiarities. For example, Valeriĭ 'male personal name in the nominative singular case or female personal name in the genitive plural case' and Valerii 'female personal name, in the genitive singular or nominative plural cases (both feminine and masculine)'; Mykhaĭla, Mykhaïla (genitive or accusative singular; the second name can be an older version used in religious literature or a graphic representation of the Russian name "Михаил"). Other patterns include male names ending in "-iĭ" in the nominative singular case, with the "-ii" ending in the nominative plural case, which makes both versions correct. Finally, even the proper name Ukraïny 'Ukraine' (the genitive singular case), which occurs in Ukrainian records with high frequency, can sometimes be used as Ukraĭny 'Ukraine' (the genitive singular case) in poetic style. In order to avoid erroneous auto-replacements, such cases should be excluded or considered individually.

3. Ukrainian Transliteration Errors: Three Case Studies in Cyrillicization

3.1 The letters "ґ" (g) and "г" (h)

According to ALA-LC RT, the Ukrainian letter "г" is transliterated as "h," and "ґ" as "g." However, we noticed that the letter "g" ("ґ") occurred much more often than it could be expected in Ukrainian records. In fact, the letter was discarded from the Ukrainian alphabet in 1933 and was rehabilitated only in 1989 (Maznichenko et al. 7). (Historically, the Ukrainian orthography adopted in 1933 was used until 1989, with a few minor revisions in the 1930s, 1946, and 1960 [Maznichenko et al.]). Therefore, the Ukrainian-language records from the Soviet period were very unlikely to contain this letter. It could only occur in Russian colophon titles, which contained the required Russian translation of all non-Russian titles published in the Soviet era. The intuition proved to be justified, as the majority of "g"s in records between 1933 and 1989 turned out to be either from Russian colophon titles or simply transliteration errors, used mistakenly instead of "h" for the Ukrainian "г."

As mentioned in Section 1 above, our working solution was to limit the chronological scope of the records to include only the materials published after 1933, and to limit the geographical scope to Ukraine as the country of publication. However, such geographical scope also covered the titles published after 1933 outside of the territory of the pre-World War II Soviet Union, where the pre-1933 orthography was still commonly used. To

address this challenge, the following rule was implemented: if the date is <= 1939 and the field 26x$a (place of publication) matches a regular expression including the most common places of publication in non-Soviet Ukraine (such as Lviv, Kolomyia, Chernivtsi, etc.), then a "g" was allowed to occur. This allowed Cyrillicization of records for books published before the beginning of World War II in these important centres of Ukrainian publishing. Yet, even with all the mentioned limitations in place, we still could not be completely confident that the letter "g" ("г") was in its right place because inside Ukraine within the abovementioned timeframe the situation at times was hectic, and the use of the post-1933 orthographic rules could not be guaranteed. Notably, during World War II a substantial part of Ukraine was occupied by Germany for some time and the materials published there were not controlled by the Soviet government. That said, we still had to be very cautious while applying transliteration to those records and make sure we manually reviewed any unusual occurrences in the records (such as the letter "g" ("г") in locations within words contradicting patterns met in most other records).

Another occurrence of "g" in transliteration turned out to be the pattern of the Ukrainian adjectival ending "-oho" (masculine, genitive singular case or accusative singular case [with animate nouns]) mistakenly represented as "-ogo," which turned out to be quite prevalent. Since letter "г" is not typically found in this ending in Ukrainian, it was decided that "-ogo" always signifies a Russian word and the record containing it should be skipped.

3.2 The Letter "ж" (ẑh)

Besides challenges of sociolinguistic and cultural character, we encountered some issues caused by the interaction of certain transliteration conventions of the Library of Congress and peculiarities of Ukrainian grammar. For example, according to the Library of Congress transliteration table, the letter "ж" is rendered as a combination of letters z+h with a ligature (zẖ). At the same time, if a ligature is omitted, the z+h combination conveys simply з+г. Erroneous omission of ligature was commonplace in plenty of records and it was important to determine which of the z+h combinations needed a ligature, especially taking into account the fact that numerous Ukrainian prefixes ending in "z," with the following stem beginning with "h," constituted a substantial part of Ukrainian vocabulary (as in words "розгадати," "розганяти," etc.). The solution could be found in keeping numeric track of patterns, where a big number of the same pattern usually meant it was correct, while rare patterns were reviewed one by one to confirm or correct them.

