Network Working Group
Internet-Draft
Expires: December 1, 2004

A. Phillips, Ed.
webMethods, Inc.
M. Davis
IBM

June 02, 2004

TOC

Tags for Identifying Languages

draft-phillips-langtags-03

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at **http://www.ietf.org/ietf/1id-abstracts.txt**.

The list of Internet-Draft Shadow Directories can be accessed at **http://www.ietf.org/shadow.html**.

This Internet-Draft will expire on December 1, 2004.

Copyright Notice

Abstract

This document describes the structure, content, construction, and semantics of language tags for use in cases where it is desirable to indicate the language used in an information object. It also describes how to register values for use in language tags and a construct for matching such language tags, including user defined extensions for private interchange. This document replaces RFC 3066 (which replaced RFC 1766).

Table of Contents

TOC

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 or requesting information.

Information about a user's language preferences commonly needs to be identified so that appropriate processing can be applied. For example, the user's language preferences in a brower can be used to select web pages appropriately. A choice of language preference can also be used to select among tools (such as dictionaries) to assist in the processing or understanding of content in different langauges.

In addition, knowledge about the particular language used by some piece of information content may be useful or even required by some types of information processing; for example spell-checking, computer-synthesized speech, Braille transcription, or high-quality print renderings.

One means of indicating the language used is by labeling the information content with a language identifier. These identifiers can also be used to specify user preferences when selecting information content, or for labeling additional attributes of content and associated resources.

These identifiers can also be used to indicate additional attributes of content that are closely related to the language. In particular, it is often necessary to indicate specific information about the dialect, writing system, or orthography used in a document or resource, as these attributes may be important for the user to obtain information in a form that they can understand, or important in selecting appropriate processing resources for the given 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. It also defines a mechanism for private use extension and how private use, registered values, and matching interact.

This document replaces RFC 3066, which replaced RFC 1766. For a list of changes in this document, see: **Section 6**.

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][10].

TOC

2. The Language Tag

2.1 Syntax

The language tag is composed of one or more parts: A primary language subtag and a (possibly empty) series of subsequent subtags. Subtags are distinguished by their length, position in the subtag sequence, and content, so that each type of subtag can be recognized solely by these features. This makes it possible to construct a parser that can extract and assign some semantic information to the subtags, even if specific subtag values are not recognized. Thus a parser need not have an up-to-date copy of the registered subtag values to perform most searching and matching operations.

The syntax of this tag in ABNF [RFC 2234][11] is:

```
 = lang *("-" extlang) ["-" script] ["-" region] *("-" variant) *("-" extension) *("-" private_use)
 =/ private_use       ; private use tag
 =/ grandfathered     ; grandfathered registrations

 lang            = 2*3 ALPHA                ; shortest ISO 639 code
                 =/ registered-lang
 extlang         = 3 ALPHA                  ; reserved for future use
 script          = 4 ALPHA                  ; ISO 15924 code
 region          = 2 ALPHA                  ; ISO 3166 code
                 =/ 3 DIGIT                 ; UN country number
 variant         =  5*15 alphanum           ; registered or private use variants
 extension       = singleton 1* ("-" (2*31 alphanum)) ; extension subtag(s)
 private_use     = "x" 1* ("-" (1*31 alphanum))       ; private use subtag(s)
 singleton       = 1 ("a"-"w"/"y"-"z")      ; single letters except x
 registered-lang = 4*15 ALPHA               ; registered language subtag
```

```
grandfathered   = ALPHA * (alphanum / "-") ; grandfathered registration
alphanum        = (ALPHA / DIGIT)          ; letters and numbers
```

Language Tag ABNF

The character "-" is HYPHEN-MINUS (ABNF: %x2D).

