

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 15, 2005

B. Lilly
March 2005

Indicating and Negotiating Text Script
draft-lilly-content-script-01

Status of this Memo

By submitting this Internet-Draft, the author represents that any applicable patent or other IPR claims of which he is aware have been or will be disclosed, and any of which he becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

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/lid-abstracts.txt>.

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

Copyright Notice

Copyright © The Internet Society (2005).

Abstract

Some written text in some languages can be represented in multiple scripts, or writing forms. This memo proposes mechanisms for identification and negotiation of script for written text.

Table of Contents

1. Introduction.....	3
1.1. Script, Language, Charset, and Content.....	3
1.1.1. Script is Distinct from Language.....	3
1.1.2. Script is Related to Charset.....	3
1.1.3. Documents in a Single Language May Have Multiple Scripts.....	3
1.1.4. Documents in Multiple Languages May Use a Single Script.....	3
2. Requirement Levels.....	3
3. ABNF References.....	3
4. Header Fields.....	4
4.1. Indicating Script; the Content-Script Header Field.....	4
4.1.1. Semantics.....	4
4.1.2. ABNF.....	4
4.1.3. Usage.....	4
4.1.4. Header Field Registration Template.....	5
4.2. Script Negotiation; the Accept-Script Header Field.....	6
4.2.1. Semantics.....	6
4.2.2. ABNF.....	6
4.2.3. Semantic Details.....	6
4.2.4. Usage.....	6
4.2.5. Header Field Registration Templates.....	7
5. Media Feature Tag.....	8
5.1. Media Feature Tag Registration Template.....	9
6. Acknowledgments.....	10
7. Security Considerations.....	10
8. Internationalization Considerations.....	10
9. IANA Considerations.....	11
Appendix A. Examples.....	11
A.1. Script Indication.....	11
A.1.1. Simple Example.....	11
A.1.2. Multiple Alternatives.....	11
A.2. Script Negotiation.....	11
Appendix B. Change History.....	11
Normative References.....	12
Informative References.....	12
Author's Address.....	13

1. Introduction

Some written text in some languages can be represented in multiple scripts, or writing forms. This memo proposes mechanisms for identification and negotiation of script.

1.1. Script, Language, Charset, and Content

1.1.1. Script is Distinct from Language

Language is a characteristic of many forms of human communication. For example, it applies to oral communication and to writing. Script, however, applies only to a subset of communication forms. Therefore, for purposes such as content negotiation, it is desirable to indicate script separately from language.

1.1.2. Script is Related to Charset

Some charsets [I1.RFC2978] apply only to a single script. For example, ANSI X3.4 applies only to Latin script, and KOI8 applies only to Cyrillic script. In other cases, such as ISO 10646, script can be inferred from the range of character codes used, provided one has access to the content and is willing to analyze it.

1.1.3. Documents in a Single Language May Have Multiple Scripts

It is desirable to specify script separately from language, as multiple scripts may be associated with a single language in a single document or piece of text. It is not uncommon for text in Japanese, for example, to contain a mix of Katakana and Hiragana, and some text also contains Latin script for some words of foreign origin.

1.1.4. Documents in Multiple Languages May Use a Single Script

It is desirable to specify script separately from language, as a text document written in a single script might contain multiple languages.

2. Requirement Levels

The key words "MUST", "MUST NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", and "MAY" in this document are to be interpreted as described in [N1.BCP14].

3. ABNF References

ABNF in this document uses grammar productions defined in [N2.RFC2234] and [N3.RFC2822].

4. Header Fields

4.1. Indicating Script; the Content-Script Header Field

4.1.1. Semantics

The Content-Script field indicates the script or scripts used in a piece of content, and (in the case of composite media including entire MIME messages) any enclosed media content.

4.1.2. ABNF

```
content-script = "Content-Script:" [CFWS] script-list [CFWS] CRLF
script-list    = script *([CFWS] "," [CFWS] script)
script        = 4ALPHA ; script tag per ISO 15924:2004; script tags
                ; are case-insensitive protocol elements
```

Note that there is no provision for linear whitespace or line-folding between the field name tag (a case-insensitive protocol element [N3.RFC2822], [I3.RFC1958]) and the colon separating the field name from the field body. Generators **MUST NOT** insert linear whitespace or line folding between the field name and the colon.

4.1.3. Usage

4.1.3.1. When

A Content-Script field **SHOULD** be used to indicate script(s) for non-trivial sequences of characters in human-readable text [I4.BCP18] where script is not unique to the language in use.

