Network Working Group   S. Ruby, Ed.
Internet-Draft   IBM
Intended status: Informational   L. Masinter
Expires: July 15, 2015   Adobe
                                                        January 11, 2015

URL Problem Statement and Directions
draft-ruby-url-problem-01

Abstract

This document lays out the problem space of possibly conflicting standards between multiple organizations for URLs and things like them, and proposes some actions to resolve the conflicts. From a user or developer point of view, it makes no sense for there to be a proliferation of definitions of URL nor for there to be a proliferation of incompatible implementations. This shouldn't be a competitive feature. Therefore there is a need for the organizations involved to update and reconcile the various Internet Drafts, Recommendations, and Standards in this area.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

This document lays out the problem space around standards for URLs and things like them, and proposes some actions to resolve the conflicts. From a user or developer point of view, it makes no sense for there to be a proliferation of definitions of URL nor for there to be a proliferation of incompatible implementations. This shouldn't be a competitive feature. Therefore there is a need for the organizations involved to update and reconcile the various Internet Drafts, Recommendations, and Standards in this area.

Possible next steps are discussed in Section 5.

Discussions have taken place on public-ietf-w3c@w3.org [1] (archive [2]) and public-ietf-w3c@w3.org [3] (archive [4]). In addition, the W3C TAG has discussed these issues in meetings and on their mailing list.

This document, as well as a test suite, reference implementation, and [WP-URL] are being developed at [5], including an issue tracker, Wiki, and related resources. Pull requests [6] for edits to doocuments or tests are most welcome. Raising issues in the GitHub tracker [7] is also helpful. Comments to the editors or on those mailing lists in email are also welcome.

2. Brief History of URL Standards

This section contains a very compressed history of URL standards, in sufficient detail to set some context. REVIEWERS: history is necessarily incomplete, but please report incorrect or missing essential facts.

The first standards-track specification for URLs was [RFC1738] in 1994. (That spec contains more background material.) It defined URLs as ASCII only. [RFC2396] later separated the generic syntax from concrete scheme definitions which are defined in separate RFCs. Many of those scheme definitions turned out not to get the attention that they needed.

When it became clear that it was desirable to allow non-ASCII characters, it was widely feared that support for Unicode by ASCII-only systems would turn out to be problematic. The tack was therefore taken to leave "URI" alone and define a new protocol element, "IRI". [RFC3987] was published in 2005 (in sync with the [RFC3986] update to the URI definition). This also turned out not to get the attention it needed.

To address issues raised both in IETF and for HTML5 (see Section 4 for more details), the IRI working group [8] was established in the IETF in 2009. However, primarily due to lack of engagement, the IRI group was closed in 2014, with the plan that the documents that had been under development in the IRI working group could be updated as individual submissions or within the IETF applications area working group. In particular, one of the IRI working group items was to update [RFC4395], which is currently under development in IETF's application area (see [appsawg-uri-scheme-reg]).

Independently, the HTML specifications in the WHATWG and W3C redefined "URL" in an attempt to match what some of the browsers were doing. This definition was later moved out into the "URL Living Standard" [URL-LS].

When W3C produced the HTML5 recommendation [9], the normative reference to the WHATWG URL standard was a gating issue, and an unusual compromised was reached [10], where the [URL] reference is given a descriptive paragraph rather than a single document reference.

The world has moved on in other ways. ICANN has approved non-ASCII top level domains, but IDNA specs ([RFC3490] and [RFC5895]) did not fully addressed IRI processing. Subsequently, the Unicode consortium produced [UTS-46], which mentions URL processing in passing.

The web security working group developed [RFC6454] ("The Web Origin Concept"), which was refined in the W3C [CORS] specification, which [URL-LS] redefines. Updates in the IETF were abandoned. Work continues in the WHATWG in the [FETCH] specification.

3. Current Organizations and Specs in Development

There are multiple umbrella organizations which have produced multiple documents, and it's unclear whether there's a trajectory to make them consistent. This section tries to enumerate currently active organizations and specs. REVIEWERS: are there important ongoing activities we've missed or gotten wrong? Who are the stakeholders whose current work might be affected? (This input will help determine the organizational coordination needed.)

