The User Interface of URLs

Paul E. Hoffman <phoffman@proper.com>

Abstract

URLs (Uniform Resource Locators) have rapidly become the standard method for specifying how to access information on the Internet. Although mostly used on the World Wide Web, URLs are also becoming more common for specifying locations for other distributed Internet services such as Gopher and anonymous FTP. Internet users see URLs both online and in print, and therefore URLs have visual interfaces. This paper gives an overview of many of the issues that concern the visual and user interfaces of URLs.

1 Introduction

To date, most people consider it sufficient for a URL to just be correct: they do not care how a particular URL looks. However, URLs are often quite intimidating to novice (and not-so-novice) users because of their structure. Further, many current URLs are quite difficult to type or even to read, making them seem mysterious and possibly scary to a large number of users.

With a little thought to their presentation, however, URLs for most Internet services can be made more inviting and useful. This is not to say that the specification for the structure of URLs must be changed: instead, people who create Internet services need to think more about how to name them so that they will appear better in URLs.

This paper describes some of the problems that are common in current URLs seen on the Internet. Solutions to these problems are not always possible, or even desirable in all cases. There are certainly no "right" or "wrong" ways to name URLs. However, if Internet users keep these interface guidelines in mind when they create Internet resources, it is likely that the URLs for those resources will be easier for other people to use.

2 History of URLs

URLs have been part of the Web since its inception in 1990. The Web was originally conceived as a way to link information in many formats (text, hypertext, pictures, sound, video, and so on). A central idea from the beginning was that a Web document could refer to any other document on the Internet as part of its content. [1]

Documents, whether they are in print or electronic form, often refer to other documents. There is usually a standard way to make references in one document to other documents. In printed books, for example, when you want to refer to another book, you usually refer to it by title. You might also refer to it by its International Standard Book Number (ISBN). At the time the Web was created, there was no standard method to refer to electronic documents. Thus, URLs were created as a reference mechanism.

It is important to understand that a URL describes a document by its location. This is somewhat akin to describing a book in a library by its location on the shelf instead of by its name. If you look at the Internet as a single, huge library, it is reasonable to describe documents by their location because there is currently no standard way to describe them by name.

3 Definition of URLs

The syntax for URLs is described in RFC 1738. [2] This document, which at the time of this writing is on the standards track within the Internet Engineering Task Force (IETF), is a product of much discussion (and a fair amount of debate) in the IETF Uniform Resource Identifiers (URI) Working Group. [3]

3.1 URL Schemes

URLs consist of two parts: a scheme, and a scheme-specific part. These two parts are separated by a colon. The schemes that are defined in RFC 1738 are:

ftp -- File Transfer Protocol

http -- Hypertext Transfer Protocol

gopher -- The Gopher protocol

mailto -- Electronic mail address

news -- Usenet news

nntp -- Usenet news using NNTP access

telnet -- Reference to interactive sessions

wais -- Wide Area Information Servers

file -- Host-specific file names

prospero -- Prospero Directory Service

Of these, "nntp," "wais," and "prospero" are rarely seen in use. Other schemes have been proposed in the URI Working Group, and the list of schemes will probably grow slowly over the coming years.

3.2 URL Syntax

Each scheme has its own rules for the scheme-specific part. Many schemes have a very similar look, based on what RFC 1738 calls the "common Internet scheme syntax." That syntax is used in the ftp, http, gopher, telnet, and wais schemes, and has the now-familiar "//" followed by a host name, another "/", and a path to the information. In addition, schemes using the common Internet scheme syntax can also optionally specify a user name, a password, and a TCP port number in the syntax.

Two schemes that do not use the common syntax are mailto and news. The scheme-specific part of a mailto URL is simply an Internet mailing address; the scheme-specific part of a news URL is simply the name of the Usenet newsgroup.

This description of the basic URL syntax starts to show some basic user interface issues for URLs. For many users, the "://" part of the common syntax reeks of technobabble. Since few users understand TCP ports, an additional ":" and number after the host name is also confusing, since there is no description of what the number means.

3.3 URL Encoding

The character set used in URLs is the US-ASCII character set, [4] minus some punctuation characters that RFC 1738 define as reserved and "unsafe" (prone to transmission errors or misinterpretation by systems on the Internet). US-ASCII is only one of the many character sets for which there are international standards, and it contains many fewer characters commonly used in human languages than other character sets. Thus, the choice of US-ASCII impacts the user's ability to name URLs if the desired names contain letters that exist in other character sets but not in US-ASCII.

