Network Working Group
Request for Comments: 3066
BCP: 47
Obsoletes: 1766
Category: Best Current Practice

H. Alvestrand
Cisco Systems

January 2001

Tags for the Identification of Languages

Status of this Memo

Copyright Notice

Abstract

This document describes a language tag for use in cases where it is desired to indicate the language used in an information object, how to register values for use in this language tag, and a construct for matching such language tags.

1. Introduction

Human beings on our planet have, past and present, used a number of languages. There are many reasons why one would want to identify the language used when presenting information.

In some contexts, it is possible to have information available in more than one language, or it might be possible to provide tools (such as dictionaries) to assist in the understanding of a language.

Also, many types of information processing require knowledge of the language in which information is expressed in order for that process to be performed on the information; for example spell-checking, computer-synthesized speech, Braille, or high-quality print renderings.

One means of indicating the language used is by labeling the information content with an identifier for the language that is used in this information content.

This document specifies an identifier mechanism, a registration function for values to be used with that identifier mechanism, and a construct for matching against those values.

The keywords "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC 2119].

2. The Language tag

2.1 Language tag syntax

The language tag is composed of one or more parts: A primary language subtag and a (possibly empty) series of subsequent subtags.

The syntax of this tag in ABNF [RFC 2234] is:

```
 Language-Tag = Primary-subtag *( "-" Subtag )

 Primary-subtag = 1*8ALPHA

 Subtag = 1*8(ALPHA / DIGIT)
```

The productions ALPHA and DIGIT are imported from RFC 2234; they denote respectively the characters A to Z in upper or lower case and the digits from 0 to 9. The character "-" is HYPHEN-MINUS (ABNF: %x2D).

All tags are to be treated as case insensitive; there exist conventions for capitalization of some of them, but these should not be taken to carry meaning. For instance, [ISO 3166] recommends that country codes are capitalized (MN Mongolia), while [ISO 639] recommends that language codes are written in lower case (mn Mongolian).

2.2 Language tag sources

The namespace of language tags is administered by the Internet Assigned Numbers Authority (IANA) [RFC 2860] according to the rules in section 3 of this document.

The following rules apply to the primary subtag:

- All 2-letter subtags are interpreted according to assignments found in ISO standard 639, "Code for the representation of names of languages" [ISO 639], or assignments subsequently made by the ISO 639 part 1 maintenance agency or governing standardization bodies. (Note: A revision is underway, and is expected to be released as

ISO 639-1:2000)

- All 3-letter subtags are interpreted according to assignments found in ISO 639 part 2, "Codes for the representation of names of languages -- Part 2: Alpha-3 code [ISO 639-2]", or assignments subsequently made by the ISO 639 part 2 maintenance agency or governing standardization bodies.

- The value "i" is reserved for IANA-defined registrations

- The value "x" is reserved for private use. Subtags of "x" shall not be registered by the IANA.

- Other values shall not be assigned except by revision of this standard.

The reason for reserving all other tags is to be open towards new revisions of ISO 639; the use of "i" and "x" is the minimum we can do here to be able to extend the mechanism to meet our immediate requirements.

The following rules apply to the second subtag:

- All 2-letter subtags are interpreted as ISO 3166 alpha-2 country codes from [ISO 3166], or subsequently assigned by the ISO 3166 maintenance agency or governing standardization bodies, denoting the area to which this language variant relates.

- Tags with second subtags of 3 to 8 letters may be registered with IANA, according to the rules in chapter 5 of this document.

- Tags with 1-letter second subtags may not be assigned except after revision of this standard.

There are no rules apart from the syntactic ones for the third and subsequent subtags.

Tags constructed wholly from the codes that are assigned interpretations by this chapter do not need to be registered with IANA before use.

The information in a subtag may for instance be:

- Country identification, such as en-US (this usage is described in ISO 639)

- Dialect or variant information, such as en-scouse

- Languages not listed in ISO 639 that are not variants of any listed language, which can be registered with the i-prefix, such as i-tsolyani

- Region identification, such as sgn-US-MA (Martha's Vineyard Sign Language, which is found in the state of Massachusetts, US)

This document leaves the decision on what tags are appropriate or not to the registration process described in section 3.