Organizations include the IETF [11], the WHATWG [12], the W3C [13], Web Platform.org [14], and the Unicode Consortium [15]. Relevant specs under development in each organization include:

3.1. IETF

[appsawg-uri-scheme-reg] has passed working group last call and entered IESG review.

New schemes and updates to old ones continue, including 'file:' [kerwin-file-scheme] and 'urn:'.

The IRI working group closed, but work can continue in the Applications Area working group. Documents sitting needing update, abandoned now, are three drafts ([iri-3987bis], [iri-comparison], and [iri-bidi-guidelines]), which were originally intended to obsolete [RFC3987].

The URNBis working group [16] has been working to update the definitions of URNs, but has difficulty with some of the wording in [RFC3986]. In particular, [17] updates [RFC3986].

3.2. WHATWG

The [URL-LS] is being developed as a living standard [18]. It primarily focuses on specifying what is important for browsers. The means by which new schemes might be registered is not yet defined. This work is based on [UTS-46], and includes an explicit goal of obsoleting both [RFC3986] and [RFC3987].

3.3. W3C

The Web Applications Working Group [19], in conjunction with the TAG [20], sporadically have been republishing the WHATWG work with no technical content differences as [W3C-URL]. There is a [url-workmode] proposal to formalize this relationship.

The W3C TAG [21] developed Best Practices for Fragment Identifiers and Media Type Definitions [22], which points out several problems with the definitions for the 'fragment' part of URLs. The TAG is working to ensure liaison exchange happens.

Note also the interim solution for the HTML5 reference to [URL] [23], which should be updated by the HTML working group [24].

3.4. WebPlatform

WebPlatform.org is an activity sponsored by W3C and web vendors [25]. [WP-URL] is being developed on a develop [26] GitHub branch based on [URL-LS]. It currently contains work that has yet to be folded back into the [URL-LS], primarily to rewrite the parser logic in a way that is more understandable and approachable. The intent is to merge this work once it is ready, and to actively work to keep the two versions in sync.

3.5. Unicode Consortium

[UTS-46] defines parameterized functions for mapping domain names. [URL-LS] builds upon this work, specifying particular values to be used for these parameters. The Unicode Consortium plans to adapt [UTS-46] as registries (e.g. DENIC) move from [RFC3490] to [RFC5895].

4. Problem Statements

This section lays out the problems we see need a coordinated solution. REVIEWERS: have we missed some things? Are any of these non-problems or not worth solving?

The main problem is conflicting specifications that overlap but don't match each other.

Additionally, the following are issues that need to be resolved to make URL processing unambiguous and stable.

- Nomenclature: over the years, a number of different sets of terminology has been used. URL / URI / IRI is not the only difference. [tantek-slice] chronicles a number of differences.

- Deterministic parsing and transformation: The IRI-to-URI transformation specified in [RFC3987] had options; it wasn't a deterministic path; in particular, which substrings of which URLs of which Unicode, for strings were to be transformed to Punycode or to %-escaped-utf8. The URI-to-IRI transformation was also heuristic, since there was no guarantee that %xx-encoded bytes in the URI were actually meant to be %xx percent-hex-encoded bytes of a UTF-8 encoding of a Unicode string.

- Parameterization: standards in this area need to define such matters as normalization forms and values for parameters such as UseSTD3ASCIIRules.

- Interoperability: even after accounting for the above, there is a demonstrable lack of interoperability across popular libraries and browsers. [whatwg-interop] identifies a number of such differences.

- Stability: Before any standard document can be marked as obsoleted, the requirements other specs that normatively reference the to-be-obsoleted standard need to be considered, to avoid dangling references.

- IDNA: [RFC3490] defines processing for 'IDN-aware domain name slots' (where "the host portion of the URI in the src attribute of an HTML <IMG> tag" is given as an example. Later, "IDNA is applicable to all domain names in all domain name slots". So in mailto:user@host, is the host a IDN-aware domain name slot? A domain name slot at all?

- Bidi URLs: The problems with writing URLs using characters from right-to-left languages are well-known among experts; what is not known is a solution for these problems. The solution given in [RFC3987] has some obvious errors (how to handle combining marks); it's general approach also probably can be improved on, but it's not sure how.

