Tracking the History of Romani Publications:

Challenges Presented by Flawed Data

Geoff Husic[1]

Abstract: Romani is a language of northern Indic origin spoken natively by an estimated 2.5 million people, primarily in Eurasia but also in North America. The history of publication patterns in Romani has not been well documented. Extracting data about this history based on available information in large bibliographic databases such as OCLC WorldCat has been hampered by unfortunate misapplication of certain language codes, making it all but impossible to efficiently filter search results using Romani language as a parameter. The author discusses how he was able to correct much of this inaccurate data in OCLC WorldCat.

Keywords: Romani (Romany) language publications, OCLC WorldCat, Cataloging databases, Language codes.

The history of publication in the Romani language or on the topic of Romani has not been very well documented.[2] This holds true especially for Internet publications, but

[1] Slavic & Near East Studies Librarian. BA Russian and German (Middlebury College), MA Slavic Languages and Literatures (University of Kansas), MS Library and Information Science (University of Illinois). Room 519, Watson Library, University of Kansas, 1425 Jayhawk Blvd, Lawrence, KS 66045-7544. (husic@ku.edu).

is also true for more traditional paper-based publications. As the Internet has been embraced as a convenient means for publishing and blogging, it has become a boon for linguistic minorities, such as the Roma, that wish to make information about their cultures and languages better known to the world at large. It has been able to serve as a convenient venue for publishing news, cultural information, literature, blogs, and chat room content in lesser-known languages in a way that would have been politically or economically unfeasible in the pre-Internet print world. One good example of this phenomenon can be found in Wikipedia, where much excellent information about lesser-known languages has been entered, curated, and edited by those who have obvious interest in and affection for these languages and the cultures of their speakers. I have been monitoring the development of Romani-language Web publishing over the last ten years in this very context. Due to the nature of Romani, it can be written in a variety of dialects and rather chaotic writing systems, so that discovering Romani sources on the Web can be challenging.

Because of the rapid movement of many aspects of publishing to the Internet, it seemed to me a good time, as a companion project, to try to create a thorough retrospective bibliography of materials published about the Romani language, or in any of the several Romani dialects, that have appeared in print in the last 300 or so years. By necessity this bibliography, when completed, will primarily index materials on the whole book or journal level. Unlike the Web, in which data is, for the most part, still very unstructured, library bibliographic databases are universally based on very structured data

---

[2] *Romani* is Library of Congress spelling of the language, but is more conventionally spelled *Romany* in English.

schemes. Theoretically this should make it very easy to extract bibliographic data concerning the publication history of a particular language, or, in my case, Romani.

However, the path to even beginning an analysis of Romani publications has been somewhat complicated. The database, based on which I wished to conduct the analysis, was OCLC WorldCat[3], the bibliographic database used by the majority of North American academic libraries for cataloging and reference purposes. Its use has also steadily expanded to libraries worldwide. Bibliographic records in this database are encoded in the MARC format, which allows for very granular encoding of information for each bibliographic entity it represents. These entities most typically represent texts (books, journals, manuscripts, etc.) but can also be scores, maps, computer files, sound files, etc.

Catalogers, who use this database for cataloging locally held library materials, the records of which then get uploaded to their library's catalog, usually do their work through a technical-services interface called Connexion. Some libraries employ another method, where they create records in their local library catalog client, and then upload their records back to OCLC WorldCat. The latter libraries, when cataloging 'copy' records (those that already have some kind of record in OCLC WorldCat and which can vary widely in quality and fullness), may choose to make corrections and enhancements in their local catalog only. This sometimes makes sense from a workflow perspective, but it does not benefit other libraries, that subsequently need to use these bibliographic records, if errors have not been corrected in the WorldCat master records. As a result of

[3] WorldCat has a variety of public and cataloging interfaces. The client version used by libraries for cataloging purposes is called Connexion.

this practice, some libraries will have somewhat more accurate data in their local catalogs then was originally imported from WorldCat.

Among the kinds of information encoded about each book or journal are the language or languages represented in the work. Each languages represented is encoded using OCLC's three-letter language designation, on which the ISO *Codes for the Representation of Names of Languages* is also based.[4] These codes are added to a dedicated portion (the fixed-field language field) of the MARC record and can be used in WorldCat and other local library catalogs, into which MARC records have been uploaded, to limit catalog search results by language. For items that are multilingual, there is a additional MARC field, the 041 field, in which further language information can be recorded, such as multilingual texts, original language of translations, languages of summaries, etc.

Most of the common language codes are transparent and easy to remember, e.g. **eng** for English, **rus** for Russian, etc. Occasionally, some codes for less-commonly encountered language must be looked up by catalogers, who can't be expected to have memorized all of the thousands of language codes available. However, due to oversight by some catalogers, a situation had developed in WorldCat that had eroded the ability to identify Romani-language materials in the database. This is the essence of the problem: The official OCLC language code for Romanian is **rum**. This code was chosen because at the time these codes were established, the spelling *Rumanian* was still the more common spelling in English. The spelling *Romanian* became more common starting in

[4] *See* http://www.loc.gov/standards/iso639-2/php/code_list.php for a full list of these language codes.

the late 1960s and is now the standard in English. Understandably, most libraries are much more likely to encounter and catalog materials in Romanian than in Romani. What has occurred over the years is that many libraries, when cataloging Romanian-language materials in the OCLC WorldCat database, have been miscoding Romanian (language code **rum**) with the language code **rom**. However **rom** is actually the language code assigned to Romani. These coding errors have resulted in many Romanian records being miscoded as Romani. As there are obviously several magnitudes more published works in Romanian than Romani, this has made the task of extracting information about Romani publications all but impossible.