ISO 639 defines a maintenance agency for additions to and changes in the list of languages in ISO 639. This agency is:

International Information Centre for Terminology (Infoterm)
P.O. Box 130
A-1021 Wien
Austria

Phone: +43 1 26 75 35 Ext. 312
Fax: +43 1 216 32 72

ISO 639-2 defines a maintenance agency for additions to and changes in the list of languages in ISO 639-2. This agency is:

Library of Congress
Network Development and MARC Standards Office
Washington, D.C. 20540
USA

Phone: +1 202 707 6237
Fax: +1 202 707 0115
URL: http://www.loc.gov/standards/iso639

The maintenance agency for ISO 3166 (country codes) is:

ISO 3166 Maintenance Agency Secretariat
c/o DIN Deutsches Institut fuer Normung
Burggrafenstrasse 6
Postfach 1107
D-10787 Berlin
Germany

Phone: +49 30 26 01 320
Fax: +49 30 26 01 231
URL: http://www.din.de/gremien/nas/nabd/iso3166ma/

ISO 3166 reserves the country codes AA, QM-QZ, XA-XZ and ZZ as user-assigned codes. These MUST NOT be used to form language tags.

2.3 Choice of language tag

One may occasionally be faced with several possible tags for the same body of text.

Interoperability is best served if all users send the same tag, and use the same tag for the same language for all documents. If an application has requirements that make the rules here inapplicable, the application protocol specification MUST specify how the procedure varies from the one given here.

The text below is based on the set of tags known to the tagging entity.

1. Use the most precise tagging known to the sender that can be ascertained and is useful within the application context.

2. When a language has both an ISO 639-1 2-character code and an ISO 639-2 3-character code, you MUST use the tag derived from the ISO 639-1 2-character code.

3. When a language has no ISO 639-1 2-character code, and the ISO 639-2/T (Terminology) code and the ISO 639-2/B (Bibliographic) code differ, you MUST use the Terminology code. NOTE: At present, all languages for which there is a difference have 2-character codes, and the displeasure of developers about the existence of 2 code sets has been adequately communicated to ISO. So this situation will hopefully not arise.

4. When a language has both an IANA-registered tag (i-something) and a tag derived from an ISO registered code, you MUST use the ISO tag. NOTE: When such a situation is discovered, the IANA-registered tag SHOULD be deprecated as soon as possible.

5. You SHOULD NOT use the UND (Undetermined) code unless the protocol in use forces you to give a value for the language tag, even if the language is unknown. Omitting the tag is preferred.

6. You SHOULD NOT use the MUL (Multiple) tag if the protocol allows you to use multiple languages, as is the case for the Content-Language: header.

NOTE: In order to avoid versioning difficulties in applications such as that of RFC 1766, the ISO 639 Registration Authority Joint Advisory Committee (RA-JAC) has agreed on the following policy statement:

> "After the publication of ISO/DIS 639-1 as an International Standard, no new 2-letter code shall be added to ISO 639-1 unless a 3-letter code is also added at the same time to ISO 639-2. In addition, no language with a 3-letter code available at the time of publication of ISO 639-1 which at that time had no 2-letter code shall be subsequently given a 2-letter code."

This will ensure that, for example, a user who implements "hwi" (Hawaiian), which currently has no 2-letter code, will not find his or her data invalidated by eventual addition of a 2-letter code for that language."

2.4 Meaning of the language tag

The language tag always defines a language as spoken (or written, signed or otherwise signaled) by human beings for communication of information to other human beings. Computer languages such as programming languages are explicitly excluded. There is no guaranteed relationship between languages whose tags begin with the same series of subtags; specifically, they are NOT guaranteed to be mutually intelligible, although it will sometimes be the case that they are.

The relationship between the tag and the information it relates to is defined by the standard describing the context in which it appears. Accordingly, this section can only give possible examples of its usage.

- For a single information object, it could be taken as the set of languages that is required for a complete comprehension of the complete object.
  Example: Plain text documents.

- For an aggregation of information objects, it should be taken as the set of languages used inside components of that aggregation.
  Examples: Document stores and libraries.

- For information objects whose purpose is to provide alternatives, the set of tags associated with it should be regarded as a hint that the content is provided in several languages, and that one has to inspect each of the alternatives in order to find its language or languages. In this case, a tag with multiple languages does not mean that one needs to be multi-lingual to get complete understanding of the document.
  Example: MIME multipart/alternative.

