Network Working Group
Request for Comments: 5645
Category: Informational

D. Ewell, Ed.
Consultant

September 2009

Update to the Language Subtag Registry

Abstract

This memo defines the procedure used to update the IANA Language Subtag Registry, in conjunction with the publication of RFC 5646, for use in forming tags for identifying languages.

Status of This Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Table of Contents

1. Introduction

[RFC4646] provides for a Language Subtag Registry and describes its format. The initial contents of the registry and rules for determining them are specified in [RFC4645].

[RFC5646] expands on [RFC4646] by adding support for approximately 7,500 new primary and extended language subtags based on [ISO639-3] and [ISO639-5] alpha-3 code elements, and seven new region subtags based on [ISO3166-1] exceptionally reserved code elements. This memo describes the process of updating the registry to include these additional subtags and to make secondary changes to the registry that result from adding the new subtags and from other decisions made by the Language Tag Registry Update (LTRU) Working Group.

In writing this document, a complete replacement of the contents of the Language Subtag Registry was provided to the Internet Assigned Numbers Authority (IANA) to record the necessary updates.

The format of the Language Subtag Registry as well as the definition and intended purpose of each of the fields are described in [RFC5646].

The registry is expected to change over time, as new subtags are registered and existing subtags are modified or deprecated. The process of updating the registry is described in Section 3 of [RFC5646].

Many of the subtags defined in the Language Subtag Registry are based on code elements defined in [ISO639-1], [ISO639-2], [ISO639-3], [ISO639-5], [ISO3166-1], [ISO15924], and [UN_M.49]. The registry is not a mirror of the code lists defined by these standards and should not be used as one.

2. Updating the Registry

This section describes the process for determining the updated contents of the Language Subtag Registry.

2.1. Starting Point

The version of the Language Subtag Registry that was current at the time of IESG approval of this memo served as the starting point for this update. This version was created according to the process described in [RFC4645] and maintained according to the process described in [RFC4646].

The source data for [ISO639-3] used for this update consisted of three files, available from the official site of the ISO 639-3 Registration Authority. (Note that this file is updated from time to time. The version used in the preparation of this memo was the one in place on February 24, 2009.)

- [iso-639-3_20090210] is a list of all language code elements in [ISO639-3], including the alpha-3 code element and reference name for each code element. For example, the entry for the Dari language contained the code element 'prs' and the name "Dari" (among other information).

- [iso-639-3_Name_Index_20090210] is a list containing all names associated with each language according to [ISO639-3], including both inverted and non-inverted forms where appropriate. An "inverted" name is one that is altered from the usual English-language order by moving adjectival qualifiers to the end, after the main language name and separated by a comma. A code element may have more than one entry in this file; the reference name and its inverted form are usually, but not always, given in the first entry. For example, this file contained an entry for the code element 'prs' with the name "Dari" (twice) and another entry with the names "Eastern Farsi" and "Farsi, Eastern".

- [iso-639-3-macrolanguages_20090120] is a list of all alpha-3 code elements for languages that are encompassed by a macrolanguage in [ISO639-3], together with the alpha-3 code element for the macrolanguage. For example, a line containing the code elements 'fas' and 'prs' indicated that the macrolanguage "Persian" encompasses the individual language "Dari". (Note that these alpha-3 code elements may not have corresponded directly to subtags in the registry, which uses 2-letter subtags derived from [ISO639-1] when possible.)

The source data for [ISO639-5] used for this update consisted of one file, available from the official site of the ISO 639-5 Registration Authority. (Note that this file is updated from time to time. The version used in the preparation of this memo was the one in place on February 24, 2009.)

- [iso639-5.tab.txt] is a list of all language code elements in [ISO639-5], including the alpha-3 code elements and English name for each code element. For example, this file includes an entry containing the code element 'ira' and the name "Iranian languages" (among other information).

Language code elements that were already retired in all of the source standards prior to IESG approval of this memo were not listed in these files and, consequently, were not considered in this update.