The tags and their subtags, including private use extensions, 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]**[4] recommends that country codes be capitalized (MN Mongolia), while **[ISO 639]**[3] recommends that language codes be written in lower case (mn Mongolian). In the language tags defined by this document, however, the tag 'mn-MN' is not distinct from 'MN-mn' or 'mN-Mn' (or any other combination) and each of these variations conveys the same meaning: Mongolian for Mongolia.

For examples of language tags, see **Appendix B** at the end of this document.

2.2 Language Tag Sources

The namespace of language tags and their subtags is administered by the Internet Assigned Numbers Authority (IANA) **[15]** according to the rules in **Section 3.1**.

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 the subtag 'Latn' in 'fr-Latn-CA'
- Code or codes refers to tags defined in external standards (and which are used as subtags in this document). For example, 'Latn' is an **ISO 15924**[2] 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**.

Note that registered subtags can only appear in specific positions in a tag. Specifically, they can only occur as primary language subtags or as variant subtags.

In addition, private use or 'Extension' subtags MUST only occur at the end of the sequence of subtags and MUST NOT be interspersed with subtags defined in the rules that follow.

Each subtag type has unique length and content restrictions that make identification of the subtag's type possible, even if the content of the subtag itself is unrecognized. This allows tags to be parsed and processed without reference to the latest version of the underlying standards or the IANA registry and it makes the associated exception handling when parsing tags simpler.

Single letter and digit subtags are reserved for current or future use. These include the following current uses:

- The single letter primary subtag "i" is reserved for grandfathered, IANA-registered primary language subtags, such as "i-hakka".
- The single letter subtag "x" is reserved to introduce a sequence of private use subtags. The interpretation of any private use subtags is defined solely by private agreement. The interpretation of private use subtags is not defined by the rules in this section or in any standard or registry defined in this document.
- All other single letter subtags are reserved to introduce standardized extension subtag sequences as described in **Section 3.3**.

The primary subtag is the first subtag in a language tag and cannot be empty. Except as noted, the primary subtag is the "language" subtag. The following rules apply to the assignment and interpretation of the primary subtag:

- All 2-character subtags are interpreted according to assignments found in ISO standard 639, "Code for the representation of names of languages" **[ISO639-1]**[3], or assignments subsequently made by the ISO 639 Part 1 maintenance agency or governing standardization bodies.
- All 3-character 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]**[1]", 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** of this document.
- ISO639-2 reserves for private use codes the range 'qaa' through 'qtz'. These codes should be used for non-registered language subtags.
- IANA registered primary language subtags of 4 to 15 characters in length may also be used to form the primary language subtag. At the time this document was created, there were no examples of this

kind of subtag.

  - Note that previously, in **RFC 3066**[16], the IANA registry contained whole tag registrations such as 'cel-gaulish', whereas this document refers to the registration of subtags, such as '-gaulish'.
  - Example: one of the grandfathered IANA registrations is "i-hakka". The subtag "hakka" could be registered as a primary language subtag (assuming that ISO 639-x does not register this language first), making tags such as "hakka-CN" and "hakka-Hant" valid.
- 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 not defined by this document.
- Other values MUST NOT be assigned to the primary subtag except by revision or update of this document.

The following rules apply to the extended language subtags:

- Future versions (or parts) of ISO 639 may add more structure to the language codes that it provides, by providing a subtag structure or information about the relationship between ISO 639 language tags. To allow for forward compatibility with such versions of these standards, three letter subtags immediately following the primary subtag are reserved for this pupose.
- Extended language subtags must follow the primary subtag and precede any other subtags.
- There may be any additional number of extended language subtags.
- Standardization of the extended language subtags for specific uses with ISO 639 will be either by revision or update of this document. This syntax is only to be used for future ISO 639 subtags. Registration of extended language subtags and non-standard use are not permitted.
- Note: The order of the extended language subtags may be important in some interpretations of language tags. See the section **Section 2.4**.
- Example: In a future revision or update of this document, 'zh-min-nan' might represent the subdialect 'nan' of the Chinese dialect 'min'.

