Tags for the identification of languages

Thu Mar 18 14:03:38 MET 1993

Harald Tveit Alvestrand
UNINETT
Harald.Alvestrand@uninett.no

Abstract

This document describes a Content-Language: header for use with body parts of MIME.

It also describes 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 IETF-822 mailing list <ietf-822@dimacs.rutgers.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-alvestrand-language-tag-00.txt

1. The Language tag

The language tag is composed of 2 parts: A language tag and a subtag.

The syntax of this header is:

```
Language-Header ::= 'Content-language:' Language [',' Language]...
Language ::= ALPHA*8 [ '-' ALPHA*8 ]
```

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.
- All 3-letter codes are reserved for the (hopefully) forthcoming revision to ISO 639
- The value "IANA" is reserved for IANA-defined subregistrations
- The value "X" is reserved for private use. Subtags of "X" will not be registered by the IANA.
- No other registration is allowed.

In the sublanguage tag:

- 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 sublanguage tag may for instance be:

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

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

- Languages not listed in ISO 639, which can be registered with the IANA prefix, such as IANA-CHEROKEE

If multiple languages are used in the MIME body part, they are listed with commas between them.

2. MEANING

The meaning of the header is:

- 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. Examples: Simple text.

- For an aggregation of information object, 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. Examples: MIME multipart/alternative.

  EXAMPLES:

  NOTE: NONE of the sublanguage codes 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-language: no-nynorsk, no-bokmaal

Voice recording from the London docks

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

Content-language: iana-sami

An English-French dictionary

Content-language: en, fr (This is a dictionary)

An official EC document

Content-language: en, fr, ge, da, gr, it

An excerpt from Star Trek dialogue

Content-language: x-klingon

3. Usage examples

Examples of protocol usage of this header are:

- WWW selection of an appropriate version of information for display, based on a profile for the user listing languages that are understood

- MIME usage of alternate body parts in E-mail

4. The difference 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: Difference
Value: One 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, Difference=Content-type is assumed.

The intent is that the MIME reader can look at this header of the message component to do an intelligent choice of what to present to the user.

(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)

MIME EXAMPLE:

```
Content-type: multipart/alternative; difference=content-language;
          boundary="limit"
Content-language: en, fr

--limit
Content-language: fr

--limit
Content-language: en

--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. Security considerations

Security considerations are not considered in this memo

6. Character set considerations

Codes are always US-ASCII. 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 (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)

7. 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.

8. References

ISO 639

ISO 3166

RFC 1521

RFC 1327