INTERNET-DRAFT
Category: Network
Title: Tags for Languages
<draft-langtags-phillips-davis-01>

Addison Phillips
webMethods Inc
Mark Davis
IBM Corp.

Status of this Memo

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, including user defined extensions for private interchange.

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. These labels can also be used to specify user preferences when selecting information content, or for labeling additional attributes of content.

In particular, it is often necessary to define additional specific information about the dialect, writing system, or orthography used in a document, as this

information may be useful to specialists or may be important in understanding the structure of the information and ways in which it should be processed.

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. It also defines a mechanism for private use extension and how private use, registered values, and matching interact.

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 sequence of subtags has a specific structure that depends on the length of the subtag to distinguish each tag type.

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

```
Language-tag         = lang ["-" script] ["-" region] *("-" variant)  [extensions]
                     =/ "x" *("-" alphanum)
                     =/ grandfathered-registrations
lang                 = shortest-alpha-ISO-639-code
                     =/ registered-lang
script               = ISO-15924-code
region               = shortest-alpha-ISO-3166-code
variant              = year
                     =/ registered-variant
registered-lang      = 5*16 alphanum
registered-variant   = 5*16 alphanum
extensions           = "-x" 1* ("-" key "=" value)
key                  = ALPHA *alphanum
value                = 1* utf8uri
alphanum             = (ALPHA / DIGIT)
utf8uri              = (ALPHA / DIGIT / 1*4 (“%” 2 HEXDIG))
year                 = non-zero-digit [*DIGIT] [“BCE”]
non-zero-digit       = %x31-39
```

The character "-" is HYPHEN-MINUS (ABNF: %x2D). The character "=" is EQUALS SIGN (ABNF: %x3D). The character “%” is PERCENT SIGN (ABNF: %x25).

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

For examples, see Appendix B: Examples of Language Tags (Informative) at the end of this document.

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.

Terminology in this section:

- Tag or tags refers to a complete language tag, such as 'fr-Latn-CA'
- Subtag refers to a specific section of a tag, separated by hyphen, such as 'Latn' in 'fr-Latn-CA'
- Code or codes refers to the tags defined in external standards. For example, 'Latn' is an ISO 15924 script code (which can be used as a script subtag in a language tag)

The rules in this section apply to the various subtags within the language tags defined in this document, excepting those "grandfathered" tags defined in section 2.2.1 of this document. Those tags should be considered as exceptions to the rules presented here.

The following rules apply to the ***primary*** (language) subtag:

- All 2-character language 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.
- All 3-character language 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, or assignments of 3-character disambiguation registrations according to Rule 7a. Ambiguity in Section 2.3 Choice of language tag of this document.
- ISO639 reserves for private use codes the range 'qaa' through 'qtz'. These codes should be used for non-registered language subtags.
- Subtags of 5 to 16 characters may be registered with IANA, according to the rules in section 3 of this document. (Note that previously, in rfc3066, the IANA registry contained whole tag registrations such as de-CH-1994, whereas this document refers to the registration of subtags such as 'klingon')
- The single character subtag "x" as the primary subtag indicates that the whole language tag is a private use tag. The value and semantic meaning of such a tag as a whole and of the subtags used within such as tag are undefined by this document.
- Other values shall not be assigned except by revision of this document.

The following rules apply to the script subtags:

- All 4-character subtags are interpreted as ISO 15924 alpha-4 script codes from [ISO 15924], or subsequently assigned by the ISO 15924 maintenance agency or governing standardization bodies, denoting the script or writing system used in conjunction with this language. These alpha4 tags may only occur as the second subtag in a tag. For example: 'de-Latn’ represents German written using Latin script.
- ISO 15924 reserves the codes Qaaa-Qtzz for private use values. These codes should be used for non-registered script values.

- Script subtags can NOT be registered using the process in Section 3 of this document. Variant subtags may be considered for registration for that purpose. Note that registered subtags can only appear in the first position or as a sequence on the end of a language tag. This makes the associated exception handling when parsing tags simpler.

