Issues to Resolve in ISO 639

(Revised, 2004-08-14)

Peter Constable, Microsoft
SIL International Liaison to ISO 639/RA-JAC
Project editor, ISO 639-3

1. Introduction

ISO 639-3 will be based on the inventory of languages in the SIL *Ethnologue*, supplemented by a catalogue of historic and artificial languages maintained by The Linguist List. A prerequisite to the development of the code table for ISO 639-3 is to establish the precise relationship between this inventory and the inventory provided in the existing parts of ISO 639.

There are numerous issues that must be resolved in establishing this relationship. The inventory provided by *Ethnologue* and The Linguist List is more granular than that of ISO 639-1/-2. It also has much more explicit documentation of denotations for entries than the English and French names provided with ISO 639-1/-2. As a result, a comparison of these brings to light several issues in ISO 639-1/-2 that need to be resolved. These include such things as the need to re-evaluate the scope of an identifier (does it denote multiple languages rather than just one individual language?) or to resolve ambiguity in the denotation where a name may be used for two or more distinct, perhaps unrelated, languages.

It is important to note that these issues *must* be resolved in order to proceed with the development of ISO 639-3. Hence, the complete and prompt cooperation of all members of the ISO 639/RA Joint Advisory Committee and associated stakeholders is requested. It is worth noting also that resolving these issues in ISO 639-1/-2 will greatly improve the usefulness and usability of these standards; thus, stakeholders should significantly benefit from this process.

It should also be noted that the resolution of some of these issues may impact MARC and its usage of ISO 639-2. While three-letter identifiers in ISO 639-2 may have originated in MARC usage, in becoming an international standard they have become available to a much wider variety of users. It is essential for inter-operability that the identifiers in these international standards are used consistently across different sectors while still serving the needs of individual user communities. As much as possible, I have attempted to propose solutions to issues that provide compatibility with existing MARC usage. All of the analyses presented here have been reviewed by Milicent Wewerka of the Library of Congress to determine that assumptions regarding MARC usage are correct, and she has also supported each of the recommendations presented. In a small number of cases, the proposed resolution of an issue has some impact on MARC usage.

1.1 Organization

In this document, problems of a similar nature are organized under level-one headings. Each individual issue is listed under separate level-two headings. Typically, a level-two section will include elements that describe the problem and that propose a solution. The problem description will provide as much detail as seemed necessary to understand the problem and the alternative

solutions. These details are drawn primarily from *Ethnologue* or the *MARC Language Code List*. Not all details in those sources are repeated here, however.

In many cases, a list of alternative solutions is also provided. These lists will include only alternatives that seemed worth mentioning, not every conceivable alternative.

1.2 Terminology and notation

The term *scope* will be used in this document to refer to the granularity or broadness of coverage associated with a given identifier. ISO 639-3 will recognize three scopes: individual language (I), "macro" language (M), and collection (C). (The macro language scope is explained in the following sub-section.) ISO 639-3 will list only entries with a scope of I. Some of the issues to be resolved for ISO 639-1/-2 involve clarifying what is the intended scope of an entry.

For clarity, language identifiers are denoted in this document in square brackets, "[…]". Lists of identifiers within prose will use a single pair of brackets to enclose the entire list, rather than separate brackets for each identifier.

Where identifiers exist in both ISO 639-1 and ISO 639-2, or where distinct ISO 639-2/T and ISO 639-2/B identifiers exist, all will be cited, separated by a slash, "/"; e.g. [af/afr].

References to the *Ethnologue* are to the 14th edition unless indicated otherwise. At some points, a language identifier from the *Ethnologue* 14th edition may be cited in order to make explicit what language is being referred to. *Ethnologue* identifiers will be cited in upper case (e.g. [BAA]);[1] in contrast, identifiers from ISO 639-1/-2 are always cited in lower case (e.g. [del]).

1.3 The "macro language" scope

The term *macro language* has been newly coined to address a specific problem: a name may be used in some contexts in which it is understood to be or treated as though referring to an individual language, while in other contexts this name may be understood to be or treated as though encompassing multiple, related but distinct languages. This may occur for various reasons. For instance,

- A large, developed language variety may have several smaller, lesser- or un-developed varieties closely-related with it that are known by the same name.
- A language may split over time into two or more distinct varieties, yet a common name or identity (possibly ethnic or political) is maintained.

In some contexts, there may be a requirement to differentiate the several varieties as distinct languages; this would be the case, for instance, for linguistic researchers. In other contexts, however, there may be a requirement to treat the entire linguistic complex as a single entity. For instance, software vendors may not want to represent the multiple varieties in their products as this may require increased maintenance costs, additional infrastructure that may be redundant if the same processing resources (fonts, input methods, etc.) are used for all varieties, and may lead to more complex user interfaces that can be confusing to users. Or because of a strong political identity among the related linguistic varieties, there may be a marketing requirement to represent only a single variety within a product (even though this may, in some cases, be somewhat artificial).

[1] Note that some identifiers from the *Ethnologue* 14th edition will be used in the draft code tables for ISO 639-3, *but not all*. Identifiers in ISO 639-3 and in the 15th edition of *Ethnologue* will align with existing identifiers in ISO 639-2.

A macro-language is like a collection in that it encompasses multiple individual languages. Macro-languages are distinct from collections, however, in that collections are (usually) based on genetic classification and include multiple languages that have distinct identities. For example, "Germanic languages" would correspond to a genetic sub-group, and the languages encompassed have distinct identities; Icelandic is never called "Dutch", for instance. In the case of Frisian, however, there are distinct varieties that are not inherently mutually intelligible, yet they share a common identity of "Frisian", and there may be usage contexts in which reference may be made to "Frisian" without differentiation between the different varieties.

A macro-language, then, must be an entry in ISO 639-1 or ISO 639-2 that corresponds to multiple closely-related entries in ISO 639-3 that share a common name and identity (as perceived either from within the language communities or by outsiders), and for which there is a need to refer to that joint identity in some usage contexts.

The code table for ISO 639-3 will include only individual language identifiers. It will also, however, include as an annex a table showing the relationship between entries in ISO 639-3 and entries in ISO 639-1/-2 that are deemed, for purposes of ISO 639-3, to be macro-languages. In this mapping, entries in ISO 639-1/-2 deemed to have macro-language scope will map to multiple individual-language entries in ISO 639-3. Entries in ISO 639-1/-2 that are considered to have individual-language scope for purposes of ISO 639-3 will be identical in ISO 639-3. Any individual-language identifier not discussed here will be considered an individual language for purposes of ISO 639-3.[2]

1.4 Note on the revised version

An initial version of this document was distributed to members of the ISO 639/RA Joint Advisory Committee for review early in 2004. Per prior JAC agreement, a careful review was conducted by Milicent Wewerka of the Library of Congress to ensure that assumptions with regard to MARC usage were correct, and that the proposed solutions were agreeable from the perspective of MARC needs. This second version was prepared after her detailed comments were received and reviewed. She recommended changes in a number of cases that have been followed, though there were a few cases in which it was felt that the original proposal needed to be retained in spite of her recommendations. Also, in the intervening period since the first version was prepared, addition information was provided for a small set of cases that were revised as a result. These are all described in the document, *Disposition of Comments from Milicent Wewerka, Library of Congress, on Issues to Resolve in ISO 639*.

2. Incomplete collections

PROBLEM: Incomplete collections, such as [afa] "Afro-Asiatic (Other)", have been problematic since they get redefined any time a member is given its own identifier, resulting in existing data becoming wrongly tagged. Also, with the development of ISO 639-3, every known member of such groups will have its own identifier; thus, these "other" collections will be empty sets (there will be no known individual languages without their own identifier) and will lose any significant purpose.

[2] Another key aspect of the relationship between ISO 639-2 and ISO 639-3 is that any given three-letter identifier will denote consistent semantics across the two standards: if an identifier is listed in both standards, it must mean exactly the same thing in both cases.

PROPOSED SOLUTION: Revise all such collections by changing the name from "… (Other)" to "… languages", making them inclusive collections. It will be up to protocols that reference ISO 639 to determine whether and when collective language identifiers can or should be used rather than more specific identifiers.

Identifiers affected: [afa, art, bat, ber, bnt, cai, cau, cel, cpe, cpf, cpp, crp, cus, dra, fiu, gem, inc, ine, ira, khi, map, mkh, nic, paa, phi, roa, sai, sem, sit, sla, smi, ssa, tai, tut]

There is no negative impact on MARC usage. There will be a benefit in that legacy data tagged with such identifiers will retain valid tagging if individual-language identifiers that would match the language of that data are added to ISO 639 (i.e., the first problem identified above will no longer be a problem).

3. Language names in ISO 639 that are non-linguistic identities

Some entries in ISO 639-1/-2 are names that may get used for language identities but do not, in fact, refer to languages. Some are ethnic cover terms; for instance, "Dayak" is a cover term used by the Muslim majority in Borneo to refer to the non-Muslim "tribal" minorities. Some are regional identities; for instance, "Himachali" simply means "language of Himachal", and must be considered a cover term for the various languages spoken in Himachal Pradesh.

In ISO 639-2, these entries have been treated like individual-language categories. In MARC, some of these were treated like collections, though they would exclude any major languages that might otherwise have been encompassed (e.g. in MARC, "Bihari" would not be used for Bhojpuri, Magahi or Maithili).

These categories in ISO 639-1/-2 are particularly problematic. Ideally, they should be withdrawn or deprecated. Alternately, they could be considered collections, though they would not be based on genetic classification. In some cases, it may also be possible to consider specifying a particular language as the denotation, though this would likely not be consistent with prior usage in MARC or elsewhere.

3.1 Bihari

PROBLEM: ISO 639 includes [bh/bih] "Bihari" as an individual language. "Bihari" is a cover term for languages spoken in Bihar, a state of India, particularly for Bhojpuri, Magahi and Maithili. This would include many smaller and less-well-known languages, but possibly also some larger languages associated with Bihar. (It would not include major, widely-used languages such as Hindhi.) As such, it could encompass numerous languages:

> Agariya, Angika, Asuri, Bhojpuri, Bijori, Birhor, Degaru, Domari, Ho, Kharia, Kharia Thar, Korwa, Kudmali, Kumarbhag Paharia, Kurux, Magahi, Mahali, Maithili, Majhi, Mal Paharia, Panchpargania, Surajpuri, Sauria Paharia, Turi; potentially also Awadhi, Bhili, Braj Bhasha, Chhattisgarhi, Kumauni, Mundari, Newari, Rabha, Sadri, Santali, Sora.

Even though some speakers may identify their language as "Bihari", this clearly does not correspond to a particular, individual language.

The MARC Language Code List has used [bih] for "Bihari" and as a collective for Angika, Kurmali ("Kudmali") and "Bajjika". "Bihari" is used to refer to a genetic sub-group of Indo-Aryan, to which Angika and Kudmali belong. (I have found references in various sources to a

language called “Bajjika” but no information on what other languages it may be related to.) This genetic sub-group also includes Bhojpuri, Magahi, Maithili and various smaller languages.

POSSIBLE SOLUTIONS:

1. Deprecate [bih], documenting the meaning of the term and the problems related to its usage.
2. Change scope of [bih] from I to C (genetic) and change name to “Bihari languages”; denotation encompasses the languages of the Bihari sub-group of Indo-Aryan..

PROPOSED SOLUTION: Option 2.

3.2 Dayak

PROBLEM: ISO has [day] “Dayak”. The term “Dayak” is an exonym used by the Muslim majority in Borneo to refer to the non-Muslim “tribal” minorities. The term used to carry pejorative connotations, though this is much less the case today. Linguistically, it cuts across major branches of Western-Malayo-Polynesian. Without some qualification, the label “Dayak” does not correspond to any useful category for purposes of language identification.

MARC has used this for various varieties from distinct branches of Western-Malayo-Polynesian, though Milicent Wewerka reports that the intent is Land Dayak.

POSSIBLE SOLUTIONS:

1. Restrict the denotation of [day] to one specific Western-Malayo-Polynesian language, such as Ngaju, changing the name accordingly.
2. Change the scope of [day] from I to C (genetic) and restrict the denotation to one specific sub-group of the Western-Malayo-Polynesian family, changing the name accordingly: either “Land Dayak languages” (encompasses 16 languages) or “Malayic-Dayak languages” (encompasses 10 languages).
3. Deprecate [day], documenting the meaning of the term and the problems related to its usage.

PROPOSED SOLUTION: Option 2, with the denotation being Land Dayak languages.

3.3 Himachali

PROBLEM: ISO 639 includes [him] as an individual language. “Himachali” is a cover term for languages spoken in Himachal Pradesh, a state of India. This would include many smaller and less-well-known languages, but possibly also some larger languages associated with Himachal Pradesh. (It would not include major, widely-used languages such as Hindi.) As such, it *could* encompass numerous languages, including all but two languages from the Western Pahari sub-group of the Indo-Aryan family:

> Western Pahari languages: Bhattiyali, Bilaspuri, Chambeali, Churahi, Dogri-Kangri, Gaddi, Harrijan Kinnarui, Hinduri, Jaunsari, Kullu Pahari, Mahasu Pahari, Mandeali, Pangwali, Sirmauri
>
> Other Indo-Aryan languages: Bauria, Chinali, Gujari, Haryanvi, Lahul Lohar, Lambadi
>
> Dravidian languages: Bazigar
>
> Tibeto-Burman languages: Bhoti Kinnauri, Chitkuli Kinnauri, Gahri, Jangshung, Kanashi, Kinnauri, Pattani, Shumcho, Stod Bhoti, Sunam, Tinani, Tukpa

Austro-Asiatic languages: Mundari

The MARC Language Code List indicates that [him] is used for “Western Pahari”, but provides no further indication of the interpretation.

POSSIBLE SOLUTIONS:

1. Deprecate [him], documenting the meaning of the term and the problems related to its usage.
2. Change scope of [him] from I to C (genetic) and change name to “Western Pahari languages”; denotation encompasses the languages of the Western Pahari sub-group of the Indo-Aryan family.
3. Change scope of [him] from I to C (ad hoc) and change name to “Himachal languages”; denotation encompasses various languages spoken in Himachal Pradesh (exact list to be determined).

