

HTTP 404 - File not found

By Mick Topping (Feb 01)

This article is mostly for folks who are somewhat new to web browsing, and are frustrated with the number of errors they get while surfing. I tried to spell-out the acronyms, and reduce jargon, but it is not easy for me, because I have been playing in this sand pile for a long time. I will not be sure how well I have succeeded, until feedback comes in. If I have left anything un-explained, drop me an email mtopping@mchsi.com or just post a note on the NuCom egroup. You should remember that I am not a guru, just persistent. I don't have any real training or credentials on any of this stuff, so use these hints at your own risk.

Remember that web page that you stumbled across some while back? It was the one that had the complete instructions on how to solve that terrible problem that you have never had, but well, just to be on the safe side, you book-marked (listed it in your favorites) it anyway. After all, it was a pretty cool page that had lots of other information somewhat related to stuff that you are interested in. Some time passes, and you finally install the new software (that you got for Christmas) that was discussed on that page. Now your desktop went all squirrely, the tool bars are chartreuse, and your icons are labeled in Chinese, and now you NEED those instructions. Thank goodness you book-marked that page.... "click... whiirrr....beep..."

...some generic **NOT FOUND** message ...
Used-to-be most times it was a "404-filenotfound" but now sometimes you get some more help.

AAARRRGUHHHH. Now, what? Even when it is someone else's link, that they created, with the thought that the linked page would be useful, even then it is annoying. When it is a page you know you want, it is particularly frustrating, and now, you can't see it. How many times have you had this happen? You click the link, or book-mark, and instead of your

information, up comes one of the dozens of varieties of little error messages, each one, explaining in impenetrable geek-speak, what has gone wrong. The Internet and world wide web are complex places, and lots of things can cause a broken link... (When I use the term "broken link", it refers to a link, that when you click it, does not cause the desired page to be displayed as intended.)

What can you do when a link is broken? Well, sometimes it is just broken, and you just can't get there from here. But often there are things that you can try. First lets see what a "link" does, then take a look at what could be wrong, and what we can do about it. First, lets take a look at what happens when everything works, just to get an appreciation of all the things that have to work right.

The link just loads a URL into the "location" or "address" window of your browser. So, what does the URL mean? OK, URL stands for Uniform Resource Locator, but that really doesn't clear up much. What a URL is really, is an instruction to the browser, Netscape, or Internet Explorer, to go somewhere, get some information, and display it. That is, display it according to its contents, and file extension. Lets break the URL into parts, and look closer at each part.

Given the following sample URL, what do the parts mean?

<http://www.visi.com/~nathan/humor/compjokes/winedlin.html>

- `http:` - This is the protocol, essentially, instructions to the browser, concerning what type of transfer to conduct. Usually (99.9% of the time) it will be `http`, but it could be something else, like `ftp:`, or `mailto:`, and maybe others. (Actually, I did think of a few more that I have used, but I think I will save that for a future article.

This one is already running too long.) HTTP stands for Hyper Text Transfer Protocol. Briefly HTTP defines how messages are formatted and transmitted, and what actions Web servers and browsers should take in response to various commands. HTTP is the basic protocol for webpages using *links*.

- `www.visi.com` -- The part of the URL just after the “http://” part and before the next “/” is generally called the *domain name*. I am not sure that it formally accurate, but what seems to work, for me, is to think of this as the name of the computer that holds the file (the web page file) that you want to retrieve. This part is NOT case sensitive.
- `~nathan, humor, compjokes` – Again, I am not sure if this is formally true, but it seems to work if you think of these terms after the domain name and before the last “/” as folders and sub-folders (or directories) on the computer specified by the domain name. This part IS case sensitive.
- `winedlin.html` – Finally, the part, after the last “/” is the name of the file that you are asking the browser to retrieve. Also case sensitive. (A web page is usually composed of multiple files, but there is always (usually) one master file that contains the instructions for getting and displaying all the others.)

To break it down just a little further, the file name has two parts. The main part of the file name (before the last period) is the author’s discretion, his name for the file. The 2-5 letters after the last period is called the extension, and it is assigned to tell computer how to interpret the file. The extension names the language that the file “speaks” so the computer is ready to translate.