Further, not all characters in the US-ASCII character set can be used in a URL. RFC 1738 lists many punctuation characters that must first be encoded as hexadecimal before being included in a URL. For example, the tilde character (~) is an "unsafe" character, and must not appear in the URL. Instead, the Web user must encode the tilde as the characters "%7E". Thus, the URL:

ftp://ftp.bigcompany.com/~chris/3q95-report

is not allowed. Instead, it would be:

ftp://ftp.bigcompany.com/%7Echris/3q95-report.

4 Common Problems With URLs

As the Web grows, more and more people who aren't familiar with user interfaces and design are setting up Internet servers. When documents at these sites are published, they get URLs, usually assigned by the creator of the document using names assigned or allowed by the server's administrator. Unless the people naming the documents and organizing the site are conscious of the ramifications of their naming, they might come up with names that make it difficult for other Internet users to access their documents.

Because a URL simply describes the location of information on the Internet, you don't really "create" a URL. Instead, you place the information on an Internet server and then derive the URL from where you placed the information. The vast majority of Internet servers base the location on file names, leading to URLs that resemble file names in a directory path. On such systems, the choices made about the names of the directories and files, and about the hierarchy, directly affect the URLs for the information on that server.

Some typical problems with URLs found on the Web today are described in this section. Note that these are not things that must be avoided at all costs, but simply guidelines for what to look out for when creating URLs. At many sites, some of these problems are unavoidable without reconfiguring the server software. However, such changes can often be made with the help of the system administrator, particularly if there is a compelling reason to do so.

4.1 Difficult to Type

Many URLs appear on paper and the user must type them into an Internet client program such as a Web browser. If the URL is particularly difficult to type, the user is prone to making typing errors. Some examples of URLs that are difficult to type include:

- Names that are unnecessarily long:

http://www.bigcompany.com/new-server-version-press-release-12-27-94.txt

http://www.bigcompany.com/PR/announcements/1994/dec/new-server-version.txt

- Names with unexpected capitalization:

ftp://ftp.bigstate.edu/pub/docs/OnTBGHill.txt

- Names with unusual punctuation:

ftp://ftp.bigstate.edu/pub/docs/moon_3+manual

It should be noted that some servers do not use the traditional directory and file names for their URLs. These sites may serve documents from a database, giving the URL simply the name of the database entry, or they use aliases at the root of the directory tree for their URLs. The URLs for these sites are often much shorter and easier to type than those that are based on full directory paths where the creator of a document does not get to specify the path or the names used on it.

4.2 Hard to Represent on Paper

Many characters are difficult to read on paper. One of these characters, the tilde (~), is unfortunately common in URLs due to a misfeature of some HTTP server software. With some printed

fonts, the tilde character can be mistaken for a quotation mark, particularly by people whose language does not include it as a diacritic. Further, as seen above, the tilde character is not even supposed to appear in URLs, since it must be encoded according to the rules of RFC 1738.

Printed URLs can contain characters that a reader may misinterpret when they type the URL into a computer program. The letter "l" looks like the digit "1"; the letter "O" looks like the digit "0"; and semicolons can be mistaken for commas. Depending on the design of the font, a double hyphen (--) can look like a single hyphen, a back quote (`) can become almost invisible, and so on.

In recent years, magazine publishers have discovered that a few characters that are normally used in typesetting also appear in URLs in such a way as to make it likely that the URL will be shown incorrectly unless the work is carefully proofread. Many typesetting systems use the tilde (~) as a special character that must be specially marked in order to appear in print; forgetting to do so causes the character to disappear from the printed text. In other typesetting systems, back quotes (`) and percent signs (%) also cause problems. Proofreaders must be careful to check the original text of a URL against what appears on the printed page when the URL contains any of these special characters.

4.3 Problems in the Host and Port

The scheme-specific part of the URL is not the only place where users can encounter difficulty. The domain name and port number of the Internet server can also cause confusion for some users, particularly those not familiar with the rules of the domain name system and TCP ports.

A pattern of domain naming has developed in the past few years, partially due to the popularity of the Web. Most hosts that serve particular Internet protocols indicate the protocol names in their domain names. For example, if there is an FTP server at bigstate.edu, the domain name for that host is usually "ftp.bigstate.edu". There have been many arguments about the logic of this arrangement, but the pattern has been established throughout the Internet.

