The Romani Language:
Cataloging Ramifications for a Language in the Process of Standardization

GEOFF HUSIC

Preprint: Final edited version is published in: *Slavic and East European Information Resources.* Volume 12, Issue 1 (January 2011), pages 37-51: DOI: 10.1080/15228886.2011.556076).

Abstract

A discussion of issues related to the cataloging of a language, Romani (or Romany), which is only in the 21st century beginning to achieve some degree of standardization. The discussion focuses on issues of Romani orthography, specifically a small number of unusual Unicode characters that may cause technical problems in certain automated cataloging environments, such as OCLC WorldCat, the OCLC cataloging client Connexion, and online library catalogs.

Keywords: Romani (Romany) language, Cataloging, MARC, Unicode, Numeric Character References

Introduction

The discussion of cataloging foreign language materials has been recently invigorated by a number of developments. The ability to add matching non-latin[1] vernacular fields in bibliographic records in online library catalogs for a variety of languages (e.g. Japanese, Arabic, Chinese, Korean, Persian, Hebrew, Yiddish, Greek, and some Cyrillic languages[2], i.e. the so-called JACKPHY languages) has somewhat ameliorated the problem of the end-user's lack of acquaintance with the romanization schemes used in library catalogs and other databases, as he now has, in some cases, the option of searching either in transliteration or in the native script. In addition, especially among the languages of the former Soviet Union, there have been changes in the orthographies of several languages that have abandoned Cyrillic in favor of romanized scripts (e.g. Moldavian, Azeri, and most recently Uzbek), which have certain ramifications for cataloging and retrieval in library catalogs. In some cases the change is unproblematic, for example, when abandoning Cyrillic the Moldovans, whose language is basically identical to Romanian, just chose to return to writing in standard Romanian, while

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

preserving the token Moldovan language designation. In the case of Azeri and Uzbek, complications arise, in that the new romanized alphabets diverge somewhat from the ALA romanized transliteration schemes for the Cyrillic that have been employed heretofore in library catalogs and many databases. In some cases this has resulted in the need to address the Library of Congress guidelines for supplying uniform titles in cases of changes in orthography that affect alphanumeric filing and retrieval. [3] The Slavic and East European Section (SEER) of Association of Research Libraries (ARL) has recently convened a taskforce to discuss these and other remaining problematic issues such as mixed orthographies, reemergence of pre-Revolutionary orthography in Russian publications, and problems with various recensions of Church Slavic. This information will be used to update the widely consulted Slavic Cataloging Manual.[4] In this article I would like to discuss a related but somewhat distinct phenomenon, which deals with the ramifications of cataloging and retrieval not of changes in an established language, but rather those of a language which is only now beginning to achieve a degree of standardization, namely Romani (commonly also spelled 'Romany', and now sometimes Rromani).

Romani, spoken by approximately 4 million Roma[5] (for a variety of reasons an exact count is impossible) in Europe, Asia Minor, and the Americas is considered to be a northern Indic language. The great bulk of Romani vocabulary is of Indic origin, and its grammar preserves some archaic Indic elements that have been lost in the other modern Indic languages. In a yet to be published article, Marcel Courthiade, a preeminent scholar of the Romani language at Institut National des Langues et Civilisations Orientales in Paris (INALCO), convincingly traces the original migration of the Roma out of the Indian city of Kannauj, in state of Uttar Pradesh, sometime after 1018, as a result of raids on the city by the Persian ruler Mahmud of Ghazni (791-1030).[6] After a lengthy peregrination though Asia Minor, the Roma first appeared in Europe in the early 1300s, a portion of the population establishing a fairly sedentary life in some areas, while others remained mainly itinerant, travelling freely throughout several countries. On its way to Europe, the Romani language was greatly influenced by other Indo-European languages spoken along the path, such as Persian, Armenian, and Greek. Because of the large populations of Roma in the Balkans, Romani has also developed sufficient features shared by the other Balkan languages (Serbian/Croatian/Bosnian, Bulgarian, Macedonian, Albanian, Greek, Romanian, and Vlax) to be also considered as part of the Balkansprachbund.[7] As is the case with all languages of Europe, Romani has been present in Europe long enough now to develop a number of definable dialects, which differ among themselves in points of grammar, vocabulary, and pronunciation, often rather substantially.[8] Until recently, orthographic variations used to write Romani have been utterly chaotic, and the most common pattern was to use the orthographic system of the host country, resulting in Romani being written in a variety of non-latin scripts (Cyrillic, Greek, and Arabic) as well as a multitude of latin scripts, based on, e.g. Czech, Croatian, Hungarian, Albanian, German, Spanish, and English, often with no internal consistency, even in writings by the same author. A survey of Romani in Internet websites, chat rooms, and Facebook shows that this situation still obtains to this very day.