Historically the file had a suffix of “html” (hypertext mark-up language) or “htm” (the DOS truncation of html.) HTML, or HTM, signify that this file is designated as being in the fundamental language of the browser. Lately extensions like “asp” (active server page), or “cgi” (common gateway interface) are very common. These newer pages are “active” in the sense that the results displayed on your computer will change with time, or depending on whether you have visited the page before. These active pages are generally associated with larger commercial sites, access large databases, and are less prone to produce 404 errors. But when you have errors associated with “asp” and “cgi” files, they are usually not fixable with the following methods.

In fact the browser will retrieve files with any extension, but if it doesn’t know what to do with it, you will have to decide. OK...where was

I going with this, (looks quickly back through notes) AH! Yes why do things break? Well, my experience one of the very common causes of broken links is the web page author, or web master changed the file name. When this happens, and you click the link, you have essentially asked for a file that no longer exists on that computer, in that folder. If there is no file of that name, what can you do? Guess what it may be renamed to? **Test Question:** (*what*, you didn’t know there was going to be a test on this material? You bet there is, and the grade is going on your permanent record.) if you had renamed a file on your computer, and couldn’t remember its new name, how would you find it? **Answer:** look in the folder where it was located for similar, meaningful names. That is also the first thing to try when encountering a broken link.

Now, you can’t always look in a folder on the other computer. Frequently, the attempt to look in a folder will prompt a notice that you are not authorized to look in that folder. But pretty commonly on smaller pages where people are not too paranoid, you will be able to just click the “address” line, and edit it to delete the file name (in our example this is `winedlin.html`), back to the last “/”, and hit return. This asks your browser to produce the folder `compjokes`, which is interpreted

by the browser into, “what he really wants is a list of what is in that folder”. Now, if all goes well, you can look down the list of files, looking particularly for “htm” or “html” files that look like candidates for the new name of the file you want.

Now, suppose the broken link was caused not by a renamed file, but a moved file. (suggested by the fact that you couldn’t find it in the first attempt to look in a folder) Again, edit the URL, this time take off the lowest level folder name (in our example this would be (compjokes),.If this does not produce a listing, delete the next level folder (e.g. humor), Usually after digging up 2 or 3 levels, if you haven’t found your file, you are not going to find what you are looking for in this way, but you will have to decide how bad you want that file to decide how long to look.

If a link was sent to you in an email, maybe the sender mangled it a bit. An interesting illustration of the complexity net space, is that sometimes the web is case sensitive, and sometimes it is not. (As near as I can figure it, if the web server is a Windows computer, it is not case sensitive, but if it is a UNIX/LINUX server, it is case sensitive. So something to try, is make sure the name that your buddy sent you is all lower case. (This is the most likely correct format.)

But what if we have tried all ways to find the misplaced or renamed file, and still can’t find it. Well, it is possible that the file is no longer there, deleted, erased, history... There is just no way to get it right? Well, maybe. But maybe, oh, the guy changed ISP’s, and moved his file to a new domain (think: another computer). Do you know anything about the file, maybe some of the text? Something you can search for? Next step google.com. You get to google.com by typing its name in the “address” window.

Google is an exceptional search engine. If you can remember a bit about the webpage, it will find it for you. (Of course it may also find another 2 million similar pages that you will have to filter out somehow—try to add more words that may have been in the file you were searching for.) OK, you get it just right, and Google finds

the file you were looking for, great! you click the link to it, and (are you ready?) there again you get the dreaded “http 404 file not found”, sure enough, the guy has taken his web page down, and gone out of business. You can’t retrieve a file that is not there. Or can you?

Like I said, Google is an exceptional search engine. An incredible percentage of the web pages that the search finds are “cached”. This means that Google has kept a copy. Very frequently, when you get a 404 error on a page that Google has located, backing up, and clicking the little green “cached” link in the last line of the Google reference will produce useful results. Maybe not all the graphics will show, and the links on the page may not work. But it is worth a try.

There you have it. Now when you find broken links, remember these steps, and you will have tools to repair a lot of the breaks.

Happy surfing.

References:

<http://webopedia.internet.com/TERM/U/URL.html>
http://www.aboutdomains.com/News/Glossary_Terms.htm
<http://www.e-mia.org/infolit/basics2.html>
<http://www.pollycyber.com/ps367/ch02-penultimate.htm>
(these all worked when this was published, may be broken now though)