3.3 The Soft Sign "ь" (') vs the Apostrophe ‘

The use of apostrophe and the soft sign in transliteration constituted another major challenge. In ALA-LC transliteration, the soft sign "ь" is rendered as a special character ' ("the prime"), which very much resembles the apostrophe character ‘. Historically, the apostrophe ‘ was indeed often mistakenly used in transliteration instead of the prime character ' for soft sign in many records, especially the older ones. This common transliteration error would lead to a high number of errors in Cyrillicization. In order to clean up the data, we considered a set of rules for an automatic algorithm to clearly distinguish between the actual apostrophe and an apostrophe used mistakenly instead of the soft sign character '. The orthographic rule for the apostrophe in Ukrainian is fairly straightforward: an apostrophe is used after letters b, p, v, m, f, and r before ia͡, iu͡, ie͡, ї; also, an apostrophe can never be followed by a consonant nor occur at the end of the word. The rules helped to correct the wrongly used apostrophes to soft signs in numerous cases, such as ‘na͡tsional'na=національна’ (from the wrongly transliterated ‘na͡tsional’na,’ which would Cyrillicize in error as ‘національ’на’). However, there were also many exceptions which were difficult to handle automatically. In particular, all proper nouns needed to be manually checked (including names of persons, place names, etc.), which was accomplished by setting aside all apostrophe-containing words that begin with an upper-case letter, unless they occurred as the first word in the title subfield.

Concluding Remarks

Figure 3 below shows an example of an automatically Cyrillicized bibliographic field generated from the transliterated field in a Ukrainian language publication (OCLC No. 796946868[3]). In addition to Field 245 "Title and statement of responsibility" in transliteration, a new, paired field has been created containing the title and the statement of responsibility in the Cyrillic script (MARC code 880). All Cyrillicized records in this project also automatically receive a note (MARC code Field 588) which reads "Non-Latin script generated programmatically."

[3] http://worldcat.org/oclc/796946868. Accessed 25 Jun. 2021.

Figure 3. A WorldCat record with auto-detransliterated Cyrillic fields in Ukrainian.

245 10 6880-01 Vcheni Ukraïny --laureaty mizh̃narodnykh premiĭ i nahorod / Vitaliĭ Ablit͡sov.
588 Non-Latin script generated programmatically.

880 10 6245-01 Вчені України --лауреати міжнародних премій і нагород / Віталій Абліцов.

The outcome of this project is not only the ability to display the bibliographic data true to its original representation. Being able to retrieve metadata in the original script is also essential to making data available to the communities it serves. While current cataloging practices call for parallel data in the original script and Latin transliteration, there is still a large backlog of older records that are missing script data. As our discussion shows, while it seems easy to simply un-transliterate the Latin back into Cyrillic, there are, in fact, many problems. Beginning with the technical side, the most common problems are missing or wrong diacritical marks. The reasons for this are many. Older systems may have made it impossible or difficult to enter the marks. Also, sometimes marks are lost when records are transferred between systems. Since text fields normally have diacritics normalized away for searching, the errors are invisible if you rely on the Latin text. However, they become immediately glaring once Cyrillicization is attempted. Detecting problems with marks and problems with using the wrong transliteration scheme (i.e., using the Russian Romanization Table instead of the Ukrainian Romanization Table) is a requirement for generating accurate Cyrillic text. Avoiding those records allows us to re-visit them later when our techniques have improved.