This discussion is intended to be somewhat forward looking. This author probably has much more contact with Romani materials than librarians in most other research libraries. Because of my close working relationship with INALCO, the University of Kansas Libraries receive most of the Romani-language publications (books, journals, and multi-media) published by this institute and others collected by the institute. Despite the lack of any authoritative cataloging documentation in how to deal with this language, as cataloger as well as bibliographer, I have had to develop my own strategies for processing Romani-language materials. We must however acknowledge that there has not traditionally been a large body of printed literature in the Romani language, which is perhaps why the problems associated with Romani in the scholarly and library context have not been raised until this point. Until most recently, educated Roma generally have deferred to the languages of their host countries for scholarly and educational purposes, whereas among the uneducated and illiterate, oral literature and song have been the usual methods of transmitting their history and culture. The environment has change dramatically in the last 20 years, as a cadre of Romani intellectuals, historians, and linguists have perceived an urgent need to try to achieve some kind of standardization of the European Romani variants, in order to promote education by creating culturally-inclusive instructional materials in Romani and to foster and preserve their rich culture.

Warsaw Agreement

In 1990, the 4th World Romani Congress, an umbrella cultural organization that represents Roma in 25 countries, convened in Warsaw Poland in order to accomplish two goals: to validate Romani as a European language on par with other European languages, and to attempt to achieve a degree of standardization, making it a viable vehicle for primary education and publishing.[9] A team of linguists forged a new standard alphabet, quite different from any that had ever been used before. There was quite a bit of debate and disagreement between the various team members. Some wished to adhere to something closer to the common latin-script usage in those countries with the largest Romani populations, i.e. the Balkan countries. Others saw an opportunity to showcase the unique features of the Romani language, and they viewed a truly distinctive alphabet (see Appendix for the complete alphabet) as a means to achieve this end. The chosen alphabet exhibits a careful attempt to avoid adhering too closely to the orthographic systems of the countries in which the majority of Roma reside, and is a rather unfortunate product from my perspective, especially from the point of view of cataloging and data retrieval. This is not purely a value judgment on my part, for practical reasons I will explain further below.

In order to expound on some of the special complications that Romani presents, I must first delve into a bit of linguistic minutiae that explain how this new alphabet became endowed with some rather exotic features that will be problematic. Romani is an Indic language, descended from Middle Indic languages that had already lost the elaborate case and verbal systems we see in the older Indic languages such as Sanskrit and Pali. It therefore shares certain syntactic features with other modern Indic languages such as Hindi, but has subsequently undergone certain innovations since the Roma left the India subcontinent. One salient feature concerns what has been traditionally viewed as a case system in Romani. This syntactic feature can indeed be interpreted quite differently when comparing Romani to its sister modern Indic languages. In Hindi, for example, case relationships, such as possession, indirect object, and location, are formed by placing a postposition after the modified noun, which, in turn, appears in the oblique form (which may or may not change from the direct form), e.g. the direct form of the Hindi word for 'boy' is **larka**, and **larke ki ankh** means ‘boy's eye', and ‘boys’ eyes’ will be **larkoŋ ke ankhe**, with the noun changing to the oblique form with the placement of the possessive postposition **ki**. We see a very similar reflex in Romani. The Romani word for boy is **raklo**,[10] and 'the boy's eye will be **rakleski yakh**, and 'the boys' eyes' will be **raklǎnge yakha**[11]. When comparing the Romani structure with Hindi, we can clearly see that what may look like a case ending can, in fact, be interpreted as a postposition.[12] What differs in Romani from Hindi is that the phonetic nature of the postpositions in Romani changes based on the consonant to the immediate left. If that consonant is voiced, then so will be the consonant of the postposition, and conversely if it is unvoiced then so will be the consonant in the postposition. This can be seen in the example above. The singular oblique stem of **raklo** is **rakles-**, while the plural oblique stem is **raklen-**. The phonetic form of the possessive postposition can thus be either **/-ke/** or **/-ge/** depending on whether the preceding constant is unvoiced or voiced. The same reflex occurs with the postpositions **-te** ‘to’ and **-tar** ‘from’, which phonetically become **/-de/** and **/-dar/** after voiced consonants. Some Romani linguists considered this syntactic feature so distinctive and significant that it should be conspicuously accounted for in any standardized orthography chosen for Romani. They therefore came up the following solution: rather than write the consonant of the postposition with two different consonants, in each instance they would choose a rather obscure grapheme to represent them. The postpositions added to the oblique stem would thus be written as below:

Pronounced	**Written as**	**Meaning**
-ke/-ge	-qe	‘of’
-te/de	-θe	‘to’
-tar/-dar	-θar	‘from’
-sa/-tsa*	-ça	‘with’

*The sound /ts/ is written Romani with the letter **c**.

Thus, the above examples would now be written as **raklesqi yakh** and **raklǎnqe yakha**.

In order to try to further distance Romani from the graphical representation of the neighboring Slavic languages, the diacritic hachek, i.e. caron (ˇ), was eschewed in favor of the acute accent, thus **ć,** **ś**, and **ź** for the aspirant sounds **ch** (as in Charley), **sh** (as in ship), and **s** (as in pleasure) for the aspirant sounds. Practically speaking, this is a rather modest modification. However somewhat more contentious was the discussion of how to represent the frequent Romani consonant corresponding to the **j** in **judge.** This particular sound (and closely related sounds) are represented by a wide variety of graphemes in the languages of Europe, e.g.: **j, g** (English), **đ**, **dž**, **dj** (Bosnian/ Croatian/ Serbian), **rz**, **ż** (Polish)[13], **c** (Turkish), **gy** (Hungarian), and **gj**, **xh** (Albanian). Previous to the 2009 Warsaw agreement, treatment of this sound varied widely, as it still does in the non-standard Romani orthographies, being variously represented as **ž, ƶ, ź,** as the number **3**, and even a very quirky square box minus the left side and an acute accent above.[14] While it would have been logically consistent to choose **ź** to represent this sound, the committee, rather inexplicably, decided to select the obscure International Phonetic Association character ‘**ezh**’ **ʒ.** The **ezh**, alongside the **theta** that appears in the Romani postpositions, will be problematic from the perspective of cataloging.

Cataloging Ramifications

The problem of languages that have undergone changes in the official orthography, e.g. Russian after the Soviet Revolution, is well known to catalogers in research libraries. However the Roma live in a number of countries and therefore are not able to claim to have any governmentally sanctioned “official orthography.” Thus, there is some uncertainty whether the AACR2 stipulation about changes in official orthography is applicable, or even desirable in the cataloging of Romani materials. Nevertheless, as more and more materials are being published in Romani, many, but not all adhering to the new standardized orthography, it is important to consider having a systematic approach to cataloging materials in the Romani language.

AACR2 cataloging records consist of two main components, the descriptive elements, transcribed from the item being cataloged, and additional controlled metadata elements assigned by the cataloger, such as names and subject headings, which are subject to authority control. In the context of Romani, the latter are generally unproblematic, as authors’ names are usually established based on the orthographies of the countries in which they reside, and therefore do not normally contain the problematic characters **ezh** and **theta**[15]. The descriptive elements of the bibliographic record, on the other hand, present very little flexibility. AACR2 clearly stipulates that the descriptive elements should ideally be transcribed exactly as they appear in the item being cataloged, although with further caveats. Until recently, for languages in non-latin scripts the only option in bibliographic utilities, such as OCLC, was a romanized transliteration scheme, which at least for North American libraries, was based on the ALA norm.[16] While we are now able to add matching vernacular fields for many scripts in OCLC WorldCat, this is not the case for all valid Unicode characters.[17] In addition, AACR2 stipulates some rather elaborate rules about how to deal with special characters such as **ezh** and **theta**, which seem inappropriate to apply in the case of a language for which these characters are part of the established orthography.[18] We are therefore left with a dilemma when it comes to Romani.

