Some Website Link Checking Tools and Utilities

Don Lancaster
Synergetics, Box 809, Thatcher, AZ 85552
copyright c2010 pub 5/10 as GuruGram #105
http://www.tinaja.com
don@tinaja.com
(928) 428-4073

A frustrated website visitor is a gone website customer. Little infuriates more than the dreaded "**404**" or **file not found** error. On a larger website, continuous vigilance is demanded to both prevent and minimize links that go nowhere or end up routing to wildly incorrect places.

I recently went through a major overhaul of our **Guru's Lair** website and ended up correcting many hundreds of broken links that accumulated over the years. In the process, I found no "best" solution. So, what I thought I'd do here is gather together some of the more useful tools and utilities useful for **404** reduction on any larger website.

Error Sources

There are often different types of 404 and related errors. Some of which you can deal with and some that may be well beyond your control...

BROKEN INTERNAL ANCHORS — These do not generate an actual 404 error but still can be quite annoying. Easy navigation is a key feature of any successful web site.

BROKEN INTERNAL LINKS — Links which go nowhere on your own site, most likely due to a typo or a file that is not online or not where you thought it was.

WRONG INTERNAL LINKS — These do not 404 but still will go to the wrong place. Usually because they were not properly updated during a cut and paste.

BROKEN EXTERNAL LINKS — These might be your own typos or the external site may no longer be available. Or they renamed some files without doing a redirect.

FUMBLE FINGERED USERS — Entry errors that largely are beyond your control. These largely should be non-repeating one time events.

DELIBERATE PHISHING ATTEMPTS — Visitors trying to access hidden or supervisory files on your website. Sometimes seeking a .PHP extension. They, of course, deserve all the 404's they can get.

The Tools

Sadly, many of the **404's** will be beyond your control. Others may have a different error number, but still need corrected. Such as an error **500** from too many w's in www. Or not enough ht's in html. Or mixing up a period with a slash.

And others may remain stuck in caches web wide. These may not go away for weeks or even months after you tried to fix them. Two reasonable goals are to…

- **Continuously minimize your known 404's.**
- **Aim for a gross error rate under ONE PERCENT.**

Here's my collection of tools and utilities that I have found more or less useful in minimizing long term **404** and similar errors on larger websites…

1. **ETERNAL VIGILANCE** — Most errors are introduced on their initial upload. So, **you should personally and manually test and verify each new website page and each and every significant page revision**.

2. **CUSTOMER COMPLAINTS** — Some errors will rarely or never receive clickthrus. But **if any visitor reports anything, stop and fix it NOW!**

3. **THIRD PARTY SEARCHES** — If you have an employee temporarily out of crucial stuff to do, **have them go through a large chunk of the website, manually checking for any and all problems**. It is important to emphasis accuracy and being meticulous. They also have to spot any website quirks, such as some files being standalone and others in archives. Naturally, looking for typos and spelling errors should also happen at the same time.

4. **FIREFOX LINK CHECKER PLUGIN** — This is the fastest and easiest link checker I have found to date. It is free and found **here**. Color coded boxes rapidly appear after each and every link after right clicking. One page at a time is handled. Negatives include it seeming to report broken anchors as valid and being disabled on .PDF files. It is, of course, **Firefox Browser** specific. Link locations are obvious.

5. **W3C TOOLS** — They have a free link checker **here** and a website validator **here**. Link checking is somewhat slow but quite thorough. They only tell you a broken link exists, not where in the file it occurred.

6. XENU — The **Xenu** link checker is blindingly fast and ends up quite convenient. It uses multithreading to check many links at once. But it is so fast my ISP throttles and hangs it, typically when it is 95 percent complete. It is quite thorough on anchor testing, but does not handle .PDF file links at all.

7. LINK CHECKER PRO — This is for-sale **shareware** that seems to limit you to only a few free trials. Its reports are detailed, thorough, and complete. It is fascinating to watch in action. But it cannot handle .PDF files. There seems to be no obvious way to find any exact error locations within a given page under test. I felt this offering seems way overpriced and demands excessively snotty user restrictions.

