

5-15-2013

# Linked Up: Using Free Tools to Maintain Access to E-resources

Amelia Mowry

Wayne State University, [amelia.mowry@wayne.edu](mailto:amelia.mowry@wayne.edu)

---

## Recommended Citation

Mowry, Amelia, "Linked Up: Using Free Tools to Maintain Access to E-resources" (2013). *Library Scholarly Publications*. Paper 66.  
<http://digitalcommons.wayne.edu/libsp/66>

This Article is brought to you for free and open access by the Wayne State University Libraries at DigitalCommons@WayneState. It has been accepted for inclusion in Library Scholarly Publications by an authorized administrator of DigitalCommons@WayneState.

Linked Up: Using Free Tools to Maintain Access to E-resources

Poster and Notes

Poster Session for MLA Academic Conference 2013

Presented May 15, 2013

Amelia Mowry

Metadata & Discovery Services Librarian, Wayne State University

[Amelia.mowry@wayne.edu](mailto:Amelia.mowry@wayne.edu)

## Presentation Notes

### How does link-checking software work?

- Link-Checkers work much like search engines. They go to all the links on a page, follow them, then check all the links on those pages and so forth.
- A good link checker also produces broken-link reports that enable you to determine where the broken link is located, and why the link is broken.
- (<http://www.techsupportalert.com/best-free-web-site-link-checker.htm>)

### What Link-Checkers Do:

- Link-Checkers let you know if a link is dead.
  - For example, if the link leads to an error
- Some link-checkers let you know if you've been redirected to a new URL
  - This is a very helpful feature because some sites will re-direct you, at least initially, when there is a URL-change. It's also common in many sites, including database providers, that if the resource you're trying to reach is no longer there you will be re-directed to a main page.

### What Link-Checkers Don't Do (at least for now)

- They will not let you know if the content of the site has changed.\*
  - For example, the resource is gone, but a new page has its URL. You may come across links to sites that are now "Domain for Sale."
- They will not let you know if you hit a paywall.\*
  - Link-checkers look for links that don't work. They won't distinguish between a site providing full-text and one telling you that you don't have access.
- \* One of the link-checking tools I tested is moving towards looking at the content of the pages. This is an exciting area of development for link-checking.

### E-resources links in Libraries

- The links that librarians need to check can be more complicated than typical URLs. This can cause problems for existing link checkers.
- A great resource for learning about how URLs are constructed is Doepud's Anatomy of a URL ( <http://doepud.co.uk/blog/anatomy-of-a-url.php> )
  - URLs for databases or other resources often include search parameters or fragments. Search parameters begin with a "?" and input options to the site, such as selecting specific databases. Fragments begin with "#" and navigate to a specific place in the page, such as a heading further down. This means that URLs can be broken by changes within the site, not just changes to the location of the site.
    - <http://guides.lib.wayne.edu/content.php?pid=62945&sid=1808666>
    - <http://www.loc.gov/standards/mods/userguide/origininfo.html#issue>

- Going through proxy servers allows off-site patrons to access materials, but it can also change URLs.
  - <http://heinonline.org.proxy.lib.wayne.edu/HOL/Welcome>
- Some subscription sites may change the URL to begin a session
  - <http://firstsearch.oclc.org.proxy.lib.wayne.edu/WebZ/FSPrefs?entityjsdetect=:javascript=true:screensize=large:sessionid=fsapp8-42077-hgf3lm7r-bov63i:entitypagenum=1:0>

### The test

- I researched link-checking tools that are available for free online. I selected three of the most highly recommended tools, and I analyzed them with libraries in mind. In particular, I looked at the following questions
  - Does the link-checker allow you to test by domain?
    - This allows you to only check links internal to the site. For example, if I want to check the pages on library services, I could limit by the domain <http://www.lib.wayne.edu/services/>. That way, the link-checker would only check links that begin with that.
  - Does the link-checker allow you to limit by depth?
    - You don't want to link-check the whole internet. Limiting depth allows you to check only the links on a given URL, or you could check the links on that URL and the links on the pages it links to.
  - Does the link-checker allow you to check a text file?
    - This can be a useful function for libraries. We have a list for our databases that can be exported as a text file. There are also ways to export URLs from the catalog. Checking links as a text file can be a good option, because that allows you to manipulate the links. (See Xenu)
- I created an extremely simple web page containing many of the type of URLs librarians are likely to come across. I tested this page using the link-checkers.