The Problematic Characters

In the context of 1990, when the new alphabet was adopted, computer automation and the Internet were not as developed as they are today, and therefore the Romani linguists did not see the introduction of new exotic letters as unduly problematic, as typewriters could be ordered with specific keys as needed. However in the current computer environment these characters present a particular problem. OCLC WorldCat is still limited to the smaller Marc-8 Unicode subset and limits what characters can be entered in certain contexts. While the vast majority of letters of the new Romani alphabet are MARC-8 compatible, the **ezh** is a character that is still unrecognized in that character set, and so we must replace this character with some alternative.[19] The introduction of the Greek character theta (**θ)** also presents some distinct problems for cataloging. This character is disallowed in OCLC cataloging in otherwise latin-script fields and is generally replaced with the bracketed name of the letter, i.e. [theta], or a phonetic equivalent. If we do attempt to input this character in OCLC Connexion (other than in a matching vernacular Greek field) it will force Connexion to assume that the record being input is a non-latin script record and will automatically code the resulting record as such. This code is system supplied and cannot be removed. Connexion will also attempt to add matching fields for romanization, as we see in OCLC WorldCat records for Cyrillic, Arabic, etc., which cannot be left blank and must be filled with text of some sort in order to validate and add the record to OCLC WorldCat.

Alongside these problematic consonants, the new orthography contains a number of vowel graphemes **ǎ, ě, ǐ, ǒ, ǔ**. These so-called pre-jotizing vowels are pronounced with a faint "y" sound after the consonant preceding them. From the descriptive aspect of cataloging these characters are not unduly problematic and are compatible in OCLC WorldCat. The most important thing to bear in mind is that these diacritics are the caron, and not the breve, which is more commonly seen above vowels in European languages.

In addressing the descriptive aspect of the bibliographic record, how then do we deal with the **theta** and **ezh**? Computer alphabetical sorting and keyword searching is normally good at regularizing variants of roman-alphabet letters occurring with various diacritics, so that, e.g. **z, ź, ż, ž,** etc., will all be retrieved when searching the stripped **z**. The **theta** and **ezh**, however, are not characters for which such a correlation is normally encoded. As far at the descriptive elements of Romani bibliographic records that contain these problematic elements are concerned, we must turn to practical solutions.

There seem to me to be two acceptable solutions for dealing with the **ezh**, and, for the sake of consistency, I believe it would be prudent to treat the **theta** as well as the **ezh** in the same fashion when cataloging in OCLC WorldCat. The approaches I will suggest below, while admittedly not extremely aesthetic, will nevertheless assure that important graphical information will be retained until catalogs and bibliographic utilities are compatible with all Unicode characters. For the time being, it is likely that these steps will be mainly necessary in the descriptive areas of the bibliographic record and will mostly affect the transcription of the titles.

Option 1:

Except for the **theta** and **ezh** all other Romani letters have ALA character equivalents and can be transcribed as they are. The **ezh** has two acceptable variants according to the Warsaw Agreement, **ƶ** and **з** (the Cyrillic /z/). However, neither of these is a valid ALA character, and therefore cannot be used in OCLC WorldCat to transcribe the title proper (MARC 245 field). Although not commonly seen in OCLC cataloging, there is a provision in the Library of Congress Rule Interpretation documentation for replacing special characters that cannot be used in the MARC record. According to this list, the **ezh,** which can also appear in some African languages, should be transcribed as **z̳** (**z** with double underscore) and **theta** as **t̳** (**t** with double underscore). Using this substitution for **ezh**, **Kote ʒàna e Kosoviaqe Rroma** would be transcribed as: **Kote z̳àna e Kosoviaqe Rroma**. Since titles are generally not unwieldy, I believe it would also be helpful to provide a note, such as "**z̳** in the word **z̳**ana appears as the IPA symbol ezh." Such notes will be helpful to users unfamiliar with the vagaries of transcribing unusual characters in OCLC cataloging.