Option 2 would appear to match most closely with MARC usage. Classification of Indo-Aryan languages at this level is not a completely-resolved matter, however. For instance, “Western Pahari” includes Pahari-Potwari (spoken primarily in Pakistan, not Himachal Pradesh), which is reportedly part of a dialect continuum with languages from a distinct branch of Indo-Aryan (in a sub-group known as “Lahnda”—see §6.8). An alternative would be to restrict the languages encompassed by [him] to languages spoken only in Himachal Pradesh (option 3), though it raises questions regarding its value as a language designation.

PROPOSED SOLUTION: Option 2.

3.4 Kachin

PROBLEM: ISO 639 has [kac] “Kachin” as an individual language. If the term “Kachin” is used as a specifically linguistic designation, it will most likely be used in reference to the Jingpho language. Most common use, however, is as an ethnonym (that is, an ethnic cover term) that refers to a collective sense of identity that crosses linguistic boundaries. The identity is based on historic and cultural affiliation and not on linguistic genetic relationship. In terms of languages, this would include several distinct language groups: Lisu, Lachit, Rawang, Zaiwa, Maru, Ngo Chang (Achang), Jingpho. These are from distinct branches at the highest level in the Tibeto-Burman family.

The 2003 version of the MARC Language Code List uses [kac] for Jingpho. (An earlier version used it for four languages from three different branches of Tibeto-Burman.)

POSSIBLE SOLUTIONS:

1. Specify denotation of [kac] as the single language Jingpho, and change the name to “Jingpho”
2. Deprecate [kac], documenting the meaning of the term and the problems related to its usage.

PROPOSED SOLUTION: Option 1.

3.5 Rajasthani

PROBLEM: ISO 639 has [raj] “Rajasthani” as an individual language. Although one can often find references to “Rajasthani” as an individual language, it is best thought of as a cover term for languages spoken in Rajasthan, a state of India.

“Rajasthani” is not a scheduled language of India, nor is it listed in the published results of the 1991 census; “Rajasthani” and “Bagri-Rajasthani” are both listed by the Central Institute of Indian Languages as “mothertongue” under “Hindi”, however. “Rajasthani” seems often to be viewed as closely related to Hindi.

“Rajasthani” is also used for a genetic sub-group of Indo-Aryan. Many of the languages of this sub-group are spoken in Rajasthan state, though some are spoken in Pakistan. The “Rajasthani” sub-group does not include Hindi.

The MARC Language Code List uses [raj] as a collective encompassing “Bagri”, “Gujari”, “Harauti”, “Jaipuri” and “Malvi”.

In other descriptions, “Rajasthani” is typically described as having several “dialects”. For example, one description lists “Bagri, Shekhawati, Mewati, Dhundhari, Harauti, Marwari, Mewari and Wagri.”[3] Comparing these lists with the inventory in *Ethnologue*, “Dhundhari”, “Jaipuri” and “Shekhawati” are listed as dialects of Marwari (*Ethnologue* [MKD]); “Mewati” is listed as an alternate name for “Mewari”.

Bagri, Gujari, Harauti, Malvi, Marwari and Mewari are listed in *Ethnologue* as languages in the Rajasthani genetic sub-group (though Gujari is not reported to be spoken in the state of Rajasthan); and “Wagri” is listed as an alternate name for Wagdi, a language spoken in Rajasthan state but not part of the Rajasthan genetic sub-group.

Marwari (*Ethnologue* [MKD]) appears to be the predominant variety among those denoted by “Rajasthani”. It should be noted, though, that "Marwari" is listed separately in ISO 639, and analysis of that item suggests that it should be considered a macro-language that encompasses Marwari varieties and Mewari as well (see §5.35).

Based on the MARC usage, the correct solution appears to be that "Rajasthani" and "Marwari" both be considered macro-languages, and that the Marwari varieties and Mewari be included in the latter but not the former. The only conflict with MARC usage is that "Jaipuri" would be removed from the scope of [raj].

POSSIBLE SOLUTIONS:

1. Specify the denotation of [raj] as “Marwari”. (Note that this would lead to issues of synonymy with [mwr]; see §5.35.)
2. Change the scope of [raj] from I to M; denotation encompasses languages associated with “Rajasthani” (including some not in the Rajasthani genetic sub-group), but *excluding* Marwari or Mewari: Bagri, Gade Lohar, Gujari, Harauti, Malvi, Wagdi.
3. Change the scope of [raj] from I to M; denotation encompasses Rajasthani languages spoken in the state of Rajasthan: Bagri, Harauti, Gade Lohar, Malvi, Marwari (*Ethnologue* [MKD]), Marwari (*Ethnologue* [MRI]), Mewari.
4. Change the scope of [raj] from I to M; denotation encompasses languages associated with “Rajasthani” (including some not in the Rajasthani genetic sub-group): Bagri, Gade Lohar, Gujari, Harauti, Malvi, Marwari (*Ethnologue* [MKD]), Mewari, Wagdi (Wagri).

[3] Interestingly, the author of that description has found it necessary to publish separate grammars for each of those eight “dialects”. This suggests that these varieties may be further apart from one another than true dialects, and that the term “dialect” is being used by this author in the non-technical sense of ‘variety that is perceived as sub-standard’ rather than the linguistic sense of ‘sub-variety within a language’.

5. Change the scope of [raj] from I to C (genetic) and change the name to “Rajasthani languages”; denotation encompasses fourteen languages of the Rajasthani genetic sub-group.
6. Deprecate [raj].

PROPOSED SOLUTION: Option 2.

4. ISO 639 “individual” languages: possible change of scope to “collection”

Where *Ethnologue* has multiple entries corresponding to a single individual-language entity in ISO 639-2, the relationship between the single entity in ISO 639-2 and the multiple entities to be added to ISO 639-3 must be resolved.

In the following cases, it is proposed that the scope of the existing ISO 639-2 entities be changed to collective-language (C). (This would entail a change of names to include “… languages”.)

4.1 [arn] “Araucanian”

Problem: ISO 639 has [arn] “Araucanian” as an individual language. *Ethnologue* lists “Araucanian” as the name of a language family that includes two languages: “Mapudungun”, also known as “Mapuche” or “Araucano”, spoken in Chile and Argentina, pop. est. 440,000; and “Huilliche”, spoken in Chile, pop. est. several thousand.

The MARC Language Code List uses [arn] for “Mapuche” and also for “Araucanian” and “Mapudungun”.

Possible solutions:

1. Specify denotation of [arn] as specifically Mapudungun; denotation does not include Huilliche. Change name to “Mapudungun”.
2. Change scope of [arn] from I to C (genetic) and change name to “Araucanian languages”; denotation encompasses languages of the Araucanian family.

PROPOSED SOLUTION: Option 1.

4.2 Banda

PROBLEM: ISO 639 has one category [bad] “Banda”. *Ethnologue* lists a genetic sub-group “Banda” (a branch of the Niger-Congo phylum) that includes 16 individual languages. Ten of these language are referred to as “Banda” or a close variation; they range in size from 3,000 to 180,000 (est.). None can be clearly identified as the denotatum for [bad].

PROPOSED SOLUTION: Change the scope of [bad] from I to C (genetic) and change the name to “Banda languages”; denotation encompasses all Banda languages.

4.3 Batak

PROBLEM: ISO 639 has [btk] “Batak”; *Ethnologue* lists seven Batak languages in the Batak sub-group of Western Malayo-Polynesian. From this alone, it is unclear which of these is the intended denotation, or whether it should encompass more than one of these. The MARC code list, however, describes this as a collective identifier corresponding to the Batak genetic sub-group.

PROPOSED SOLUTION: Change the scope of [btk] from I to C (genetic) and change the name to "Batak languages"; denotation encompasses all seven Batak languages.

4.4 Gondi

PROBLEM: ISO 639 has [gon] "Gondi". *Ethnologue* lists two languages called by this name, "Northern" and "Southern Gondi". These languages belong to a genetic sub-group of the Dravidian phylum known as "Gondi". The MARC Language Code List indicates that [gon] is also used as a collective that includes "Abujhmaria". This is another of the ten languages in the Gondi sub-group.[4]

It might have been appropriate to consider [gon] a macro language that includes Northern and Southern Gondi. Yet it has been used in MARC as a collective that includes languages from the Gondi sub-group other than these two. Milicent Wewerka has indicated that this change would be acceptable, however.

POSSIBLE SOLUTIONS:

1. Change scope of [gon] from I to M; denotation encompasses Northern and Southern Gondi.
2. Change scope of [gon] from I to C (genetic) and change name to "Gondi languages"; denotation encompasses the ten languages of the Gondi sub-group.

PROPOSED SOLUTION: Option 1.

4.5 Grebo

PROBLEM: ISO 639 has [grb] "Grebo". *Ethnologue* lists five languages spoken in Liberia that use this name. These languages are considered the Liberian genetic sub-group of a larger genetic classification known as the "Grebo" sub-group. The other four languages in the Grebo sub-group are not referred to using the name "Grebo".

MARC usage was intended to represent a single language, "Grebo".

POSSIBLE SOLUTIONS:

1. Change scope of [grb] from I to M; denotation encompasses five languages, "Barclayville Grebo", "Central Grebo", "Gboloo Grebo", "Northern Grebo" and "Southern Grebo".
2. Change scope of [grb] from I to C (genetic) and change name to "Grebo languages"; denotation encompasses the nine languages of the Grebo sub-group.

PROPOSED SOLUTION: 1.

4.6 Ijo

PROBLEM: ISO 639 has [ijo] "Ijo" as an individual language. *Ethnologue* lists three languages that use that name: "Izon", also known as "Ijo" or "Central Western Ijo", Nigeria, pop. est. 1,000,000; "Biseni", also known as "Northeast Central Ijo", Nigeria, pop. est. 4,800; "Southeast Ijo", Nigeria, pop. est. 71,500. These three languages are from distinct sub-groups of the Ijoid

[4] The 14th edition of *Ethnologue* lists eleven Gondi languages; since publication, one of these, Abujmaria, has been discovered to be included in another, Maria; this is documented in the change history data file that is published online semi-annually at http://www.ethnologue.com/codes/ChangeHistory.tab.

branch of the Niger-Congo phylum. The Ijoid branch includes ten languages; it is divided into two sub-groups, one known as “Ijo”, which includes nine of the ten languages.

The MARC Language Code List uses [ijo] for several similar language names. Most of these appear to be alternates or variations for one or more of the languages mentioned above. (MARC lists one use of [ijo] as for “Ido (African)”. *Ethnologue* lists “Ido” as an alternate name for a language from a distinct branch of the Niger-Congo phylum; we could assume, though, that this is intended to be just one more phonologically-similar variant of “Ijo”.) MARC also uses [ijo] for Nembe, which *Ethnologue* lists as a dialect of Southeast Ijo; and for Ibani, which is another language from the Ijoid branch of Niger-Congo (and in yet another sub-group distinct from those of Izon, Biseni and Southeast Ijo).

POSSIBLE SOLUTIONS:

1. Specify denotation of [ijo] as the single language “Izon”.
2. Change scope of [ijo] from I to M; denotation encompasses Izon (“Central-Western Ijo”), Biseni (“Northeast-Central Ijo”) and Southeast Ijo.
3. Change scope of [ijo] from I to C (genetic) and change name to “Ijo languages”; denotation encompasses nine languages of the Ijo sub-group of the Ijoid branch of Niger-Congo.
4. Change scope of [ijo] from I to C (genetic) and change name to “Ijoid languages”; denotation encompasses the ten languages of the Ijoid branch of Niger-Congo.

MARC usage appears to require options 3 or 4.

PROPOSED SOLUTION: 3.

4.7 Karen

Problem: ISO 639 has [kar] “Karen” as an individual language. *Ethnologue* lists 19 languages that are referred to as “Karen”. “Karen” is not the only name used for these languages, and may not be the preferred name in some cases. These languages belong to a branch of the Tibeto-Burman family that is known as “Karen”. There is only one language in the Karen branch for which “Karen” is not listed in *Ethnologue* as a primary or alternate name: Wewaw. There are significant divisions within the languages that comprise the Karen branch, linguistically, socio-linguistically and culturally.

The MARC Language Code List uses [kar] for “Karen” but also as a collective code for “Pwo Karen” or “Sgaw Karen”.

POSSIBLE SOLUTIONS:

1. Change the scope of [kar] from I to M; denotation encompasses 19 languages that are called “Karen”.
2. Change the scope of [kar] from I to C (genetic) and change the name to “Karen languages”; denotation encompasses all 20 languages of the Karen branch.

PROPOSED SOLUTION: 2

4.8 Kru

PROBLEM: ISO 639 has [kro] “Kru”. “Kru” is a genetic sub-group of the Niger-Congo phylum that encompasses some thirty-nine languages. This sub-group has an internal taxonomy that

includes a sub-group known as “Grebo”, which is also listed in ISO 639 as an individual language ([grb]; see §4.5).

The MARC Language Code List associates [kro] with the name “Kru (Other)”; it uses [kro] only as a collective.

PROPOSED SOLUTION: Change scope of [kro] from I to C (genetic) and change name to “Kru languages”; denotation encompasses thirty-nine languages of the Kru sub-group.[5]

4.9 Nahuatl

PROBLEM: ISO 639 has [nah] “Nahuatl”. *Ethnologue* lists twenty-eight languages called “Nahuatl”. They vary significantly in size and degree of development, but none is substantially larger or more developed than all the others. These languages comprise the “Aztec” genetic sub-group of Uto-Aztecan.

The MARC Language Code List uses [nah] for “Aztec” or “Mexican”, but also as a collective that encompasses “Pipil”. Pipil is an off-shoot Aztec language spoken in El Salvador and Honduras rather than Mexico. Other Aztec languages are spoken in Mexico no further south than Oaxaca, Veracruz and Tabasco, and are more closely-related to one another than they are to Pipil. A higher-level sub-group, “General Aztec”, encompasses the Nahuatl varieties plus Pipil.