The following rules apply to the script subtags:

- All 4-character subtags are interpreted as ISO 15924 alpha-4 script codes from **[2]**, 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 subtags may only occur

following the primary language subtag and any extended language subtags and must occur before any other type of subtag described below.
- Example: 'de-Latn' represents German written using the Latin script.
- ISO 15924 reserves the codes Qaaa-Qabx for private use values. These codes should be used for non-registered script values.
- Script subtags MUST NOT be registered using the process in **Section 3.1** of this document. Variant subtags may be considered for registration for that purpose.

The following rules apply to the region subtags:

- The region subtag must follow any language, extended language, or script subtags and must precede all other subtags.
- All 2-character subtags following the primary subtag denote the region or area to which this language variant relates, and are interpreted according to assignments found in ISO 3166 alpha-2 country codes from **[4]**, assignments subsequently made by the ISO 3166 maintenance agency, or governing standardization bodies.
- All 3-character codes consisting of digit (numeric) characters denote the region or area to which this language tag relates, and are interpreted according to the assignments found in **UN Standard Country or Area Codes for Statistical Use**[5] or assignments made thereto by the governing standards body.
- The ISO3166 code MUST be used to form the subtag except for countries with ambiguous ISO 3166 alpha-2 codes as defined in Rule 7a in **Section 2.3**.
- If the ISO 3166 alpha-2 code for a country is ambiguous as defined in Rule 7a, then the UN numeric code MUST be used to form the region subtag.
- UN numeric codes assigned to 'macro-geographical (continental)' or sub-regions not associated with an assigned ISO3166 alpha-2 code MAY be used to form the region subtag.
- UN numeric codes for 'economic groupings' or 'other groupings' MUST NOT be used to form language tags.
- Note: Generally there will be an informative IANA registration for valid UN numeric country codes. Continental codes will not be listed in the IANA registry, but may be used.
- Note: the alphanumeric codes in Appendix X of the UN document must not be used. (At the time this document was created these values match the ISO 3166 alpha-2 codes.)
- ISO 3166 reserves the country codes AA, QM-QZ, XA-XZ and ZZ as user-assigned codes. These codes should be used for private use region subtags.
- Region subtags MUST NOT be registered using the process in **Section 3.1** of this document. Variant

subtags may be considered for registration for this purpose.
- Region subtags must occur after any script subtags and before any variant subtags or extensions.
- Example: 'de-Latn-CH' represents German written using Latin script for Switzerland.
- Example: 'sr-Latn-891' represents Serbian written using Latin script for Serbia and Montenegro, whose ISO3166 alpha2 code is ambiguous.
- Example: 'es-419' represents Spanish as spoken in the UN-defined 'Latin America and Caribbean' region.

The following rules apply to the variant subtags:

- Variant subtags must follow all of the other defined subtags, but precede any private use extensions.
- Alphanumeric subtags of 5 to 15 characters may be registered with IANA, according to the rules in **Section 3.1** 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')
- Alphanumeric subtags of 5 to 15 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. For example the tag 'en-US-xzzyx' contains a single private use variant subtag.

The following rules apply to extensions:

- Extension subtags are separated from the other subtags defined in this document by a single-letter subtag ("singleton"). The singleton MUST either be one allocated to a registration authority via the mechanism described in **Section 3.3**. Private use subtags starting with the singleton subtag "x-" are described below.
- Each singleton subtag MUST appear at most one time in each tag. That is, singleton subtags MUST NOT be repeated.
- Extension subtags MUST meet whatever requirements are set by the RFC that defines their singleton prefix and whatever requirements are provided by the maintaining authority.
- Each extension subtag MUST be from two to thirty-one characters long and consist solely of letters or digits, with each subtag separated by a single "-".
- Each singleton MUST be followed by at least one subtag.
- Extensions MUST follow at least a valid primary language subtag. That is, the tag "a-myExtension" is

not valid.
- Extension subtags MUST follow all language, extended language, script, region and variant subtags in a tag.
- In the event that more than one extension appears in a single tag, the tag SHOULD be canonicalized as described in **Section 2.4.3**.
- When comparing or matching language tags, extensions MAY be ignored.
- For example, if the prefix "-r" and the shown subtags were defined, then the following tag might be a valid example: "en-Latn-GB-boont-r-anExtension-another-x-privatetag"

