



## MicroModule: vanishing

### What is the Vanishing Web? How Can I Search it?

#### REVIEW Page

Below is the entire module on one page.

#### Gone, but not forgotten!

Have ever returned to a favorite site, only to discover the dreaded **404 page not found** error message? Have you ever clicked through a Google result to find the page missing or altered since it was last indexed?

The web is changing all of the time; pages are added or *removed* everyday. Top websites are regularly overhauled. Some sites move to new addresses or are taken down when their webmasters find other interests. **Millions of pages of information are vanishing from the web every day.**



This lost content can be thought of as the Vanishing Web. *Once a page disappears is it really gone? Is it possible to find something that is no longer on the net?* Web pages that vanish are gone but not *quite* forgotten. There are steps you can take to retrieve those missing pages.

We'll look at three strategies that might save your day by pulling the missing rabbit from the proverbial hat:

- Google's Cache feature
- Researching the Internet Archives
- Creating a personal archive for offline browsing.

#### Using Google's Cache

The Google Cache is a powerful feature offered by the world's largest search engine. Typically when you click on one of Google's recommended links you go to the actual website. If you find the page is missing or has changed and no longer contains the information you seek, you can drop back on Google's 'Cached' feature. You can use this feature *if* the word 'Cached' appears at the end of the 'snippets' paragraph on the Google hit list.

Clicking 'Cached' will take you to Google's index copy of the page (rather than to the actual website). The cached page will appear with your keywords highlighted, making it easier for you to skim to the pertinent information. Additionally, on the first line of the page, you'll see the date Google retrieved the page!

**Example:**



**The Google cache: Operator**

Google also provides a specific operator that will reveal the current version of the webpage in the Google index. To use this operator type it into the Google search bar. The syntax is: cache: www.domainname.com Remember, there is no space between the operator cache: and the URL. The feature won't work if you insert a space.

**Example:**



If this cached information is crucial, consider making an archive copy of the page. After all, next time Google

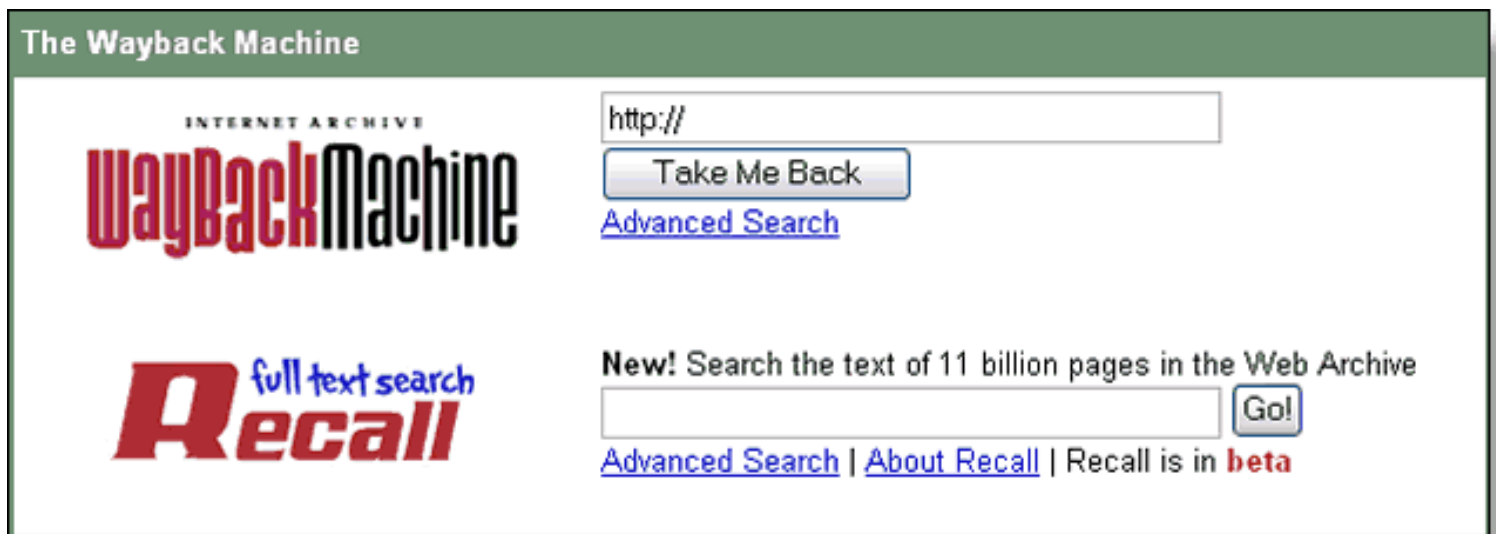
updates the index, it's likely that the Cache copy will be removed or replaced.