POSSIBLE SOLUTIONS:

1. Change scope of [nah] from I to M; denotation encompasses the twenty-eight “Nahuatl” languages.
2. Change scope of [nah] from I to C (genetic) and change name to “Nahuatl languages”; denotation encompasses the twenty-eight “Nahuatl” languages.
3. Change scope of [nah] from I to C (genetic) and change name to “Aztec languages”; denotation encompasses the twenty-eight “Nahuatl” languages plus Pipil.

PROPOSED SOLUTION: 3

4.10 Occitan

PROBLEM: ISO 639 has [oc/oci] “Occitan (post 1500); Provençal”. *Ethnologue* lists six languages from the Oc genetic sub-group of Indo-European: Auvergnat, Gascon, Languedocian, Limousin, Provençal and Shaudit. It lists “Occitan” as an alternate name for the first four of these languages, but not Provençal or Shaudit.

The MARC Language Code List uses [oci] for “Occitan (post-1500)” and also for “Langue d’oc (post-1500) and “Provençal, Modern (post-1500)”; it also uses it as a collective that encompasses “Béarnais (post-1500)” and “Gascon (post-1500)”. “Béarnais” is listed in *Ethnologue* as a dialect of Gascon. The MARC list, then, refers to three of the six languages listed in *Ethnologue*.

Input I received from a representative of the software industry in France suggested a likelihood of implementations that would treat Auvergnat, Gason, Languedocian and Limousin as a single entity using the name “Occitan” that would be distinguished from Provençal and Shaudit.

POSSIBLE SOLUTIONS:

1. Select one modern Oc language to be the denotation of [oc/oci].

[5] An inclusive collection is proposed rather than a collection using the name “Kru (Other)”, in keeping with the recommendation in §2.

2. Change scope of [oc/oci] from I to M; denotation encompasses Auvergnat, Gascon, Languedocian and Limousin.
3. Change scope of [oc/oci] from I to C (genetic) and change name to “Oc languages”; denotation encompasses the six languages of the Oc sub-group.

PROPOSED SOLUTION: Opiton 2.

4.11 Quechua

PROBLEM: ISO 639 has [qu/que] “Quechua”. *Ethnologue* lists thirty-four “Quechua” languages, plus another ten languages that use the alternate pronunciation “Quichua”. These languages are spoken in Peru, Ecuador, Bolivia and Argentina.

“Quechua” is also the name of a language family that includes these forty-four languages plus two others, “Inga” or “Highland Inga”, and “Jungle” or “Lowland Inga”, both of which are spoken in Colombia.

Of the various Quechuan languages, a few are spoken by populations on the order of one to three million (Ayacucho, Cuzco, and South Bolivian Quechua; Chimborazo Highland Quichua) with the majority spoken by populations in the thousands or ten-thousands.

The MARC Language Code List uses [que] for “Quechua”, “Quichua”, “Inca” and “Runasimi” (another alternate for “Quechua”).

The linguistic diversity within Quechua stands in contrast with emerging nationalist sentiment that promotes a single identity. There is some potential that, over time, a small number of shared, written forms could emerge. One indicator is a current project involving development of Cuzco Quechua for use in localized software products (including terminology development), which is being presented to users as “Quechua” (i.e. without further distinction). There is no way to predict how the sociolinguistic situation will evolve, however.

POSSIBLE SOLUTIONS:

1. Change the scope of [que] from I to M; denotation encompasses thirty-four “Quechua” languages.
2. Change the scope of [que] from I to M; denotation encompasses fourty-four “Quechua” and “Quichua” languages.
3. Change the scope of [que] from I to C (genetic) and change the name to “Quechua languages”; denotation encompasses all forty-six languages of the Quechua family.

PROPOSED SOLUTION: Option 2, based on the promotion of Quechua as a single identity and language-development activities promoting particular varieties as the common “Quechua” language.

4.12 Romany

PROBLEM: ISO 639 has [rom] “Romany”. *Ethnologue* lists seven Romani languages: “Balkan Romani”, “Baltic Romani”, “Carpathian Romani”, “Kalo-Finnish Romani”, “Sinte Romani”, “Vlax Romani” and “Welsh Romani”. These languages comprise a genetic sub-group of the Indo-European phylum that is also known as “Romani”.

The MARC Language Code List uses [rom] for “Romany” and “Gypsy”, and also as a collective that encompasses “Caló (Romany)”.

POSSIBLE SOLUTIONS:

1. Change the scope of [rom] from I to M; denotation encompasses the seven languages of the Romani sub-group.
2. Change the scope of [rom] from I to C (genetic) and change the name to “Romany languages”; denotation encompasses the seven languages of the Romani sub-group.

PROPOSED SOLUTION: Option 1.

4.13 Zapotec

PROBLEM: ISO 639 has [zap] “Zapotec”. *Ethnologue* lists fifty-eight “Zapotec” languages. These comprise a genetic sub-group that is also known as “Zapotec”.

There is a high degree of linguistic diversity within Zapotec: the language family is reportedly comparable to Romance in historic depth and internal diversity. Given the size and status of Zapotec communities, it is unlikely that this diversity will survive in the long term. If significant language promotion ever occurs, especially if combined with nationalism, it is likely that a small number of shared, developed varieties would emerge. In spite of the linguistic diversity, there is a common cultural identity and history, and situations in which the language network is referred to as a single variety.

POSSIBLE SOLUTIONS:

1. Change the scope of [zap] from I to M; denotation encompasses the fifty-eight “Zapotec” languages.
2. Change the scope of [zap] from I to C (genetic); denotation encompasses the fifty-eight “Zapotec” languages of the Zapotec sub-group.

PROPOSED SOLUTION: Option 1.

4.14 Zande

PROBLEM: ISO 639 has [znd] “Zande”. The MARC Language Code List uses [znd] for “Zande”, for “Naym-Nyam” and also as a collective for “Nzakara”. The reference to “Nyam-Nyam” is reportedly based on information from Voegelin's Classification and Index of the World's Languages, where this term is given as an alternative name for Zande.

Ethnologue lists “Zande”, pop. est. 1,142,000; and also “Nzakara”, pop. est. 50,000. Both are spoken in the Democratic Republic of Congo and also in Central African Republic.

“Zande” is also the name of a genetic sub-group of the Niger-Congo phylum. This sub-group contains within it a sub-group known as “Zande-Nzakara”, which contains the Zande and Nzakara languages.

Ethnologue also lists “Nyam-Nyam” as an alternate name for “Nimbari”, a language spoken in Cameroon. It belongs to a different branch of Niger-Congo than Zande and Nzakara. This appears to be unrelated to the reference to “Nyam-Nyam” in MARC.

PROPOSED SOLUTION: Change scope of [znd] from I to C (genetic) and change name to “Zande languages”; denotation encompasses the six languages of the Zande genetic sub-group; it does not encompass “Nimbari”.

5. ISO 639 “individual” languages: possible change of scope to “macro-language”

Where *Ethnologue* has multiple entries corresponding to a single individual-language entity in ISO 639-2, the relationship between the single entity in ISO 639-2 and the multiple entities to be added to ISO 639-3 must be resolved. In the following cases, it is proposed that the scope of the existing ISO 639-2 entities be considered “macro-languages” (M) for purposes of defining relationship to ISO 639-3.

5.1 Albanian

PROBLEM: ISO 639 has one category [sq/sqi/alb] “Albanian”, whereas *Ethnologue* has four entries for Albanian languages: Gheg, Tosk, Arbëreshë and Arvanitika. Tosk is the national language of Albania; Gheg is the language of Kossovars; it is spoken by a similarly-sized population Tosk, is also a developed language, and is an official language of Serbia and Montenegro. Arbëreshë and Arvanitika are spoken in Italy and Greece by much smaller populations than Gheg or Tosk, and are less- or un-developed. All four varieties comprise a genetic sub-group; Gheg and Tosk alone do not.

POSSIBLE SOLUTIONS:

1. Specify denotation of [sq/sqi/alb] as Tosk.
2. Change the scope of [sq/sqi/alb] from I to M; denotation encompasses Tosk and Gheg.
3. Change the scope of [sq/sqi/alb] from I to M; denotation encompasses all four Albanian varieties.
4. Change the scope of [sq/sqi/alb] from I to C and change the name to “Albanian languages”; denotation encompasses all four varieties.

PROPOSED SOLUTION: alternative 3

5.2 Arabic

PROBLEM: ISO has one category [ar/ara] whereas *Ethnologue* has several: Standard Arabic, and over thirty un- or less-developed Arabic vernacular languages. These varieties alone do not comprise a genetic sub-group. Note that five of these are Judeo-Arabic varieties, for which there is also an ISO 639 identifier, and which do not appear to be known typically as “Arabic” (see §5.23).

In situations that have major, developed variety and several other undeveloped varieties, the normal recommendation would be to equate the ISO 639 category with the developed variety. In this case, however, because of the sociolinguistic nature of Arabic varieties, and because of existing use of [ar/ara] in IT implementations, it may make better sense to give [ar/ara] a scope of M, and have the denotation encompass all Arabic varieties in ISO 639-3.

POSSIBLE SOLUTIONS:

1. Specify denotation of [ar/ara] as Standard Arabic.
2. Change scope of [ar/ara] from I to M; denotation encompasses all Arabic varieties (including Judeo-Arabic varieties).
3. Change scope of [ar/ara] from I to M; denotation encompasses Arabic varieties other than Judeo-Arabic varieties.

PROPOSED SOLUTION: 3

5.3 Aymara

PROBLEM: ISO has one category [aym] "Aymara" whereas *Ethnologue* has "Central Aymara" (pop. est. 2,200,000) and "Southern Aymara" (no pop. est.). Central Aymara is developed, whereas Southern Aymara evidently is not. These two languages alone do not comprise a genetic sub-group.

POSSIBLE SOLUTIONS:

1. Specify denotation of [aym] as "Central Aymara".
2. Change scope of [aym] from I to M; denotation encompasses both Central and Southern Aymara.

PROPOSED SOLUTION: 2

5.4 Azerbaijani

PROBLEM: ISO has one category [az/aze], whereas *Ethnologue* has "North Azerbaijani" (national language of Azerbaijan) and "South Azerbaijani" (language of wider communication in Iran). These two languages alone do not comprise a genetic sub-group.

PROPOSED SOLUTION: Change scope of [az/aze] from I to M; denotation encompasses both "North" and "South Azerbaijani".

5.5 Baluchi

PROBLEM: ISO has [bal] "Baluchi"; *Ethnologue* has three entries: "Eastern Balochi", "Southern Balochi" and "Western Balochi", all are spoken primarily in Pakistan, have populations ranging from 1.8 to 3.4 million, and appear to have similar levels of development. These three languages alone do not comprise a genetic sub-group.

PROPOSED SOLUTION: Change scope of [bal] from I to M; denotation encompasses all three varieties.

5.6 Bikol

PROBLEM: ISO has [bik] "Bikol". *Ethnologue* lists "Bikol" as an alternate name for Central Bicolano. "Bikol" is also used to refer to a genetic sub-group of the Central Philippine family, which includes three "Agta" and five "Bicolano" languages. Of the varieties known as "Bicolano", Central Bicolano is by far the largest. It is not clear whether "Bikol" might be used to refer to other "Bicolano" languages, or only Central Bicolano. The Bicolano varieties alone do not comprise a genetic sub-group.

POSSIBLE SOLUTIONS:

1. Specify denotation of [bik] as Central Bicolano.
2. Change scope of [bik] from I to M; denotation encompasses all five "Bicolano" languages.
3. Change the scope of [bik] from I to C (genetic) and change the name to "Bikol languages"; denotation encompasses all eight Agta and Bicolano languages.

PROPOSED SOLUTION: 2

5.7 Buriat

PROBLEM: ISO 639 has [bua] "Buriat". *Ethnologue* 14 lists three languages. "Russia Buriat" is a literary variety (Cyrillic script) spoken west of Lake Irkutsk, with influences from Russian; pop. est. 318,000. "Mongolia Buriat" and "China Buriat" are distinct, with influences from Halh Mongolian and various other languages, respectively, each spoken by approximately 65,000. These three languages comprise a genetic sub-group also known as "Buriat".

POSSIBLE SOLUTIONS:

1. Specify denotation of [bua] as Russia Buriat.
2. Change scope of [bua] from I to M; denotation encompasses all three varieties.
3. Change scope of [bua] from I to C (genetic) and change the name to "Buriat languages"; denotation encompasses all three varieties.

PROPOSED SOLUTION: 2

5.8 Chinese

PROBLEM: ISO 639 has [zh/zho/chi] "Chinese"; *Ethnologue* lists thirteen languages that are referred to as "Chinese" and that are spoken in China; these are all the languages of the Chinese sub-group of Sino-Tibetan except for Dungan, which is spoken in Kyrgyzstan and other countries of southwestern Asia. These thirteen varieties are clearly distinct languages, yet "Chinese" is used in some contexts as though this was a single language. (This is reinforced in part by the fact that a shared ideographic writing system makes written material, at least to a significant extent, intelligible across languages, even though the spoken form of the language would not be.)

Due to existing implementations, in particular, IANA registrations (zh-guoyo, zh-yue, etc.), [zh/zho/chi] must encompass these various Chinese languages.

PROPOSED SOLUTION: change scope of [zh/zho/chi] from I to M; denotation encompasses all thirteen Chinese languages spoken in China (these may also be spoken in other countries; it excludes Dungan).

5.9 Cree

PROBLEM: ISO 639 has [cre] "Cree". *Ethnologue* lists six Cree languages; these represent all of the languages from the Cree-Montagnais-Naskapi sub-group of Algonquian that are known as "Cree"; other languages from that sub-group are Atikamekw, Montagnias and Naskapi. The MARC Language Code List indicates that [cre] is used for "Cris", "Kristineaux" and "Maskegon" (as an alternate name for Swampy Cree, not the language known as "Muskogee" or "Creek"), and also as a collective code for "Montagnais" and "Naskapi".