It **MAY** be omitted for short texts where script may be determined from the charset and character codes used, or where only a single script is used for the language(s) applicable to the text.

It **MAY** be used for image data representing text, such as facsimile image data.

It **MUST NOT** be used where no script is applicable, such as in audio data of spoken language, or image or video media where no script is applicable to the content.

It **SHOULD NOT** be used where visible text is merely incidental to the content, as may be the case with some content using image, video, or model media types.

4.1.3.2. Where

The Content-Script field **MAY** be used in the message header [N3.RFC2822] of a MIME message [I5.RFC2045], in a MIME-part header [I6.RFC2046], or in the header of a protocol which uses MIME header

fields to indicate content characteristics such as [I7.RFC1945] and [I8.RFC2616].

The field MAY be used in the MIME-part header of a composite media type [I6.RFC2046], if and only if it is equally applicable to each part of the composite media type. When used with composite media types, each component piece of content acquires the semantics associated with the Content-Script field(s) in the enclosing composite media type MIME-part headers, plus those of any Content-Script fields in the MIME message header, plus those of any Content-Script fields in that individual component media type's MIME-part header. There is no mechanism to remove the semantics associated with an enclosing composite media type, therefore a script code MUST NOT be specified in a Content-Script field in a composite media type MIME-part header if the concept of script is not applicable to some enclosed media type or if some enclosed media type does not use that script.

Its use is RECOMMENDED with media type message/external-body as it may help to reduce wasted resources that might otherwise be expended on retrieval of unintelligible content.

4.1.3.3. Who

The Content-Script field MAY be set by a message or content author or a user agent acting on the author's behalf.

It MUST NOT be inserted, modified (except for non-protocol elements), or deleted by submission, transport, or delivery agents [I9.Crocker05].

It SHOULD, when present, be used by recipient user agents to assist in presentation of human-readable content (presentation includes display as well as text-to-speech conversion and similar technologies).

4.1.3.4. How Many

It is RECOMMENDED that a single Content-Script field be used in the header associated with a piece of content.

Multiple Content-Script fields MAY be used, and if present in a single piece of content MUST be interpreted identically to a single field listing all scripts listed in all Content-Script fields applicable to the content.

4.1.4. Header Field Registration Template

[I10.BCP90] requires a registration template. The template is provided in this section.

Header field name: Content-Script

Applicable protocol: mime

Status: standards track

Author/Change controller: IESG

Specification document(s): This document (when approved and an RFC number assigned)

Related information: none

4.2. Script Negotiation; the Accept-Script Header Field

4.2.1. Semantics

The Accept-Script field indicates a set of preferences related to script. See below for details of interpretations of preference values.

4.2.2. ABNF

Note that there is no provision for linear whitespace or line-folding between the field name tag (a case-insensitive protocol element [N3.RFC2822], [I3.RFC1958]) and the colon separating the field name from the field body. Generators **MUST NOT** insert linear whitespace or line folding between the field name and the colon.

4.2.3. Semantic Details

Each script may have an associated preference value, indicated as a decimal floating-point number with at most three decimal places. An asterisk matches any script not explicitly listed. The default preference value associated with a script or asterisk is 1. Scripts with larger preference values are preferable to scripts with lower preference values. A script **SHOULD NOT** be named more than once in an Accept-Script field; if it is, however, the preference value associated with the script is the last one presented with that script in left-to-right order in the field body. If an Accept-Script field is presented, any scripts not explicitly named have an implicit preference value associated with an asterisk if one is presented in the field; if there is no asterisk, the preference value for unnamed scripts is implicitly zero. If no Accept-Script field is presented, all scripts are to be presumed to be equally preferred.

4.2.4. Usage

4.2.4.1. When

An Accept-Script field **MAY** be used to indicate script preferences where a suitable negotiation method, such as [I11.RFC2295] is available, and the requester has a preference, and script is potentially relevant to one or more media types under consideration. It **SHOULD NOT** be used if any of those conditions is not met.

4.2.4.2. Where

Usage of an Accept-Script field is dictated by the negotiation protocol and is outside of the scope of this document.

4.2.4.3. Who

The Accept-Script field MAY be set by a message or content requester or a user agent acting on the requester's behalf.

It MUST NOT be inserted, modified (except for non-protocol elements), or deleted by transport protocols.

