IETF
Internet-Draft
Obsoletes: 3987 (if approved)
Intended status: Standards Track
Expires: January 10, 2013

J. Klensin
S. Moonesamy

July 9, 2012

An XML-based Simple Resource Identifier
draft-klensin-iri-sri-00.txt

Abstract

While the URI specification has been widely deployed, it has long been recognized that many valid URIs, especially those that contain extensive information in the "tail" are unsuitable for user presentation, especially for internationalized environments. IRIs have been proposed as a solution for that problem but inherit (and are constrained by) the complex and sometimes method-dependent syntax model of URIs as well as positional and ordering assumptions that make them more difficult to localize than is desirable.

This specification illustrates a way to define an "above URI" model for a localization-friendly simple reference identifier (SRI) that explicitly identifies fields and is more appropriate than IRIs to support localization. The current version is intended simply to initiate a discussion. In particular, while it is written to use an XML element syntax model, variations using JSON or some other system with explicitly-labeled data fields might be as, or more, appropriate.

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 10, 2013.

Copyright Notice

Table of Contents

1. Introduction

While the URI specification [RFC3986] has been widely deployed, it has long been recognized that many valid URIs, especially those that contain extensive information in the "tail" are unsuitable for user presentation, especially for internationalized environments. IRIs [RFC3987] have been proposed as a solution for that problem but inherit (and are constrained by) the complex and sometimes method-dependent syntax model of URIs as well as positional and ordering assumptions that make them more difficult to localize than is desirable.

This specification illustrates a way to define a localization-friendly "above URI" simple syntax (a "SRI") that explicitly identifies fields and is more appropriate than IRIs to support localization.

[[anchor2: Note in Draft: "Simple" is chosen in the grand tradition of "simple" protocols like SMTP and SIP". Certainly the parsing of the compound identifier into components is simpler than the URI model. But suggestions for alternate terms would be welcome if "simple" turns into flame-bait.]]

This specification obviates most, if not all, of the perceived need for IRIs and hence obsoletes the specification of them in RFC 3087. A discussion of the reasons for that action appears in Appendix A.

1.1. Terminology

The terms "i18n" and "l10n" are liberally used as abbreviations for "internationalization" and "localization", respectively, in this specification.

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].

1.2. Status and Discussion

[[anchor5: RFC Editor: Please remove this subsection.]]

This draft is a pre-proposal to stimulate discussion of the IRI approach and alternatives to it. While it is deliberately incomplete, the path to an actual proposal should be clear. Also, the choice of an XML element syntax model [XML] structure was fairly arbitrary. It would probably be equally reasonable to support a JSON [RFC4627] or other structure instead (or additionally) as long as the basic syntax chosen supports clear identification of data elements

and a very precise and context-independent syntax for element values.

Discussion of this draft should occur on the IRI WG mailing list. Details about subscription and archives for the list may be found at XXXXX.

2. Tagged Elements

Much of the complexity in the URI specification lies in trying to identify and extract the various parts of a URI. That process is complicated by scheme-dependent elements and the associated delimiters which may be reserved or not depending on the scheme. That work may be appropriate if some system actually needs to parse and execute a URI -- an activity that requires understanding the scheme in any event-- but may be less appropriate for an i18n / l10n overlay.

This specification overcomes that problem and the associated complexities introduced by characters outside the ASCII repertoire, URI escaping conventions, and so on by eliminating the constraint of forward compatibility with URIs in favor of a more international format that can be easily localized and equally easily be mapped into that URI syntax.

3. Data Element Description

This section maps the various components of URIs into XML elements. For purposes of this specification, the URI syntax is discarded; only the data elements are retained. The mapping from an XML-structured document using these elements to URI syntax should be fairly obvious [[anchor8: ...and possibly covered in more detail in a future version of this spec]]. It is obviously possible to specify a collection of elements with this specification that, when mapped back into URI syntax, will be invalid or confusing for a particular scheme. If that is perceived as an issue, specific lists of what elements are valid for which schemes should be easy to compile.

The basic structure starts with a localization-friendly element that contains all other elements (and has no direct textual content):
<SRI>

[[anchor9: Note in Draft: Each of the subsections that follow can probably benefit from some fleshing-out. For this version, the general intent should be clear. It is likely that several more subsidiary elements are needed, but that is a topic for future discussion.]]

3.1. scheme Element

<scheme> SchemeName </scheme>

The Scheme element has no subsidiary elements.

3.2. authority Element

<authority>
Authority elements as below.
</authority>

The Authority element has the subsidiary elements listed in the subsections below.

3.2.1. user-info Element

3.2.2. host Element

Domain names are subject to special rules because of IDNA considerations, so the normal content of the host element is a domain element. [Domain-]relative URIs do not use the domain element.

3.2.2.1. domain Element

<domain> Fully-qualified-domain-name </domain>

3.2.3. port Element

<port> NN <port>
NN is a numeric port number.

3.3. path Element

<path> PathString </path>

[[anchor16: Subsidiary elements here, including <domain> and/or <SRI> when appropriate.]]

3.4. query Element

<query> QueryString </query>

[[anchor18: Subsidiary elements here, including <domain> and/or <SRI> when appropriate.]]

3.5. fragment Element

<fragment> FragmentName or other identifier </fragment>

The Fragment element has no subsidiary elements.