The use of **t̳** for **theta** seems a little less satisfactory, since this letter has two phonetic outcomes, /**t**/ and /**d**/ depending on the preceding consonant. Nevertheless, it makes sense to consider employing this substitute character. In the case of **theta** there will also be cases where added title entries will be helpful. Regrettably, the treatment of **theta** will require a bit of knowledge of Romani phonology as revealed in the discussion of postpositions above. However, with this knowledge in hand, when constructing added title entries (MARC 246 field), I would recommend, in cases where the **theta** is indeed pronounced as the voiced **d**, that the **theta** be transcribed according to its phonetic value, i.e. **d** or **t** depending on whether the preceding consonant is voiced or unvoiced respectively, e.g. **Sar me vastesθar xutǐlav tu** would be transcribed as **Sar me vastestar xutǐlav tu**, while **But lenθar ʒivdinèna** would be transcribed as **But lendar z̳ivdinèna**.

Option 2:

The second option involves the use of a so-called "lossless" solution. Such a proposal for dealing with problematic characters in MARC cataloging records already exists in the MARC documentation.[20] Although accepted in 2006, its use with characters that fall outside the MARC-8 character repertoire has not be widely used in OCLC cataloging, the reasons for which might become evident from the examples below. In short, this lossless method recommends replacing a Unicode character that cannot be mapped to MARC-8 with a placeholder that contains the Numeric Character Reference (NCR) for the character. The NCR consists of the hexidecimal representation of the code point of the character (four ASCII characters), preceded by **#x** and all surrounded by **&** and **;**

Using this method yields the following:

The lower case **theta** can be substituted with **&#x03B8;** and the capital letter as **&#x0398;**

The lower case **ezh** will be **&#x0292;** and the upper case **&#x01B7**

There are advantages and disadvantages to each of these options.

The first option will allow for searching and alphanumeric filing based on a more natural representation of what a speaker of Romani, or one researching Romani might expect, in the case of the **ezh** because **z̳** will regularize out with the many forms of the letter **z** frequently used to represent this sound in the various common manifestations of the Romani alphabet. However it is not ideal in this respect, as diagraphs, such as **dž**, **dj, xh** are also frequently used to represent this sound in Romani texts. In regards to the **theta**, this approach is somewhat better suited, as it represents a form used by the majority of Romani writers, many of which still shun the use of the **theta**. On the other hand, the second option will allow for eventual automatic conversion of the characters back to the proper graphic representation when these Unicode characters become available in the MARC format. Some newer web interfaces are already able to render these characters correctly when they are manually input in the local catalog database, even if it is not yet possible to input these non-MARC-8 characters in OCLC WorldCat directly.[21] Nevertheless there are obvious drawbacks to this approach as well. First, since these NCRs are mostly likely to occur conspicuously in the title of the bibliographic record, any user, other than those familiar with cataloging limitations and hexidecimal coding of Unicode characters, will likely assume that a coding error has occurred, resulting in garbled characters.[22] Second, this method will profoundly affect keyword searching and alphanumeric sorting of main titles. Finally, even if the characters are rendered correctly in the catalog and browser, another unfortunate result is that the catalog user may not have the ability to input the correct characters for the purpose of searching.

If the second option is chosen, as in the case of the first option, it will still be desirable to provide some alternate title headings to the bibliographic record, based on a reasonable assumption of what the user might expect. This is clearly not a trivial issue. As in the case of cataloging all foreign language materials, the cataloger must have at an adequate working knowledge of the language being cataloged in order to be able to discern which added entries are most appropriate and useful for the end user. In the Appendix I discuss my personal preference for treating these problematic characters but I feel both approaches are equally legitimate.

Authority Control Implications

Turning now to the portion of the bibliographic record subject to authority control, i.e. the assignment of controlled vocabulary and standardized names, we are left with a further dilemma.[23] Many Romani writers are either unaware of the standardized alphabet adopted by the 1990 World Romani Congress or find the system so bizarre that they simply choose to disregard it, opting for the orthographic system of their home country or an idiosyncratic one of their own design. However, as a cataloger of these materials, I must make a decision, one way or another, how to approach this issue. As a linguist, I recognize that Romani is very different from most other European languages in that it is represented by no nation state and therefore has no governmentally-sanctioned official orthography, and is represented by a wide variety of dialects. However as a result of my experience as a cataloger I recognize the importance of collocation in the context of the library catalog and strive to find a means to make Romani materials as accessible as possible. This raises the question of whether it is desirable to accept the 1990 Romani Congress decision as authoritative enough to apply the AACR2 guidelines concerning change in orthography to Romani. I would argue that this is indeed desirable, at least in certain cases where possible, and where there a desire to bring together related manifestations of a specific work.[24]