POSSIBLE SOLUTIONS:

1. Change scope of [cre] from I to M; denotation encompasses six Cree varieties.
2. Change scope of [cre] from I to C (genetic); change name to "Cree-Montagnais-Naskapi languages"; denotation encompasses all nine languages of the Cree-Montagnais-Naskapi sub-group of Algonquian (six Cree languages, Atikamekw, Montagnais and Naskapi).

Note that both of these options would differ from MARC usage in the exclusion of "Maskegon".

PROPOSED SOLUTION: Option 1 (differs from past MARC usage, but Milicent Wewerka has indicated that MARC usage can be revised with regard to Montagnais and Naskapi.).

5.10 Delaware

PROBLEM: ISO 639 has [del] "Delaware". *Ethnologue* lists two languages that are referred to by this name: "Munsee" and "Unami". MARC uses [del] for "Lenape" and "Lenni Lenapi" (altnernate names for Unami), and also for "Munsee". Munsee and Unami are two out of ten languages from the Eastern sub-group of the Algonquian family; other languages from this family are not known by the name "Delaware".

PROPOSED SOLUTION: Change scope of [del] from I to M; denotation encompasses Munsee and Unami (may also encompass other extinct varieties if later identified).

5.11 Dinka

PROBLEM: ISO 639 has [din] "Dinka"; *Ethnologue* lists five languages known as "Dinka". These languages comprise a genetic sub-group of the Nilo-Saharan phylum known as Dinka.

POSSIBLE SOLUTIONS:

1. Change scope of [din] from I to M; denotation encompasses all five Dinka languages.
2. Change scope of [din] from I to C (genetic); change name to "Dinka languages"; denotation encompasses all five Dinka languages.

PROPOSED SOLUTION: 1

5.12 Mari (Chemeris)

PROBLEM: ISO 639 has [chm] "Mari". *Ethnologue* lists two languages: "High Mari", or "Hills Mari", which is spoken by a smaller population (est. 66,000) and is not a highly developed language; and "Low Mari", or "Woods Mari", which spoken by a much larger population (est. 526,000) and is the developed variety (taught in schools, used in mass media). These two languages are considered a genetic sub-group.

The MARC Language Code List uses [chm] for "Mari" and for "Cheremissian".

POSSIBLE SOLUTIONS:

1. Specify denotation of [chm] as "Low Mari".
2. Change scope of [chm] from I to M; denotation encompasses both varieties.
3. Change scope of [chm] from I to C (genetic) and change name to "Mari languages"; denotation encompasses both varieties.

PROPOSED SOLUTION: 2

5.13 Slavey

PROBLEM: ISO 639 has [den] "Slave (Athapascan)". *Ethnologue* lists two languages, North Slavey and South Slavey. These two languages alone do not comprise a genetic sub-group.

PROPOSED SOLUTION: Change scope of [den] from I to M; denotation encompasses both North and South Slavey.

5.14 Fijian

PROBLEM: ISO 639 has [fij] "Fijian". *Ethnologue* lists two languages: Fijian, also known as "Fiji", "Eastern Fijian", "Standard Fijian" or "Nadroga"; this is the developed, standard variety

spoken in Fiji, Nauru, New Zealand and Vanuatu; pop. est. 350,000. The second is Western Fijian, also known as “Fiji” or “Nadroga”; it is a non-standard variety spoken only in Fiji; pop. est. 57,000. These languages come from distinct branches of Eastern Malayo-Polynesian.

POSSIBLE SOLUTIONS:

1. Specify denotation of [fij] as “Standard Fijian” (or “Eastern Fijian”).
2. Change scope of [fij] from I to M; encompasses both Eastern/Standard Fijian and Western Fijian.

PROPOSED SOLUTION: The proximity and common names of these languages provide motivation for the use of the macro-language scope, i.e. solution 2. However, the fact that these languages are from distinct branches of Eastern Malayo-Polynesian suggests that they should be quite distinct. (In fact, the classification of these languages suggests that Standard Fijian should be closer to languages such as Hawaiian and Maori than it is to Western Fijian.). For this reason, solution 1 is proposed.

5.15 Frisian

PROBLEM: ISO 639 has [fy/fry] “Frisian; *Ethnologue* lists three Frisian languages: Eastern Frisian, Western Frisian, Northern Frisian. Eastern and Northern Frisian are spoken in Germany; Western Frisian is spoken in the Netherlands by a much larger population than the other two varieties, and is the only one of the three that has any level of current development. These three languages are considered a genetic sub-group.

POSSIBLE SOLUTIONS:

1. Specify denotation of [fy/fry] as Western Frisian.
2. Change scope of [fy/fry] from I to C (genetic) and change the name to “Frisian languages”; denotation encompasses all three varieties.
3. Change scope of [fy/fry] from I to M; denotation encompasses all three varieties.

PROPOSED SOLUTION: 3

5.16 Fulah

PROBLEM: ISO 639 has [ful] “Fulah”. This name is used for Fulani languages, of which *Ethnologue* lists 9. These languages are spoken in sub-Saharan Africa from Cameroon to Senegal with speaker populations ranging from 150,000 to 7,500,000. No one of these is significantly larger or more highly developed than all the others. These nine languages together comprise the Fulani genetic sub-group.

POSSIBLE SOLUTIONS:

1. Change scope of [ful] from I to C (genetic) and change the name to “Fulani languages”; denotation encompasses all nine Fulani languages.
2. Change scope of [ful] from I to M; denotation encompasses all nine languages.

PROPOSED SOLUTION: Option 2 (assumes there are contexts in which these languages are treated as though a single language).

5.17 Gbaya

PROBLEM: ISO 639 has [gba] "Gbaya". The *Ethnologue* lists four related languages that use this name: Northwest Gbaya, Southwest Gbaya, Gbaya-bossangoa and Gbaya-bozoum It is also used in a dialect name (Gbaya de Boda) for another related language, Bokoto, and various sources include Bokoto as part of the Gbaya ethnic and linguistic identity. The Ngbaka language (also related) is sometimes known as "Ngbaka Ngaya", though it appears to have a distinct identity. These languages alone do not comprise a genetic subgroup.

POSSIBLE SOLUTIONS:

1. Change scope of [gba] from I to M; denotation encompasses the four Gbaya languages listed above.
2. Change scope of [gba] from I to M; denotation encompasses the four Gbaya languages listed above plus Bokoto.
3. Change scope of [gba] from I to M; denotation encompasses the four Gbaya languages listed above plus Bokoto and Ngbaka.

PROPOSED SOLUTION: 2

5.18 Guarani

PROBLEM: ISO 639 has [gn/grn] "Guarani". *Ethnologue* lists four languages that use this name: "Paraguayan Guaraní", a national language of Paraguay, also spoken in Argentia, pop. est. 5,000,000; "Eastern Bolivian Guaraní", spoken in Bolivia, Paraguay and Argentina, pop. est. 32,000; "Western Bolivian Guaraní", pop. est. 5,000; and "Mbyá Guaraní", spoken in Paraguay, Argentina and Brazil, pop. est. 12,000. "Ava Guaraní" is also cited as an alternate name for the Chiripá language (pop. est. 12,000), and in MARC usage [grn] encompasses Chiripá. These languages belong to a branch of the Tupi language family that is also known as Guaraní.

POSSIBLE SOLUTIONS:

1. Specify denotation of [gn/grn] as specifically "Paraguayan Guaraní".
2. Change scope of [gn/grn] from I to M; denotation encompasses the four languages "Paraguayan Guaraní", "Eastern Bolivian Guaraní", "Western Bolivian Guaraní" and "Mbyá Guaraní".
3. Change scope of [gn/grn] from I to M; denotation encompasses the five languages "Paraguayan Guaraní", "Eastern Bolivian Guaraní", "Western Bolivian Guaraní", "Mbyá Guaraní" and "Chiripá".
4. Change scope of [gn/grn] from I to C (genetic) and change name to "Guarani languages"; denotation encompasses the 10 languages of the Guaraní sub-group of the Tupi family.

PROPOSED SOLUTION: 3

5.19 Haida

PROBLEM: ISO 639 has [hai] "Haida". *Ethnologue* lists two languages, "Northern Haida" and "Southern Haida". These two languages comprise the Haida genetic sub-group of the Na-Dene language phylum.

POSSIBLE SOLUTIONS:

1. Change scope of [hai] from I to C (genetic) and change name to “Haida languages”; denotation encompasses both Haida languages.
2. Change scope of [hai] from I to M; denotation encompasses both Haida languages.

PROPOSED SOLUTION: 2

5.20 Hmong

PROBLEM: ISO 639 has [hmn] “Hmong”. *Ethnologue* lists twenty-one languages that are known as “Hmong”. These languages are from three different branches of the Hmong-Mien family (also known as Miao-Yao). They do not alone comprise a genetic sub-group of Hmong-Mien. No one of these twenty-one languages is significantly larger or more developed than all the others.

PROPOSED SOLUTION: Change scope of [hmn] from I to M; denotation encompasses all twenty-one Hmong languages.

5.21 Inuktitut

PROBLEM: ISO 639 has [iku] “Inuktitut”. *Ethnologue* lists three languages known by that name: “Eastern Canadian Inuktitut”, “Western Canadian Inuktitut” and “Greenlandic Inuktitut”. The latter is also known as “Kalaallisut” and has its own identifier in ISO 639, [kl/kal].

The MARC Language Code List uses [iku] for “Inuktitut” and also for “Inuit”. The latter is covered in MARC by three terms, “Inuktitut”, “Inupiaq” and “Kalaallisut”. (That is, MARC requires one to choose between these three varieties; it does not have an identifier that encompasses all “Inuit” varieties).

POSSIBLE SOLUTIONS:

1. Change scope of [iku] from I to M; denotation encompasses the Inuktitut varieties spoken in Canada, but not that spoken in Greenland (Kalaallisut).
2. Change scope of [iku] from I to M; denotation encompasses all three Inuktitut varieties.

PROPOSED SOLUTION: 1

5.22 Inupiaq

PROBLEM: ISO 639 has [ik/ipk] “Inupiaq”. *Ethnologue* lists two languages, “North Alaskan Inupiatun” and “Northwest Alaskan Inupiatun”. It cites “Inupiaq” as only used in Canada, for “North Alaskan Inupiatun”, but there is other evidence for the use of “Inupiaq” in relation to varieties spoken in Alaska.

POSSIBLE SOLUTIONS:

1. Specify denotation of [ik/ipk] as “North Alaskan Inupiatun”.
2. Change scope of [ik/ipk] from I to M; denotation encompasses both Inupiatun varieties.

PROPOSED SOLUTION: 2

5.23 Judeo-Arabic

PROBLEM: ISO 639 lists [jrb] “Judeo-Arabic”. *Ethnologue* lists five languages that are known as “Judeo-Arabic”. These are Arabic varieties spoken by Jewish communities and written using the Hebrew script.

PROPOSED SOLUTION: Change scope of [jrb] from I to M; denotation encompasses the five Judeo-Arabic languages listed in *Ethnologue*.

5.24 Kanuri

PROBLEM: ISO 639 lists [kau] "Kanuri". *Ethnologue* lists three languages: "Central Kanuri", a national language of Nigeria, also spoken in neighboring countries, pop. est. 3,500,000; "Manga Kanuri", spoken in Niger and Nigeria, pop. est. 480,000; and "Tumari Kanuri", spoken in Niger, pop. est. 40,000.

There is also a genetic sub-group of the Nilo-Saharan phylum that is known as "Kanuri". It is comprised of these three languages plus Kanembu.

The MARC Language Code List uses [kau] for "Kanuri" and for "Bornu", which *Ethnologue* lists as an alternate name for Central Kanuri (but not the other Kanuri languages).

POSSIBLE SOLUTIONS:

1. Specify denotation of [kau] as specifically Central Kanuri.
2. Change scope of [kau] from I to M; denotation encompasses three Kanuri languages.
3. Change scope of [kau] from I to C (genetic) and change name to "Kanuri languages"; denotation encompasses the three Kanuri varieties plus Kanembu.

PROPOSED SOLUTION: 2

5.25 Khmer

PROBLEM: ISO 639 has [km/khm] "Khmer". *Ethnologue* lists two languages: "Central Khmer", the national language of Cambodia, also spoken by a diaspora outside Southeast Asia, pop. est. 7,000,000; and "Northern Khmer", spoken in Surin province of Thailand, pop. est. 1,000,000. Northern Khmer is recognized by regional administrations as a distinct but closely-related language. It is not spoken in Cambodia; any literature has used Thai script, not Khmer script.

The MARC Language Code List uses [khm] for "Cambodian" and also as a collective that encompasses "Surin Khmer".

POSSIBLE SOLUTIONS:

1. Specify denotation of [km/khm] as specifically Central Khmer.
2. Change scope of [km/khm] from I to M; denotation encompasses both Khmer languages.
3. Change scope of [km/khm] from I to C (genetic) and change name to "Khmer languages"; denotation encompasses both Khmer languages.

MARC usage suggests the need to adopt option 2 or 3. This is potentially risky in relation to language resources, however: in existing usage in software implementations and on the Internet, [km/khm] would likely be used only for Central Khmer, and confusion with Northern Khmer, which is recognized as distinct, could lead to significant concerns on the part of general users or government bodies. Also, since a script distinction coincides with the language distinction, text-based software processes or resources intended for Central Khmer would not be used for Northern Khmer as well.

PROPOSED SOLUTION: Option 1 (differs from past MARC usage, though Milicent Wewerka has indicated that this solution would be acceptable).

5.26 Konkani

PROBLEM: ISO 639 has [kok] "Konkani". *Ethnologue* lists two languages: "Konkani" or "Standard Konkani", spoken along the west of India from northern Maharashtra to Kerala, pop. est. 4,000,000; and "Goanese Konkani" ("Gomtaki", "Goan"), spoken from southern Maharashtra down to Kerala, pop. est. 2,000,000.