The following rules apply to private use subtags:

- Private Use subtags are separated from the other subtags defined in this document by the reserved single-character subtag "x".
- Private Use subtags MUST follow all language, extended language, script, region, variant, and extension subtags in the tag. Another way of saying this is that all subtags following the singleton 'x' MUST be considered private use. Example: The subtag 'US' in the tag 'en-x-US' is a private use subtag.
- Unlike Extensions, a tag MAY consist entirely of private use subtags.
- No source is defined for private use subtags. Use or standardization of the private use subtags is by private agreement and should not be considered part of this document.

For example: Users who wished to utilize SIL Ethonologue for identification might agree to exchange tags such as 'az-Arab-x-AZE-derbend'. This example contains two private use subtags. The first is "AZE" and the second is "derbend".

2.2.1 Pre-Existing RFC3066 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'

is now 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 'zh-min-nan' 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 appropriate subtags using the new format (but are not required to).

2.2.2 Possibilities for Registration

Possibilities for registration of subtags include:

- Languages not listed in ISO 639 that are not variants of any listed or registered language, can be registered. At the time this document was created there were no RFC 3066 examples of this form of subtag. 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, which are under consideration by the ISO 639 maintenance or registration authorities, or which have never been attempted for registration with those authorities. If ISO 639 has previously rejected a language for registration, it is reasonable to assume that there must be additional compelling evidence of need before it will be registered under this document.
- 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) may be registered as variants.

This document leaves the decision on what subtags are appropriate or not to the registration process described in **Section 3.1**.

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)
Aichholzgasse 6/12, AT-1120
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
c/o International Organization for Standardization
Case postale 56
CH-1211 Geneva 20 Switzerland
Phone: +41 22 749 72 33 Fax: +41 22 749 73 49
URL: http://www.iso.org/iso/en/prods-services/iso3166ma/index.html

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

The Statistics Division of the United Nations Secretariat maintains the Standard Country or Area Codes for Statistical Use and can be reached at:

Statistical Services Branch

Statistics Division
United Nations, Room DC2-1620
New York, NY 10017, USA

Fax: +1-212-963-0623
E-mail: statistics@un.org
URL: http://unstats.un.org/unsd/methods/m49/m49alpha.htm

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 when all users use the same tag in order to represent the same language. If an application has requirements that make the rules here inapplicable, then that application risks damaging interoperability. Users of this document are strongly discouraged against defining their own rules for language tag choice and matching.

Standards, protocols and applications that reference this document normatively but apply different rules to the ones given in this section MUST specify how the procedure varies from the one given here.

1. Use as precise a tag as possible, but no more specific than is justified. For example, 'de' might suffice for tagging an email written in German, while 'de-CH-1996' is probably unnecessarily precise for such a task.
2. Avoid using subtags that add no distinguishing information about the content. For example, the script subtag in 'en-Latn-US' is generally unnecessary, since nearly all English texts are written in the Latin script.
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**[8], 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.
    Beginning with these standards as of 1 January 2003 (or, in the case of ISO 15924, 1 May 2004), in the event that ISO 639, ISO 3166, or ISO 15924 assigns a code a new meaning or reassigns a deprecated code, the new use of the code is not permitted in language tags defined by this document.
    In the event that either ISO 639 or ISO 15924 assigns a new meaning to an existing code, 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 or registered language subtag as a surrogate value for the new code. The form of the registered language subtag or variant subtag will be at the discretion of the language subtag reviewer and must conform to other restrictions on language or variant subtags in this document.
    In the event ISO 3166 assigns a new meaning to an existing code, then the language subtag reviewer, as described in **Section 3**, shall prepare a proposal for entering the appropriate numeric UN country code as an informative entry in the IANA registry.
    The normal registration process described in **Section 3.1** of this document applies to the review and registration of the registered subtags 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-891 (Serbian for Serbia and Montenegro, using the UN country code)
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. Some examples, current at the time this document was drafted, are listed below. 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, some deprecated ISO 639 codes:

iw ~ he
in ~ id
ji ~ yi
sh

For example, some 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 is typically "narrower" or "more specific" than A. For example, 'zh-Hant-TW' is more specific than 'zh-Hant'.

This relationship is not guaranteed in all cases: specifically, languages that begin with the same sequence of subtags are NOT guaranteed to be mutually intelligible, although they may be. For example, the tag 'az' shares a prefix with both 'az-Latn' (Azerbaijani written using the Latin script) and 'az-Cyrl' (Azerbaijani written using the Cyrillic script). A person fluent in one script may not be able to read the other, even though the text might be identical. Content tagged as 'az' most probably is written in just one script and thus might not be intelligible to a reader familiar with the other script.

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 <span lang="FR">C'est la vie.</span> 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.

2.4.1 Language Range

A Language Range is 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**[14].

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

That is, a language-range has the same syntax as a language-tag, or is the single character "*" and implicitly assumes that there is a semantic relationship between tags that share the same subtag prefixes.

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.

As noted above, not all languages or content denoted by a specific language-range may be mutually intelligible and 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.

2.4.2 Matching Language Tags

Implementations that are searching for content or otherwise matching language tags to a language-range [**Section 2.4.1**] may choose to assume that there is a semantic relationship between two tags that share common prefixes. This is called 'language tag fallback'. The most common implementations follow this pattern:

1. When searching for content using language tag fallback, the language tag is progressively truncated from the end until a match is located. For example, starting with the tag 'en-US-boont', searchs or matches would first be performed with the whole tag, then with 'en-US', and finally with 'en'. This allows some flexibility in finding content in accordance with Rules 1 and 2 in **Section 2.3**; allows better maintenance; and usually provides better results when data is not available at a specific level of tag granularity or is sparsely populated (than if the default language for the system or content were used). Any implementation that uses this technique should ensure that appropriate data is available on each level.

```
Tag to match: en-US-boont
```

```
1. en-US-boont
2. en-US
3. en
```

Default Fallback Pattern Example

2. Private Use and Extension subtags are orthogonal to language tag fallback. By default, implementations SHOULD ignore both private use and extension subtags and follow the default fallback pattern (above). Thus a request to match the tag "en-US-boont-x-traditional" would produce exactly the same information content as the example above.
3. Implementations that choose to interpret one or more private use or extension subtags can choose a different fallback pattern or use the private use or extension subtags to interpret content in a different fashion.
4. Implementations that choose not to interpret one or more private use or extension subtags SHOULD NOT remove or modify these extensions in content that they are processing.

2.4.3 Canonicalization of Extensions in Language Tags

As noted above, by default extension and private use subtags are ignorable when matching.

Some applications of language tags may want or need to consider extensions and private use subtags when comparing tags. In order to produce tags that can be matched using simple string comparisons or operations, extension subtag sequences SHOULD be ordered into alphabetical (ASCII) order by singleton subtag.

For example, the tag "en-A-ext1-B-ext3-ext2" is correctly ordered, while "en-B-ext3-ext2-A-ext1" is not.

If the order of the extension subtags is not important to their meaning, then the RFC that defines the extension mechanism should define an alternate canonicalization mechanism for processes that generate these tags. For example, the subtags could be placed into alphabetical (ASCII) order: "en-a-aaa-bbb-ccc".