- In markup languages, such as HTML and XML, language information can be added to each part of the document identified by the markup structure (including the whole document itself). For example, one could write C'est la vie. inside a Norwegian document; the Norwegian-speaking user could then access a French-Norwegian dictionary to find out what the marked section meant. If the user were listening to that document through a speech synthesis interface, this formation could be used to signal the synthesizer to appropriately apply French text-to-speech pronunciation rules to that span of text, instead of misapplying the Norwegian rules.

2.5 Language-range

Since the publication of RFC 1766, it has become apparent that there is a need to define a term for a set of languages whose tags all begin with the same sequence of subtags.

The following definition of language-range is derived from HTTP/1.1 [RFC 2616].

```
language-range  = language-tag / "*"
```

That is, a language-range has the same syntax as a language-tag, or is the single character "*".

A language-range matches a language-tag if it exactly equals the tag, or if it exactly equals a prefix of the tag such that the first character following the prefix is "-".

The special range "*" matches any tag. A protocol which uses language ranges may specify additional rules about the semantics of "*"; for instance, HTTP/1.1 specifies that the range "*" matches only languages not matched by any other range within an "Accept-Language:" header.

NOTE: This use of a prefix matching rule does not imply that language tags are assigned to languages in such a way that it is always true that if a user understands a language with a certain tag, then this user will also understand all languages with tags for which this tag is a prefix. The prefix rule simply allows the use of prefix tags if this is the case.

3. IANA registration procedure for language tags

The procedure given here MUST be used by anyone who wants to use a language tag not given an interpretation in chapter 2.2 of this document or previously registered with IANA.

This procedure MAY also be used to register information with the IANA about a tag defined by this document, for instance if one wishes to make publicly available a reference to the definition for a language such as sgn-US (American Sign Language).

Tags with a first subtag of "x" need not, and cannot, be registered.

The process starts by filling out the registration form reproduced below.

LANGUAGE TAG REGISTRATION FORM

Name of requester :

E-mail address of requester:

Tag to be registered :

English name of language :

Native name of language (transcribed into ASCII):

Reference to published description of the language (book or article):

Any other relevant information:

The language form must be sent to <ietf-languages@iana.org> for a 2-week review period before it can be submitted to IANA. (This is an open list. Requests to be added should be sent to <ietf-languages-request@iana.org>.)

When the two week period has passed, the language tag reviewer, who is appointed by the IETF Applications Area Director, either forwards the request to IANA@IANA.ORG, or rejects it because of significant objections raised on the list. Note that the reviewer can raise objections on the list himself, if he so desires. The important thing is that the objection must be made publicly.

The applicant is free to modify a rejected application with additional information and submit it again; this restarts the 2-week comment period.

Decisions made by the reviewer may be appealed to the IESG [RFC 2028] under the same rules as other IETF decisions [RFC 2026]. All registered forms are available online in the directory http://www.iana.org/numbers.html under "languages".

Updates of registrations follow the same procedure as registrations. The language tag reviewer decides whether to allow a new registrant to update a registration made by someone else; in the normal case, objections by the original registrant would carry extra weight in such a decision.

There is no deletion of registrations; when some registered tag should not be used any more, for instance because a corresponding ISO 639 code has been registered, the registration should be amended by adding a remark like "DEPRECATED: use <new code> instead" to the "other relevant information" section.

Note: The purpose of the "published description" is intended as an aid to people trying to verify whether a language is registered, or what language a particular tag refers to. In most cases, reference to an authoritative grammar or dictionary of the language will be useful; in cases where no such work exists, other well known works describing that language or in that language may be appropriate. The language tag reviewer decides what constitutes a "good enough" reference material.

4. Security Considerations

The only security issue that has been raised with language tags since the publication of RFC 1766, which stated that "Security issues are believed to be irrelevant to this memo", is a concern with language ranges used in content negotiation - that they may be used to infer the nationality of the sender, and thus identify potential targets for surveillance.

This is a special case of the general problem that anything you send is visible to the receiving party; it is useful to be aware that such concerns can exist in some cases.

The evaluation of the exact magnitude of the threat, and any possible countermeasures, is left to each application protocol.

