URIs, URLs, and URNs…
Web Naming and Addressing

Zhou, Xin

Course: INFSCI 2560
Instructor: M. Spring
Date: Jan. 29, 2007
A dozen years ago, web locations and addresses were relatively obscure. Even nowadays, people still easily getting confused with these terms: URIs, URLs and URNs. There are often misunderstandings of these three abbreviations. First, some people use URLs instead of URIs as if they are the same, while, as a matter of fact, they are not. Second, some people still consider the upper letter $U$ stands for Universal, which, to be honest, is correct before Dec. 1994. Third, nobody seemed to know what URNs is, however we almost use it everyday. Following will discuss these three terms and their related standards and history.

URIs stands for Uniform Resource Identifiers. As it was said by Dan Connolly (Chair of URI Interest Group, W3C, URI Activity Lead) in 2005, “My role in that history goes back at least as far as the Hypertext conference in 1991, where I met both Douglas Engelbart (inventor of the mouse and pioneer of networked computers and hypertext) and Tim Berners-Lee (inventor of the World Wide Web). In a 1990 summary of his 20-plus years of research, Engelbart listed among the requirements for an Open Hyperdocument System, ‘in principle, every object that someone might validly want or need to cite should have an unambiguous address.’ In his 1991 design document on naming, Berners-Lee wrote: This is probably the most crucial aspect of design and standardization in an open hypertext system. It concerns the syntax of a name by which a document or part of a document (an anchor) is referenced from anywhere else in the world.” The reason I used this quotation is because it perfectly explained the initial motivation of the invention of URIs. The “Web” we are using now is equal to an “Open Hyperdocument System”, and Engelbart briefly described the whole requirements of URIs with three demands: every object on the Web should be available for valid citation, this citation should be done with the address of the object resource, and the address should be unambiguous. And Tim’s words just reinforced these ideas.

The first demand represents the core advantage of the Web; this also can be called “sharing” in my understanding. The Web is designed to benefit the users all around the world through sharing the Hyperdocument, which means the resources should be viewable by valid users. This, consequently, leads to the second demand. This view and citation should base on some kind of address name, instead of some physical transportation or data copying. Simply imagine if we have to copy every single document to our disk instead of including a simple link pointing to that document in ours, the Web will becoming useless because of the redundancy. And this finally leads to the last demand: unique. In Tim Berners-Lee’s original writings on Document Naming, part of Design Issues for the Web, 1991, he proposed that: “A hypertext link to a document ought to be specified using the most logical name as opposed to a physical address.” And this was probably “the only way” to locate the document regardless the physical movement. He also mentioned the concepts of the abstract
structure and naming schemes, and the relationship between abstract and physical. This document he wrote could be considered as the first description of URI.

Three years later, in March 1994, the Internet Engineering Task Force (IETF) URI working group is formed, the group is chartered to define a set of standards for the encoding of system-independent resource location and identification information for the use of Internet information services. It is expected to produce a set of documents that specify standard representations of Uniform Resource Locators (URLs) for encoding location and access information across multiple information systems, Unique Resource Serial Numbers (URSNs) which specify a standardized method for encoding unique resource identification information for Internet resources, and Uniform Resource Identifiers (URIs) which specify a standardized method for encoding combined resource identification and location information systems to be used for resource discovery and access systems in an Internet environment. The Chairs of this group is Jim Fullton, Director of Research & Development, the Clearinghouse for Networked Information Discovery and Retrieval (CNIDR), and Alan Emtage, the creator of Archie, the world's first Internet search engine for locating material in public FTP archives, and carried out “Uniform Resource Locators (URL) A Syntax for the Expression of Access Information of Objects on the Network”, 04/05/1994, T. Berners-Lee.

In June 1994, the Request for Comments (RFC) 1630: Universal Resource Identifiers in WWW: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web was released. It was an Informational RFC, which means it did not carry any endorsement from the community. It was written by Tim Berners-Lee and discussed all aspects of URLs (relative, etc.). This probably becomes the reason why some professionals still consider the from URLs or URIs stands for Universal. The reason why I mentioned above that after December 1994, this term should not be used any more is because a Proposed Standard (which means it was the result of a consensus process, though it was not yet tested and mature enough to be a full Internet Standard) – RFC 1738: Uniform Resource Locators (URL) was carried out by T. Berners-Lee, M. McCahill and L. Masinter (from Xerox Corporation). The Universal had been changed into Uniform, which is more precise. This document specified a Uniform Resource Locator (URL), the syntax and semantics of formalized information for location and access of resources via the Internet. It discussed each of the known at that time URL schemes, but its grammar has some mistakes, and it doesn't cover several aspects of URL syntax, such as relative URLs and fragment identifiers, which important to nowadays especially in Extensible Markup Language (XML). At the same time, L. Masinter together with K. Sollins carried out another Informational RFC 1737: Functional Requirements for Uniform Resource Names, which specified a minimum set of requirements for a kind of Internet resource identifier known as Uniform Resource Names (URNs), this is the born of this term.

