# Tags for the identification of languages

Fri Aug 5 14:21:43 MET DST 1994

**Harald Tveit Alvestrand**
**UNINETT**
**Harald.T.Alvestrand@uninett.no**

Abstract

This document describes a language tag for use in cases where it is desired to indicate the language used in an information object.

It also defines a Content-language: header, for use in the case where one desires to indicate the language of something that has RFC-822-like headers, like MIME body parts or Web documents, and a new parameter to the Multipart/Alternative type, to aid in the usage of the Content-Language: header.

Status of this Memo

This draft document is being circulated for comment.

If consensus is reached it may be submitted to the RFC editor as a Proposed Standard protocol specificiation.

Please send comments to the author, or to the MAILEXT mailing list <mailext@cs.wisc.edu>

The following text is required by the Internet-draft rules:

This document is an Internet Draft. 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. Internet Drafts may be updated, replaced, or obsoleted by other documents at any time. It is not appropriate to use Internet Drafts as reference material or to cite them other than as a "working draft" or "work in progress."

Please check the I-D abstract listing contained in each Internet Draft directory to learn the current status of this or any other Internet Draft.

The filename of this document is draft-mailext-lang-tag-01.txt

## 1. Introduction

There are a number of languages spoken by human beings in this world.

A great number of these people would prefer to have information presented in a language that they understand.

In some contexts, it is possible to have information in more than one language, or it might be possible to provide tools for assisting in the understanding of a language (like dicationaries).

A prerequisite for any such function is a means of labelling the information content with an identifier for the language in which is is written.

In the tradition of solving only problems that we think we understand, this document specifies an identifier mechanism, and one possible use for it.

## 2. The Language tag

The language tag is composed of 1 or more parts: A main language tag and a (possibly empty) series of subtags.

The syntax of this tag in RFC-822 EBNF is:

```
Language-Tag =3D Tag-List
Tag-List =3D Tag-Component *[ '-', Tag-List ]
Tag-Component =3D 1*8ALPHA
```

Whitespace is not allowed within the tag.

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.

The namespace of language tags and subtags is administered by the IANA. The following registrations are predefined:

In the language tag:

- All 2-letter codes are interpreted according to ISO 639.

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

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

- Other values cannot be assigned except by updating 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 requirements.

In the first subtag:

- All 2-letter codes are interpreted as ISO 3166 country codes, according to the rules laid down in ISO 639.

- Codes of 3 to 8 letters may be registered with the IANA by anyone who feels a need for it. IANA has the right to reject registrations that are felt to be misleading.

The information in the subtag may for instance be:

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

- Dialect or variant information, such as no-NYNORSK or en-COCKNEY

- 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-cherokee

- Script variations, such as az-arabic and az-cyrillic

In the second and subsequent subtag, any value can be registered.

NOTE: The ISO 639/ISO 3166 convention is that language names are written in lower case, while country codes are written in upper case. This convention is recommended, but not enforced; the tags are case insensitive.

NOTE: ISO 639 defines a registration authority for additions to and changes in the list of languages in ISO 639. This authority 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

The following codes have been added in 1989 (nothing later): ug (Uigur), iu (Inuktitut, also called Eskimo), za (Zhuang), he (Hebrew, replacing iw), yi (Yiddish, replacing ji), and id (Indonesian, replacing in).

NOTE: The registration agency for ISO 3166 (country codes) is:

ISO 3166 Maintenance Agency Secretariat
c/o DIN Deutches Institut F=FCr Normung
Burggrafenstrasse 6
Postfach 1107
D-10787 Berlin
Germany
Phone: +49 30 26 01 320
Fax: +49 30 26 01 231

The codes AA, QM-QZ, XA-XZ and ZZ are reserved by ISO 3166 as user-assigned codes.

## 2.1. Meaning of the language tag

The language tag always defines a language as spoken (or written) by human beings for communication of information to other human beings. Computer languages are explicitly excluded.

There is no guaranteed relationship between languages that start

out with the same series of tags; especially, they are NOT gurauranteed to be mutually comprehensible, although this will sometimes be the case.

Applications should always treat language tags as a single token; the division into subtags is an administrative mechanism, not a navigation aid.

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

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

- 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 in life is providing alternatives, it should be regarded as a hint that the material inside 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, multiple languages need not mean that one needs to be multilingual to get complete understanding of the document. Example: MIME multipart/alternative.

- It would be possible to define (for instance) an SGML DTD that defines a <LANG xx> tag for indicating that following or contained text is written in this language, such that one could write "<LANG FR>C'est la vie</LANG>"; the Norwegian-speaking user could then access a French-Norwegian dictionary to find out what the quote meant.

## 3. The Content-language header

The RFC-822 ABNF of the Language header is:

Language-Header =3D "Content-Language" ":" 1#Language-tag

Note that the Language-Header is allowed to list several languages in a comma-separated list.

Whitespace is allowed, which means also that one can place parenthesized comments anywhere in the language sequence.

### 3.1. Examples of Content-language values

NOTE: NONE of the subtags shown in this document have actually been assigned; they are used for illustration purposes only.