- o Specific scheme definitions: some UR* scheme definitions are woefully out of date, incomplete, or don't correspond to current practice, but updating their definitions is unclear. This includes 'file:', for which there is a current effort, but there are others which need review (including 'ftp:', 'data:').

5. Next Steps, Solutions

Many of the problem above require some cross-organizational collaboration. This section outlines alternatives and possible next steps, both in terms of documents and possible updates and also procedural issues.

REVIEWERS: Neccessary? Sufficient? What are we missing, what did we get wrong?

5.1. Working Groups and Discussion Venues

The XML Signature WG [27] is an example of a joint IETF/W3C Working Group. Perhaps a joint working group covering the topics of URL and URI could be formed. Elements of the [url-workmode] proposal could be incorporated into the charter of this new WG, and thereby establishing the WHATWG as a third joint participant in this activity.

Failing that, it may be desirable to have some organizational assignment of responsibility in IETF and W3C to working groups in each organization.

There has been discussion of IETF/W3C liaison getting involved, with the proposal that W3C liaison to IETF making a formal liaison request to which IETF would respond. Perhaps the liaison request might reference this document.

In IETF, the scope of changes proposed may determine how IETF consensus can best be obtained. It seems unlikely that the scope of necessary changes to IETF documents could be managed through individual submissions. Some opinions have been that updating [RFC3986] and/or obsoleting [RFC3987] would require a full IETF working group. Unless and until another group is chartered (perhaps using this document as the Problem Statement / scope), discussion is occuring in the IETF apps area. Previous venues for related topics (public-iri@w3.org [28], uri@w3.org [29]) are old enough that there is likely poor representation of important communities, unless a concerted effort is made to revive them.

In W3C, either W3C WebApps, TAG, HTML or some new activity might be necessary to manage changes, but the nature of the group necessary to review depends on the extent of changes needed.

At the moment, the most reliable way of giving feedback on this document is to raise or comment on issues in the GitHub issue list [30].

5.2. Leave, Update or Obsolete RFC 3986 (URI)

At various times, many have called for replacing the IETF URI standard [RFC3986], or updating it. How to approach this is controversal, but at a minimum the following are needed:

- Make it clear that ASCII-only URIs (as now defined by [RFC3986]) are not what is mainly used on the web.

- Incorporate updates for URN.

- Incorporate updates for fragment identifier semantics.

- Note terminology issue and resolution.

More controversial is whether this can be done on a strictly "need-to" basis, or whether the merger of URI from [RFC3986] and IRI from [RFC3987] would result in clearer specifications for implementors.

5.3. Obsolete RFC 3987 (IRI)

There is some sentiment to restart the work of updating [RFC3987] by starting again, fixing errors and integrating errata. However, this path doesn't seem to satisfy the desire for a single spec that lays out deterministic processing for URLs and references for browser and operating-system handling of both.

After ensuring that topics covered in [RFC3987] are also covered by a W3C URL recommendation, mark [RFC3987] as obsolete with a short RFC noting the conditions laid out in this document.

5.4. Obsolete RFC 6454 (Origin)

Replaced by [CORS], [URL-LS], and/or [FETCH].

5.5. file: URI scheme

Coordinate 'file:' syntax in [URL-LS] and [kerwin-file-scheme], possibly moving the 'file:' part of URL-LS into a separate document.

5.6. Other actions

- Update [RFC5895] to be consistent with [URL-LS] and [UTS-46]. This may involve working to get the other specifications updated, if only to clarify nomenclature.

- Obsolete any previous definition of x-url-encoded.

- Change the [URL-LS] goals to only obsolete specifications listed above that are not updated. Presuming that [RFC3986] is updated, explicitly state that conforming URLs are a proper subset of valid URIs, and further state that canonical URLs (i.e., the output of the URL parser) not only round trip, but also are valid URIs.

- Update and incorporate (or reference) the content currently present in [tantek-slice], probably as an appendix to [URL-LS], so that readers will understand what terms are in use and how they map.

- Reconcile how [appsawg-uri-scheme-reg] and [URL-LS] handle currently unknown schemes, update [appsawg-uri-scheme-reg] to state that registration applies to both URIs and URLs, and update [URL-LS] to indicate that [appsawg-uri-scheme-reg] is how you register schemes.

- Have the W3C adopt [url-workmode].