valid simple link	<a href="http://nces.ed.gov">http://nces.ed.gov</a>
valid complex link - query with parameters	<a href="http://guides.lib.wayne.edu/content.php?pid=62945&amp;sid=1808666">http://guides.lib.wayne.edu/content.php?pid=62945&amp;sid=1808666</a>
valid complex link - fragment	<a href="http://doepud.co.uk/terms.php#quote-calculator">http://doepud.co.uk/terms.php#quote-calculator</a>
invalid link - no access	<a href="http://chem.sis.nlm.nih.gov/nci3d/">http://chem.sis.nlm.nih.gov/nci3d/</a>
invalid link - typo	<a href="http://www.whitehuse.gov/about/presidents/">http://www.whitehuse.gov/about/presidents/</a>
invalid proxy - resource no longer exists	<a href="http://proxy.lib.wayne.edu/login?url=http://firstsearch.oclc.org/FSIP?dbname=HeritagePrintedBook&amp;done=referer">http://proxy.lib.wayne.edu/login?url=http://firstsearch.oclc.org/FSIP?dbname=HeritagePrintedBook&amp;done=referer</a>
invalid proxy - typo	<a href="http://proxy.lib.wayne.edu/login?url=http://firstsearch.oclc.org/FaIP?dbname=HeritagePrintedBook&amp;done=referer">http://proxy.lib.wayne.edu/login?url=http://firstsearch.oclc.org/FaIP?dbname=HeritagePrintedBook&amp;done=referer</a>
valid proxy	<a href="http://proxy.lib.wayne.edu/login?url=http://heinonline.org/HOL/Welcome">http://proxy.lib.wayne.edu/login?url=http://heinonline.org/HOL/Welcome</a>
site redirected	<a href="http://etext.virginia.edu/japanese">http://etext.virginia.edu/japanese</a>
session begins	<a href="http://proxy.lib.wayne.edu/login?url=http://firstsearch.oclc.org/FSIP?dbname=ERIC">http://proxy.lib.wayne.edu/login?url=http://firstsearch.oclc.org/FSIP?dbname=ERIC</a>

### The results

- Xenu
  - Download from: <http://home.snafu.de/tilman/xenulink.html>
  - Limit by domain: Yes. You can enter a URL and not choose to check external links. All links will show up on the report, but Xenu will only test the ones that begin with that URL.
  - Limit by depth: Yes. However, to check just the URLs on one page by adding a URL, the depth is one. When checking a .txt list, the depth is zero.

- Check text file: Yes.
- Xenu allows you to change settings. One of the most useful settings is to mark redirections as errors.
- Issues:
  - When testing by entering a URL, it strips fragments.
  - It doesn't do well with URLs that begin sessions.
- Summary
  - I tested Xenu multiple times using different settings and processes, including using a URL vs. .txt file, marking redirects as errors, going through proxy.
  - The most accurate results were achieved by using a text file with proxy prefixes stripped off and marking redirects as errors. In this case, the only issues were with URLs that began sessions.
  - If you don't have a .txt file to use, there are two options
    - Multiple Xenu Tests
      - Run Xenu treating redirections as errors and select not to check anything beginning with the proxy prefix.
      - Run Xenu again, this time not treating redirections as errors. This will identify proxied URLs that lead to errors, but it *won't* identify cases where the site redirects because a particular resource/database is not there.
    - Use Xenu to create .txt list
      - Run Xenu on a URL and then export the list as tab delimited.
      - Open list in excel, import as tab delimited, and choose not to import any field except URL
      - Do a find and replace for proxy prefix and replace with nothing.
      - Save this as a .txt file. You can then test this as described above, marking redirects as errors.
- Initial results for testing URLs with sessions were confusing, so I creating a separate test for sessions including the following URLs:
  - Invalid URLs that go through proxy and begin sessions
  - Valid URLs that go through proxy and begin sessions
  - Invalid URLs that begin sessions
  - Valid URLs that begin sessions
  - Test 1: basic settings
    - Identified as wrong only invalid URLs that did not go through proxy, missed the ones that did
  - Test 2: changed setting to mark redirects as errors
    - Correctly identified non-proxied invalid and valid links
    - Flagged invalid proxied links as "object temporarily moved"
    - Identified one valid, proxied URL as "temporarily moved," other OK
  - Summary
    - Results for this type of URL are inconsistent
    - A contributing problem could be that sites that start sessions are very good about redirecting if a database no longer exists.