Norwegian official document, with parallel text in both official versions of Norwegian. (Both versions are readable by all Norwegians).

```
  Content-Type: multipart/alternative; differences=3Dcontent-
language
  Content-Language: no-nynorsk, no-bokmaal
```

Voice recording from the London docks

```
  Content-type: audio/basic
  Content-Language: en-cockney
```

Document in Sami, which does not have an ISO 639 code, and is spoken in several countries, but with about half the speakers in Norway, with six different, mutually incomprehensible dialects:

```
  Content-type: text/plain; charset=3Diso-8859-10
  Content-Language: i-sami-no (North Sami)
```

An English-French dictionary

```
  Content-type: application/dictionary
  Content-Language: en, fr (This is a dictionary)
```

An official EC document (in a few of its official languages)

```
  Content-type: multipart/alternative
  Content-Language: en, fr, de, da, el, it
```

An excerpt from Star Trek

```
  Content-type: video/mpeg
  Content-Language: x-klingon
```

## 4. Use of Content-Language with Multipart/Alternative

When using the Multipart/Alternative body part of MIME, it is possible to have the body parts giving the same information content in different languages. In this case, one should put a Content-Language header on each of the body parts, and a summary Content-Language header onto the Multipart/Alternative itself.

### 4.1. The differences parameter to multipart/alternative

As defined in RFC 1541, Multipart/Alternative only has one parameter: boundary.

The common usage of Multipart/Alternative is to have more than one format of the same message (f.ex. PostScript and ASCII).

The use of language tags to differentiate between different alternatives will certainly not lead all MIME UAs to present the most sensible body part as default.

Therefore, a new parameter is defined, to allow the configuration of MIME readers to handle language differences in a sensible manner.

```
Name: Differences
Value: One or more of
     Content-Type
     Content-Language
```

Further values can be registered with IANA; it must be the name of a header for which a definition exists in a published document. If not present, Differences=3DContent-Type is assumed.

The intent is that the MIME reader can look at these headers of the message component to do an intelligent choice of what to present to the user, based on knowledge about the user preferences and capabilities.

(The intent of having registration with IANA of the fields used in

this context is to maintain a list of usages that a mail UA may expect to see, not to reject usages)

(NOTE: The MIME specification [RFC 1521], section 7.2, states that headers not beginning with "Content-" are generally to be ignored in body parts. People defining a header for use with "differences=3D" should take note of this)

The mechanism for deciding which body part to present is outside the scope of this document.

MIME EXAMPLE:

```
Content-Type: multipart/alternative; differences=3DContent-Language;
          boundary=3D"limit"
Content-Language: en, fr, de

--limit
Content-Language: fr

Le renard brun et agile saute par dessus le chien paresseux
--limit
Content-Language: de
Content-Type: text/plain; charset=3Diso-8859-1
Content-Transfer-encoding: 8bit

Der schnelle braune Fuchs h=FCpft =FCber den faulen Hund
--limit
Content-Language: en

The quick brown fox jumps over the lazy dog
--limit--
```

When composing a message, the choice of sequence may be somewhat arbitary. However, non-MIME mail readers will show the first body part first, meaning that this should most likely be the language understood by most of the recipients.

## 5. IANA registration procedure for language tags

Any language tag must start with an existing tag, and extend it.

This registration form should be used by anyone who wants to use a language tag not defined by ISO or IANA.

---

LANGUAGE TAG REGISTRATION FORM

Name of requester :
E-mail address of requester:
Tag to be registered :

English name of language :

Native name of language (in ASCII):

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

---

The language form must be sent to language-review@uninett.no for a 2-week review period before submitting it to IANA. (This is an open list. Requests to be added should be sent to language-review-request@uninett.no. General language discussions are not appropriate for this list)

The completed form should then be sent to IANA@ISI.EDU; all registered forms are available online in the directory ftp://ftp.iana.isi.edu/registrations/languages/

(NOTE: The IANA may suggest alternative text here).

The IANA is free to reject registrations where it feels, based on list feedback, that information is lacking, or that the tag name suggests something different from the language referenced.

## 6. Security considerations

Security considerations are not considered in this memo

## 7. Character set considerations

Codes are always expressed using US-ASCII (a-z).

The issue of deciding upon the rendering of a character set based on the language encoding is not addressed in this memo; however,

the author cautions against thinking that such a decision can be made 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 will fail to work when a mixed Japanese-Chinese text is encountered)

## 8. Gatewaying considerations

RFC 1327 defines a Language: header. This header is not recommended now, because it is defined to be a single 2-letter language code, and the X.400 header it is supposed to gateway is a list of language codes.

It is suggested that RFC 1327 be updated to produce the Content-Language: header, and to turn this header into the ISO/CCITT specified Language components rather than the RFC-822-headers heading extension.

## 9. References

[ISO 639]
ISO 639:1988 (E/F) - Code for the representation of names of languages - The International Organization for Standardization, 1st edition, 1988 17 pages Prepared by ISO/TC 37 - Terminology (principles and coordination)

[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 1521]
MIME Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies - Borenstein and Freed - September 1993

[RFC 1327]
Mapping between X.400(1988) / ISO 10021 and RFC 822 - Kille -

May 1992

10. Change log Changes from draft-ietf-lang-tag-02.txt:

Clarified that a language tag is a single token

Changes from draft-alvestrand-language-tag-00:

IANA registration form added

IANA-reserved tag changed from "IANA" to "I", in order to avoid clashing with possible ISO 4-letter codes

Separated "tag" definition from "header" definition

Info on ISO 639 registration office added

Created a multi-level tag, rather than strict two-level

Added examples of SGML usage

Lots of small nits fixed

=0C