One problem, until very recently, has been the lack of any authoritative reference source to resolve questions of orthography for Romani. Such reference sources are frequently consulted by foreign-language catalogers, for example, when cataloging Arabic to ensure a consistent transcription of the short vowels, which are normally not written. In 2009, a dictionary, of which I was coeditor of the English content, was published and which may be very helpful in resolving problems. This

UNESCO sponsored dictionary is the first international dictionary of the proposed standardized form of the Romani and it employs the new orthography discussed above. [25]

While it is difficult at this stage of Romani publication to point to many examples to support this argument, I can present at least one real-life example where assigning uniform titles based on the current form of the orthography can bring together related works. The Romani scholar, Rajko Đurić, has published many books on the topic of the Roma, several in Romani. His earlier Romani publications were written in an orthography that mirrored Serbian latin-script orthography, whereas he later chose to adopt the 1990 Romani Congress orthography. Đurić published two related books treating the history of the Roma, the first, in 1988 under the title Bibahtale breša [Unhappy years]. Then in, 1996, he published a companion volume, using the new orthography, under the title: Bibaxtale berśa]. [26] Clearly these works will regularize differently in a title browse and a keyword search. In cases such as this, it seems that it would be prudent to provide uniform titles (MARC 130 or 240 fields) based on the new orthography on the bibliographic record for the title in the old orthography, in order to provide access to the new established orthography. An authoritative source, such as the dictionary mentioned above, can facilitate in determining the appropriate form.

Other than examples such as above, where there is a desire to bring together related manifestations of a work, I fear that it may impractical to try to routinely construct uniform titles for Romani works for a number of practical reasons. First, the new orthography is only a recommended standard, with no legal weight. While being used by several Romani organizations and scholars in Europe, others may either totally disregard it, or use it only partially (especially avoiding the **q**, **ç**, **ʒ**, and **θ**). While above I have concentrated on the problem of the consonants, similar problems exist with the new Romani vowel characters. To give just an example for **ǎ** and **ǒ**, these have been written as **ja**, **ia**, **ya, я** and **yo**, **io**, **yo**, **ë** respectively in various kinds of orthographies. While books being printed by publishers that specialize in Romani materials may be more likely to adhere to the new orthography, there will continue to be a great variety in how Romani is written in books and journals, the Internet, etc. The burden will therefore be on the cataloger to attempt to construct uniform titles for a language that is generally little known outside the Romani community and among Balkanologists.

Conclusion

As a non-territorial and transnational language, it is very difficult to envision a time in the near future when even European variants of Romani will achieve the degree of standardization of other

European languages. Nevertheless, considering that there are now UN-affiliated and other cultural groups that are attempting to promote the new standard orthography and are introducing it more and more in printed publications and Internet resources, it should be approached by the cataloger with the same degree of attention that we pay to other complicated issues of orthography. The goal of our diligence is always to keep the potential end user in mind, and to anticipate how in the future automated algorithms may be able to bring together related works, as envisioned in FRBR models[27]. Even if we conclude that certain issues of authority control, such as assigning uniform titles, are too complicated or ambiguous to be applied systematically, we are nevertheless left with the problematic characters that presently prevent their exact transcription in some cataloging environments. It is my hope that this discussion will be of help to other catalogers who may be presented with Romani materials to understand the complicated cataloging issues involving this fascinating language. Please feel free to contact me at **husic@ku.edu** if you should need any assistance in cataloging Romani materials.

Appendix: The Unified Romani Alphabet (Based on the 1990 Warsaw Agreement) and Summary of Suggested Cataloging Approaches

Most Romani letters are pronounced very similarly to Croatian. The exceptions are: **ç** (as **s** or **ts, q** (as **k** or **g**), and **θ** (as **t** or **d**); **x** (as the Russian **x**); **kh**, **ph**, and **th** (aspirated **k**, **p**, and **t**), **rr** (uvular **r**), and **ʒ** (as the **j** in **jam**.) The letters **ć**, **ś**, **ź** are very close the Croatian **č**, **š**, and **ž** in most dialects.