The following rules apply to the region subtags:

- All 2-character and 3-character subtags following the primary subtag denote the area to which this language variant relates, and are interpreted according to assignments found in ISO 3166 alpha- 2 (or alpha-3) country codes from [ISO 3166], assignments subsequently made by the ISO 3166 maintenance agency or governing standardization bodies, or assignments of 3-character disambiguation registrations according to Rule 7a. [Ambiguity] in Section 2.3 [Choice of language tag] of this document.
- Region subtags must occur after any script subtags and before any variant subtags or extensions. The shortest form ISO3166 code must be used to form the subtag. At the time this document was written, all alpha3 codes had a corresponding alpha2 code. Use of alpha3 codes is provided for a foreseeable future in which alpha2 codes have been exhausted. Example: 'de-Latn-CH' represents German written using Latin script for Switzerland.
- ISO 3166 reserves the country codes AA, QM-QZ, XA-XZ and ZZ (plus any three-character sequences starting with these codes) as user-assigned codes. These codes should be used for private use region subtags.
- No region subtags can be registered using the process in Section 3 of this document. Variant subtags may be considered for registration for this purpose.

The following rules apply to the variant subtags:

- Numeric subtags are variant subtags interpreted to be Gregorian calendar year numbers. The year number must not start with a zero and is interpreted to be a year number in the Common Era (C.E.) unless a suffix of "BCE" is appended to indicate a Before Common Era date. By definition there is no year "0" and the implied calendar is proleptic (that is, the year numbers extend back to a period before the actual adoption of the Gregorian calendar and thus may not be historically correct). Dates are useful to distinguish events such as official spelling reforms. For example: de-CH-1904.
- Additional alphanumeric subtags of 5 to 16 characters may be registered with IANA, according to the rules in section 3 of this document. Registered subtags must not begin with the character 'x', which is reserved for private use subtags. (Note that previously, in rfc3066, the IANA registry contained whole tag registrations such as 'en-boont', whereas this document refers to the registration of subtags such as 'boont')
- Additional alphanumeric subtags of 5 to 16 characters starting with 'x' are reserved for private use. The semantics of these subtags must be defined by the end users of such subtags and the semantic meaning should be considered external to this document.

The following rules apply to the extensions:

- Extension subtags are separated from the other subtags defined in this document by the single character subtag ‘x’.
- Extensions must follow all language, script, region, and variant subtags.
- Extensions consist of key-value pairs. Each key-value pair is separated using the HYPHEN-MINUS character. The key and value are separated by EQUALS SIGN.

- The key must consist of an alphanumeric sequence of any length starting with a letter (alpha character). It may not be empty.
- The value must consist of a non-empty sequence of UTF-8 [RFC 3629] characters, encoded using the URI encoding rules in Section 2.1 of [RFC 2396]. Note that alphanumeric values will be encoded as themselves. So the value "abc" is encoded as "abc", while the character LATIN SMALL LETTER A WITH GRAVE (U+00E0, ABNF %xE0) is encoded as "%c3%a0".
- No source is defined for extensions. External agreement or standardization of extension subtags is by private agreement and should not be considered part of this document.

For example: 'az-Arab-x-SIL=AZE-dialect=derbend' contains two extension subtags. The first is "SIL=AZE" and the second is "dialect=derbend".

2.2.1 Pre-Existing RFC 3066 Tag Registrations

Existing IANA registered language tags from RFC1766/RFC3066 that are not defined by additions to this document maintain their validity. IANA will maintain these tags, adding a notation that they are "grandfathered from RFC 3066".

The rules governing existing RFC 1766 and RFC 3066 registered tags are:

- If the formerly registered tag would now be defined by this document, then the existing tag is marked as superseded by this document and no subtag will be registered as a result. For example, zh-Hans would now be defined by the addition of ISO 15924 script codes.
- If the registered tag contained one or more subtags that follow the guidelines for registered language or variant subtags, and all of the subtags are either now defined by this document or would be valid to register, then each subtag not already covered by this document will be registered automatically by IANA without further review and the existing tag marked as superseded by this document. For example: the tag 'en-boont' fits the pattern for a registered variant. The variant subtag "boont" will be registered automatically and 'en-boont' marked as superseded.
- If the registered tag contains any subtags that are not otherwise valid for registration according to the rules in this document, then the tag as a whole is maintained as an exceptional case (that is, it is "grandfathered"). This includes special cases of Sign Language tags. For example, the tag 'i-klingon' is not covered by any addition and is grandfathered, as is sgn-BE-fr (Belgian French Sign Language).