These languages are among seven that form a genetic sub-group of Indo-Aryan that is also known as "Konkani". Existing software implementations treat [kok] as an individual language, though libraries have numerous records that use [kok] to include both Standard and Goanese Konkani.

POSSIBLE SOLUTIONS:

1. Specify denotation of [kok] as specifically Standard Konkani.
2. Change scope of [kok] from I to M; denotation encompasses both Standard Konkani and Goanese Konkani.

PROPOSED SOLUTION: 2

5.27 Komi

PROBLEM: ISO 639 has [kv/kom] "Komi". *Ethnologue* lists two languages: "Komi-Permyak" and "Komi-Zyrian". These two languages alone do not comprise a genetic sub-group.

The MARC Language Code List uses [kom] for "Zyrian" and also as a collective that encompasses "Komi-Permyak".

PROPOSED SOLUTION: Change scope of [kom] from I to M; denotation encompasses "Komi-Permyak" and "Komi-Zyrian".

5.28 Kongo

PROBLEM: ISO 639 has [kon] Kongo. *Ethnologue* lists two languages: "Kongo" or "Kikongo", a national language of the Democratic Republic of Congo, pop. est. 3,200,000; and "San Salvador Kongo", also known as "Kikongo" or "Congo", pop. est. 1,500,000. These are among seven languages that comprise a genetic sub-group that is also known as "Kongo".

The *MARC Language Code List* uses [kon] for "Kongo" and other alternative names for the same language, and for "Fiote", which is a dialect of Kongo. It also uses [kon] as a collective that encompasses "Laadi" and "Ntaandu", which are dialects of Kongo, and also "Kituba". One can find references to "Kituba" as a dialect of Kongo or even an alternate name for Kongo. But Kituba is, in fact, a separate language, a Kongo-based creole, used as a language of wider communication.

There are two problems regarding denotation, then. The first is whether [kon] should be considered a macro-language that encompasses San Salvador Kongo as well as Kongo. The second has to do with the relationship between [kon] and "Kituba".

POSSIBLE SOLUTIONS:

1. Specify denotation of [kon] as Kongo, excluding San Salvador Kongo and Kituba.
2. Change scope of [kon] from I to M; denotation encompasses Kongo and San Salvador Kongo.
3. Change scope of [kon] from I to M; denotation encompasses Kongo, San Salvador Kongo and Kituba.

PROPOSED SOLUTION: Option 2 (differs from MARC usage—on the assumption that the creole language Kituba must be quite distinct from the two Kongo languages; Milicent Wewerka has indicated that MARC can be revised accordingly).

5.29 Kpelle

PROBLEM: ISO 639 has [kpe] "Kpelle". *Ethnologue* lists two languages: Guinea Kpelle, spoken in southeast Guinea, pop. est. 308,000; and Liberia Kpelle, spoken in Liberia, pop. est. 487,400.

PROPOSED SOLUTION: Change scope of [kpe] from I to M; denotation encompasses both Kpelle varieties.

5.30 Kurdish

PROBLEM: ISO 639 has [ku/kur] "Kurdish". *Ethnologue* lists two languages: "Kurmanji", or "Northern Kurdish", spoken in Turkey and numerous other countries, pop. est. 8,000,000; and "Kurdi", or "Southern Kurdish", spoken in Iraq and Iran, pop. est. 6,000,000.

PROPOSED SOLUTION: Change scope of [ku/kur] from I to M; denotation encompasses both Kurdi and Kurmanji.

5.31 Mandingo

PROBLEM: ISO 639 has [man] "Mandingo". *Ethnologue* lists "Mandingo" as an alternate name for "Mandinka" (*Ethnologue* [MNK]), also known as "Mande", which is spoken in Senegal and also in Gambia and Guinea-Bissau.

The MARC Language Code List uses [man] for "Mandingo" and "Mandeka", but also for "Malinka", and for "Maninka and "Meninka".

Regarding the names cited by MARC, Malinka is listed in *Ethnologue* as a distinct language (*Ethnologue* [MLQ]), though closely-related to Mandinka. "Malinka" is listed in the constitution of Senegal as a national language. It is also known as "Northwestern Maninka" or "Western Maninkakan". It is also spoken in Mali, Gambia and Guinea.

The name "Maninka" could apply to "Malinka", but also to other languages from the Manding sub-group of the Niger-Congo phylum. One of these, Kita Maninkanan (also known as "Kita Maninka" or "Malinke") is from the same division within the Manding sub-group as Malinka and Mandinka; it is spoken in Mali. The others are also known as "Maninka" but are from a distinct branch within the Manding subgroup (alternate names are shown in parentheses):

> Forest Maninka, Kankan Maninka (Southern Maninka, Mande), Konyanka Maninka (Konya, Konyakakan), Sankarkan Maninka (Faranah, Sankarkan),

These languages are spoken in Côte d'Ivoire, Guinea and Sierre Leone.

These languages alone do not comprise a genetic sub-group.

POSSIBLE SOLUTIONS:

1. Change scope of [man] from I to M; denotation encompasses Mandinka (*Ethnologue* [MNK]) and Malinka (*Ethnologue* [MLQ]).
2. Change scope of [man] from I to M; denotation encompasses the seven languages listed in Ethnololgue that use the names "Mandinka", "Malinka" or "Maninka" (viz. Mandinka, Malinka, Kita Maninkanan, Forest Maninka, Kankan Maninka, Konyanka Maninka and Sankarkan Maninka).

3. Change scope of [man] from I to C (ad hoc) and change the name to "Maninka/Mandinka languages"; denotation encompasses the seven languages listed in Ethnololgue that use the names "Mandinka", "Malinka" or "Maninka" (viz. Mandinka, Malinka, Kita Maninkanan, Forest Maninka, Kankan Maninka, Konyanka Maninka and Sankarkan Maninka).

PROPOSED SOLUTION: Option 2.

5.32 Malay

PROBLEM: ISO 639 has [ms/msa/may] "Malay". *Ethnologue* lists "Malay", also known as "Bahasa Malaysia". It also lists several other languages that use the name "Malay". Most are non-standard languages closely related to Bahasa Malaysia:

> Varieties spoken in Sabah: Banjar (Banjar Malay), Cocos Islands Malay, Sabah Malay (Bazaar Malay),
>
> Varieties spoken in Indonesia: Banjar (Banjar Malay), Berau Malay, Bukit Malay, Jambi Malay, Kota Bangun Kutai Malay, Menadonese Malay, North Moluccan Malay, Tenggarong Kutai Malay
>
> Varieties spoken in Thailand: Kedah Malay, Pattani Malay

(Additional note on "Sabah Malay": This is a variety used primarily by speakers of Malayic languages other than Bahasa Malaysia as second language for communication with speakers of other varieties. It is not fully developed, and usage is diglossic, with speakers shifting to other languages for lexica in certain domains.)

There are other languages in the same immediate genetic sub-group as Bahasa Malaysia (the Local Malay sub-group) that are not listed in *Ethnologue* as being known by the name "Malay".

Ethnologue also lists "Negeri Sembilan Malay" (spoken in Peninsular Malaysia), which is not from the Local Malay sub-group but rather from a different sub-group at the same level in the classification taxonomy. Speakers refer to themselves as "Orang Negeri".

Ethnologue also lists several Malay-based Creoles that use the name "Malay"

> Ambonese Malay (Melayu Ambon), Baba Malay (Straits Malay, Chinese Malay), Betawi (Betawi Malay, Jakarta Malay), Kupang Malay, Malaccan Creole Malay (Chitties Creole Malay), Sri Lankan Creole Malay

Of course, Bahasa Indonesia is also closely related to Bahasa Malaysia.

The MARC Language Code List uses [may] for Bahasa Malaysia and other languages from the Local Malay sub-group: Enim, Kaya Agung, Lembak, Palembang, Semendo and Siladang (Lubu), all of which are spoken in Sumatra. MARC also uses [may] for Urak Lawoi', a Malayic language from a different sub-group than Local Malay that is spoken in Thailand. It also uses [may] for the Malay-based creoles Ambonese Malay and Betawi.

There are existing software implementations that use [ms/msa/may] for "Malay" (Bahasa Malaysia) of Malaysia and of Brunei.

POSSIBLE SOLUTIONS:

1. Specify denotation of [ms/msa/may] as specifically Bahasa Malaysia; denotation does not encompass non-standard varieties or Malay-based creoles.

2. Change scope of [ms/msa/may] from I to M; denotation encompasses Bahasa Malaysia and Sabah Malay (the Local Malay varieties used primarily in Malaysia; would encompass Bahasa Malaysia as used in Brunei).
3. Change scope of [ms/msa/may] from I to M; denotation encompasses the thirteen languages from the Local Malay sub-group that use the name "Malay".
4. Change scope of [ms/msa/may] from I to C (ad hoc); denotation encompasses varieties that are called "Malay", whether Local Malay, other Malayic or Malay-based Creoles, and whether spoken in Malaysia, Brunei, Indonesia or Thailand.
5. Change scope of [ms/msa/may] from I to C (genetic) and change name to "Malay languages"; denotation encompasses the thirty-eight languages of the Local Malay sub-group.

PROPOSED SOLUTION: Option 3.

5.33 Malagasy

PROBLEM: ISO 639 has [mg/mlg] "Malagasy". The 14th edition of *Ethnologue* lists four "Malagasy" languages. Since the publication of the 14th edition, a language survey of Madagascar has been undertaken, the results of which indicate ten distinct "Malagasy" languages:

> "Antankarana Malagasy", "Bara Malagasy", "Masikoro Malagasy", "Northern Betsimisaraka Malagasy", "Plateau Malagasy", "Sakalava Malagasy", "Southern Betsimisaraka Malagasy", "Tandroy Malagasy", "Tanosy Malagasy", "Tsimihety Malagasy"

Plateau Malagasy is what would be considered the standard variety.

"Malagasy" is also used as the name of a genetic sub-group of Austronesian. As cited in the 14th edition of *Ethnologue*, the Malagasy sub-group includes the various "Malagasy" varieties as well as a language not referred to using the name "Malagasy" and spoken on the island of Mayotte: "Bushi", also known as "Sakalava".

The MARC Language Code List uses [mlg] for Standard Malagasy and also as a collective that includes "Tsimihety" and "Sakalava". In terms of correspondence with *Ethnologue* , the latter appears to corresponds to "Northern Sakalava Malagasy" rather than to "Bushi".

POSSIBLE SOLUTIONS:

1. Specify denotation of [mg/mlg] as specifically Standard Malagasy (i.e. "Plateau Malagasy").
2. Change scope of [mg/mlg] from I to M; denotation encompasses the ten "Malagasy" varieties.
3. Change scope of [mg/mlg] from I to C (genetic) and change name to "Malagasy languages"; denotation encompasses the languages of the Malagasy genetic sub-group.

PROPOSED SOLUTION: Option 2.

5.34 Mongolian

PROBLEM: ISO 639 has [mn/mon] "Mongolian". *Ethnologue* lists two languages: "Halh Mongolian", also known as "Central Mongolian" or "Khalkha", spoken primarily in Mongolia and the official language of that country, pop. est. 2,329,000; and "Peripheral Mongolian", also

known as “Southern-Eastern” or “Inner Mongolian”, spoken primarily in China, pop. est. 4,807,000.

“Mongolian” is also used for one of the main branches of the Altaic language phylum.

There are also other languages from the Mongolian family—varieties of Buriat or Kalmyk-Oirat—that have alternate names using “Mongolian”; e.g. “Northern Mongolian” for Mongolia Buriat, or “Western Mongol” for Kalmyk-Oirat.

MARC usage encompasses both Halh and Peripheral Mongolian.

PROPOSED SOLUTION: Change scope of [mn/mon] from I to M; denotation encompasses Halh Mongolian and Peripheral Mongolian.

5.35 Marwari

PROBLEM: ISO 639 has [mwr] “Marwari”. *Ethnologue* lists three languages: Marwari (*Ethnologue* [MKD]), spoken in India and also Nepal, pop. est. 12,963,000; “Marwari”, also known as “Marwari Meghwar” or “Jaiselmer”, spoken in Pakistan, pop. est. 220,000; and “Mewari” (*Ethnologue* [MTR]), spoken in India, pop. est. 1,220,000. The entries in *Ethnologue* note uncertainty as to whether “Mewari” is, in fact, distinct from Marwari (India) or not.

“Marwari” is also used for a genetic sub-group of Indo-Aryan that includes the three languages list above plus three others with small speaker populations.

The MARC Language Code List has [mwr] used for “Marwari” and also as a collective that encompasses “Dingal” and “Mewari”. The former refers to a literary form used in the court of Marwar, which existed prior to unification with India in 1948.

PROPOSED SOLUTION: Change scope of [mwr] from I to M; denotation encompasses both “Marwari” varieties and “Mewari”.

5.36 Norwegian. Nynorsk, Bokmal

PROBLEM: ISO 639 has three entries: [no/nor] “Norwegian”, [nn/nno] “Norwegian Nynorsk” and [nb/nob] “Norwegian Bokmål. *Ethnologue* lists two languages: “Nynorsk Norwegian” and “Bokmaal Norwegian”.

ISO 639-3 must have entries for “Nynorsk” ([nn/nno]) and “Boksmål” ([nb/nob]). The question is whether it should also have an entry for “Norwegian” ([no/nor]).

PROPOSED SOLUTION: Change status of [no/nor] from I to M; denotation encompasses [nn/nno] and [nb/nob].

5.37 Ojibwa

PROBLEM: ISO 639 has [oji] “Ojibwa”. *Ethnologue* lists seven languages that are referred to as “Ojibwa”: “Central Ojibwa”, “Eastern Ojibwa”, “Northwestern Ojibwa”, “Severn Obijwa”, “Western Ojibwa” (also known as “Saulteaux” or “Plains Ojibwa”), “Ottawa” (also known as “Ojibwa”), and “Chippewa” (also known as “Southwestern Ojibwa”). These belong to a genetic sub-group also known as “Ojibwa”.