Unproblematic characters: The following letters (in Romani order) present no special problems to the cataloger in any cataloging environment. The only possible point of confusion is whether **rr** (pronounced as a uvular sounds, such as **r** in French **Pa<u>r</u>is**), when initial, is written as **Rr or RR** and is of little concern, although **Rr** is preferred.

a, **b**, **c**, **ç**, **d**, **e**, **f**, **g**, **h**, **x**, **i**, **j**, **k**, **kh**[28], **l**, **m**, **n**, **o**, **p**, **ph**, **r**, **rr**, **s**, **t**, **th**, **u**, **v**, **z.**

Vowels with the acute (representing unpredictable stress: **à**, **è**, **ì**, **ò**, **ù**)

Possibly problematic characters: The pre-jotizing vowels, i.e. vowels with a caron (hachek) written above. For bibliographic utilities such as OCLC Connexion these characters no longer present a problem. Each vowel can be input with the necessary caron above. Library catalogs may display them correctly or strip out the diacritics, depending on the system, but searching and filing will not usually be affected. Nevertheless, it behooves the cataloger to input these letters with the proper caron, rather than to accidentally substitute the diacritic breve.

ǎ, ě, ǐ, ǒ, ǔ

The Latin characters with the acute diacritic: The characters representing these letters are available in OCLC Connexion and should search and display correctly in any Unicode-compliant library catalog. The characters may be downgraded when imported into some bibliographic utilities, such as Endnote, to their stripped forms **c**, **s**, **z**, or may be replaced with blank or filler characters, depending on how character importing is set. However this particular problem is beyond the scope of this article.

ć, ś, ź

The Postpositions q and **ç**,

In the examples with the **ezh** and **theta**, I have recommended making title added entries in certain cases where problematic characters occur in a title proper. I would also recommend constructing such added entries in the cases of the characters, **q** and **ç**, used in the postpositions, based on their phonetic outcome, i.e. **k**/**g** and **s**/**c**. Again, while requiring a bit of knowledge of Romani phonology, these added entries will aid a user by providing keyword searching based on a more natural form.

Problematic Characters and their Numerical Character Reference Codes. These characters may present considerable problems for the reasons discussed above.

Theta **θ** and **ʒ**

Although visually unaesthetic, my preferred recommendation for transcribing these characters in OCLC is to use the lossless method discussed above by replacing these characters with their appropriate Numerical Character References, i.e. **&#x03B8;** (θ) and **&#x0398;** (Θ) **&#x0292;** (ʒ) and **&#x01B7** (Ʒ). I am hopeful that in the near future we will have all Unicode character sets available to us in MARC cataloging, and therefore this problematic situation should not last indefinitely. Until then some catalogs can already display these characters correctly when encoded in this way.

If this solution is locally considered to be unacceptable, then the use of the**t̳** and **z̳** (t and z with double underscore) is a good pragmatic alternative. At some future point these replacement characters could be searched and replaced in the database with the proper Unicode characters.

One further note: if attempting to search OCLC WorldCat for examples of records for items in Romani you will discover that a large number of records have the MARC fixed-field language code for Romanian (**rum**) miscoded as **rom**, which is in fact the language code for Romani. This will result in many false hits when attempting to filter by language.

Romani Samples on the WWW

Below I have listed a few examples “in the wild” Internet sites that use the new orthography or a permutation of it, that may interest the reader.

A website of news from Kosova, using the new orthography in totality:

http://rroma.courriers.info/spip.php?article394

A website devoted to Romani rights. Some articles are in the new orthography, which others are in a variety of the idiosyncratic variants:

http://romarights.wordpress.com/category/nevipenewsvijesti/nevipe/page/3/

Website of the European Roma Rights Center. Except for the notation of the aspirated consonants, i.e. **kh**, **ph**, and **th**, the orthography is the same as Croatian.

http://www.errc.org/cikk.php?cikk=2255

NOTES

1 . In cataloging literature and documentation, the terms latin and roman are often used interchangeably when speaking of scripts. However the process of transliterating a non-latin script into latin script is almost always called “romanization.”