The values of the File-Date field, the Added date for each new subtag record, and the Deprecated date for each existing grandfathered or redundant tag deprecated by this update were set to a date as near as practical to the date this memo was approved for publication by IESG.

2.2. New Language Subtags

For each language in [ISO639-3] that was not already represented by a language subtag in the Language Subtag Registry, a new language subtag was added to the registry, using the [ISO639-3] code element as the value for the Subtag field and using each of the non-inverted [ISO639-3] names as a separate Description field. The [ISO639-3] reference name is represented by the first Description field.

If the language was encompassed by one of the [ISO639-3] macrolanguages 'ar' (Arabic), 'kok' (Konkani), 'ms' (Malay), 'sw' (Swahili), 'uz' (Uzbek), or 'zh' (Chinese), as determined by [iso-639-3-macrolanguages_20090120], an extended language subtag was also added, with the primary language subtag of the macrolanguage as the value for the Prefix field. These macrolanguage subtags were already present in the Language Subtag Registry and were chosen because they were determined by the LTRU Working Group to have been used to represent a single dominant language as well as the macrolanguage as a whole, making the extended language mechanism suitable for languages encompassed by the macrolanguage.

If the name of the language included the word "Sign", an extended language subtag was added, with the string "sgn" as the value for the Prefix field. This is a special case that treats the existing primary language subtag for "Sign languages" as if it were a macrolanguage encompassing all sign languages.

All extended language subtags were added with a Preferred-Value equal to the corresponding primary language subtag.

If the language was encompassed by a macrolanguage, as determined by [iso-639-3-macrolanguages_20090120], a Macrolanguage field was added for the encompassed language, with a value equal to the subtag of the macrolanguage. (Note that 'sgn' is defined as a "collection code" by [ISO639-3] and hence is not included in that standard; therefore, no Macrolanguage field was added for sign language subtags.)

If the language was assigned a "Scope" value of 'M' (Macrolanguage) in [iso-639-3_20090210], a Scope value of "macrolanguage" was added

for the language. Otherwise, if the language was assigned a "Scope" value of 'S' (Special), a Scope value of "special" was added. Most languages in [ISO639-3] have scope 'I' (Individual) and thus were not assigned a Scope value in the registry.

For each language in [iso639-5.tab.txt] that was not already represented by a language subtag in the Language Subtag Registry, a new language subtag was added to the registry, using the [ISO639-5] code element as the value for the Subtag field and using the "English name" field as the Description field. Each of these languages was assigned a Scope value of "collection" in the registry.

All subtags were added to the registry maintaining alphabetical order within each type of subtag: all 2-letter "language" subtags first, then all 3-letter "language" subtags, and finally all "extlang" subtags. Some existing records were moved to ensure this order.

2.3. Modified Language Subtags

For each language in [ISO639-3] that was already represented by a language subtag in the Language Subtag Registry, Description fields were added as necessary to reflect all non-inverted names listed for that language in [iso-639-3_Name_Index_20090210]. Any existing Description fields that reflected inverted names or that represented a strict subset of the information provided by the [ISO639-3] name were deleted. An example of the latter was the name "Ainu" for the subtag 'ain', which provided less information than the [ISO639-3] name "Ainu (Japan)".

The order of Description fields was adjusted to ensure that the reference name from [ISO639-3] was listed first, followed by other names from [ISO639-3] in the order presented by that standard, followed by any other names already existing in the registry. In some cases, this resulted in a reordering of Description fields for existing entries, even when no new values were added.

For each language that was encompassed by a macrolanguage in [ISO639-3], a Macrolanguage field was added, with a value equal to the subtag of the macrolanguage.

For each language in [iso639-5.tab.txt] that was already represented in the Language Subtag Registry, the Description field was adjusted as necessary to match the "English name" field in [iso639-5.tab.txt]. Names in inverted form were rearranged to remove the inversion. Each of these languages was assigned a Scope value of "collection". Existing language subtags whose code elements were assigned prior to the publication of [ISO639-3] or [ISO639-5] and that were identified by the [ISO639-3] Registration Authority as representing collections

were also assigned a Scope value of "collection", even though they are not listed as such in [iso639-5.tab.txt].