After the fundamental procedures, in June 1995, the RFC 1808: Relative Uniform Resource Locators released. This Proposed Standard wrote by R. Fielding, and
introduced an important concept which reduced the information contained in the absolute URL while the base URL is well-defined and known to the parser inheriting instead of re-specifying. This RFC is both authoritative and accurate on the matter of the syntax of URLs. Its grammar is complete and consistent. However, it does not discuss any of the actual URL schemes which been addressed later. The IETF URI Working Group closed in July of the same year. In the summary of the meeting they mentioned that a new Working Group to be formed specifically for the purpose of determining procedures for IANA acceptance of new URNs for registration and revise the RFC 1736: Functional Recommendations for Internet Resource Locators. This is really a significant one among all the others in my opinion. It describes the minimum set of requirements for Internet resource locators, which convey location and access information for resources; and specifies that: “A resource locator is a kind of resource identifier.” This document declares the main purpose of the locator and claims the usage and needs of this kind of locator. All the designs should follow the needs, so this RFC 1736 is the impetus of the further implementations.

In May 1997, R. Moats from AT&T carried out the URN Syntax (RFC 2141), which specified the URN syntax using sets of regular expressions. And emphasized two attributes of URNs: persistence and location-independency. These two attributes are the original purpose of the URNs, as said by Dan Connolly, 2005; the persistence is fighting against availability. If the definition of the resource is persistent, the transporting of the resource could cause the breaking down of the availability, which means the URNs should be location-independently. Like the International Standard Book Number (ISBN), regardless the position of the real book, the ISBN can locate the specific book with a specific edition. This is really important and useful. Since the transmission of the resource happens all the time, the real location of the resource could not be persistent, so the persistent name of the resource can help to track the resource and find it whenever the user needs or wants it.

In August 1998, based on RFC 1738 and RFC 1808, the RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax has been carried out by Tim Berners-Lee and R. Fielding and L. Masinter. This is the first well-formed standard of URIs, and it specified the definition of URI: “A Uniform Resource Identifier (URI) is a compact string of characters for identifying an abstract or physical resource.” The document defines “the generic syntax of URI, including both absolute and relative forms, and guidelines for their use” and “a grammar that is a superset of all valid URI, such that an implementation can parse the common components of a URI reference without knowing the scheme-specific requirements of every possible identifier type.” What the document did not defined are a generative grammar for URI, and the issues and recommendations for dealing with characters out side of the US-ASCII character set, which later becoming the motivation and unsolved problem of Internationalized Resource Identifiers (IRIs).

The document summarized the benefits of uniformity: “it allows different types of resource identifiers to be used in the same context, even when the mechanisms used to access those resources may differ”, this helps the different protocols being
developed after that; “it allows uniform semantic interpretation of common syntactic conventions across different types of resource identifiers”, this is the main purpose of uniformity which join all kinds of resource identifiers to the same representation; “it allows introduction of new types of resource identifiers without interfering with the way that existing identifiers are used”, this reserves the possibilities that the new forms of resources and resource identifiers development; and, “it allows the identifiers to be reused in many different contexts”, which represents the core of Web and Internet discussed above. Also, the document defined the resource as “the conceptual mapping to an entity or set of entities, not necessarily the entity which corresponds to that mapping at any particular instance in time.” This supported the possibility of uniformity and its persistence.

There are two documents of URL Schemes came out in November 1999: The RFC 2717: Registration Procedures for URL Scheme Names by R. Petke (UUNET Technologies) and I. King (Microsoft Corporation); and RFC 2718: Guidelines for new URL Schemes by L. Masinter (Xerox Corporation), H. Alvestrand (Maxware, Pirsenteret), D. Zigmond (WebTV Networks, Inc.) and R. Petke. The former described the procedures of the registration of the new URL scheme names, including resource and non-resource schemes. Although most of the schemes have been defined, there always been new demands together with the development of the technology, this produced the way of future extensions. The later described the rules and restrictions of creating new URL Schemes. Till then, the full structure of URIs, URLs, and URNs have been set up.

In a minute of “Future of URI (FURI) BOF” of the meeting of IETF/W3C coordination on March 20, 2001, I found following conversation in the open discussion section: Dave Crocker said “Do people use and hear others use these terms (UR{N,L,I}) with consistency?” Jeff Hodges said “URLs are well understood in literature, but not URNs.” This definitely represents some of the problems. Also, Micheal Mealling said that “what is UR*” is one of the biggest topics on these conferences. This really shocked me. Consider the URI have been developed almost ten years till then, people still need to find out what is UR*. This probably is the most conceptual and theoretical term of the most practical use. Also, in the same section, Larry Masinter proposed a draft, which “In January, the name changed. We don’t call them URIs. We are defining something new, IRIs (Internationalized Resource Identifiers). I would like to get this document out.” This should be the origin of IRIs.