From the socio-cultural perspective, the project specifically, and largely successfully, focused on publications from Ukraine as the first step. However, many centres of Ukrainian life in diaspora were quite prolific in publications during the years of the Soviet rule in Ukraine. The growing interest in the Ukrainian émigré publications among scholars make tackling the Cyrillicization issues in records for materials from outside of Ukraine (especially from cities like Munich, Baltimore, Toronto or Winnipeg—well-known centres of Ukrainian cultural activity in emigration) a potential focus for future projects. Understanding the cultural context for metadata and the work described is essential to understanding the transliteration present in records. Where works are published, when and where the metadata is created—all of these factors must be studied to produce the best possible Cyrillic data.

Future Work

From June 2019 to November 2020, the number of records with Cyrillic text in WorldCat increased from 1.6M to 3M—almost double![4] This is the result of two passes at Russian language records and one pass over Ukrainian language records. OCLC Research is continuing to work with Slavic language experts to add Cyrillic text to WorldCat. Other languages such as Bulgarian are under review at this time. In addition to the ALA-LC Romanization Tables for English as the language of cataloguing, other transliteration schemes for other languages of cataloging are being explored. At this stage, the focus is on identifying questionable Latin data and skipping those records so we can focus on quick wins with the most promising data, as successfully demonstrated by the Ukrainian Kyrylytsia project. Later stages of the project will focus on ways to handle poorly transliterated text.

[4] Toves, J. [Calculated from a research copy of the data from worldcat.org] [Unpublished raw data]. OCLC.

Works Cited

"Character Sets Present." *OCLC,* 13 May 2021, https://help.oclc.org/Librarian_Toolbox/Searching_WorldCat_Indexes/Bibliographic_records/Bibliographic_record_indexes/Indexes_A_to_C/Character_Sets_Present. Accessed 25 June 2021.

Fletcher, Peter. "Ukrainian Help Needed by OCLC." *SlavLibs / SlavCats*, 11 June 2020.

"Index Labels and Examples of an Expert Search in WorldCat." *OCLC,* 29 Jul. 2020, https://help.oclc.org/Discovery_and_Reference/FirstSearch/Search/Expert_search_in_WorldCat_indexes/Index_labels_and_examples_of_an_expert_search_in_WorldCat?sl=en. Accessed 25 June 2021.

Jacobs, Jane W., et al. "Cyril: Expanding the Horizons of MARC21." *Library Hi Tech*, vol. 22, no. 1, Jan. 2004, pp. 8-17. DOI: 10.1108/07378830410524459.

Maznichenko, Ie. I., et al., editors. *Ukrainskyi pravopys*. Naukova dumka, 2019.

Non-Latin Script Materials Affinity Group, ALA ALCTS CaMMS Committee on Cataloging: Asian & African Materials. "Linked Data for Production: Pathways to Implementation (LD4P2) Survey on Romanization: Report." *ALA Connect,* 2 March 2020, http://connect.ala.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=4199970e-0b71-4182-a5fc-1d53c1731458. Accessed 25 June 2021.

"PCC Guidelines for Creating Bibliographic Records in Multiple Character Sets." *Library of Congress,* 28 June 2016, rev. 7 Sept. 2017, https://www.loc.gov/aba/pcc/bibco/documents/PCCNonLatinGuidelines.pdf. Accessed 25 June 2021.

Summers, Ed. "MARC-Detrans-1.41 - De-Transliterate Text and MARC Records." *Metacpan.org*, 16 Nov. 2009. https://metacpan.org/dist/MARC-Detrans. Accessed 15 Dec. 2020.

Toves, Jenny, et al. "Kirillitsa v WorldCat." *Hanging Together: The OCLC Research Blog,* 15 April 2020, https://hangingtogether.org/?p=7868. Accessed 15 Dec. 2020.

"Ukrainian (2011)." *Library of Congress ALA-LC Romanization Tables.* https://www.loc.gov/catdir/cpso/romanization/ukrainia.pdf. Accessed 15 Dec. 2020.

WorldCat. https://www.worldcat.org. Accessed 15 Dec. 2020.