This has led to users expecting these domain names, even though they are voluntary and not terribly meaningful. If a host has the "wrong" name, it can confuse users who expect names to follow the pattern. For instance, if the host name part of an http URL was "web.bigcompany.com" or "http.bigcompany.com", the user might not notice that the name was different than the expected "www" and type the URL with "www.bigcompany.com". Fortunately, this problem is easy to remedy: system administrators can give all the expected names to a single host computer.

URLs that include a TCP port number are also likely to cause confusion and typing mistakes. Most Internet users understand a bit about domain names and understand directory trees, but few understand what TCP ports are. This is a particularly thorny issue for a single host that serves information differently from different ports.

For example, assume that a user sees the URL:

http://admin.bigstate.edu:8001/docs/thesis/jones

They may at first type the URL without the port number. If there is also an HTTP server at the standard port, port 80, the user will access the wrong server, and will probably get an error message stating that the document doesn't exist. Similarly, if they type "8000" instead of "8001" and there is a server on that port as well, there may be no indication that they are on the wrong server.

5 Human Language Issues in URLs

Most URL names have words and names that are based on human languages (as compared to numbers or other sets of symbols). Even domain names have a strong relationship to human languages. This becomes important when thinking about the user interface of URLs because it is unwise to assume that everyone using a URL will speak the language(s) embodied in that URL. Even if they do understand the language, they may not understand it well enough to derive meaning that the URL creator might expect.

Throughout this paper, the URLs given as examples have all used English as a base language. In fact, the vast majority of the Internet is English-centric, even though many people on the Internet speak little or no English. This emphasis on English is pervasive on the Internet, not just in URLs.

Naming URLs with languages other than English brings up the character set issue discussed earlier in this paper. In many languages, diacritic characters are essential for differentiating words, yet these characters don't exist in the US-ASCII character set. In many cases, the lack of international characters prevents people from using "logical" names for URLs when those names contain characters not represented in the US-ASCII character set.

The issue of human languages also goes a bit deeper in that some names used in URLs have different meanings in different languages. A word or name in one language can have quite different meanings in other languages. Further, there are also cultural meanings for the same word in the same language used in different countries. For example, there are words commonly used in American English that would be considered rude or vulgar in British English and vice versa.

6 URLs That Help Users Hunt

It is common for Internet users to glean information from URLs in order to help them hunt

on the Internet for additional information. Of course, this practice often leads to unsuccessful searches, but modifying parts of a URL and submitting the changed URL to a server is a common search technique. The creator of a URL can help foster this kind of search with the names they choose.

For example, assume that a user sees the following URL:

http://www.bigstate.edu/math/yee/course-notes.html

There are many guesses that a user can make from this URL:

- A person named Yee is somehow affiliated with the Math Department at Big State University
- That person has a directory which may contain files other than his or her course notes
- The Math Department may have other directories for other people in that department
- The Physics Department may have its own directory with a similar structure

Of course, any of these guesses could be wrong, but if you wanted to find out information about someone with the last name "Smith" in the Physics Department at Big State University and you had seen the URL above, it would be worth a try to request:

http://www.bigstate.edu/physics/smith

7 URLs That Dissuade Users From Hunting

On the other hand, someone publishing URLs may not want people to infer anything from those URLs. Some sites want to dissuade people from using this kind of guessing and prevent people from aimlessly poking around their Internet server. It is just as easy to create "opaque" URLs, ones that make the structure of the information on the server difficult to guess.

Instead of the URL shown previously for Yee's course notes, the server might instead make them available with the URL:

http://www.bigstate.edu/j37od

The course notes would be the same in both cases, but someone seeing this URL would not have any clues how to find information about anyone else in the Math Department, or any other department for that matter.

8 Conclusion

This paper has raised many topics that people creating URLs should consider when they name their Internet documents. Some problems are simple to avoid, others take more planning and possibly the assistance of a system administrator or site manager.

References

[1] A complete history of the Web can be found at <http://www.w3.org/hypertext/WWW/History.html>

[2] Berners-Lee, et. al., "Uniform Resource Locators (URL)," RFC 1738. <ftp://ds.internic.net/rfc/rfc1738.txt>

[3] For more infomation, see: http://www.ietf.cnri.reston.va.us/html.charters/uri-charter.html

[4] "Coded Character Set -- 7-bit American Standard Code for Information Interchange," ANSI X3.4-1986.

Author Information

Paul E. Hoffman, President
Proper Publishing
127 Segre Place
Santa Cruz, CA 95060 USA
Tel: 408-426-6222
phoffman@proper.com