Network Working Group
Internet Draft
Intended status: Standards Track
Expires: August 2014

O. Mazahir
D. Thaler
M. Cox
G. Montenegro
Microsoft Corporation
14 February 2014

Deterministic URI Encoding
draft-montenegro-httpbis-uri-encoding-00

Abstract

The "http" and "https" URI schemes do not have a fixed character encoding. This document defines HTTP headers to enable an explicit indication of the character encoding.

Copyright

Table of Contents

1. Introduction

The "http" and "https" URI schemes don't have a fixed character encoding. The URI RFC [RFC3986] talks about the generic syntax for URI components:

. Legacy URI components (before 2005) tend to use UTF-8 "or some other superset of the US-ASCII character encoding"
. New schemes (after 2005) use UTF-8 with percent encoding for reserved characters.

The first bullet explains why the character encoding for "http" and "https" URIs is not deterministic. This is particularly

problematic when parsing URIs at the server side or at intermediate proxies (e.g., when looking for a cache hit).

URI's have different components with different character encoding issues.

Per the IDNA rules in [RFC5890], the host component is encoded using A-labels.

There is more non-determinism with respect to the path and query components. Furthermore, these two components are not necessarily encoded the same way [Handbook].

This document defines HTTP headers that explicitly state the character encoding for the path and query components.

1.1. Requirements Language

The key words "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 [RFC2119].

2. URI Path and Query Encoding Headers

The URI Path encoding is conveyed in the following header:

```
URI-Path-Encoding   = "URI-Path-Encoding" ":" 1charset
```

The URI Query encoding is conveyed in the following header:

```
URI-Query-Encoding   = "URI-Query-Encoding" ":" 1charset
```

charset is defined in section 3.4 of [RFC2616]. The expected value indicates the character encoding for the path or query component in the URI prior to percent encoding. (A value of UTF-8 does not mean that the URI carries raw UTF-8.)

If the user agent is certain that the path component was formed from percent-encoded UTF-8, it sets the header as follows:

```
URI-Path-Encoding: UTF-8
```

Similarly, for the query component:

```
URI-Query-Encoding: UTF-8
```

This signals that the query component in the URI is in UTF-8 with percent encoding.

Absence of the URI-Path-Encoding or URI-Query-Encoding header is equivalent to the legacy situation of non-determinism with respect to the path or query component, respectively, as mentioned above in section 1.

Likewise, if the URI-Path-Encoding or URI-Query-Encoding header is set to an invalid value or unrecognized charset, this is equivalent to the legacy situation of non-determinism with respect to the path or query component, respectively, mentioned above in section 1.

3. IANA Considerations

IANA is requested to add these headers to the "Permanent Message Header Field Names" registry. Per [RFC3864], the template for these headers is specified below.

3.1. URI-Path-Encoding

Applicable protocol: http

Status: standard

Author/change controller:

IETF (iesg@ietf.org)

Specification document(s):

This document.

3.2. URI-Query-Encoding

Applicable protocol: http

Status: standard

Author/change controller:

IETF (iesg@ietf.org)

Specification document(s):

This document.

4. Security Considerations

Due to the non-deterministic character encoding of URI's, URI parsing at servers or proxies currently may involve trying different possible character encodings searching for a match. This represents a potential attack vector [RFC6943]. The headers proposed in this document could be used to reduce the attack surface by enabling a more explicit interpretation of the data within a URI, thus preventing unintended consequences.

5. Acknowledgments

Thanks to Ivan Pashov and Wade Hilmo for useful discussions in this space.

This document was prepared using 2-Word-v2.0.template.doc.

6. References

6.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[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.

[RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, January 2005.

6.2. Informative References

[Handbook] Zalewski, M., "Browser Security Handbook, part 1", http://code.google.com/p/browsersec/wiki/Part1

March 2011.

[RFC3864] Klyne, G., Nottingham, M., and J. Mogul, "Registration Procedures for Message Header Fields", BCP 90, RFC 3864, September 2004.

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, August 2010.

[RFC6943] Thaler, D., Ed., "Issues in Identifier Comparison for Security Purposes", RFC 6943, May 2013.

7. Author's Addresses

Osama Mazahir
Microsoft Corporation

Email: OsamaM@microsoft.com

Dave Thaler
Microsoft Corporation

Email: DThaler@microsoft.com

Matthew Cox
Microsoft Corporation

Email: MaCox@microsoft.com

Gabriel Montenegro
Microsoft Corporation

Phone:
Email: gabriel.montenegro@microsoft.com