Note in particular that the change from [ISO639-2] names such as "Afro-Asiatic (Other)" to [ISO639-5] names such as "Afro-Asiatic languages" implies a broadening of scope for some of these subtags, designated "remainder groups" in [ISO639-5]. While [iso639-5.tab.txt] includes a field indicating which code elements are designated as "groups" or "remainder groups" in [ISO639-2], [RFC5646] does not make this distinction, and consequently this field was not used in updating the Language Subtag Registry.

A Scope value of "private-use" was added for the unique record with Subtag value 'qaa..qtz'. This record has a Description of "Private use" (changed from "PRIVATE USE") and corresponds to a range of code elements that is reserved for private use in [ISO639-2]. The Description fields for script and region private-use subtags were also capitalized as "Private use".

2.4. New Region Subtags

[RFC5646] expands the scope of region subtags by adding subtags based on code elements defined as "exceptionally reserved" in [ISO3166-1]. These code elements are reserved by the ISO 3166 Maintenance Agency "at the request of national ISO member bodies, governments and international organizations". At the time of IESG approval of this memo, ISO 3166/MA had defined nine exceptionally reserved code elements, all of which were added to the Language Subtag Registry except for the following:

- 'FX' (Metropolitan France) was already present in the Language Subtag Registry because it was an assigned [ISO3166-1] code element from 1993 to 1997, but was deprecated with a Preferred-Value of "FR".

- 'UK' (United Kingdom) was not added because it is associated with the same UN M.49 code (826) as the existing region subtag 'GB'. [RFC5646], Section 3.4, item 15 (D) states that a new region subtag is not added to the Language Subtag Registry if it carries the same meaning as an existing region subtag.

2.5. Grandfathered and Redundant Tags

As stated in [RFC5646], "grandfathered" and "redundant" tags are complete tags in the Language Subtag Registry that were registered under [RFC1766] or [RFC3066] and remain valid. Grandfathered tags cannot be generated from a valid combination of subtags, while redundant tags can be.

Under certain conditions, registration of a subtag under [RFC5646] may cause a grandfathered tag to be reclassified as redundant. It may also enable the creation of a generative tag with the same meaning as a grandfathered or redundant tag; in that case, the grandfathered or redundant tag is marked as Deprecated, and the generative tag (including the new subtag) becomes its Preferred-Value.

As a result of adding the new subtags in this update, each of the following grandfathered tags became composable, were reclassified as redundant, and were deprecated with the indicated generative tag serving as the Preferred-Value:

zh-cmn (Preferred-Value: cmn)

zh-cmn-Hans (Preferred-Value: cmn-Hans)

zh-cmn-Hant (Preferred-Value: cmn-Hant)

zh-gan (Preferred-Value: gan)

zh-wuu (Preferred-Value: wuu)

zh-yue (Preferred-Value: yue)

The following grandfathered tags were deprecated, with the indicated generative tag serving as the Preferred-Value:

i-ami (Preferred-Value: ami)

i-bnn (Preferred-Value: bnn)

i-pwn (Preferred-Value: pwn)

i-tao (Preferred-Value: tao)

i-tay (Preferred-Value: tay)

i-tsu (Preferred-Value: tsu)

zh-hakka (Preferred-Value: hak)

zh-min (no Preferred-Value; see below)

zh-min-nan (Preferred-Value: nan)

zh-xiang (Preferred-Value: hns)

The tag "zh-min", originally registered under [RFC1766], is a special case: it represents a small class of Chinese languages, but is not a true macrolanguage. The string "min" could not ever be used to tag these languages since the [ISO639-3] code element 'min' is assigned to an individual language (Minangkabau) that is not related to Chinese ('zh'). Because it is not believed to represent a useful linguistic entity for tagging purposes, it was deprecated without a Preferred-Value.