The Ojibwa sub-group includes one other language, “Algonquin”. Algonquins are aware of a close relationship with Ojibwas, but there is not a strong perception of just *how* they are related, and the name “Ojibwa” is not typically used for this language. Thus, the “Ojibwa” identity

appears to be somewhat more tenuous in this case than with the other varieties, including Ottawa and Chippewa.

The MARC Language Code List uses [oji] for “Ojibwa” and also for “Cheppewa” and “Saulteaux”.

POSSIBLE SOLUTIONS:

1. Change the scope of [oji] from I to M; denotation encompasses those languages for which “Ojibwa” is the primary name: Central Ojibwa, Eastern Ojibwa, Northwestern Ojibwa, Severn Obijwa and Western Ojibwa.
2. Change the scope of [oji] from I to M; denotation encompasses the seven languages for which the name “Ojibwa” can be used: Central Ojibwa, Eastern Ojibwa, Northwestern Ojibwa, Severn Ojibwa, Western Ojibwa, Ottawa, and Chippewa.
3. Change the scope of [oji] from I to C (genetic) and change the name to “Ojibwa languages”; denotation encompasses the eight languages of the Ojibwa sub-group (including Algonquin).

PROPOSED SOLUTION: 2

5.38 Oromo

PROBLEM: ISO 639 has [om/orm] “Oromo”. *Ethnologue* lists three languages: “Borana-Arsi-Guji Oromo”, also known as “Afan Oromo”, “Boran”, “Southern Oromo”, “Galla” or “Gallinya”; “Eastern Oromo”; and “West-Central Oromo”. It also lists a further language from the same genetic sub-group, “Orma” which is referred to in some sources as “Orma-Oromo”.

The MARC Language Code List uses [orm] for “Oromo”, for “Afan”, “Galla” and “Gallinya”, and also as a collective that encompasses “Boran”.

PROPOSED SOLUTION: Change scope of [om/orm] from I to M; denotation encompasses all four Oromo languages (including Orma).

5.39 Persian

PROBLEM: ISO 639 has [fa/fas/per] “Persian”. *Ethnologue* lists two languages: “Eastern Farsi”, also known as “Persian” or “Dari”, spoken primarily in Afghanistan and also in Pakistan, pop. est. 7,000,000; and “Western Farsi”, also known as “Persian”, spoken primarily in Iran, pop. est. 24,280,000.

The MARC Language Code List uses [per] for “Persian” and “Farsi”, and also as a collective that encompasses “Dari”.

PROPOSED SOLUTION: Change scope of [fa/fas/per] from I to M; denotation encompasses both Eastern and Western Persian.

5.40 Pushto

PROBLEM: ISO 639 has [ps/pus] “Pushto”. *Ethnologue* lists three languages: “Northern Pashto”, also known as “Pakhto” or “Afghan”, spoken in Pakistan and Afghanistan, pop. est. 9,685,000; “Central Pashto”, spoken in Pakistan; and “Southern Pashto”, spoken in Afghanistan and also in Iran and Pakistan, pop. est. 9,204,000.

The MARC Language Code List uses [pus] for “Pushto” and also for “Afghan” and “Pakhto”.

PROPOSED SOLUTION: Change scope of [ps/pus] from I to M; denotation encompasses all three Pashto languages.

5.41 Songhai

PROBLEM: ISO 639 has [son] "Songhai". *Ethnologue* lists three languages: "Humburi Senni Songhay", or "Central Songai", spoken in Mali and Burkina Faso, pop. est. 140,000; "Chiini Koyra Songhay", or "West Songhoy", spoken in Mali, pop. est. 200,000; and "Senni Koyraboro Songhay", or "East Songhay", spoken in Mali, pop. est. 400,000. Songhai varieties are being actively developed by the Mali government.

The MARC Language Code List uses [son] for "Songhai" but also as a collective that encompasses "Dendi" and "Zarma". These languages together with the three listed above comprise the Southern sub-group of the Songhai sub-group of the Nilo-Saharan phylum. Dendi is spoken in Benin and Nigeria, pop. est. 31,000. Zarma, or "Djerma", is a national language of Niger, and is also spoken in Mali, Burkina Faso, Benin and Nigeria, pop. est. 2,100,000.

One finds occasional references to Zarma as "Songhay"; documents I have seen from the Mali government make reference to "Djerma" but now "Songhay", however. I have not encountered references to Dendi as "Songhai".

POSSIBLE SOLUTIONS:

1. Change scope of [son] from I to M; denotation encompasses three "Songhai" languages listed above.
2. Change scope of [son] from I to M; denotation encompasses three "Songhai" languages plus Zarma and Dendi.
3. Change scope of [son] from I to C (genetic) and change name to "Songhai languages"; denotation encompasses the nine languages of the Songhai genetic sub-group.

Given the status of Songhai varieties in Mali and the status of Zarma in Niger, it may be appropriate for these to have separate entries in ISO 639-2.

PROPOSED SOLUTION: Option 3 – this is consistent with MARC usage and copes with the fact that Zarma is larger than all the others.

5.42 Sardianian

PROBLEM: ISO 639 has [sc/srd]. *Ethnologue* lists four languages: "Campidanese Sardinian", "Gallurese Sardinian", "Logudorese Sardinian" and "Sassarese Sardinian". These four languages comprise a genetic subgroup of Indo-European known as "Sardinian".

POSSIBLE SOLUTIONS:

1. Change scope of [sc/srd] from I to M; denotation encompasses the four "Sardinian" languages.
2. Change scope of [sc/srd] from I to C (genetic) and change name to "Sardinian languages"; denotation encompasses the languages of the Sardinian genetic sub-group.

PROPOSED SOLUTION: 1

5.43 Swahili

PROBLEM: ISO 639 has [sw/swa] "Swahili". *Ethnologue* lists two languages: "Swahili", a language of wider communication across several countries of central and southern Africa, pop. est. 30,000,000 (including second-language users); and "Congo Swahili", spoken in Congo primarily as a second language used for wider communication. "Swahili" is also the name of a genetic sub-group of the Niger-Congo phylum.

The MARC Language Code List uses [swa] for "Swahili" but also as a collective that encompasses "Comorian", "Kae" and "Kingwana". The latter is listed in *Ethnologue* as a dialect of Congo Swahili. "Comorian" and "Shingazidja Comorian" are other languages of the Swahili sub-group and are spoken in Comoros Islands. I am not aware of these languages being referred to as "Swahili".

POSSIBLE SOLUTIONS:

1. Change scope of [sw/swa] from I to M; denotation encompasses Swahili and Congo Swahili.
2. Change scope of [sw/swa] from I to M; denotation encompasses Swahili, Congo Swahili, Comorian and Shingazidja Comorian.
3. Change scope of [sw/swa] from I to C (genetic) and change name to "Swahili languages"; denotation encompasses the six languages of the Swahili genetic sub-group.

PROPOSED SOLUTION: Option 1 (differs from MARC usage). To avoid having different entries that use the same name, the name for the existing entry (the macro-language entry) would be changed to "Swahili (generic)"

5.44 Tamashek

PROBLEM: ISO 639 has [tmh] "Tamashek". The MARC Language Code List uses [tmh] for "Tamashek" and for "Tuareg". *Ethnologue* lists four languages: "Tamasheq", also known as "Tuareg", spoken in Mali and also in Algeria and Burkina Faso, pop. est. 270,000; "Tawallammat Tamajaq", also known as "Tamasheq", "Tuareg" or "Amazigh", spoken in Niger, Mali and Nigeria, pop. est. 640,000; "Tayart Tamajeq", also known as "Tamachek", "Tuareg" or "Amazigh", spoken in Niger, pop. est. 250,000; and "Tahaggart Tamahaq", also known as "Tamachek" or "Tuareg", spoken in Algeria, Libya and Niger, pop. est. 62,000. These four languages comprise a genetic sub-group of the Afro-Asiatic phylum also known as "Tamasheq".

POSSIBLE SOLUTIONS:

1. Change scope of [tmh] from I to M; denotation encompasses the four "Tamasheq"/"Tuareg" languages.
2. Change scope of [tmh] from I to C (genetic) and change the name to "Tamashek languages"; denotation encompasses the languages of the Tamasheq genetic sub-group.

PROPOSED SOLUTION: Option 1 (on the assumption that these languages are perceived in some contexts as one, as suggested by the common name "Tuareg" and similiarity in alternate names "Tamasheq", "Tamajaq", etc.)

5.45 Ukranian/Rusyn

PROBLEM: ISO 639 has [uk/ukr] "Ukranian". The MARC Language Code List uses [ukr] for "Ukranian" but also for "Ruthenian" and as a collective encompassing "Carpatho-Rusyn". *Ethnologue* describes "Ruthenian" and "Carpatho-Rusyn" as alternate names for "Rusyn".

Rusyn is sometimes referred to as a dialect of Ukranian, but speakers are reported to consider themselves distinct from Ukrainians. Rusyn has become a focus of development in Slovakia, being taught in schools and used for textbooks and other publications.

POSSIBLE SOLUTIONS:

1. Specify denotation of [uk/ukr] as specifically Ukranian; denotation does not encompass Rusyn.
2. Change scope of [uk/ukr] from I to M; denotation encompasses Ukranian and Rusyn.

PROPOSED SOLUTION: 1

5.46 Uzbek

PROBLEM: ISO 639 has [uz/uzb] "Uzbek". *Ethnologue* lists two languages: "Northern Uzbek", the official language of Uzbekistan and also spoken in neighboring countries, pop. est. 18,466,000; and "Southern Uzbek", spoken primarily in Afghanistan, pop. est. 1,403,000.

PROPOSED SOLUTION: Change scope of [uz/uzb] from I to M; denotation encompasses Northern and Southern Uzbek.

5.47 Yiddish

PROBLEM: ISO 639 has [yi/yid] "Yiddish". *Ethnologue* lists two languages: "Western Yiddish", spoken in various countries of central and western Europe; and "Eastern Yiddish", spoken in Israel and also various countries of northeastern Europe and by a diaspora around the world.

PROPOSED SOLUTION: Change the scope of [yi/yid] from I to M; denotation encompasses both Eastern and Western Yiddish.

5.48 Zhuang

PROBLEM: ISO 639 has [za/zha] "Zhuang". *Ethnologue* lists two languages: "Northern Zhuang", pop. est. 10,000,000; and "Southern Zhuang", pop. est. 4,000,000. Both are spoken in China. "Zhuang" is considered an official minority in China. The two are reportedly fairly distinct (65% lexical similarity), and are classified in distinct sub-groups of the Tai family.

PROPOSED SOLUTION: Change the scope of [za/zha] from I to M; denotation encompasses both Northern and Southern Zhuang.

5.49 Dogri

PROBLEM: ISO 639 has [doi] "Dogri". MARC uses this for "Dogri", but also as a collective for Kangri. This corresponds directly to a single entry in the 14th edition of *Ethnologue*, "Dogri-Kangri". As of June 2003, however, *Ethnologue* has split this into two languages: "Dogri" and "Kangri". This change will be reflected in the 15th edition of *Ethnologue*.

PROPOSED SOLUTION: Change the scope of [doi] from I to M; denotation encompasses both Dogri and Kangri.

6. Other uncertain denotations

6.1 [arc] "Aramaic"

PROBLEM: ISO 639 has [arc] "Aramaic". The MARC Language Code List uses [arc] for "Aramaic", "Biblical Aramaic" and "Chaldean". Milicent Wewerka reports that MARC usage is for ancient languages rather than any modern languages.

The Linguist List's catalog of historic and artificial languages lists "Aramaic" (spoken 7^{th} – 4^{th} centuries BC) and also "Old Aramaic".

Ethnologue lists "Babylonian Talmudic Aramaic" as well as four modern "Neo-Aramaic" languages.

The issues to be resolved pertain to how [arc] relates to entries in the *Ethnologue* or the Linguist List's catalog of historic languages. Particularly important in this is whether [arc] refers only to ancient languages, or also encompasses modern varieties. Information provided regarding MARC usage suggests the former.

Given that, the question remains whether *Ethnologue*'s "Babylonian Talmudic Aramaic" is one of the varieties within [arc] or a separate language. Reportedly, Talmudic Aramaic is a variety that was used by Jewish scribes from a period post-dating Imperial Aramaic by around 1000 years and that differs from the latter as a result of the different context of its usage.

PROPOSED SOLUTION: The denotation of [arc] encompasses ancient varieties of Aramaic, but not Talmudic Aramaic. The name will be changed to "Aramaic, Ancient". The scope of [arc] is left as I. A separate entry will be added for Talmudic Aramaic.

6.2 [bas] "Basa"

PROBLEM: ISO 639 lists [bas] "Basa". *Ethnologue* cites several uses of this name, either as a primary or alternate name for an individual language, or as the name used for a genetic sub-group, giving rise to the possible denotations listed below.

POSSIBLE SOLUTIONS:

1. Basaa (*Ethnologue* [BAA]), a Bantoid language spoken in Cameroon, pop. est. 230,000, also spelled "Basa", also known as Bisaa, Bicek, Mbele, etc.?
2. Basa (*Ethnologue* [BZW]), a non-Bantoid Niger-Congo language spoken in Nigeria, pop. est. 100,000, also known as: Basa-Benue, Abatsa, Rubasa, etc.?
3. The Basa genetic sub-group of the Niger-Congo phylum, comprised of Basa (previous item) plus three languages with small speaker populations, Basa-Gumna, Basa-Burmana, Bassa-Kontagora.
4. Basa (*Ethnologue* [BQA]), a non-Bantoid Niger-Congo language spoken in Benin, pop. est. 1,000?
5. The Basa dialect of Ngwo, a Bantoid language spoken in Cameroon, pop. est. 31,000?