Private use subtags introduced with the singleton subtag "x-" SHOULD always be treated as if their order matters.

TOC

3. IANA Considerations

This section deals with the registration of subtags for use in language tags defined by this document, in accordance with the requirements of RFC2434[13].

3.1 Registration Procedure for Subtags

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 only be used to register primary language (registered-lang in the ABNF) or variant subtags, with the exception noted in Section 2.3, Rule 7.

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), or additional information about a registration previously made via this procedure.

Variant subtags MUST NOT be registered using the pattern 2 ALPHA * DIGIT to accommodate the provisions in Section 2.3, rule 7a of this document. That is, the subtag yx1234 can NOT be registered except under the aforementioned provisions.

Subtags MUST NOT be registered that start with the letter 'x', since this prefix is reserved for Private Use subtags.

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

```
LANGUAGE SUBTAG REGISTRATION FORM
Name of requester:
E-mail address of requester:
Subtag to be registered:
Type of Subtag:
  [ ] language
  [ ] variant
  [ ] region (informative; for use by language subtag reviewer only)
```

```
 Full English name of subtag:
 Intended meaning of the subtag:
 If variant subtag, the intended prefix(es) of subtag:
 If extended language 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. This information MUST be provided in the registration form.

Any registered subtag MAY be incorporated into a variety of language tags, according to the rules of **Section 2.1**. This makes 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 IESG, 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]**[9] 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.

3.2 Conversion of the RFC 3066 Language Tag Registry

Upon publication of this document as a BCP, the existing IANA language tag registry must be converted into the new subtag registry. This section defines the process for performing this conversion.

When this document is published, an email request will be sent to the list ietf-languages@iana.org requesting the conversion of the registry. The language subtag reviewer will then take all existing registered tags and evaluate them for conversion to the new format.

Tags that are currently obsolete will be maintained as grandfathered entries.

Tags that consist entirely of subtags that are valid under this document and which have the correct form and format for tags defined by this document are marked as obsoleted by this document. For example, zh-Hant is now defined by this document.

Tags that contain subtags which are consistent with registration under the guidelines in this document will have a new subtag registration created for each eligible subtag. If all of the subtags in the original tag are fully defined by the resulting registrations or by this document, then the original tag is marked as obsoleted by this document. For example, en-boont will result in a new subtag "-boont".

Tags that contain subtags that do not match the valid registration pattern and which are not otherwise defined by this document are marked as grandfathered.

The language subtag reviewer will have a reasonable period in which to evaluate the existing registry and complete any necessary registrations, notifying IANA@IANA.ORG and the ietf-languages mail lists when the task it complete or at a reasonable interval during the process.

Registrations that are in process under the rules defined in RFC 3066 MAY be completed under the former rules, at the discretion of the language tag reviewer. Any new registrations submitted after the request for conversion of the registry MUST be rejected.

All existing RFC 3066 language tag registrations will be maintained in perpetuity.

3.3 Extensions and Extensions Namespace

Extension subtags are those introduced by single-letter subtags other than 'x-'. They are reserved for the generation of identifiers which contain a language component, and are compatible with applications that process language tags according to this specification. For example, they might be used to define locale identifiers, which are generally based on language.

The structure and form of extensions are defined by this document so that implementations can be created that are forward compatible with applications that may be created using single-letter subtags in the future. In addition, defining a mechanism for maintaining single-letter subtags will lend to the stability of this document by reducing the likely need for future revisions or updates.