The following grandfathered and redundant sign language tags were deprecated, with the indicated generative tag serving as the Preferred-Value:

sgn-BE-FR (Preferred-Value: sfb)

sgn-BE-NL (Preferred-Value: vgt)

sgn-BR (Preferred-Value: bzs)

sgn-CH-DE (Preferred-Value: sgg)

sgn-CO (Preferred-Value: csn)

sgn-DE (Preferred-Value: gsg)

sgn-DK (Preferred-Value: dsl)

sgn-ES (Preferred-Value: ssp)

sgn-FR (Preferred-Value: fsl)

sgn-GB (Preferred-Value: bfi)

sgn-GR (Preferred-Value: gss)

sgn-IE (Preferred-Value: isg)

sgn-IT (Preferred-Value: ise)

sgn-JP (Preferred-Value: jsl)

sgn-MX (Preferred-Value: mfs)

sgn-NI (Preferred-Value: ncs)

sgn-NL (Preferred-Value: dse)

sgn-NO (Preferred-Value: nsl)

sgn-PT (Preferred-Value: psr)

sgn-SE (Preferred-Value: swl)

sgn-US (Preferred-Value: ase)

sgn-ZA (Preferred-Value: sfs)

No change was made to the Description field(s) for any of the grandfathered or redundant tags. For example, the redundant tag "sgn-US" continues to carry the Description "American Sign Language". The sign language tags registered prior to [RFC4646] remain an exception to the general principle that the meaning of a non-grandfathered tag can be derived from its component subtags.

In previous versions of the registry, grandfathered tags that had been deprecated as a result of adding an ISO 639-based language subtag included a Comments field, with a value of the form "replaced by ISO code xxx", where 'xxx' represented the new language subtag. These comments duplicated the information contained within the Preferred-Value field and were deleted as part of this update. No changes were made to other Comments fields.

2.6. Preferred-Value Changes

[RFC5646], Section 3.1.7 provides for the value of Preferred-Value fields to be updated as necessary to reflect changes in one of the source standards. Accordingly, the Preferred-Value fields for the following deprecated tags were changed:

i-hak (changed from zh-hakka to hak)

zh-guoyu (changed from zh-cmn to cmn)

This makes it unnecessary for consumers of the Language Subtag Registry to follow a "chain" of Preferred-Values in order to arrive at a non-deprecated subtag.

2.7. Additional Changes

For consistency with the handling of alternative names in language subtags, Description fields for script subtags taken from [ISO15924] that represent alternative names were converted to multiple Description fields. For example, the Description "Han (Hanzi, Kanji, Hanja)" was converted to four separate Description fields. Some Description fields for script subtags contained parenthetical material that was explanatory, rather than identifying alternative names; these fields were not altered.

This situation does not apply to region subtags taken from [ISO3166-1] and [UN_M.49] because those standards do not provide freely available alternative names for code elements.

Description fields in inverted form for script and region subtags were rearranged to remove the inversion, for consistency with the handling of language subtags in Sections 2.2 and 2.3. For example, the Description field "Korea, Republic of" was changed to "Republic of Korea".

The capitalization of the Subtag fields for certain grandfathered and redundant tags (sgn-BE-fr, sgn-BE-nl, sgn-CH-de, and yi-latn) was modified to conform with the capitalization conventions described in [RFC5646], Section 2.1.1. This has no effect on the validity or meaning of these tags.

The Description field for subtag ’sgn’ was capitalized as "Sign languages" to match the capitalization used for other languages in [ISO639-5], even though this capitalization does not exactly match that used for code element ’sgn’ in any of the ISO 639 parts.

The Deprecated field for the region subtag TP was modified from 2002-11-15 to 2002-05-20, to correct a clerical error. The corrected date reflects the actual date the code element TP was withdrawn in [ISO3166-1].