PROPOSED SOLUTION: 1

6.3 [bin] "Bini"

PROBLEM: ISO 639 lists [bin] "Bini". This name is used as an alternate name for Edo (Nigeria, pop. est. 1,000,000), an alternate spelling for Pini (Australia, nearly extinct), an alternate name

for the Bunu dialect of Yoruba (Nigeria, 18,850,000), a dialect of Anyin (Côte d'Ivoire, 610,000), and possibly also a Bantu variety (Edo, Yoruba and Anyin are not Bantu languages).

The MARC Language Code List uses [bin] for "Bini" and also for "Benin", "Do" and "Edo".

PROPOSED SOLUTION: Specify the denotation of [bin] as Edo.

6.4 [bra] "Braj"

PROBLEM: ISO 639 has [bra] "Braj". "Braj" is described by some as a western Hindi dialect. *Ethnologue* lists Braj as alternate names for Braj Bhasha and Kanauji (both in the Western Hindi genetic sub-group). Grierson also lists several kinds of "Braj" as dialects of Bundeli (a Western Hindi language). "Braj" does not correspond to any genetic sub-group. The Central Institute of Indian Languages lists "Brij Bhasha" as a "mothertongue" under "Hindi". The term "Braj" is also used to refer to medieval-era Hindi.

The MARC Language Code List uses [bra] for "Braj" and also for "Pingal". The latter appears to be an Indic literary term referring to a traditional poetic form.

POSSIBLE SOLUTIONS:

1. Specify denotation of [bra] as Braj Bhasha.
2. Specify denotation of [bra] as Kanauji.
3. Specify denotation of [bra] as a historic variety, a pre-cursor to modern Hindi and other related languages.
4. Change scope of [bra] from I to M; denotation encompasses both Braj Bhasha and Kanauji.
5. Change scope of [bra] from I to M; denotation encompasses Braj Bhasha, Kanauji and also other Western Hindi varieties, such as Bundeli.

PROPOSED SOLUTION: 1

6.5 [car] "Carib"

PROBLEM: ISO 639 lists [car] "Carib". *Ethnologue* lists four usages: three individual languages known as "Carib", and the Carib language phylum (29 indivudual languages). One of the individual languages (*Ethnologue* [CRB]) is also known as "Galib" or "Kalinya", pop. est. 10,000, a Carib language spoken in Venezuela and other countries.

The MARC Language Code List uses [car] for "Carib" and also for "Galibi".

PROPOSED SOLUTION: Specify the denotation of [car] as specifically "Carib, Galibi" (*Ethnologue* [CRB]).

6.6 [ewe] "Ewe"

PROBLEM: ISO 639 has [ewe] "Ewe", which *Ethnologue* also lists as an individual language. Ewe is in the Gbe genetic sub-group of the Niger-Congo phylum.

The MARC Language Code List uses [ewe] for "Ewe". Earlier versions also used [ewe] for "Gbe" and as a collective encompassing "Aja", "Gen-Gbe", "Gun-Gbe" and "Tofingbe". "Aja" is a name used for a genetic sub-group within the Gbe sub-group, and also for an individual language within the Aja sub-group. Gub-Gbe is also part of the Aja sub-group. Gen-Gbe is from a different sub-group under Gbe. Ewe is not considered part of either of these sub-groups.

POSSIBLE SOLUTIONS:

1. Specify denotation of [ewe] as specifically Ewe; denotation does not encompass any other Gbe languages.
2. Change scope of [ewe] from I to M; denotation encompasses some selection of Gbe languages.
3. Change scope of [ewe] from I to C (genetic) and change name to "Gbe languages"; denotation encompasses twenty-one languages of the Gbe sub-group.

PROPOSED SOLUTION: Option 1 (differs from past MARC usage, but is consistent with current MARC usage).

6.7 [ewo] "Ewondo"

PROBLEM: ISO 639 has [ewo] "Ewondo, which *Ethnologue* also lists. The MARC Language Code List, however, also uses [ewo] for Beti. The complication here is that Beti is a cover term for a group of mutually intelligible varieties corresponding to distinct ethnic groups: Bebele, Bebil, Bulu, Eton, Ewondo, Fang and Mengisa. (ISO 639 also has an identifier [fan] for Fang.) The MARC documentation gives the impression that Beti is a sub-variety of Ewondo, but it is, in fact, the other way around, and each of these seven languages can equally be considered "Beti".

On purely linguistic criteria, the seven varieties listed above should be considered the same language, but for other sociolinguistic reasons each is considered an individual language. Beti, then, would fit our definition of macro language (a variety that is considered for some purposes an individual language that encompasses several varieties that are also considered individual languages), even though in the typical case it is the cluster that is, for *non-linguistic* reasons, considered a single language, whereas here it is the members.

Therefore, if there is data that, for whatever reason, is to be considered "Beti", it would not be best practice to tag that data using [ewo]. It would be better, rather, to have an identifier for a macro-language category "Beti".

POSSIBLE SOLUTIONS:

1. Specify denotation of [ewo] as specifically Ewondo; the denotation does not include Beti. Add a macro-language entry to ISO 639-2 for Beti if there is a need to represent such semantics using a language identifier.
2. Change scope of [ewo] from I to M; denotation encompasses both Ewondo and Beti.

PROPOSED SOLUTION: 1

6.8 [lah] "Lahnda" and [pa/pan] "Panjabi"

PROBLEM: ISO 639 has entries [lah] "Lahnda" and [pa/pan] "Panjabi". These two must be considered together.

Corresponding to "Panjabi", *Ethnologue* lists three languages: "Eastern Panjabi", or "Gurmukhi" is spoken primarily in the Punjab state of India, pop. est. 27,125,000; "Mirpur Panjabi", or "Mirpuri", is spoken primarily in Kashmir, pop. est. 30,000; and "Western Panjabi", or "Lahnda", spoken primarily in the Punjab province of Pakistan, pop. est. 45,000,000. "Lahnda" is also used as the name of a genetic sub-group that includes Western and Mirpur Panjabi, but not Eastern Panjabi.

Grierson used the term “Panjabi” for varieties spoken in “Eastern Panjab” (what is now the Punjab state of India plus the eastern fringe of the Punjab province of Pakistan). This would match *Ethnologue*’s “Eastern Panjabi”.

Grierson introduced the term “Lahnda” (a Punjabi word meaning ‘western’) for varieties in “Western Panjab” (roughly what is now the Punjab province of Pakistan) due to their distinctness from “Panjabi”, having significant differences from the latter while also much in common with Sindhi. “Landha” is not used by speakers of these varieties, though the term caught on among many linguists.

Following Grierson’s usage, “Lahnda” has beeen described as a cover term for a dialect chain between Sindhi in the south and various northern varieties including Western Punjabi, Pahari-Potwari and Hindko varieties. In terms of the classification scheme referred to in *Ethnologue*, this would likely include the languages of the “Lahnda” sub-group (possibly excluding Khetrani or Jakati), plus Pahari-Potwari;[6]

Ethnologue lists “Lahnda” as an alternate name for “Western Panjabi”. “Western Panjabi” appears to correspond to “Shahpuri”, which Grierson considered to be “standard Lahnda” (Masica 1991, p. 18).

The *MARC Language Code List* uses “Lahnda” for Western Panjabi, but also as a collective that appears similar to Grierson’s usage.[7] Comments from Milicent Wewerka suggest that MARC usage of [pa/pan] should be equated with *Ethnologue*’s “Eastern Panjabi”.

POSSIBLE SOLUTIONS: The following are possibilities for [pa/pan] “Punjabi”:

1. Equate [pa/pan], Grierson’s “Panjabi” and *Ethnologue*’s “Eastern Panjabi”.
2. Change the scope of [pa/pan] from I to M; denotation encompasses Eastern Punjabi and Western Punjabi.
3. Change the scope of [pa/pan] from I to M; denotation encompasses Eastern Punjabi, Western Punjabi, and Mirpur Panjabi.

It should be noted that existing software implementations use [pa/pan] for “Punjabi” in both India and Pakistan, though this appears to be in reference to the same variety, “Eastern Panjabi”.

The following are possibilities for [lah] “Lanhda”.

1. Specify the denotation of [lah] as “Western Panjabi” (as described in the *Ethnologue*).
2. Change the scope of [lah] from I to M; denotation encompasses various “Lahnda” languages (exact list to be determined).
3. Change the scope of [lah] from I to C (ad hoc) and change name to “Lahnda languages”; denotation encompasses various “Lahnda” languages (exact list to be determined).

MARC usages suggests the need for the second alternative. If option 2 is followed, it would seem appropriate to include the languages assumed by MARC, but not “Siraiki Sindhi” (see note 7).

---

[6] In the classification cited by *Ethnologue*, Pahari-Potwari belongs to the Western Pahari sub-group, which is from a distinct branch of Indo-Aryan to that to which the “Lahnda” sub-group belongs. Western Pahari is relevant for some possible interpretations of [him] “Himachali”—see §3.3.

[7] One of the languages encompassed by “Lahnda” is known as “Saraiki”. That term is also used, however, for a dialect of Sindhi; The *MARC Language Code List* makes the error of incorporating “Siraidi Sindhi” into the “Lahnda” collective.

PROPOSED SOLUTION: For [pa/pan], option 1. For [lah], option 2, with the denotation encompassing: the seven languages of the Lahnda sub-group described in *Ethnologue* plus Pahari-Potwari.

6.9 [lam] "Lamba"

PROBLEM: ISO 639 has [lam] "Lamba". *Ethnologue* lists two languages: "Lamba" (*Ethnologue* [LAB]), a Bantu language spoken in Zambia and also Democratic Republic of Congo, pop. est. 211,000; and "Lama", also known as "Lamba" or "Losso", a Niger-Congo language (from a branch other than Bantu) spoken in Togo and also in Benin and Ghana; pop. est. 177,400.

PROPOSED SOLUTION: Specify the denotation of [lam] as Lamba (*Ethnologue* [LAB]).

6.10 [lua] "Luba-Lulua"

PROBLEM: ISO 639 has [lua] "Luba-Lulua". *Ethnologue* lists "Luba-Lulua" as an alternate name for "Luba-Kasai". The problem lies in the MARC Language Code List, which uses [lua] for various names. Most are alternates for Luba-Lulua, but one is "Kalebwe (Luba-Lulua)". "Kalebwe" is an alternate name for Songe (also known as "Luba-Songi" or "Northeast Luba"), a distinct language related to Luba-Lulua. This raises a question as to whether MARC usage has actually been with a wider scope covering several "Luba" languages, or whether this is just an anomaly within MARC usage..

PROPOSED SOLUTION: Specify denotation of [lua] as specifically "Luba-Lulua" (*Ethnologue* [LUB]); denotation does *not* encompass Songe. (Differs from MARC usage.)

6.11 [nzi] "Nzima"

PROBLEM: ISO 639 has [nzi] "Nzima". *Ethnologue* lists "Nzima" as an alternate for "Nzema", a Kwa language spoken in Ghana and Côte d'Ivoire. The MARC Language Code List describes [nzi] as used for "Nzima" and also for "Amanaya", "Nsima", "Zema" and "Zimba". Except for instances of the MARC Language Code List, I have not found references to "Amanaya" or "Zema", but "Zimba" is used to refer to a Bantu language, distinct from Nzima, spoken in Democratic Repulic of Congo. Milicent Wewerka has reported that MARC usage corresponds to the single entry in *Ethnologue*.

PROPOSED SOLUTION: Specify denotation of [nzi] as "Nzima", a Kwa language of Ghana and Côte d'Ivoire, and not the Bantu language Zimba. (Differs from MARC usage.)

6.12 [rm/roh] "Raeto-Romance"

PROBLEM: ISO 639 has [rm/roh] "Raeto-Romance". *Ethnologue* lists "Raeto-Romance" as an alternate name for Romansh. Romansh is named in the constitution of Switzerland as one of the official languages of that country.

The problem arises when reviewing MARC usage: the MARC Language Code List uses [roh] for "Raeto-Romance", but also as a collective that encompasses "Ladin", which is spoken in Italy. Romansh and Ladin are two of the languages from the Rhaetian sub-group of Indo-European, the third being Friulian, also spoken in Italy. MARC usage covers two of these three languages, then; there is no indication that it extends to cover the Rhaetian sub-group, though this is possible.

There are existing implementations, however, that use [rm] specifically for Romansh.

POSSIBLE SOLUTIONS:

1. Specify the denotation of [rm/roh] as specifically Romansh; change name from “Rhaeto-Romance” to “Romansh”.
2. Change the scope of [rm/roh] from I to C (generic) and change the name to “Rhaetian languages”; the denotation encompasses all three languages of the Rhaetian sub-group.
3. Do not equate [rm] and [roh]. Specify the denotation of [rm] as specifically Romansh. Change the scope of [roh] from I to C (generic) and change the name to “Rhaetian languages”; the denotation of [roh] encompasses the three languages of the Rhaetian sub-group. ISO 639-3 will include an identifier for Romansh other than [roh].

Option 3 is unorthodox in dispensing with an assumption of equivalence between entries in ISO 639-1 and ISO 639-2 for the same name, but it preserves existing usage of [rm] in implementations such as Internet protocols, and existing usage of [roh] in bibliographic records. Unfortunately, it hinges on the assumption that no existing implementations assume that equivalence between [rm] and [roh], an assumption that cannot be made with confidence. While there may be a cost in the maintenance of bibliographic records to restricting the denotation of [roh] to just Romansh, as in option 1, there are many other sectors that potentially stand to face costs if the denotation of [rm] is broadened to encompass other Rhaetian languages beyond Romansh, as in option 2.

PROPOSED SOLUTION: Option 1. While this differs from MARC usage, Milicent Wewerka has recommended that this recommendation be adopted and that MARC usage be revised accordingly.

6.13 [sah] “Yakut”

PROBLEM: ISO 639 has [sah] “Yakut”. *Ethnologue* has “Yakut”, also known as “Sakha”, an Altaic language spoken in Russia, pop. est. 363,000.

The MARC Language Code List uses [sah] for “Yakut” and “Sakha”, but also as a collective that encompasses “Dolgan”. Dolgan is a distinct language from the same genetic sub-group and spoken in the same region, pop. est. 5,000.