In late 2011, I contacted OCLC to alert them to this problem and to solicit their cooperation in correcting it. I informed them that my goal was to extract information from the database about Romani materials in order to construct a thorough retrospective bibliography and chronology, as well as my desire, as a cataloger, to see these coding errors corrected. OCLC technical staff was very cooperative and eager to help. They provided me with an initial spreadsheet of all records in OCLC that had the language code **rom** (Romani), either in the MARC language fixed field or the MARC 041 field. This spreadsheet was helpful for getting an initial overview of the scope of the problem.

For a variety of reasons, I soon decided to abandon the spreadsheet approach to scrutinizing the data and to do my work directly through the OCLC Connexion interface. The main impediment was that the spreadsheet bibliographic records also included hundreds of duplicate records for many books based on cataloging done by mainly European national libraries. In these records, the language of cataloging, i.e. the description, notes, subject headings, etc., are not in English but rather in the language of

the national cataloging agency. I felt that it was unmanageable for me to attempt to correct all these records. A recent enhancement to the OCLC Connexion software allowed me to easily limit to bibliographic records produced by English-language cataloging agencies. These are the records that will be used by North American and British libraries, so I felt my efforts were best placed in correcting these.

Sifting through and correcting the number of records that needed to be reviewed (over 2600) required being familiar with Romanian, Romani, and a number of other languages. Fortunately, as a specialist in Eastern European languages, being proficient in Romanian and very knowledgeable in several Romani dialects, I was eager to lend my assistance to the task. Scouring through the records encoded **rom**, I was able to eliminate records that clearly had some Romani content and thus eliminate them from the problematic set. I had to scrutinize the remainder with care. While the problem originated in the confusion of the language codes for Romanian and Romani, there are in fact many works that contain both Romanian and Romani content, so I needed to assure I didn't eliminate works purely based on, say, a Romanian title. Ultimately I ended up with approximately 1400 items that required further scrutiny. I then downloaded the OCLC records for these items so that I could view the full bibliographic information, such as subject headings to help me ascertain the content.

The following is a brief overview of the kinds of errors I identified when examining the problematic set: Approximately 1200 were actually **Romanian** language materials that had been miscoded as **rom**. Forty or so were **Romansh** language materials, the correct code for which is **roh**. In addition, there were quite a few other oddities, such as **rom** being coded for Latin texts, or apparently as a result of confusion with the place

of publication, such as an example of an English text published in Rome. Several dozen more were actually texts in Hungarian and other languages, incorrectly coded **rom**, that were printed in Romania. In these cases, a cataloger, unfamiliar with the languages, presumably extrapolated the incorrect language based on the place of publication. Or perhaps they were artifacts from an automatic conversion project. A small number were coded **rom** because a cataloger apparently thought this was the proper way to indicate something in the roman script. Finally, in a few cases, not only was the record coded with **rom** but the record also contained textual language notes such as "In Hungarian and Romanian." This is rather curious and perhaps was the result of automatically generated notes based on the fixed field or 041 languages codes. It is difficult to tell for certain.

In those cases where I found the language code **rom** to be incorrectly assigned, as a cataloger in an OCLC Enhance Program library, I was able, in most cases, to correct the OCLC WorldCat master record to reflect the correct language.[5] In many cases I also fix incorrect language notes as appropriate. There were quite a few cases where I was unable to correct the codes. These were mainly cases in which there was other incorrect coding in the records that made it impossible to update the master record. Some errors also appeared in records for music scores. I am not familiar with the scores format, so I was reluctant to make corrections to these records for fear of causing unforeseen problems.

The bulk of this project has now been completed. Users of WorldCat will now be able to filter results using the Romani language as a search parameter much more

---

[5] The Enhance Program allows qualified member libraries to correct and added additional information to bibliographic records in OCLC WorldCat.

reliably, if not yet perfectly, than before. There are a number of items I will need to contact OCLC or other libraries to fix in the master records. I intend to monitor new items added to WorldCat periodically in order to catch new incorrectly coded records that are sure to be added. I would encourage other WorldCat users to also correct these mistaken language codes, especially for minority language, such as Romani, when encountered. A few major academic libraries have also made corrections in their local catalogs based on my personal communications with them about this issue.

Now begins the hard part! Having cleaned up much of the data in WorldCat, I have begun to examine how best to extract the data. I will likely import the data either into Endnote (a citation management program), Zotero (a bibliographic tool plugin for the Mozilla Firefox browser) or work with both tools. There are certain character encoding issues that occur when importing bibliographic data into these software tools that will also have to be addressed to minimize the amount of manual editing I must do. When the majority of the data is imported satisfactorily, I can begin to correct any additional errors and add additional useful metadata as well as my annotations. I can then finally begin an analysis of the actually publication patterns based on time period, dialect, place of publication, genre, and other parameters of interest. Results will be published upon completion in a venue to be determined.