Users of tags that are grandfathered should consider registering the appropriate subtags using the new format (but are not required to).

2.2.2 Possibilities for registration consideration include:

- Languages not listed in ISO 639 that are not variants of any listed language, can be registered, such as tsolyani. Before attempting to register a language subtag, there should be a good faith attempt to register the language with ISO 639. No language subtags will be registered for codes that exist in ISO 639-1 or ISO 639-2.
- Dialect or other divisions or variations within a language, its orthography, writing system, regional variation, or historical usage, such as the "scouse" subtag (the Scouse dialect of English).

This document leaves the decision on what subtags 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/

The registration authority for ISO 15924 (script codes) is:

Unicode Consortium Box 391476
Mountain View, CA 94039-1476, USA
URL: http://www.unicode.org/iso15924

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. In general, one should choose the tag with the fewest number of subtags for a given context. For example, "de" might suffice for most tagging of emails, while "de-Latn-DE-1903-x-collation=phonebook" is probably unnecessarily precise for such a task.

2. Use the most precise tag known to the sender that can be ascertained and is useful within the application context. Although this appears to be a contradiction of the first rule, it is important to consider the application context and be as precise as possible within reasonable limitations.

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

4. 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) codes differ, you MUST use the Terminology code. NOTE: At present all languages that have both kinds of 3-character code also are assigned a 2-character code, and the displeasure of developers about the existence of two different code sets has been adequately communicated to ISO. So this situation will hopefully not arise.

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 in HTTP.

NOTE: In order to avoid versioning difficulties in applications such as that experienced in 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 "haw" (Hawaiian), which currently has no 2-character code, will not find his or her data invalidated by eventual addition of a 2-character code for that language."

7. To maintain backwards compatibility, there are two provisions to account for potential instability in ISO 639, 3166, and 15924 codes.

   a. **Ambiguity.** In the event that one of these standards assigns a code a new meaning or reassigns a deprecated code, the new use of the code will not be permitted in language tags defined by this document.

      The language subtag reviewer, as described in Section 3, shall prepare a proposal for entering in the IANA registry, as soon as practical, a variant subtag as a surrogate value for the new code. The form of the registered variant should be a 3-character tag that is not otherwise permitted by the ISO registration authority, such as one containing a sequence number (e.g. CS1). However, should the ISO registration authority or standard expand the allowable 3-character tags so that this is not possible, then this may be any other valid registration (such as CS2003, marking the year of introduction).

      The normal registration process described in Section 3 of this document applies to the review and registration of such variant subtags, except that they may be 3 characters long, as described above. Note that these subtags should never be used in combination with the subtag type for which they are a surrogate. For example, a "region" variant subtag should not be used with a region subtag.

      For example:
       cs-CS (Czech for Czechoslovakia)
       sr-CS1 (Serbian for Serbia and Montenegro, using a registered variant)
       sr-CS-CS1 (Incorrect usage)

       qx-Latn (hypothetical reassigned value 'qx')
       qx2003-Latn (hypothetical registered language subtag)

   b. **Stability.** All other ISO codes are valid, even if they have been deprecated. At the time of this writing, this includes the following list. Where a new equivalent code has been defined (given below on the right side after a tilde), implementations should treat these tags as identical.

      For example, deprecated ISO 639 codes:
            iw ~ he
            in ~ id
            ji ~ yi

```
Deprecated ISO 3166 codes
      FX
      TP ~ TL
      YU
```

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.

If a language tag B contains language tag A as a prefix, then B should typically be "narrower" or "more specific" than A. However, this relationship is not guaranteed. There is also no guaranteed relationship between languages whose tags begin with the same series of subtags; specifically, they are NOT guaranteed to be mutually intelligible, although they may be.

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 3066, 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 subtag not given an interpretation in section 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 or subtag 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).

Variant subtags may not be registered using the pattern 2 ALPHA 4 DIGIT to accommodate the provisions in Section 2.3, rule 7a of this document. That is, the subtag xx1234 can NOT be registered except under the aforementioned provisions.

Subtags that start with the letter 'x' also many not be registered, since this prefix is reserved for unregistered, private use subtags.