5. Character set considerations

Language tags may always be presented using the characters A-Z, a-z, 0-9 and HYPHEN-MINUS, which are present in most character sets, so presentation of language tags should not have any character set issues.

The issue of deciding upon the rendering of a character set based on the language tag is not addressed in this memo; however, it is thought impossible to make such a decision correctly for all cases unless means of switching language in the middle of a text are defined (for example, a rendering engine that decides font based on Japanese or Chinese language may produce suboptimal output when a mixed Japanese-Chinese text is encountered)

6. Acknowledgements

This document has benefited from many rounds of review and comments in various fora of the IETF and the Internet working groups.

Any list of contributors is bound to be incomplete; please regard the following as only a selection from the group of people who have contributed to make this document what it is today.

In alphabetical order:

Glenn Adams, Tim Berners-Lee, Marc Blanchet, Nathaniel Borenstein, Eric Brunner, Sean M. Burke, John Clews, Jim Conklin, Peter Constable, John Cowan, Mark Crispin, Dave Crocker, Mark Davis, Martin Duerst, Michael Everson, Ned Freed, Tim Goodwin, Dirk-Willem van Gulik, Marion Gunn, Paul Hoffman, Olle Jarnefors, Kent Karlsson, John Klensin, Alain LaBonte, Chris Newman, Keith Moore, Masataka Ohta, Keld Jorn Simonsen, Otto Stolz, Rhys Weatherley, Misha Wolf, Francois Yergeau and many, many others.

Special thanks must go to Michael Everson, who has served as language tag reviewer for almost the complete period since the publication of RFC 1766, and has provided a great deal of input to this revision.

7. Author’s Address

Harald Tveit Alvestrand
Cisco Systems
Weidemanns vei 27
7043 Trondheim
NORWAY

Phone: +47 73 50 33 52
EMail: Harald@Alvestrand.no

8. References

[ISO 639] ISO 639:1988 (E/F) - Code for the representation of names of languages - The International Organization for Standardization, 1st edition, 1988-04-01 Prepared by ISO/TC 37 - Terminology (principles and coordination). Note that a new version (ISO 639-1:2000) is in preparation at the time of this writing.

[ISO 639-2] ISO 639-2:1998 - Codes for the representation of names of languages -- Part 2: Alpha-3 code - edition 1, 1998-11-01, 66 pages, prepared by a Joint Working Group of ISO TC46/SC4 and ISO TC37/SC2.

[ISO 3166] ISO 3166:1988 (E/F) - Codes for the representation of names of countries - The International Organization for Standardization, 3rd edition, 1988-08-15.

[RFC 1327] Kille, S., "Mapping between X.400 (1988) / ISO 10021 and RFC 822", RFC 1327, May 1992.

[RFC 1521] Borenstein, N., and N. Freed, "MIME Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1521, September 1993.

[RFC 2026] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

[RFC 2028] Hovey, R. and S. Bradner, "The Organizations Involved in the IETF Standards Process", BCP 11, RFC 2028, October 1996.

[RFC 2119] Bradner, S."Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC 2234] Crocker, D. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, November 1997.

[RFC 2616] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", RFC 2616, June 1999.

[RFC 2860] Carpenter, B., Baker, F. and M. Roberts, "Memorandum of Understanding Concerning the Technical Work of the Internet Assigned Numbers Authority", RFC 2860, June 2000.

Appendix A: Language Tag Reference Material

The Library of Congress, maintainers of ISO 639-2, has made the list of languages registered available on the Internet.

At the time of this writing, it can be found at http://www.loc.gov/standards/iso639-2/langhome.html

The IANA registration forms for registered language codes can be found at http://www.iana.org/numbers.html under "languages".

The ISO 3166 Maintenance Agency has published Web pages at

http://www.din.de/gremien/nas/nabd/iso3166ma/

Appendix B: Changes from RFC 1766

- Email list address changed from ietf-types@uninett.no to ietf-languages@iana.org
- Updated author’s address
- Added language-range construct from HTTP/1.1
- Added use of ISO 639-2 language codes
- Added reference to Library of Congress lists of language codes
- Changed examples to use registered tags
- Added "Any other information" to registration form
- Added description of procedure for updating registrations
- Changed target category for document from standards track to BCP
- Moved the content-language header definition into another document
- Added numbers to the permitted characters in language tags

Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.