UNIFORM RESOURCE IDENTIFIER

Wikipedia
http://en.wikipedia.org/wiki/Uniform_Resource_Identifier

September 29, 2009

In computing, a Uniform Resource Identifier (URI) consists of a string of characters used to identify or name a resource on the Internet. Such identification enables interaction with representations of the resource over a network (typically the World Wide Web) using specific protocols. Schemes specifying a specific syntax and associated protocols define each URI.

1 Relationship to URL and URN

Computer scientists may classify a URI as a locator (URL), or a name (URN), or both. A Uniform Resource Name (URN) functions like a person's name, while a Uniform Resource Locator (URL) resembles that person's street-address. In other words: the URN defines an item's identity, while the URL provides a method for finding it.

The ISBN system for uniquely identifying books provides a typical example of the use of typical URNs. ISBN 0486275574 (urn:isbn:0-486-27557-4) cites unambiguously a specific edition of Shakespeare's play Romeo and Juliet. In order to gain access to this object and read the book, one would need its location: a URL address. A typical URL for this book on a unix-like operating system might look like the file path file:///home/username/RomeoAndJuliet.pdf, identifying the electronic book saved in a file on a local hard disk. So URNs and URLs have complementary purposes.

1.1 Technical View

One can define a URL as a URI that, in addition to identifying a resource, provides a means of acting upon or obtaining a representation of the resource by describing its primary access mechanism or network "location". For example, the URL http://www.wikipedia.org/ identifies a resource (Wikipedia's home page) and implies that a user can get a representation of that resource (such as the home page's current HTML code, as encoded characters) via HTTP from a network host named www.wikipedia.org. A Uniform Resource Name (URN) comprises a URI that identifies a resource by name in a particular namespace. One can use a URN to talk about a resource without implying its location or how to access it. For example, the URN urn:isbn:0-395-36341-1 is a URI that specifies

the identifier system, i.e. International Standard Book Number (ISBN), as well as the unique reference within that system and allows one to talk about a book, but doesn't suggest where and how to obtain an actual copy of it.

Technical publications, especially standards produced by the IETF and by the W3C, normally no longer use the term URL, as the need to distinguish between URLs and URIs rarely arises.[1] However, in non-technical contexts and in software for the World Wide Web, the term URL remains widely used. Additionally, the term web address (which has no formal definition) often occurs in non-technical publications as a synonym for URL or URI, although it generally refers only to the "http" and "https" URL schemes.

1.2 RFC 3305

Much of this discussion comes from RFC3305, titled "Report from the Joint W3C/IETF URI Planning Interest Group: Uniform Resource Identifiers (URIs), URLs, and Uniform Resource Names (URNs): Clarifications and Recommendations". This RFC outlines the work of a joint W3C/IETF working group set up specifically to normalize the divergent views held within the IETF and W3C over the relationship between the various "UR*" terms and standards. While not published as a full standard by either organization, it has become the basis for the above common understanding and has informed many standards since then.

2 Syntax

The URI syntax essentially offers a URI scheme name such as: "HTTP",
"FTP",
"mailto",
"URN",
"tel",
"rtsp",
"file",
followed by a colon character, and then by a scheme-specific part. The specifications that govern the schemes determine the syntax and semantics of the scheme-specific part, although the URI syntax does force all schemes to adhere to a certain generic syntax that, among other things, reserves certain characters for special purposes (without always identifying those purposes). The URI syntax also enforces restrictions on the scheme-specific part, in order to, for example, provide for a degree of consistency when the part has a hierarchical structure. Percent-encoding is an often-misunderstood[by whom?] aspect of URI syntax.

3 History

3.1 Naming, Addressing, and Identifying Resources

URIs and URLs have a shared history. In 1990, Tim Berners-Lees proposals for HyperText[2] implicitly introduced the idea of a URL as a short string representing a resource as the target of a hyperlink. At the time people referred to it as a "hypertext name"[3] or "document name".

Over the next three-and-a-half years, as the World Wide Web's core technologies of HTML (the HyperText Markup Language), HTTP, and web browsers developed, a need to distinguish a string that provided an address for a resource from a string that merely named a resource emerged. Although not yet formally defined, the term Uniform Resource Locator came to represent the former, and the more contentious Uniform Resource Name came to represent the latter.

During the debate over how to best define URLs and URNs, it became evident that the two concepts embodied by the terms merely displayed aspects of the fundamental, overarching notion of resource identification. So in June 1994, the IETF published Berners-Lee's RFC 1630: the first RFC that (in its non-normative text) acknowledged the existence of URLs and URNs, and, more importantly, defined a formal syntax for Universal Resource Identifiers URL-like strings whose precise syntaxes and semantics depended on their schemes. In addition, this RFC attempted to summarize the syntaxes of URL schemes in use at the time. It also acknowledged, but did not standardize, the existence of relative URLs and fragment identifiers.