It SHOULD, when present, be used by content-serving protocols to supply preferred content to requesters when content in multiple scripts otherwise meeting requests is available. This memo does not address how content-serving protocols should balance preferences for multiple characteristics of requested content; that is left to content-serving protocol specifications and/or implementations.

4.2.4.4. How Many

At most one Accept-Script field may be presented.

4.2.5. Header Field Registration Templates

[I10.BCP90] requires separate templates for different "protocols". Since the Accept-Script field is not a MIME field, and may be used by a number of protocols which support content negotiation, templates are provided in this section for such protocols using header fields known at the time of writing.

4.2.5.1. HTTP Header Field Registration Templates

There are two Hyper text transfer protocols (HTTP): [I7.RFC1945], [I8.RFC2616]. The registration templates for those protocols are provided in this section.

4.2.5.1.1. HTTP/1.0 template

Header field name: Accept-Script

Applicable protocol: [I7.RFC1945]

Status: informational

Author/Change controller: IESG

Specification document(s): This document (when approved and an RFC number assigned)

Related information: none

4.2.5.1.2. HTTP/1.1 template

Header field name: Accept-Script

Applicable protocol: http

Status: standards track

Author/Change controller: IESG

Specification document(s): This document (when approved and an RFC number assigned)

Related information: none

4.2.5.2. RFC 2295 protocol template

Header field name: Accept-Script

Applicable protocol: RFC 2295 [I11.RFC2295]

Status: experimental

Author/Change controller: IESG

Specification document(s): This document (when approved and an RFC number assigned)

Related information: none

4.2.5.3. HTCPCP template

Header field name: Accept-Script

Applicable protocol: RFC 2324 [I12.RFC2324]

Status: informational

Author/Change controller: IESG

Specification document(s): This document (when approved and an RFC number assigned)

Related information: none

5. Media Feature Tag

[I13.BCP31] provides a registration template for registration of media feature tags. Media feature tags may be used for content negotiation such as in Content-Alternative, Content-Features, and Media-Accept-Features fields [I14.RFC2912], [I15.RFC3297], [I16.RFC2533], [I17.RFC2738]. The media feature tag registration appears below.

5.1. Media Feature Tag Registration Template

Media feature tag name: script

Summary of the media feature indicated by this feature tag:
Indication of script(s) used in a text document using ISO
standard script name tags

Values appropriate for use with this feature tag:

- [] 1. The feature tag is Boolean and may have values of TRUE or FALSE. A value of TRUE indicates an available capability. A value of FALSE indicates the capability is not available.
- [X] 2. The feature has an associated numeric or enumerated value.
 - [] 2a. Signed Integer
 - [] 2b. Rational number
 - [] 2c. Token (equality relationship)
 - [] 2d. Token (ordered)
 - [] 2e. String (equality relationship)
 - [X] 2f. String (defined comparison) Comparison is as case-insensitive strings. Strings are compared for equality only (no ordering). The special value "*" matches any script.

The feature tag is intended primarily for use in the following applications, protocols, services, or negotiation mechanisms:
MIME

Examples of typical use: script=Latn

Related standards or documents: [N4.ISO15924], [I2.15924Lists]

Considerations particular to use in individual applications, protocols, services, or negotiation mechanisms: none

Interoperability considerations: Applications developed prior to registration of this tag cannot be expected to recognize the tag. Such applications will be unable to participate in script content negotiation.

Security considerations:

Privacy concerns, related to exposure of personal information:
While script may identify an author as belonging to an ethnic group, and that information might be abused, script information can be determined from content. Negotiation of script may reveal a preference for script, and that information also has potential for abuse.

Denial of service concerns related to consequences of specifying incorrect values: none known.

Other: none known.

Additional information: none

Keywords: none

Related feature tags: charset, language

Related media types or data formats: all subtypes of the text media type.

Related markup tags: none known

Name(s) & email address(es) of person(s) to contact for further information:

Bruce Lilly
blilly@erols.com

Intended usage: COMMON

Author/Change controller: IESG

Requested IANA publication delay: none

Other information: none

6. Acknowledgments

The author gratefully acknowledges discussions on this topic which took place in December 2004 and January 2005 on the IETF discussion mailing list.

7. Security Considerations

While script may identify an author as belonging to an ethnic group, and that information might be abused, script information can be determined from content as noted in section 1.1.2. Negotiation of script may reveal a preference for script, and that information also has potential for abuse.

8. Internationalization Considerations

This memo raises no new internationalization considerations.

9. IANA Considerations

IANA shall register the header field names defined in this document (on approval by the IESG) in the permanent header field registry.