IANA will maintain a registry of allocated single-letter subtags. This registry contain the following information: letter identifier; name; purpose; RFC defining the subtag namespace and its use; and the name, URL, and email address of the maintaining authority.

Allocation of a single-letter subtag shall take the form of an RFC defining the name, purpose, processes, and procedures for maintaining the subtags. The maintaining or registering authority, including name, contact email, discussion list email, and URL location of the registry must be indicated clearly in the RFC. The RFC MUST reference this section of this document. The RFC MUST specify each of the following:

- The specification of valid subtags MUST be available over the Internet and at no cost.
- The specification MUST be in the public domain or available via a royalty-free license acceptable to

the IETF and specified in the RFC.

- The specification MUST be versioned and each version of the specification MUST be numbered, dated, and stable.
- The specification MUST be stable. That is, extension subtags, once defined by a specificaiton, MUST NOT be retracted or change in meaning in any substantial way.
- IANA MUST be informed of changes to the contact information and URL for the specification.

The determination of whether an Internet-Draft meets the above conditions and the decision to grant or withhold such authority rests solely with the IESG, and is subject to the normal review and appeals process associated with the RFC process.

TOC

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.

TOC

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 sub-optimal output when a mixed Japanese- Chinese text is encountered)

TOC

6. Changes from RFC3066

The main goals were to maintain backward compatibility (so that all previous codes would remain valid); reduce the need for large numbers of registrations; to provide a more formal structure to allow parsing into subtags even where software does not have the latest registrations; to provide stability in the face of potential instability in ISO 639, 3166, and 15924 codes (*demonstrated instability* in the case of ISO 3166); and to allow for external extension mechanisms.

- Allows ISO15924 script code subtags and allows them to be used generatively.
- Adds the concept of a variant subtag and allows variants to be used generatively.
- Adds an extension mechanism including both private use and standardized extensions.
- Defines the private use tags in ISO639, ISO15924, and ISO3166 as the mechanism for creating private use language, script, and region subtags respectively
- Defines a syntax for private use variant subtags which can be used without registration.
- Defines an extended language subtag for use with certain anticipated features of ISO 639-3.
- Defines a process for handling reuse of values by ISO639, ISO15924, and ISO3166 in the event that they register a previously used value for a new purpose.
- Changes the IANA language tag registry to a language subtag registry
- Makes the choice of UN or ISO3166 codes explicit in Section 2.2 and modified the text of Rule 7a. It also sets a start date for ambiguity resolution.
- Expands the maximum size of some of the subtags, including the primary subtag. In RFC 3066 the maximum subtag length was eight characters long. In this document the maximum length for most subtags is fifteen characters (and is thirty-one characters for extensions and private use subtags).

Substantive changes between draft-02 and this version are:

- Added references to RFC's obsoleted by this document in abstract and introduction. The abstract was slightly rewritten.
- Clarified that single-letter subtags can also be reserved by update to this document.
- Modified the text in section 2.3 to clarify that the text means that users should use the same tag for the same language, not that everyone should just use (i.e. speak) the same language.
- Modified the text in section 2.3 to clarify that applications which normatively reference this document are strongly discouraged from defining their own tag choice and matching rules. The text was written in such a way as to be consistent with this document's eventual status as a BCP.
- Per IESG request, modified section 3.1 to change the appointment of the language subtag reviewer from the applications area director to the IESG itself.
- Added section 3.2, which defines how the existing registry is converted.
- Tightened up the ABNF to be more succinct. Several rules were removed defining things like 'alphanumdash'. Other rules were reworked to make them more consistent.
- Removed the definition of i- as the prefix for registered language subtags.
- Defined singleton subtag handling in the ABNF and the body of the document, allowing parsers to take future standardization into account. This includes significant new text related to the creation and maintenance of extensions. This syntax should be stable for a long time.