3.2 Refinment of Specifications

In December 1994, RFC 1738 formally defined relative and absolute URLs, refined the general URL syntax, defined how to resolve relative URLs to absolute form, and better enumerated the URL schemes then in use. The agreed definition and syntax of URNs had to wait until the publication of RFC 2141 in May 1997.

The publication of RFC 2396 in August 1998 saw the URI syntax become a separate specification: http://www.faqs.org/rfcs/rfc2396.html, and the revision and expansion of most of the parts of RFCs 1630 and 1738 relating to URIs and URLs in general. The new RFC changed the significance of the "U" in "URI": it came to represent "Uniform" rather than "Universal". The sections of RFC 1738 that summarized existing URL schemes migrated into a separate document[4]. IANA keeps a registry of those schemes[5]; RFC 2717 first described the procedure to register them.

In December 1999, RFC 2732 provided a minor update to RFC 2396, allowing URIs to accommodate IPv6 addresses. Some time later, a number of shortcomings discovered in the two specifications led to the development of a number of draft revisions under the title rfc2396bis. This community effort, coordinated by RFC 2396 co-author Roy Fielding, culminated in the publication of RFC 3986 in January 2005. This RFC, as of 2009 the current version of the URI syntax recommended for use on the Internet, renders RFC 2396 obsolete. It does not, however, render the details of existing URL schemes

obsolete; RFC 1738 still governs those, except where otherwise superseded RFC 2616 for example, refines the "http" scheme. Simultaneously, the IETF published the content of RFC 3986 as the full standard STD 66, reflecting the establishment of the URI generic syntax as an official Internet protocol.

In August 2002, RFC 3305 pointed out that the term "URL" has, despite its widespread use in the vernacular of the Internet-aware public at large, faded into near-obsolescence. It now serves only as a reminder that some URIs act as addresses because they have schemes that imply some kind of network accessibility, regardless of whether systems actually use them for that purpose. As URI-based standards such as Resource Description Framework make evident, resource identification need not suggest the retrieval of resource representations over the Internet, nor need imply network-based resources at all.

On November 1, 2006, the W3C Technical Architecture Group published "On Linking Alternative Representations To Enable Discovery And Publishing", a guide to best practices and canonical URIs for publishing multiple versions of a given resource. For example, content might differ by language or by size to adjust for capacity or settings of the device used to access that content.

The Semantic Web uses the HTTP URI scheme to identify both documents and concepts in the real world: this has caused confusion as to how to distinguish the two. The Technical Architecture Group of W3C (TAG) published an e-mail in June 2005 on how to solve this problem. The e-mail became known as the httpRange-14 resolution.[6] To explain this (rather brief) email, W3C published in March 2008 the Interest Group Note Cool URIs for the Semantic Web[7]. This explains the use of content negotiation and the 303-redirect code in more detail.

4 URI reference

A URI reference another type of string represents a URI, and, in turn, the resource identified by that URI. Informal usage does not often maintain the distinction between a URI and a URI reference, but protocol documents should not allow for ambiguity.

A URI reference may take the form of a full URI, or just the scheme-specific portion of one, or even some trailing component thereof even the empty string. An optional fragment-identifier, preceded by "#", may appear at the end of a URI reference. The part of the reference before the "#" indirectly identifies a resource, and the fragment identifier identifies some portion of that resource.

In order to derive a URI from a URI reference, software converts the URI reference to "absolute" form by merging it with an absolute "base" URI according to a fixed algorithm. The system treats the URI reference as relative to the base URI, although in the case of an absolute reference, the base is irrelevant. The base URI typically identifies the document containing the URI reference, although this can be overridden by declarations made within the document or as part of an external data transmission protocol. If the base URI includes a fragment identifier, it is ignored during the merging process. If a fragment identifier is present in the URI reference, it is preserved during the merging process.

Web-document markup languages frequently use URI references in places where they need to point to other resources, such as to external documents or to specific portions of the same logical document.

4.1 Uses of URI references in markup languages

In HTML, the value of the src attribute of the img element functions as a URI reference, as does the value of the href attribute of the a or link element. In XML, the system identifier appearing after the SYSTEM keyword in a DTD is a fragmentless URI reference.

In XSLT, the value of the href attribute of the xsl:import element/instruction is a URI reference; likewise the first argument to the document() function.

