Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 4, 2012

J. Reschke
greenbytes

July 3, 2011

Processing potentially invalid URI and IRI References
draft-reschke-ref-parsing-00

Abstract

The parsing of Uniform Resource Identifiers (URIs, RFC 3986) and Internationalized Resource Identifiers (IRIs, RFC 3987) is defined in terms of Augmented Backus-Naur Form (ABNF). The ABNF grammars are defined in terms of valid identifiers, and thus technically do not address how to handle invalid ones.

The URI specification however includes a note how to use Regular Expressions for parsing, and this note applies to invalid identifiers as well. This document introduces terminology referring to potentially invalid identifiers, and demonstrates how the rules in the URI specification can be applied to them.

Editorial Note (To be removed by RFC Editor before publication)

Distribution of this document is unlimited. Although this is not a work item of the IRI Working Group, comments should be sent to the IRI mailing list at public-iri@w3.org [1], which may be joined by sending a message with subject "subscribe" to public-iri-request@w3.org [2].

Discussions of the IRI Working Group are archived at <http://lists.w3.org/Archives/Public/public-iri/>.

XML versions and latest edits for this document are available from <http://greenbytes.de/tech/webdav/#draft-reschke-ref-parsing>.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on January 4, 2012.

Table of Contents

1. Introduction

The parsing of Uniform Resource Identifiers (URIs, [RFC3986]) and Internationalized Resource Identifiers (IRIs, [RFC3987]) is defined in terms of Augmented Backus-Naur Form (ABNF). The ABNF grammars are defined in terms of valid identifiers, and thus technically do not address how to handle invalid ones.

The URI specification however includes a note how to use Regular Expressions for parsing, and this note applies to invalid identifiers as well. This document introduces terminology referring to potentially invalid identifiers, and demonstrates how the rules in the URI specification can be applied to them.

2. Terminology

In addition to the terms defined in the URI specification, namely the Syntax Components (see Section 3 of [RFC3986]), this document defines:

Candidate URI Reference

> A string that may or may not be a valid URI-reference according to Section 4.1 of [RFC3986].

Candidate Scheme Component

> A string that may or may not be a valid URI scheme component according to Section 3.1 of [RFC3986].

Candidate Authority Component

> A string that may or may not be a valid URI authority component according to Section 3.2 of [RFC3986].

Candidate Path Component

> A string that may or may not be a valid URI path component according to Section 3.3 of [RFC3986].

Candidate Query Component

> A string that may or may not be a valid URI query component according to Section 3.4 of [RFC3986].

Candidate Fragment Component

A string that may or may not be a valid URI fragment component according to Section 3.5 of [RFC3986].

3. Processing

3.1. Parsing a Candidate URI Reference into Components

The regular expression given in Appendix B of [RFC3986] will parse any input string into a Candidate Scheme Component, a Candidate Authority Component, a Candidate Path Component, a Candidate Query Component, and a Candidate Fragment Component. Note that of these five components, all components except for the Path Component can be undefined.

If each of the defined components is valid according to the related URI component definition, the input was a valid URI reference.

3.2. Resolution of Candidate References

Section 5 of [RFC3986] defines Reference Resolution based on the five components. This algorithm works both for components obtained from valid and invalid references. The result will be a valid URI Reference if and only if the components used by the algorithm were valid themselves.

4. Security Considerations

[[anchor3: TBD]]

5. IANA Considerations

There are no IANA Considerations related to this specification.

6. References

6.1. Normative References

[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

[RFC3987] Duerst, M. and M. Suignard, "Internationalized Resource Identifiers (IRIs)", RFC 3987, January 2005.

URIs

[1] <mailto:public-iri@w3.org>

[2] <mailto:public-iri-request@w3.org?subject=subscribe>

Appendix A. Implementations

<http://greenbytes.de/tech/tc/uris/> shows results for the parsing/resolution processing described above, based on a test implementation written in XSLT 2.0.

Appendix B. Open issues (to be removed by RFC Editor prior to publication)

B.1. edit

Type: edit

julian.reschke@greenbytes.de (2011-07-02): Umbrella issue for editorial fixes/enhancements.

B.2. iri

Type: change

julian.reschke@greenbytes.de (2011-07-02): Expand for IRIs.

B.3. proc

Type: change

julian.reschke@greenbytes.de (2011-07-02): Re-state the parsing algorithm as a procedural algorithm, maybe in JS?

B.4. pre

Type: change

julian.reschke@greenbytes.de (2011-07-02): Define pre-processing steps for extraction of candidate references from content (WS stripping)?

B.5. post

Type: change

julian.reschke@greenbytes.de (2011-07-02): Define post-processing

steps, such as query component rewriting based on document encoding.

Author's Address

Julian F. Reschke
greenbytes GmbH
Hafenweg 16
Muenster, NW 48155
Germany

EMail: julian.reschke@greenbytes.de
URI: http://greenbytes.de/tech/webdav/