- Defined extended language subtags as being reserved for use with ISO 639 (and presumably ISO639-3) tags only. Removed registration mechanism for extended language subtags.
- Defined a canonicalization mechanism for extension blocks and a recommended canonicalization for extension subtags in **Section 2.4.3**.
- Added the note about the length-of-subtag expansion, which might be incompatible with some implementations.

TOC

References

[1] International Organization for Standardization, "ISO 639-2:1998 - Codes for the representation of names of languages -- Part 2: Alpha-3 code - edition 1", August 1988.

[2] ISO TC46/WG3 and M. Everson, Ed., "ISO 15924:2003 (E/F) - Codes for the representation of names of scripts", March 2003.

[3] International Organization for Standardization, "Code for the representation of names of languages, 1st edition", ISO Standard 639, 1988.

[4] International Organization for Standardization, "Codes for the representation of names of countries, 3rd edition", ISO Standard 3166, August 1988.

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

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

[7] **Borenstein, N.** and **N. Freed**, "**MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies**", RFC 1521, September 1993.

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

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

[10] **Bradner, S.**, "**Key words for use in RFCs to Indicate Requirement Levels**", BCP 14, RFC 2119, March 1997 (**HTML**, **XML**).

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

[12] **Berners-Lee, T.**, **Fielding, R.** and **L. Masinter**, "**Uniform Resource Identifiers (URI): Generic Syntax**", RFC 2396, August 1998 (**HTML**, **XML**).

[13] **Narten, T.** and **H. Alvestrand**, "**Guidelines for Writing an IANA Considerations Section in RFCs**", BCP 26, RFC 2434, October 1998 (**HTML**, **XML**).

[14] **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 (**HTML**, **XML**).

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

[16] Alvestrand, H., "**Tags for the Identification of Languages**", BCP 47, RFC 3066, January 2001.

[17] Yergeau, F., "**UTF-8, a transformation format of ISO 10646**", STD 63, RFC 3629, November 2003.

TOC

Authors' Addresses

Addison Phillips (editor)
webMethods, Inc.
432 Lakeside Drive
Sunnyvale, CA 94088
US
EMail: **aphillips@webmethods.com**

Mark Davis

IBM
EMail: mark.davis@us.ibm.com

TOC

Appendix A. 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, Jeremy Carroll, John Clews, Jim Conklin, Peter Constable, John Cowan, Mark Crispin, Dave Crocker, Martin Duerst, Michael Everson, Doug Ewell, Ned Freed, Tim Goodwin, Dirk-Willem van Gulik, Marion Gunn, Paul Hoffman, Richard Ishida, Olle Jarnefors, Kent Karlsson, John Klensin, Alain LaBonte, Eric Mader, Keith Moore, Chris Newman, Masataka Ohta, George Rhoten, Markus Scherer, Keld Jorn Simonsen, Thierry Sourbier, 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.

TOC

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-891 (Serbian, Latin script, Serbia and Montenegro)

Language-Script-Region-Variant:

en-Latn-US-boont (Boontling dialect of English)

Language-Region:

de-DE (German for Germany)
zh-SG (Chinese for Singapore)
cs-CS (Czech for Czechoslovakia)
sr-891 (Serbian for Serbia and Montenegro, UN country code, see 7a in **Section 2.3**

Other Mixtures:

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

Private Use mechanism:

de-CH-x-phonebook
az-Arab-x-AZE-derbend

Extended language subtags (examples ONLY: extended languages must be defined by revision or update to this document):

zh-min
zh-min-nan-Hant-CN

Private Use tags:

qaa-Qaaa-QM-xsouthern (all private tags)
de-Qaaa (German, with a private script)
de-Latn-QM (German, Latin-script, private region)
de-Qaaa-DE (German, private script, for Germany)

Tags that use extensions (examples ONLY: extensions must be defined by revision or update to this document or by RFC):

- en-US-u-gregorianCalendar
- zh-CN-x-privateUse
- en-a-anExtension-b-another

Some Invalid Tags:

de-891-DE (two region tags)
a-DE (use of a single character tag in primary position)
zh-xsouthern-DE (private use variant followed by another tag)
ar-a-aaa-b-bbb-a-ccc (two extensions with same single letter prefix)

TOC

Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on the IETF's procedures with respect to rights in standards-track and standards-related documentation can be found in BCP-11. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementors or users of this specification can be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

Full Copyright Statement

Acknowledgment

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