4.2 Examples of absolute URIs

```
http://example.org/absolute/URI/with/absolute/path/to/resource.txt
ftp://example.org/resource.txt
urn:issn:1535-3613
```

4.3 Examples of URI references

```
http://en.wikipedia.org/wiki/URI#Examples_of_URI_references
```

("http" specifies the 'scheme' name, "en.wikipedia.org" is the 'authority', "/wiki/URI" the 'path' pointing to this article, and "#Examples_of_URI_references" is a 'fragment' pointing to this section.)

```
http://example.org/absolute/URI/with/absolute/path/to/resource.txt
//example.org/scheme-relative/URI/with/absolute/path/to/resource.txt
/relative/URI/with/absolute/path/to/resource.txt
relative/path/to/resource.txt
../../../resource.txt
./resource.txt#frag01
resource.txt
#frag01
(empty string)
```

5 URI resolution

To "resolve" a URI means either to convert a relative URI reference to absolute form, or to dereference a URI or URI reference by attempting to obtain a representation of the resource that it identifies. The "resolver" component in document-processing software generally provides both services.

One can regard a URI reference as a same-document reference: a reference to the document containing the URI reference itself. Document-processing software is encouraged[by

whom?] to use its current representation of the document to satisfy the resolution of a same-document reference without fetching a new representation. This is only a recommendation, and document processing software is free to use other mechanisms to determine whether it should obtain a new representation.

The current URI specification as of 2009, RFC 3986, defines a URI reference as a same-document reference if, when resolved to absolute form, it equates exactly to the base URI in effect for the reference. Typically, the base URI is the URI of the document containing the reference. XSLT 1.0, for example, has a document() function that, in effect, implements this functionality. RFC 3986 also formally defines URI equivalence, which can serve to determine that a URI reference, while not identical to the base URI, still represents the same resource and thus can be considered to be a same-document reference.

RFC 2396 prescribed a different method for determining same-document references; RFC 3986 made RFC 2396 obsolete, but RFC 2396 still serves as the basis of many specifications and implementations. This specification defines a URI reference as a same-document reference if it is an empty string or consists of only the "#" character followed by an optional fragment.

6 Relation to XML namespaces

XML has a concept of a namespace, an abstract domain to which a collection of element and attribute names can be assigned. The namespace name, a character string which must adhere to the generic URI syntax, identifies an XML namespace. However, the namespace name is not considered[by whom?] to be a URI because the "URI-ness" of strings is, according to the URI specification, based on their intended use, not just their lexical components. A namespace name also does not necessarily imply any of the semantics of URI schemes; a namespace name beginning with "http:", for example, likely has nothing to do with the HTTP protocol. XML professionals have debated this intensively on the xml-dev electronic mailing list; some feel that a namespace name could be a URI, since the collection of names comprising a particular namespace could be considered to be a resource that is being identified, and since a version of the "Namespaces in XML" specification says that the namespace name is a URI reference.[8] But the consensus seems to suggest that a namespace name is just a string that happens to look like a URI, nothing more.

Initially, the namespace name could match the syntax of any non-empty URI reference, but an erratum to the "Namespaces In XML Recommendation" later deprecated the use of relative URI references. A separate specification, issued for namespaces for XML 1.1, allows IRI references, not just URI references, to serve as the basis for namespace names.

In order to mitigate the confusion that began to arise among newcomers to XML from the use of URIs (particularly HTTP URLs) for namespaces, a descriptive language called RDDL (Resource Directory Description Language) developed, though the specification of RDDL (http://www.rddl.org/) has no official standing and no relevant organization (such as W3C) has considered or approved it. An RDDL document can provide machine-

and human-readable information about a particular namespace and about the XML documents that use it. Authors of XML documents were encouraged[by whom?] to put RDDL documents in locations such that if a namespace name in their document somehow becomes de-referenced, then an RDDL document would be obtained, thus satisfying the desire among many developers for a namespace name to point to a network-accessible resource.

7 References

1. URI Planning Interest Group, W3C/IETF (21 September 2001). "URIs, URLs, and URNs: Clarifications and Recommendations 1.0". Retrieved 2009-07-27.

2. Palmer, Sean B.. "The Early History of HTML". Retrieved 2009-04-30.

3. "W3 Naming Schemes". W3. Retrieved 2009-07-24. "The format of a hypertext name consists of the name of the naming sub-scheme to be used, then a name in a format particular to that subscheme, then an optional anchor identifier within the document. For example, the format is for all internet-based access methods: `scheme : // host.domain:port / path / path # anchor`.

4. This separate document is not explicitly linked[by whom?], RFC 2717 and RFC 4395 point to the IANA registry as the official URI scheme registry.

5. IANA registry of URI schemes.

6. The httpRange-14 resolution consists of three bullet points: see Fielding, Roy T. (2005-06-18). "[httpRange-14 Resolved]". Retrieved 2009-07-24., and did not help much to reduce the confusion.

7. http://www.w3.org/TR/cooluris/

8. World Wide Web Consortium (1999-01-14). "Namespaces in XML" (PDF). W3C. Retrieved 2009-09-14. "[Definition:] The attribute's value, a URI reference, is the namespace name identifying the namespace."