4. Internationalization and Escapes

Part of the goal for the format specified here is to express the abstract components of a URI as naturally as possible. Consequently, any text component of any element can be expressed in UTF-8 in normalization form NFC. Escapes ("%" or otherwise) are prohibited except as required by XML. If "%" appears, it must be doubled in mapping to URI syntax.

5. Examples

[[anchor22: There should be several of these, each showing a URI and the matching XRI form.]]

The URI that would appear as
http://example.com/test?sri=http://example.net/
Would appear in this form as:

```
<uri>
   <sri>
     <scheme>http</scheme>
     <authority>
       <host>
         <domain>example.com</domain>
       </host>
     </authority>
     <path>/test</path>
     <query>
       <sri>
         <scheme>http</scheme>
         <authority>
           <host>
             <domain>example.net</domain>
           </host>
          </authority>
       </sri>
     </query>
   </sri>
</uri>
```

[[anchor23: Note in draft: RFC (and I-D) constraints prohibit showing one of these data structures with characters in it outside the ASCII repertoire. If the document ever progresses to RFC, an alternate form that can show such examples including such characters should be a requirement.]]

6. Acknowledgements

Some of the structuring information for this document was derived from a W3C working draft on URLs [W3C-URL] as well as the URI specification. The thinking that led to this work started with a discussion many years ago with James Seng in which he pointed out that the "natural" ordering of components of compound identifiers differed by culture.

7. IANA Considerations

[[anchor24: RFC Editor: Please remove this section before publication.]]

This memo includes no requests to or actions for IANA.

8. Security Considerations

The model introduced in this specification does not raise any security issues not already present in the URI specification that would not be caught by a URI processor. Because it is less subtle and complex than the URI specification, it may actually lead to a reduction in vunerabilities.

9. References

9.1. Normative References

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

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

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

[XML] Bray, T., Ed., Paoli, J., Ed., Sperberg-McQueen, C., Ed., and E. Maler, Ed., "Extensible Markup Language (XML) 1.0 (Second Edition), W3C=20 Recommendation", October 2000, <http://www.w3.org/TR/REC-xml>.

9.2. Informative References

[IRI-Charter] IETF, "Internationalized Resource Identifiers (iri)", Captured 2012-07-05, 2019, <http://datatracker.ietf.org/wg/iri/charter/>.

[RFC3490] Faltstrom, P., Hoffman, P., and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, March 2003.

[RFC4627] Crockford, D., "The application/json Media Type for JavaScript Object Notation (JSON)", RFC 4627, July 2006.

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

[RFC6055] Thaler, D., Klensin, J., and S. Cheshire, "IAB Thoughts on Encodings for Internationalized Domain Names", RFC 6055, February 2011.

[W3C-URL] W3C, "URL", Captured 2012-07-03, 2012, <http://www.w3.org/TR/url/>.

Appendix A. This Specification and the IRI Approach

The original IRI specification [RFC3987] was intended as a strict superset of the URI syntax [RFC3986] with all URI forms being permitted but with the use of non-escaped UTF-8 strings also being allowed. IRIs were not separate protocol identifiers or intended for use "on the wire". Instead, they were intended as an overlay for URIs that was more convenient for users. In part because of the interaction with the original [RFC3490] and revised [RFC5890] versions of the IDNA specification, the mapping from IRIs to URIs was not unique: one could map a domain name expressed as a UTF-8 string into either a URI escape sequence or into a set of IDNA A-labels. That choice interacted badly with the domain name encoding considerations discussed by the IAB [RFC6055] and, more importantly, with URI comparisons in caches and similar contexts.

Based on those and other considerations, an IETF WG charged with IRI

revision [IRI-Charter] concluded that IRIs should be treated as a separate protocol identifier, primarily for use in new protocols, rather than as a strictly-forward-compatible URI overlay. That decision immediately raised the question of whether it was more valuable to preserve a URI-like syntax or depart from it entirely. This specification resulted from the desire to explore the possibilities that would be opened up by abandoning the constraint of apparent similarity to the URI syntax. But, just as the decision to move to a separate protocol identifier essentially recognizes that the IRIs defined in RFC 3987 was not feasible and an IRI variation that defined a new protocol element while retaining the general form of the URI syntax would obsolete 3987, this specification does as well: whether the underlying syntax model is changed or not, the WG has concluded that IRIs as defined in RFC 3987 are inappropriate for general use on the public Internet.

Appendix B. XML DTD

```
<!ELEMENT uri (sri)>

<!-- Simple Resource Identifier  -->
<!ELEMENT sri (scheme, authority, path?, query?,
                   fragment?)>

<!ELEMENT authority ( user-info?, host, port?)>

<!ELEMENT scheme (#PCDATA)>

<!ELEMENT user-info (#PCDATA)>

<!ELEMENT host (domain | ip-address)>

<!ELEMENT port ((#PCDATA)>

<!ELEMENT path (#PCDATA | domain | sri)*>

<!ELEMENT query (#PCDATA | domain | sri)*>

<!ELEMENT fragment (#PCDATA)*>

<!-- This contains a FQDN  -->
<!ELEMENT domain (#PCDATA)>

<!ELEMENT ip-address (#PCDATA)>
```

Authors' Addresses

John C Klensin
1770 Massachusetts Ave, Ste 322
Cambridge, MA 02140
USA

Phone: +1 617 491 5735
Email: john-ietf@jck.com

Subramanian Moonesamy
76, Ylang Ylang Avenue
Quatre Bornes
Mauritius

Email: sm+ietf@elandsys.com