January 2005, the two latest documents have been carried out: RFC 3986: Uniform Resource Identifier (URI): Generic Syntax and RFC 3987: Internationalized Resource Identifiers (IRIs). The definition of IRIs including the basic syntax and normalization rule, the most important thing here is the explanation of the relationship between IRIs and URIs. “IRIs are meant to replace URIs in identifying resources for protocols, formats, and software components that use a UCS-based character repertoire. These protocols and components may never need to use URIs directly, especially when the resource identifier is used simply for identification purposes.” So IRIs replace part of the URIs, where it can, and where it needs. RFC 3986 is updated
from RFC 1738, replace the RFC 2396. It revised the former documents, introduced the Augmented Backus-Naur Form (ABNF) notation of [RFC2234], and made the final determination of the concepts and syntax of the URI. However, it should be modified sometime later.

Through RFC 2396 to RFC 3986, the definitions and distinctions among URI, URL and URN have always been the main problem of the workgroup. In RFC 3986, there is an official description:

“A URI can be further classified as a locator, a name, or both. The term "Uniform Resource Locator" (URL) refers to the subset of URIs that, in addition to identifying a resource, provide a means of locating the resource by describing its primary access mechanism (e.g., its network "location"). The term "Uniform Resource Name" (URN) has been used historically to refer to both URIs under the "urn" scheme [RFC2141], which are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable, and to any other URI with the properties of a name.

An individual scheme does not have to be classified as being just one of "name" or "locator". Instances of URIs from any given scheme may have the characteristics of names or locators or both, often depending on the persistence and care in the assignment of identifiers by the naming authority, rather than on any quality of the scheme. Future specifications and related documentation should use the general term "URI" rather than the more restrictive terms "URL" and "URN" [RFC3305].

This, to my point of view, is the clearest explanation of these three terms, and the final judgment of using URI instead of other two to represent this sort of concept.

January 2002, the W3C/IETF URI Planning Interest Group has turned out, and becoming the URI Interest Group, chaired by Dan Connolly (W3C) and Norman Walsh (Sun Microsystems) on 2005-02-16. The group reviews ongoing work related to Uniform Resource Identifiers (URIs) and Internationalized Resource Identifiers (IRIs) and helps to deploy quality implementations by maintaining testing materials, and chartered through 28 February 2007. The upcoming activity is to update RFC 2718.

The standards of URIs are created from the demands of information sharing based on the Internet and World Wide Web, they defines a set of rules of the representation of the identifiers, locators and names. However, it still left out some of the definitions for instance the URI to the real world or the relocation of the resource. The IETF and W3C carried out these standards to reinforce the WWW technologies. The standards are good, but not perfect; there are still so many problems to solve: the IRI, the persistence of URN, and the length of URL. Together with the development of the Web technology, more and more protocols would be involved and circumstance could changes. Keep on updating is the best way to catch up with the evolution.

UR* is a set of important terms, which related to almost all the Web technologies
and activities. For instance in XML, retrieving specific piece of document could directly related to URI, not to mention any other location based applications. As a matter of fact, the whole Web is built on Hyperdocument structure as I mentioned in the beginning of this article, as the main implementation of hypertext structure, the significance of URI is very obvious.

People always mistaken these three terms as I discussed above, however, together with the original motivation he provided, Dan Connolly, 2005, proposed his understanding of them which I agreed: “In information management, persistence and availability are in constant tension. This tension has led to separate technologies for Uniform Resource Names (URNs) and Uniform Resource Locators (URLs). Meanwhile, Uniform Resource Identifiers (URIs) are designed to serve as both persistent names and available locations.” It is about persistence and availability, URN is more persistent while URL has more availability.