The order of fields within records in the registry was adjusted as necessary to match the order in which these fields are described in [RFC5646], Section 3.1.2. This ordering is not required by [RFC5646] and may not necessarily be reflected in future additions or modifications to the registry.

3. Updated Registry Contents

IANA has updated the Language Subtag Registry according to the provided replacement contents. The replacement content was listed in the working draft of this document, but was deleted prior to publication as an RFC to avoid potential confusion with the registry itself. The Language Subtag Registry is available from the IANA website, <http://www.iana.org>.

4. Security Considerations

For security considerations relevant to the Language Subtag Registry and the use of language tags, see [RFC5646].

5. IANA Considerations

IANA has updated the Language Subtag Registry, which can be found via <http://www.iana.org>. For details on the procedures for the format and ongoing maintenance of this registry, see RFC 5646.

6. References

6.1. Normative References

[ISO639-3] International Organization for Standardization, "ISO 639-3:2007. Codes for the representation of names of languages - Part 3: Alpha-3 code for comprehensive coverage of languages", February 2007.

[ISO639-5] International Organization for Standardization, "ISO 639-5:2008. Codes for the representation of names of languages -- Part 5: Alpha-3 code for language families and groups", May 2008.

[RFC5646] Phillips, A., Ed. and M. Davis, Ed., "Tags for Identifying Languages", RFC 5646, September 2009.

[iso-639-3-macrolanguages_20090120] International Organization for Standardization, "ISO 639-3 Macrolanguage Mappings", January 2009, <http://www.sil.org/iso639-3/iso-639-3-macrolanguages_20090120.tab>.

[iso-639-3_20090210] International Organization for Standardization, "ISO 639-3 Code Set", February 2009, <http://www.sil.org/iso639-3/iso-639-3_20090210.tab>.

[iso-639-3_Name_Index_20090210] International Organization for Standardization, "ISO 639-3 Language Names Index", February 2009, <http://www.sil.org/iso639-3/iso-639-3_Name_Index_20090210.tab>.

[iso639-5.tab.txt] International Organization for Standardization, "ISO 639-5 code list, Tab-delimited text", February 2009, <http://www.loc.gov/standards/iso639-5/iso639-5.tab.txt>.

6.2. Informative References

[ISO15924] International Organization for Standardization, "ISO 15924:2004. Information and documentation -- Codes for the representation of names of scripts", January 2004.

[ISO3166-1] International Organization for Standardization, "ISO 3166- 1:2006. Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes", November 2006.

[ISO639-1] International Organization for Standardization, "ISO 639-1:2002. Codes for the representation of names of languages -- Part 1: Alpha-2 code", July 2002.

[ISO639-2] International Organization for Standardization, "ISO 639-2:1998. Codes for the representation of names of languages -- Part 2: Alpha-3 code", October 1998.

[RFC1766] Alvestrand, H., "Tags for the Identification of Languages", RFC 1766, March 1995.

[RFC3066] Alvestrand, H., "Tags for the Identification of Languages", RFC 3066, January 2001.

[RFC4645] Ewell, D., "Initial Language Subtag Registry", RFC 4645, September 2006.

[RFC4646] Phillips, A. and M. Davis, "Tags for Identifying Languages", BCP 47, RFC 4646, September 2006.

[UN_M.49] Statistics Division, United Nations, "Standard Country or Area Codes for Statistical Use", Revision 4 (United Nations publication, Sales No. 98.XVII.9, June 1999.

Appendix A. Acknowledgements

This memo is a collaborative work of the Language Tag Registry Update (LTRU) Working Group. All of its members have made significant contributions to this memo and to its predecessor, [RFC4645].

Specific contributions to this memo were made by Stephane Bortzmeyer, John Cowan, Mark Davis, Martin Duerst, Frank Ellermann, Debbie Garside, Kent Karlsson, Gerard Lang, Addison Phillips, Randy Presuhn, and CE Whitehead.

Author's Address

Doug Ewell (editor)
Consultant

EMail: doug@ewellic.org
URI: http://www.ewellic.org