Xenu can't tell difference between a redirection to session and a redirection to another part of the site.

- LinkChecker
  - Download from: <http://wummel.github.io/linkchecker/>
  - Limit by domain: Yes.
  - Limit by depth: Yes.
  - Check text file: No.
  - Any issues:
    - Difficult to use and personalize without experience working in command line.
    - It won't check pages with robots.txt.
  - Summary
    - This test ended quickly. LinkChecker does not test any page with a robots.txt, which is a file that tells web crawlers not to enter. Lots of propriety resources have this, which means LinkChecker would refuse to test many library resources.
  
- Link Evaluator
  - Download from: <http://evaluator.oclc.org/>
  - Limit by domain: No, but you can select URLs not to test.
  - Limit by depth: Not applicable. Only tests current page.
  - Check text file: No.
  - Any issues:
    - Analyzing results in the interface is confusing. However, it also highlights the links on the page, which is much easier to understand.
    - The customization is exciting, but would require significant time to maintain.
  - Link Evaluator is a firefox plug-in that can only evaluate the page you are currently on. Despite this limitation, it's the one to watch. Rather than crawl like traditional link checkers it actually looks at the html of the page for words that imply a good link or a bad link. For example, if it sees something like "domain for sale" that will be a mark towards bad. It highlights the links on the page green for good or red for bad. Colors range between green and red for pages that are mixed. Link Evaluator also allows you to add phrases to either the good list or the bad list. I conducted additional tests to see how taking advantage of this option would affect results.
    - I tested it on the "A" portion of our Article Databases page. I went through the whole list manually checking each link. It contains 43 databases, and five should have been suspicious. Three had changed or required purchase and two were law school only and required log in.
      - Results of standard Link Evaluator settings
        - 10 bright green
          - Included law school sources
        - 32 light green
          - Included questionable links
        - 1 yellow

- Phase 1: Changed settings added:
  - Bad list
    - buy/Buy
    - purchase
  - Good list
    - log out
    - Wayne State
    - WSU
    - MCLS
    - Midwest Consortium
  - Results
    - All bright green except 1 light green and 1 yellow
    - Didn't identify problem sites at all
- Phase 2: changed settings added:
  - Bad list
    - assword
    - ember Number
    - ubscriber log
    - ubscriber Log
    - ubscribe now
    - (404)
    - Page Not Found
    - ontinue to the new
  - Results
    - Identified a lot more yellows, oranges, and a red
    - Only 2 were from questionable list, rest false positives
    - Could not identify some like law because they didn't make it clear you had to log in initially.
- Summary
  - At the moment, this tool is not very effective, but I'm excited to see how it develops in the future.
- Next Steps
  - Continue to test additional link checking tools and follow the development of Xenu and Link Evaluator

# Linked Up: Using Free Tools to Maintain Access to E-resources

## Using Free Tools to Maintain Access to E-resources

### How Linkcheckers Work

([www.techsupportalert.com/best-free-web-site-link-checker.htm](http://www.techsupportalert.com/best-free-web-site-link-checker.htm))

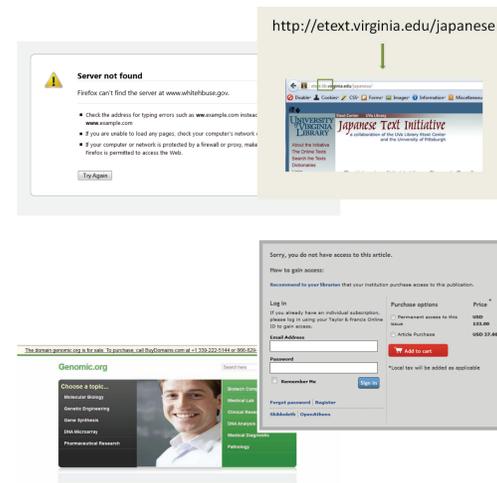
- o Link-Checkers work much like search engines. They go to all the links on a page, follow them, then check all the links on those pages and so forth.
- o A good link checker also produces broken-link reports that enable you to determine where the broken link is located, and why the link is broken.