2. The complete character sets for certain Central Asian and other Cyrillic languages are not yet available in OCLC. Further information on available character sets can be found at: http://www.oclc.org/support/documentation/worldcat/records/subscription/4/4.pdf

3. Library of Congress Rule Interpretation based on *Anglo-American Cataloging Rules. 2nd ed., 2002 Revision.* (Chicago: American Library Association), 2002 via Catalogers Desktop: http://desktop.loc.gov: ‘Section 25.3A’: “For monographs, on the bibliographic record for any edition of a work whose title proper contains a word in the old orthography provide a uniform title reflecting the new orthography, although no edition with the reformed orthography has been received.”

4. *Slavic Cataloging Manual* (http://www.indiana.edu/~libslav/slavcatman/smtocs.html)

5. The term ‘Gypsy’ is considered pejorative and is no longer used in scholarly writing. In English both Roma and Romanies are used for the plural.

6. This information has been included with the kind permission of Prof. Courthiade.

7. A linguistic concept that examines the syntactic, morphological, and lexicological commonalities that have arisen in the territorially contiguous, but linguistically quite different Indo-European language groups spoken in the Balkans.

8. The Romani dialects are broadly grouped into Vlax, Balkan, Carpathian, and Sinti.

9. Courthiade, Marcel, ed. *Morri angluni rromane ćhibăqi evroputni lavustik*. Budapest: Fővárosi Onkormányzat Cigány Ház-- Romano Kher, 2009, pg. 496-497.

10. Both **raklo** and **yakh** are cognate with the Hindi **larka** and **ankh**.

11. I am intentionally not using the new standardized alphabet in these examples, as it would obscure my argument.

12. Victor Friedman. “Case in Romani: old grammar, new affixes. *Journal of the Gypsy Lore Society*, 5th ser., vol. 1, no. 2 (1991, pg. 85-102.

13. In fact in standard Bosnian/Croatian/Serbian and Polish there are two distinct “zh” phonemes, one more palatalized than

the other. In Romani there is generally just one such phoneme, although there is a great variety in the actually pronunciation of the phoneme.

14. Courthiade, Marcel. *Gramatika e gjuhës rrome*. Tirana: [s.n.], 1989, pg. 19.

15. Since **theta** only exists in the postposition, it of course cannot occur in the non-oblique form of a name.

16. http://www.loc.gov/catdir/cpso/roman.html

17. As of this writing OCLC Worldcat is limited to the Marc-8 character set, plus four additional scripts (Thai, Bengali, Devanagari, and Tamil), which use the UTF-8 character sets.

18. *Anglo-American Cataloging Rules*. 2nd ed., 2002 Revision. (Chicago: American Library Association), 2002 via Catalogers Desktop: http://desktop.loc.gov: ‘Section 1.0.E: Language and script of the description.’

19. In order to not become overly technical in this article and to concentrate on the current cataloging implications, I have chosen not to go into specific details about the variety of MARC formats and ALA, MARC-8, and Unicode character sets, as they are out of scope.

20. http://www.loc.gov/marc/marbi/2006/2006-09.html

21. This is the case in, e.g., with the Voyager catalog.

22. An example can be seen on the OCLC WorldCat record #_233230871 (an example with Mongolian and Kazakh in Cyrillic script).

23. An excellent explanation of authority control in the context of bibliographic records can be found at: http://en.wikipedia.org/wiki/Authority_control

24. The new requirements for how related manifestations are to be handled will be more clear when RDA, the cataloging code that is to replace AACR2, is published in 2010, but this is not likely to greatly affect the suggestions made in this article.

25. Courthiade, Marcel, ed. *Morri angluni rromane ćhibăqi evroputni lavustik*. Budapest: Fővárosi Onkormányzat Cigány Ház-- Romano Kher, 2009. This dictionary has not yet been widely circulated, but questions concerning its availability can be directed to: Mr Zsigo Jenö, Roma Parlament, 1084 Budapest, Tavaszmezö ut 6. (Roma.parlament@chello.hu) or to Mr Lakatos Laszlo (lakatoslaszlo@frokk.hu).

26. Admittedly this is an imperfect example, as these two items, although they share the same main title, are not variant editions, but different books covering different time periods. The discrepancy between breša and berśa is not a typo. It is a dialect variation for this word.

27. FRBR (Function Requirements for Bibliographic Records) is an entity-relationship model for displaying related works in a catalog in a more holistic manner.

28. The **h** after another consonant indicates aspiration. These represent separate phonemes and are thus also considered separate digraph letters.