Till now, all the purposes driven by the original demands have been achieved. While the demand evolved with the development of the technology, further improvements should be made, for instance the IRI. Although the IRI is being used all over the world, it still full of flaws. There usually being some trouble with other languages especially in languages other than English/Latin family, like Chinese. It is interesting to see that almost no one fighting against URIs, URLs and URNs, however, only a few knows systematically how they works.
References

Naming and Addressing: URIs, URLs, ...
   http://www.w3.org/Addressing/#background

Connolly D. (2005). Untangle URIs, URLs, and URNs

Uniform Resource Identifier (URI) Activity Statement
   http://www.w3.org/Addressing/Activity

URI Interest Group
   http://www.w3.org/2001/12/URI/

URI Interest Group Charter
   http://www.w3.org/2004/07/uri-ig-charter.html

URIs, URLs, and URNs: Clarifications and Recommendations 1.0
   http://www.w3.org/TR/2001/NOTE-uri-clarification-20010921/

Tim Berners-Lee's original writings on Document Naming, part of Design Issues for the Web
   http://www.w3.org/DesignIssues/Naming

Uniform Resource Identifiers (URI) Charter

Hypertext Style: Cool URIs don't change
   http://www.w3.org/Provider/Style/URI

Future of URI (FURI) BOF
   http://www2.ietf.org/proceedings/01mar/minutes/furi-m.txt

RFC 1630: http://www.w3.org/Addressing/rfc1630.txt
RFC 1737: http://www.w3.org/Addressing/rfc1737.txt
RFC 1738: http://www.w3.org/Addressing/rfc1738.txt
RFC 1808: http://www.w3.org/Addressing/rfc1808.txt
RFC 2141: http://www.ietf.org/rfc/rfc2141.txt
RFC 2396: http://www.ietf.org/rfc/rfc2396.txt
RFC 2717: http://www.ietf.org/rfc/rfc2717.txt
RFC 2718: http://www.ietf.org/rfc/rfc2718.txt
RFC 3987: http://www.ietf.org/rfc/rfc3987.txt
Appendix: Chapter 1.3: URLs, URIs, URNs, and IRIs from:


URLs, URIs, URNs, and IRIs

HTML documents and other Web resources are denoted by Uniform Resource Locators (URLs), which typically have a familiar structure:

http://www.w3.org/TR/html4/ 

This URL consists of a scheme (http, which identifies the HTTP communication protocol), a server (www.w3.org, the machine where the resource is located), and a path (TR/html4/, typically corresponding to a file on the server).

However, URLs are just a subset of the more general concept of Uniform Resource Identifiers (URIs), which are meant to describe all points in the information space, even those that do not have a physical presence. The actual syntax of URIs is very permissive and includes many other schemes besides http, such as ftp, file, mailto, imap, or https. The official register of schemes is maintained by The Internet Assigned Numbers Authority (IANA).

In fact, a general URI essentially has the form

scheme: scheme-specific-part

So everything following the first colon depends entirely on the choice of the scheme. However, all schemes satisfy certain informal rules: the slash character (/) always implies a hierarchical structure (as in directory paths on operating systems), a question mark (?) separates query strings from queryable resources, and the hash character (#) is reserved as a delimiter separating a URI from a fragment identifier. Special symbols are escaped using the notation %NN with character codes written in hexadecimal; and only US-ASCII characters are permitted.

The concept of a relative URI is used to specify a URI that should be resolved relatively to a base URI. For example, in an HTML document with URI

http://www.w3.org/TR/html4/ 

the relative URI

sgml/dtd.html

is resolved to mean the absolute URI

http://www.w3.org/TR/html4/sgml/dtd.html
For HTML documents, the base tag can be used to specify a base URI different from the document URI. A fragment of a resource can be identified by appending a # character and a fragment identifier to a URI:

http://www.w3.org/TR/html4/#minitoc

This reference identifies the fragment named minitoc in the resource located at http://www.w3.org/TR/html4/. A special class of URIs, called Uniform Resource Names (URNs), is defined by the urn scheme. A URN is a pointer to a resource, but without a reference to a particular location. As an example, the URN

urn: isbn: 0-471-94128-X

Denotes a particular book title identified by an ISBN number, which is clearly a specific piece of information that does not have a physical location on the Web. In contrast, a URL specifies the location of a resource.

There has been great confusion about the terms URL, URI, and URN. One reason for this is that URNs are not widely used, and ‘URI’ and ‘URL’ are often used interchangeably. In the description above, we apply what is called the contemporary view on this issue.

As mentioned above, the syntax for URIs is restricted to use US-ASCII characters, which precludes the use of international characters. To remedy this shortcoming, the concept of Internationalized Resource Identifiers (IRIs) is being developed. An IRI is defined to be a string that maps to a URI when subjected to a specific encoding function. Unfortunately, this function is rather complicated, since different encoding schemes currently are used for domain names and other parts of the URI string. For example, the IRI

http://www.blabaergrod.dk/blabaergrod.html

corresponds to the following URI:

http://www.xn--blbrgrd-fxak7p.dk/bl%E5b%E6rgr%E8d.html

The intention is that tools perform this encoding behind the scenes and then work with the encoded URIs. Since the encoding scheme allows arbitrary subsets of characters to be encoded, there are now combinatorially many IRIs that are displayed the same by complying browsers, which is a potential security problem since clients can no longer be sure which server their browser will contact.