PROPOSED SOLUTION: Specify the denotation of [sah] as specifically Yakut; denotation does not encompass Dolgan. (Differs from MARC usage.)

6.14 [syr] “Syriac”

PROBLEM: The MARC Language Code List uses [syr] for “Syriac” and for “Neo-Syriac”. *Ethnologue* lists “Classical Syriac”, the ancient language of the Syriac church. It also describes “Neo-Syriac” as used for the modern Neo-Eastern Aramaic languages spoken by Christians; this would include two principle languages: Assyrian Neo-Aramaic and Chaldean Neo-Aramaic.

The classical language is still in use by modern communities; both the classical language and modern varieties are written using the Syriac script. As a result, there are software implementations that support classical and modern varieties without differentiation.

POSSIBLE SOLUTIONS:

1. Specify denotation as Classical Syriac (differs from MARC usage).
2. Change scope from I to M; denotation encompasses Assyrian Neo-Aramaic and Chaldean Neo-Aramaic.
3. Change scope from I to M; denotation encompasses Classical Syriac, Assyrian Neo-Aramaic and Chaldean Neo-Aramaic.

4. Change scope from I to C (genetic) and change name to “Syriac languages”; denotation encompasses the languages of the Eastern Aramaic sub-group, including Classical Syriac, Classical Mandaic, plus some fifteen modern languages.

PROPOSED SOLUTION: 3

6.15 [bo/bod/tib] “Tibetan”

PROBLEM: The MARC Language Code List uses [tib] for “Tibetan” and for “Bhotanta”, “Bhutan” and “Boutan”, but also as a collective that encompasses “Kagete” and “Sherpa”.

The language typically referred to as “Tibetan” (*Ethnologue* [TIC]) is from the Central sub-group of the Tibetan sub-group of Tibeto-Burman. Kagete is also from the Central sub-group. Sherpa, however, is from Southern sub-group of Tibetan, the same sub-group as Dzongkha.

“Bhotanta” and “Bhotia” appear to be names used for many of the Tibetan languages.“Bhutan” is, of course, the name of a country in the region, and most of the languages of Bhutan are in the Tibetan sub-group. The national language, Dzongkha, is also known as “Bhotia of Bhutan”, and may possibly get referred to as “Bhutan”. Dzongkha has its own identifier in ISO 639, however: [dz/dzo].

POSSIBLE SOLUTIONS:

1. Specify denotation of [bo/bod/tib] as specifically Tibetan (*Ethnologue* [TIC]).
2. Change scope of [bo/bod/tib] from I to M; denotation encompasses Tibetan, Kagete, Sherpa, and possibly other languages of the Tibetan sub-group.
3. Change scope of [bo/bod/tib] from I to C (genetic) and change name to “Tibetan languages”; denotation encompasses the fifty-two languages of the Tibetan sub-group.

PROPOSED SOLUTION: Option 1 (differs from MARC usage).

6.16 Kalmyk/Oirat

PROBLEM: ISO 639 has [xal] “Kalmyk”. *Ethnologue* lists a single language for which it uses the reference name “Kalmyk-Oirat”. “Kalmyk” is reported to be the name used in Russia; “Oirat” is used in China and Mongolia. “Oirat” is a name used by non-speakers, and is applied to other languages as well.

The Kalmyks migrated from the region of western China and Mongolia to the west side of the Caspian Sea several centuries ago. Some later returned to the east. There is on-going contact between the two sub-communities with people moving back and forth. At least until recently, those living in China or Mongolia were regarded by Russian authorities as “Kalmyk” and permitted to enter without restriction.

Linguists I have consulted who are familiar with these communities, including some who have lived within the Kalmyk community for some time, report that “Kalmyk-Oirat” is a single language, not two languages.

POSSIBLE SOLUTIONS:

1. Specify the denotation of [xal] as being equal to that of the *Ethnologue* entry “Kalmyk-Oirat”. Preferably, the name would be changed to “Kalmyk; Oirat”.

2. Specify the denotation of [xal] as “Kalmyk”, i.e. what is spoken in Kalmykia, Russia, but not what is spoken in Mongolia or China. ISO 639-3 would list separate entries for “Kalmyk” and “Oirat”.

It should be noted that no linguistic or sociolinguistic evidence has been seen that would validate option 2.

PROPOSED SOLUTION: 1

7. Multiple entries in ISO 639 with one corresponding entry in Ethnologue 14th edn.

7.1 Akan/Fanti/Twi

PROBLEM: There are three ISO 639 entries that, per *Ethnologue* data, correspond to a single individual language: [aka, fat, tw / twi]. *Ethnologue*. describes Fanti and Twi as dialects of Akan.

Milicent Wewerka reports that Fanti and Twi are distinct *ethnic* groups that share a common language, but the names “Fanti” and “Twi” are not acceptable as language names to the opposite communities. In some respects, then, there is similarity with other situations such as Serbo-Croatian in which there is a single language yet ethnic distinctions impose artificial language distinctions on various IT applications.

POSSIBLE SOLUTIONS:

1. Change scope of [aka] from I to M, and have it encompass [fat] and [tw/twi]. This would entail that ISO 639-1 list a member of a macro-language but not the macro-language itself. This would be an exceptional situation, but not necessarily problematic (and if a problem, better to deal with in ISO 639-1 than ISO 639-3).
2. Deprecate [fat] and [tw/twi] and document that they are subsumed by [aka].
3. Deprecate [aka] and [fat], document that they are subsumed by [tw/twi], and change name of [tw/twi] to include all three names: “Akan, Fanti, Twi”
4. If evidence were available to support the analysis of Akan, Fanti and Twi being three distinct languages, all three could be kept as they are in ISO 639-1/-2, and they could all be listed in ISO 639-3.

PROPOSED SOLUTION: Option 1.

7.2 Serbo-Croatian

PROBLEM: ISO 639 has three categories, [bs/bos] “Bosnian”, [hr/hrv/scr] “Croatian” and [sr/srp/scc] “Serbian”. *Ethnologue 14th edn*. has only one entry, “Serbo-Croatian”. *Ethnologue 15th edn.* will include three entries, Bosnian, Croatian and Serbian, however.

ISO 639-3 must list three entries for Bosnian, Serbian and Croatian. The only issue to resolve is whether ISO 639 should include a macro-language category for “Serbo-Croatian”.

PROPOSED SOLUTION: A macro-language category for “Serbo-Croatian” will be included in ISO 639-3.

7.3 Moldovian/Romanian

PROBLEM: ISO 639 has two categories: [mo/mol] "Moldavian", and [ro/ron/rum] "Romanian". *Ethnologue* has only one entry, "Romanian".

"Moldavian" was an artificially-distinct linguistic identity created for political purposes during the era of Soviet control. To support this distinction, Cyrillic script was imposed, and attempts were made to introduce archaic Romanian forms as well as Russian loanwords. Today, there is no significant linguistic difference between "Romanian" and "Moldavian"; the only possible distinction between [mo/mol] and [ro/ron/rum] in ISO 639 would appears to be the distinction between Cyrillic and Latin scripts.

POSSIBLE SOLUTIONS:

1. Deprecate [mo/mol]; document it as a synonym for [ro/ron/rum].
2. Document [mo/mol] as a synonym for [ro/ron/rum]; do not deprecate it, but allow this isolated case of synonymy to exist within the coding system.
3. Specify the denotation of [mo/mol] as referring to the artificial linguistic identity created by the Soviets; change the name to "Moldavian (Soviet-era Romanian of Moldavian SSR)"; document relationship to [ro/ron/rum].
4. Specify the denotation of [mo/mol] as "Moldovian, =. Romanian (Cyrillic script)"; denotation of [ro/ron/rum] would be "Romanian (Latin Script)"
5. Contrary to fact, assume two distinct linguistic identities denoted by [mo/mol] and by [ro/ron/rum]; ISO 639-3 would list both entries; a new macro-language entry for "Romanian-Moldavian" would be added to ISO 639-2. (*Ethnologue* would continue to list only one language, which would correspond to this macro-language entry in ISO 639-2.)
6. Change scope of [ro/ron/rum] from I to M; denotation would encompass [mo/mol]. This would be exceptional in that, in every other case, the macro-language notion is used because there is a single identity associated with multiple distinct identities.

It is strongly recommended that options 4 and 5 be avoided as it is considered unwise to allow ISO 639 to make distinctions based purely on script,[8] or to knowingly present artificial language distinctions for no reason.

PROPOSED SOLUTION: Option 1. The impact on MARC is that records should ideally be updated to use [rum] rather than [mol], though continued usage of [mol] will still be valid and will continue to have the same semantics.

7.4 Turkish/Ottoman Turkish

PROBLEM: ISO 639 has two categories, [tr/tur] "Turkish" and [ota] "Turkish, Ottoman (1500-1928)". *Ethnologue* has one corresponding entry, "Turkish".

In other cases in which ISO 639-1/-2 has identifiers for historic varieties, the time period 1500 to the present is treated as a whole, to which the modern form belongs. Turkish is the exception. The year of division (1928) correspondis with the orthographic reform in which Arabic script was replaced with Latin script. This makes it appears that these identifiers are making a distinction in script, which is not a purpose for which ISO 639 is intended.

[8] As indicated in clause 4.1.3 of ISO 639-2, it should be left to a separate standard to designate information concerning the script or writing system of a language.

In addition to the orthographic change, it has been reported to me that Turkish underwent significant linguistic change during the 20th century.

POSSIBLE SOLUTIONS:

1. Deprecate [ota] and document that it is subsumed by [tr/tur]; where there is a need to record a distinction in script, protocols should reference a derivative tagging system that supports distinctions based on written form.
2. Continue to use both [ota] and [tr/tur], allowing them to distinguish historic varieties at a finer level of granularity than the 1500-to-present norm; denotation of [tr/tur] would be Turkish from 1928 to present. Document the denotations accordingly.
3. Specify denotation of [ota] as modern-era (i.e., 1500-to-present) Turkish written in Arabic script, and that of [tr/tur] as modern-era Turkish written in Latin script, allowing a distinction based on script in this isolated case. Document the denotations accordingly.
4. Specify denotation of [ota] as modern-era (1500-to-present) Turkish written in Arabic script; denotation of [tr/tur] would be modern-era Turkish, with no indication of the script used. (Thus, [ota] would denote a subset of [tr/tur].) ISO 639-3 would list only [tur].
5. Specify denotation of [ota] as Turkish written in Arabic script from 1500 to 1928; denotation of [tr/tur] would be modern-era Turkish, with no indication of the script used. (Thus, [ota] would denote a subset of [tr/tur].) ISO 639-3 would list only [tur].

PROPOSED SOLUTION: Option 2.

8. Denotation of geographically-defined collections

ISO 639-2 includes some entries for language collections that are geographically defined. Given the nature of the geographic distribution of language communities, there is a general problem of ambiguity in the denotation of such identifiers; i.e., it is not entirely clear precisely what languages are intended to be within their scope. Some particular issues are described below.

It is not a pre-requisite for the development of ISO 639-3 that these issues be resolved, and that the exact denotation of geographically-defined collections be determined. Doing so, however, would allow a complete mapping between ISO 639-2 and ISO 639-3 to be defined.

8.1 [cai] "Central American Indian"/[nai] "North American Indian"

The MARC Language Code List defines [cai] as encompassing languages of Central America and Mexico, and the languages of the "Azteco-Tanoan phylum". This would encompass languages not-only from the region corresponding to the classical notion of "Meso-America", but also as far north as Idaho and Wyoming. That is beyond common interpretations of "Central America". Excluding Northern Uto-Aztecan and Kiowa-Tanoan from [cai] (including them, instead, within [nai]) would provide a boundary corresponding roughly with the US-Mexico border. Excluding the Sonoran sub-group of Uto-Aztecan from [cai] as well would provide a northern limit for [cai] that is closer to the notion of "Meso-America".

On the other hand, Milicent Wewerka has shown that some sources do treat all of Uto-Aztecan and Oto-Manguean as members of Central-Amerind. She is reluctant to revise MARC usage.

PROPOSED SOLUTION: Independent of this particular item, it is recommended that the relationship between collections and member languages be made clear in ISO 639-5. This will

resolve any uncertainty with respect to this particular case. It is proposed, then, that the MARC usage be left, and clearly documented in ISO 639-5.

8.2 [paa] "Papuan"

PROBLEM: ISO 639 has [paa] "Papuan (Other)".[9] The classification system referred to in *Ethnologue* includes a genetic sub-group of the Austronesian phylum known as "Papuan Tip". There are also two distinct language families, "East Papuan" and "West Papuan".

"Papuan" can also be used in a geographic sense to refer to any of the non-Austronesian languages of Papua New Guinea and Irian Jaya. This would encompass "East" and "West Papuan" plus some eleven other genetic language groups.

PROPOSED SOLUTION: Specify denotation of [paa] as non-Austronesian families spoken in the region, together with any isolates or unclassified languages. (This encompasses some 840 languages listed in *Ethnologue*).

8.3 [sai] "South American Indian"

Some language families spoken primarily in South America have member languages that are not themselves spoken in South America. Similarly, there are languages spoken in South America that do not belong to language families spoken primarily in South America. It may be unclear to users which of these are or are not included.

Milicent Wewerka has clarified that in MARC usage the denotation is based strictly on geography. Languages not spoken in South America are not included, regardless of their genetic relationships.

PROPOSED SOLUTION: The geographic basis of the denotation should be made clear in ISO 639-5.

9. References

Grimes, Barbara F., ed. 2000. *Ethnologue. 14th edn.* Dallas: SIL International.

Library of Congress Network Development and MARC Standards Office. 2003. *MARC Code List for languages.* (Web version). Published online at http://www.loc.gov/marc/languages/.

Masica, Colin P. 1991. *The Indo-Aryan languages.* (Cambridge languages surveys.) Cambridge: Cambridge University Press.

[9] See §2 on the proposed change to "Other" collections.