IANA shall register the media feature tag defined in this document (on approval by the IESG) in the IETF tree of the media feature tag registry.

Appendix A. Examples

A.1. Script Indication

A.1.1. Simple Example

```
MIME-Version: 1.0
Content-Type: text/plain ; charset=iso-2022-jp-2
Content-Language: ja
Content-Script: Hira, Kana
```

```
<Japanese language text in a mix of Katakana and Hiragana,
ISO 2022-JP-2 charset goes here>
```

A.1.2. Multiple Alternatives

```
MIME-Version: 1.0
Content-Type: multipart/alternative ; boundary=next
Content-Language: ja
```

```
--next
```

```
Content-Type: text/plain ; charset=iso-2022-jp-2
Content-Script: Kana
```

```
<Japanese language text in Katakana, ISO 2022-JP-2 charset goes here>
--next
```

```
Content-Type: text/plain ; charset=iso-2022-jp-2
Content-Script: Hira
```

```
<Japanese language text in Hiragana, ISO 2022-JP-2 charset goes here>
--next--
```

A.2. Script Negotiation

```
Accept-Script: Latn ; q = (foo) 1, Cyrl ; q = 0.5, * ; q = 0.001
```

The example expresses a strong preference for Latin script, followed in preference by Cyrillic script, but accepting any script with a low but non-zero preference value.

Appendix B. Change History

[[This change history will not be part of a published RFC]]

-00 to -01

- added this change history
- fixed ABNF bug in script production
- reformatted ABNF
- added media feature tag description and registration template; revised title accordingly

Normative References

- [N1.BCP14] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [N2.RFC2234] Crocker, D. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, November 1997.
- [N3.RFC2822] Resnick, P., "Internet Message Format", RFC 2822, April 2001.
- [N4.ISO15924] International Organization for Standardization (ISO), "ISO 15924:2004 -- Codes for the representation of names of scripts", March 2003.

Informative References

- [I1.RFC2978] Freed, N. and J. Postel, "IANA Charset Registration Procedures", BCP 19, RFC 2978, October 2000.
- [I2.15924Lists] ISO has designated The Unicode Consortium as the ISO 15924 Registration Authority. Lists of ISO 15924 codes may be obtained free of charge from <http://www.unicode.org/iso15924/codelists.html>
- [I3.RFC1958] Carpenter, B., "Architectural Principles of the Internet", RFC 1958, June 1996.
- [I4.BCP18] Alvestrand, H., "IETF Policy on Character Sets and Languages", BCP 18, RFC 2277, January 1998.
- [I5.RFC2045] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies", RFC 2045, November 1996.
- [I6.RFC2046] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types", RFC 2046, November 1996.
- [I7.RFC1945] Berners-Lee, T., Fielding, R., and H. Frystyk, "Hypertext Transfer Protocol -- HTTP/1.0", RFC 1945, May 1996.

- [I8.RFC2616] 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.
- [I9.Crocker05] Crocker, D., "Internet Mail Architecture", Work in progress (February 2005).
- [I10.BCP90] Klyne, G., Nottingham, M., and J. Mogul, "Registration Procedures for Message Header Fields", BCP 90, RFC 3864, September 2004.
- [I11.RFC2295] Holtman, K. and A. Mutz, "Transparent Content Negotiation in HTTP", RFC 2295, March 1998.
- [I12.RFC2324] Masinter, L., "Hyper Text Coffee Pot Control Protocol (HTCPCP/1.0)", RFC 2324, April 1998.
- [I13.BCP31] Holtman, K., Mutz, A., and T. Hardie, "Media Feature Tag Registration Procedure", BCP 31, RFC 2506, March 1999.
- [I14.RFC2912] Klyne, G., "Indicating Media Features for MIME Content", RFC 2912, September 2000.
- [I15.RFC3297] Klyne, G., Iwazaki, R., and D. Crocker, "Content Negotiation for Messaging Services based on Email", RFC 3297, July 2002.
- [I16.RFC2533] Klyne, G., "A Syntax for Describing Media Feature Sets", RFC 2533, March 1999.
- [I17.RFC2738] Klyne, G., "Corrections to "A Syntax for Describing Media Feature Sets"", RFC 2738, December 1999.

Author's Address

Bruce Lilly

Email: blilly@erols.com

Full Copyright Statement

Copyright © The Internet Society (2005).

This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the author retains all his rights.

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE

INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights 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; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat 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 implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Acknowledgment

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