Uniform Resource Locators (URLs) and Uniform Resource Identifiers (URIs)

William Stallings

Uniform Resource Locator

A key concept in the operation of the World-Wide Web (WWW) is that of URL. In the defining documents (RFC 1738, RFC 1808), the URL is characterized as follows:

A **Uniform Resource Locator (URL)** is a compact representation of the location and access method for a resource available via the Internet. URLs are used to `locate' resources, by providing an abstract identification of the resource location. Having located a resource, a system may perform a variety of operations on the resource, as might be characterized by such words as 'access', 'update', 'replace', 'find attributes'. In general, only the 'access' method needs to be specified for any URL scheme.

A **resource** is any object that can be accessed by the Internet, and includes directories, files, documents, images, audio or video clips, and any other data that may be stored on an Internet-connected computer. The term *resource* in this context also includes electronic mail addresses, the results of a finger or archie command, USENET newsgroups, and individual messages in a USENET newsgroup.

With the exception of certain dynamic URLs, such as the email address, we can think of a URL as a networked extension of a filename. The URL provides a pointer to any object that is accessible on any machine connected to the Internet. Furthermore, because different objects are accessible in different ways (e.g., via Web, FTP, Gopher, etc.), the URL also indicates the access method that must be used to retrieve the object.

The general form of a URL is as follows:

<scheme>:<scheme-specific-part>

The URL consists of the name of the access scheme being used, followed by a colon, and then by an identifier of a resource whose format is specific to the scheme being used.

Although the scheme-specific formats differ, they have a number of points in common, as we will see. In particular, many of the access schemes support the use of hierarchical structures, similar to the hierarchical directory and file structures common to file systems such as UNIX. For the URL, the components of the hierarchy are separated by a "/", similar to the UNIX approach.

RFC 1738 defines URL formats for the following access schemes:

ftp	File Transfer Protocol
http	Hypertext Transfer Protocol
gopher	The Gopher Protocol
mailto	Email address
news	USENET news
nntp	USENET news using NNTP access
telnet	Reference to interactive sessions
wais	Wide-Area Information Servers
file	Host-specific file name
prospero	Prospero Directory Service

We discuss the more important ones in the remainder of this section.

Hypertext Transfer Protocol (HTTP)

The HTTP URL scheme designates Internet resources accessible using the HTTP protocol and, in particular, designates Web sites. In its simplest form, an HTTP URL has the following format:

```
"http://" <host> [ ":" <port>] ["/" <path>]
```

The syntactic notation on the preceding line is known as Backus-Naur Form (BNF) and is used in the RFCs that define URLs and URIs. A separate document at this book's Web site discusses BNF. Briefly, the above definition states that the URL begins with the string of characters http://. This is followed by a host name; the variable `host` is to be replaced by a specific host name. Items in square brackets are optional.

In the above definition, `host` is the name of an Internet host or a dotted decimal IP address of the host.

The default port number for HTTP is 80. Most access schemes, including HTTP, designate protocols that have a default port number; for HTTP, it is 80. Another port number may be optionally used and, if so, is supplied by a colon and the port number following the <host> value.

If the `path` portion is omitted, then the URL points to the top-level resource, such as a home page. For example,

```
http://ietf.org
```

points to the home page of the IETF Web site. A more complex path points to hierarchically subordinate pages. For example,

```
http://ietf.org/rfc.html
```

points to the RFC repository page on the IETF Web site. An HTTP URL can also point to a document available via the Web, such as

```
http://www.w3.org/Addressing/rfc1738.txt
```

This is the URL of RFC 1738, available through the Web site of the WWW Consortium.

An HTTP URL can also include, after the remainder of the URL, a *search part*, giving the URL the form:

```
"http://" host [ ":" port] ["/" path] [ "?" search]
```

When present, the `search` part designates a query that will be invoked when the resource is accessed.

File Transfer Protocol

The FTP URL scheme designates files and directories accessible using the FTP protocol. In its simplest form, an FTP URL has the following format:

```
"ftp://" host "/" directoryname [ "/" filename]
```

As an example, consider the document `index` on the anonymous FTP server rtfm.mit.edu in directory `pub`. The URL for this file is

```
ftp://rtfm.mit.edu/pub/index
```

The URL for the directory is

```
ftp://rtfm.mit.edu/pub
```

And the URL for the FTP site itself is simply

```
ftp://rtfm.mit.edu
```

A more general form of the FTP URL is

"ftp://" [user ":" password "@" host] [":" port] *["/" directoryname] ["/" filename]

Some FTP sites require that the user provide a user id and a password; these are provided in the form `user ":" password` and the @ symbol preceding the `host` value. The next new item in the format is `port`. The default port number; for FTP, is 21, but another port number may be optionally used.

After the specification of the host, with an optional user-ID and password, and an optional port number, a slash indicates the beginning of the directory/file designation. There may be zero or more directory names in the designation. In effect, each directory name is an argument to a cwd (change working directory) command, such as is used in UNIX. The `filename` value, if present, is the name of a file. Here is an example of an FTP URL that involves multiple levels of directory:

```
ftp://ftp.intel.com/pub/ietf/ippm/README
```

This designates the file `README`, in the directory `ippm`, which is in the directory `ietf`, which is in the directory `pub`, which is at the ftp site `ftp.intel.com`.

Electronic Mail Address

The email URL scheme designates the Internet mailing address of an individual or service. When invoked by a Web client, it triggers the creation of an email message to be sent by Internet electronic mail. For example,

```
mailto:web-human@w3.org
```

designates the email address of the Webmaster for the WWW Consortium Web site.

Uniform Resource Identifier

URI is a term for a generic WWW identifier. The URI specifications (RFC 1630, RFC 2396) define a syntax for encoding arbitrary naming or addressing schemes, and provide a list of such schemes. The concept of a URI, and in particular its details are still evolving. The URL is type of URI, in which an access protocol is designated and a specific Internet address is provided.

The potential advantage of the URI is that it decouples the name of a resource from its location and even from its access method. With the URL, a specific instance of a resource at a specific location is designated. If there are multiple instances, and that specific instance is unavailable at the time of a request, then the requestor must determine an alternative URL and try that. In principle, with a URI this process could be automated. In practice, documents such as the HTTP specification refer to the use of URIs, but are currently implemented using only URLs.

Note: The WWW Consortium, which develops interoperable technologies and specifications for the Web, states that the URL is an informal term no longer used in standards specifications and that URI should be used instead. However, the latest IETF standard on the matter, RFC 2396, makes the distinction between URI and URL explained in the preceding paragraphs.

Note: In some earlier RFCs, URI was considered to stand for universal resource identifier. This latter term is still sometimes used; the meaning is the same as that for uniform resource identifier.

TO LEARN MORE

The WWW Consortium page on URL and URI at http://www.w3.org/Addressing is useful. The following RFCs are referenced in this document:

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*. June 1994.

RFC 1738 *Uniform Resource Locators (URLs)*. December 1994.

RFC 1808 *Relative Uniform Resource Locators (URLs)*. June 1995.

RFC 2396 *Uniform Resource Identifiers (URI): Generic Syntax*. August 1998.