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

LANGUAGE TAG SUBTAG REGISTRATION FORM

Name of requester:

E-mail address of requester:

Subtag to be registered:

Type of Subtag: [ ] language [ ] region [ ] variant

Full English name of subtag:

Intended meaning of the subtag:

If variant subtag, the intended prefix(es) of subtag:

Native name of language (transcribed into ASCII):

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

Any other relevant information:

The subtag registration form must be sent to <ietf-languages@iana.org> for a two 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>.)

Variant subtags are generally registered for use with a particular prefix or set of prefixes. For example, the subtag ‘boont’ is intended for use with the prefix 'en-', since Boontling is a dialect of English.

Any registered subtag can be incorporated into a variety of language tags, according to the rules of 2.1 Language tag syntax. This makes the validation simpler and thus more uniform across implementations, and does not require new registrations for different intended prefixes.

However, the intended prefixes for a given registered subtag will be maintained in the IANA registry as a guide to usage. If it is necessary to add an additional intended prefix to that list for an existing language tag, that can be done by filing an additional registration form. In that form, the "Any other relevant information: " field should indicate that it is the addition of an additional intended prefix.

When the two week period has passed, the subtag 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 or she 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 two 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 subtag reviewer decides whether to allow a new registrant to update a registration made by someone else; normally objections by the original registrant would carry extra weight in such a decision.

Registrations are permanent and stable. When some registered subtag should not be used any more, for instance because a corresponding ISO 639 code has been created, 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 subtag refers to. In most cases, reference to an authoritative grammar or dictionary of that 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 subtag reviewer decides what constitutes "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, EQUALS SIGN, 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 sub-optimal output when a mixed Japanese- Chinese text is encountered)

6. Acknowledgements

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.

The contributors to RFC 3066 and RFC 1766, the precursors of this document, made enormous contributions directly or indirectly to this document and are generally responsible for the success of language tags.

The following people (in alphabetical order) contributed to this document or to RFCs 1766 and 3066:

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

Very special thanks must go to Harald Tveit Alvestrand, who originated RFCs 1766 and 3066, and without whom this document would not have been possible. 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.

7. Authors' Addresses

Addison P. Phillips
webMethods, Inc.
432 Lakeside Drive
Sunnyvale, CA, 94088, USA
Phone: +1 408 962-5487
EMail: aphillips@webmethods.com

Mark Davis
IBM
Email: mark.davis@us.ibm.com

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.

[ISO 15924] ISO 15924:2003 (E/F) - Codes for the representation of names of scripts - The International Organization for Standardization, 2003-03-04, prepared by ISO TC46/WG3 and Michael Everson.

[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.

[RFC 3629] Yergeau, F., "UTF-8, a transformation format of ISO 10646”, RFC 3629, November 2003

[RFC 2396] Berners-Lee, T., Fielding, R., Masinter, L., “Uniform Resource Identifiers (URI): Generic Syntax”, RFC 2396, August 1998

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: Examples of Language Tags (Informative)

Simple language code:

```
      de (German)
      fr (French)
```

ja (Japanese)

Language code plus Script code :
zh-Hant (Traditional Chinese)
en-Latn (English written in Latin script)
sr-Cyrl (Serbian written with Cyrillic script)

Language-Script-Region:
zh-Hans-CN (Simplified Chinese for the PRC)
sr-Latn-CS1 (Serbian, Latin script, Serbia and Montenegro)

Language-Script-Region-Variant:
en-Latn-US-boont (Boontling dialect of English)
de-Latn-CH-1904 (Swiss German with the 1904 spelling reform)

Language-Region:
de-DE (German for Germany)
zh-SG (Chinese for Singapore)
cs-CS (Czech for Czechoslovakia)
sr-CS1 (Serbian for Serbia and Montenegro, IANA registered variant, see 7a in Section 2.3)

Language-Year Variant:
grc-700BCE (Ancient Greek, circa 700 BCE, e.g. during Homer’s lifetime)
frm-1400 (Middle French, circa 1400 C.E.)
ang-1066 (Old English, circa the Norman Conquest)

Other Mixtures:
zh-CN (Chinese for the PRC)
en-boont (Boontling dialect of English)

Extension mechanism:
de-CH-x-collation=phonebook
az-Arab-x-SIL=AZE-dialect=derbend

EXPIRATION: This document expires in November 2004.