8. YOUR OWN LOG FILES — Your web log files will continuously report all 404's and similar internal errors for you. If their log **referral** field is activated, the source of the errors can also be reported or at the very least hinted at. **Always DEMAND log file access from your ISP!** With full referral field activation! Your log files can easily become your single most important web improvement tool.

9. CUSTOM LOG FILE PROCESSING — A number of pricey services can process log files for you. But your own custom routines can do even more. My **ANALOGEB.PDF** and **HISTOLOG.PDF** are a tutorial pair that describes my **LOGRPTX2.PSL** and its updated **NEWEBAY1.PSL** utilities. These give you detailed 404 reports and an amazing amount of other info, even including **eBay** image theft detection! They can get easily customized to extract an utterly amazing amount of info on all your website visitors.

10. HAND WRITTEN CODE — Your own personal routines can greatly ease the 404 discovery problems. Ferinstance, my **ERRSRCHA.PSL** can search a host-based list of website pages named **curgroup** for a list of phrases named **curerr**. And **ERR404A.PSL** searches a host-based list of log pages named **listoflogstoprocess**, and seeks out any 404 errors. **ERR500A.PSL** similarly deals with 500 errors most often typo caused. Accumulated errors emphasize the more urgent ones. A customized search-and-destroy routine is offered as my **PSSEARCH.PDF**.

10. DEALING WITH BROKEN EXTERNALS — Firstoff, don't become part of the problem. **Never let a URL go dark! ALWAYS provide a referral link instead!** Should an external link go dark, try to replace it with its revision or update. Or downgrade the link to their main url. Or replace the link with a nonworking color emphasis. Or you can rewrite your entire description. Or simply delete any and all references to their problem link. Details will vary with the importance of the link to your message. The **Wayback Machine** can sometimes be of help here.

11. ACROBAT .PDF HASSLES — Those **Acrobat** .PDF files are typically compressed and hold their links in a **URI** format described in Adobe's **PDFMark Reference Manual**. The linking process is also detailed in my **TWOWAYLK.PDF** tutorial. Most URL checkers can **not** check a .PDF file, so **a manual verify often may be needed**. Regular **Google** and the **Google Custom Site Search can** do .PDF searches. At one time **Adobe** had useful link checkers (some ready for use and some requiring a C compiling from a **SDK**), but these seem absent in their latest software. I have an older and very primitive .PDF link extractor **here**. It demands an uncompressed .PDF file, so a print to disk may be required. A HTML link list can optionally be generated. My self auto-tracking method to **insert** links into a **PostScript** program appears **here**. And a general tutorial on our **Gonzo Utilities here**.

12. WATCH THE .ASP TRAP! — The .ASP file you upload to your ISP will typically **not** be the same .ASP file as your client downloads. Any links provided by an **include** "subroutine" will need separately tested. As will any bells and whistles such as a **banner rotator**.

13. THE XP SEARCH DOG — Copies of your web pages on a host computer can be searched for strings using the XP search dog. Restrict to a subfolder and then do a **control-F**. This can be quite handy to find a really obscure error. It works offline only. But cannot find any phrases in .PDF or any other encrypted or compressed file.

14. WAITING — Even when immediately and properly connected, 404 errors may continue for quite some time. This may be caused by external web caches retaining links. Or by third party cross links that retain typos or other errors. Thus, **it may take weeks or even months for even the most obvious website errors to get picked up**.

15. SOME ALSO RANS — An ideal link checker would combine the ease and convenience of the **Firefox Plugin**, routinely handle .PDF files with aplomb, be throttle proof, handle anchors properly, work either on or off line, and show exact problem source locations. Besides, of course, being free. Several other link checkers that I have yet to check appear **here** and **here**.

For More Help

Further details on web tools and techniques can be picked up in our **Webmastering**, **Acrobat**, and **PostScript** libraries.

Additional consulting services are available per our **Infopack** services on a contract or hourly basis. More **GuruGrams** can be found **here**. Seminars available. Email **don@tinaja.com** or call **(928) 428-4073**.