- Other than keeping on top of [UTS-46] and responding to any feedback that may be provided, no changes to any Unicode Consortium product is required.

6. Acknowledgements

Helpful comments and improvements to this document have come from Anne van Kesteren, Bjoern Hoehrmann, Graham Klyne, Julian Reschke, and Martin Duerst.

7. IANA Considerations

This memo currently includes no request to IANA, although an updated [appsawg-uri-scheme-reg] might add some additional requirements and information to IANA URI scheme registry [31] to make clear that the schemes serve as URL schemes and IRI schemes as well as URI schemes.

8. Security Considerations

In addition to the security exposures created when URLs work differently in different systems, all of the security considerations defined in [RFC3490], [RFC3986], [RFC3987], and [RFC5895] apply to URLs.

9. Informative References

[CORS]   van Kesteren, A., "Cross-Origin Resource Sharing", 2014, <http://www.w3.org/TR/cors/>.

[FETCH]   van Kesteren, A., "Fetch Living Standard", 2014, <https://fetch.spec.whatwg.org/>.

[RFC1738] Berners-Lee, T., Masinter, L., and M. McCahill, "Uniform Resource Locators (URL)", RFC 1738, December 1994.

[RFC2396] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifiers (URI): Generic Syntax", RFC 2396, August 1998.

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

[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.

[RFC4395] Hansen, T., Hardie, T., and L. Masinter, "Guidelines and Registration Procedures for New URI Schemes", BCP 35, RFC 4395, February 2006.

[RFC5895] Resnick, P. and P. Hoffman, "Mapping Characters for Internationalized Domain Names in Applications (IDNA) 2008", RFC 5895, September 2010.

[RFC6454] Barth, A., "The Web Origin Concept", RFC 6454, December 2011.

[URL-LS]   van Kesteren, A. and S. Ruby, "URL Living Standard", 2014, <https://url.spec.whatwg.org/>.

[UTS-46] Davis, M. and M. Suignard, "Unicode IDNA Compatibility Processing", 2014, <http://unicode.org/reports/tr46/>.

[W3C-URL] van Kesteren, A. and S. Ruby, "URL Working Draft", 2014, <http://www.w3.org/TR/url/>.

[WP-URL] van Kesteren, A. and S. Ruby, "URL Standard", 2014, <https://specs.webplatform.org/url/webspecs/develop/>.

[appsawg-uri-scheme-reg]
Thaler, D., Hansen, T., Hardie, T., and L. Masinter, "Guidelines and Registration Procedures for New URI Schemes", draft-ietf-appsawg-uri-scheme-reg-04 (work in progress), October 2014.

[iri-3987bis]
Duerst, M., Suignard, M., and L. Masinter, "Internationalized Resource Identifiers (IRIs)", draft-ietf-iri-3987bis-13 (work in progress), October 2012.

[iri-bidi-guidelines]
Duerst, M., Masinter, L., and A. Allawi, "Guidelines for Internationalized Resource Identifiers with Bi-directional Characters (Bidi IRIs)", draft-ietf-iri-bidi-guidelines-03 (work in progress), October 2012.

[iri-comparison]
Masinter, L. and M. Duerst, "Comparison, Equivalence and Canonicalization of Internationalized Resource Identifiers", draft-ietf-iri-comparison-02 (work in progress), October 2012.

[kerwin-file-scheme]
Kerwin, M., "The file URI Scheme", draft-kerwin-file-scheme-13 (work in progress), September 2014.

[tantek-slice]
Celik, T., "How many ways can you slice a URL and name the pieces?", 2011, <http://tantek.com/2011/238/b1/many-ways-slice-url-name-pieces>.

[url-workmode]
Ruby, S., "URL WorkMode", 2014, <https://github.com/webspecs/url/blob/develop/docs/workmode.md#preface>.

[whatwg-interop]
Ruby, S., "URL test results", 2014, <https://url.spec.whatwg.org/interop/test-results/>.

Authors' Addresses

Sam Ruby (editor)
IBM
Raleigh, NC
USA

Email: rubys@intertwingly.net
URI: http://intertwingly.net/

Larry Masinter
Adobe
345 Park Ave
San Jose, CA 95110
USA

Email: masinter@adobe.com
URI: http://larry.masinter.net/