### What Link-Checkers Do

- o Link-Checkers let you know if a link is dead.
- o Some link-checkers let you know if you've been redirected to a new URL

### What Link-Checkers Don't Do (For Now)

- o They will not let you know if the content of the site has changed.
- o They will not let you know if you hit a paywall.



### What to look for

- Does the link-checker allow you to test by domain?
  - This allows you to only check links internal to the site. For example, to check the pages on library services, I could limit by the domain [www.lib.wayne.edu/services/](http://www.lib.wayne.edu/services/). That way, the link-checker would only check links that begin with that.
- Does the link-checker allow you to limit by depth?
  - You don't want to link-check the whole internet. Limiting depth allows you to check only the given URL and the links on the pages it links to.
- Does the link-checker allow you to check a text file?
  - This can be a useful function for libraries. We have a list for our databases that can be exported as a text file. There are also ways to export URLs from the catalog.



I researched link-checking tools that are available for free online. I selected three of the most highly recommended tools, and I analyzed them with libraries in mind.

	Download from	Limit by domain	Limit by depth	Check text file	Issues
Xenu	<a href="http://home.snafu.de/tilman/xenulink.html">http://home.snafu.de/tilman/xenulink.html</a>	Yes. You can opt not check external links, then add domains to consider internal.	Yes. To check just the URLs on one page by adding a link, the depth is one. When checking a .txt list, the depth is zero.	Yes.	When testing by entering a URL, it strips fragments.  It doesn't do well with URLs that begin sessions.
LinkChecker	<a href="http://wummel.github.io/linkchecker/">http://wummel.github.io/linkchecker/</a>	Yes.	Yes.	No.	Difficult to use and personalize without experience working in command line.  It won't check pages with robots.txt.
Link Evaluator	<a href="http://evaluator.oclc.org/">http://evaluator.oclc.org/</a>	No, but you can select URLs not to test.	Not applicable. Only tests current page.	No.	Analyzing results in the interface is confusing.  The customization is exciting, but would require significant time to maintain.

**Xenu's Link Sleuth**

Best results were from testing the links on-campus as a text file, stripping off the proxy prefix, and treating redirections as errors.

Xenu's Best Results	
Valid simple URL	X
Valid complex URL with parameters	X
Valid complex URL with a fragment	X
Invalid URL with no access	X
Invalid URL with a typo	X
Valid URL with proxy prefix	X
Proxied URL to resource that no longer exists	X
Proxied URL with a typo	X
URL that redirects to a new URL	X
URL to a site that begins a session	

**LinkChecker**

LinkChecker should only be used for open access materials. It respects robots.txt restrictions, which are commonly used on subscription materials.

Xenu's Best Results	
Valid simple URL	X
Valid complex URL with parameters	X
Valid complex URL with a fragment	X
Invalid URL with no access	X
Invalid URL with a typo	X
Valid URL with proxy prefix	X
Proxied URL to resource that no longer exists	
Proxied URL with a typo	
URL that redirects to a new URL	
URL to a site that begins a session	

**Link Evaluator**

Link Evaluator is a Firefox plugin. It is ideal checking individual pages as you create them.

Xenu's Best Results	
Valid simple URL	X
Valid complex URL with parameters	X
Valid complex URL with a fragment	X
Invalid URL with no access	X
Invalid URL with a typo	X
Valid URL with proxy prefix	X
Proxied URL to resource that no longer exists	X*
Proxied URL with a typo	X
URL that redirects to a new URL	
URL to a site that begins a session	X

\* Can sometimes identify if the page is not what was intended. Function can be personalized.

Amelia Mowry,  
Metadata Librarian