## Researching the Internet Archives

When a print publication is issued an ISBN number in the United States, an archive copy of the publication is sent to the Library of Congress. Similar procedures for creating a permanent record of print publications are in place in other countries. This isn't the case for web-publications. Indeed, some fear we've stumbled into an Internet Dark Ages, with our digital cultural history vanishing as we speak. However the need to archive Internet information is being recognized around the world. Increasingly national libraries are making copies of culturally significant web pages. Of particular note is Egypt's multilingual archive effort The Library of Alexandria. <http://www.bibalex.org/website>.

## The Internet Archive

In the United States a partnership between Alexa and the Internet Archive is creating a huge collection that documents the World Wide Web back to 1996. The Internet Archive claims to have 30 billion pages in the vault. The Internet Archive includes web pages, moving images, texts, and audio files. These archived pages can be accessed via The Wayback Machine: <http://www.archive.org>



Unlike a search engine you cannot search the Wayback Machine by concept, keyword, or popularity ranking. You'll need at least the domain name of the website you are seeking. The Internet Archive has released Recall, a new tool that will perform a full text search of about a third (11 billion) of the pages in the archive. Sample searches demonstrate how this promising new tool will work.

## Creating a Personal Archive of a Website

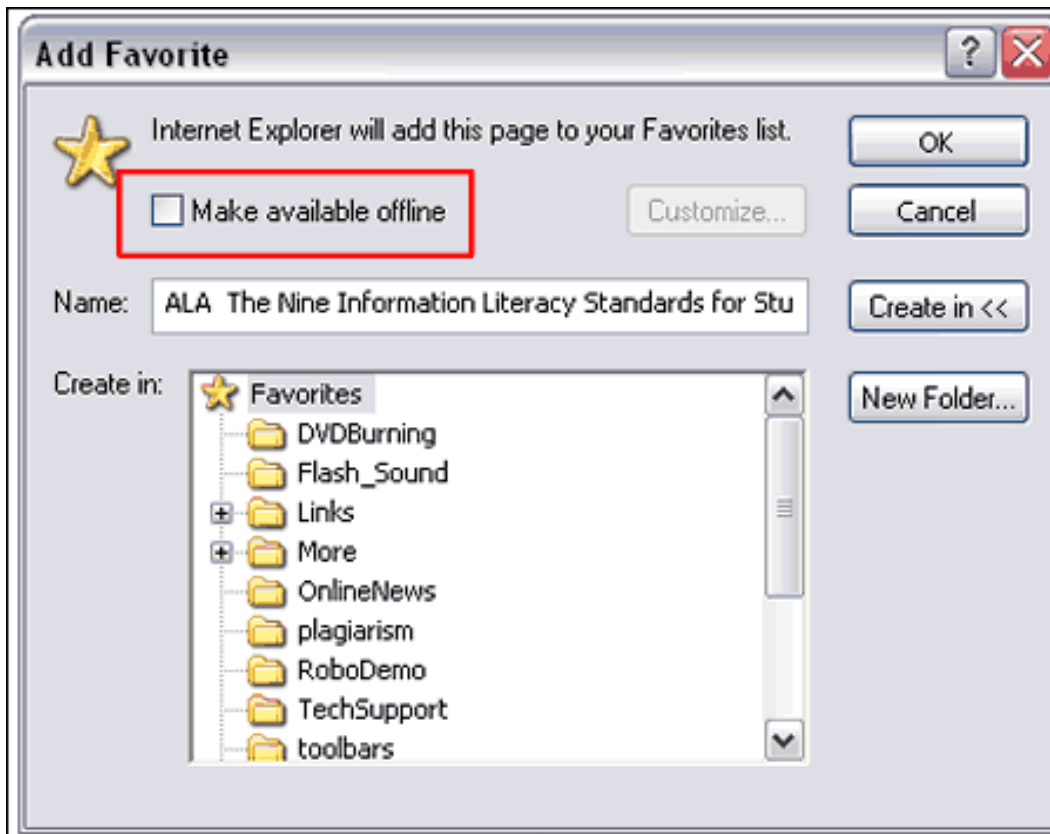
You can create a copy of your favorite web materials on your hard drive by manipulating your browser's 'save as' feature. This allows you to keep an 'offline' copy of the web pages you really value. Creating a personal archive will also allow you to use the information without going online. This is a particularly useful technique if you use a portable computer, since you can save now, and read later. This method is often called 'offline browsing'. This is also a good option if you have narrow bandwidth or limited access to the Internet. Of course offline copies won't offer database interactivity, or remain current unless you update them regularly.

## Internet Explorer

**Internet Explorer** : IE makes it easy to save multiple levels on a website via their 'work offline' feature. Once you have created an 'offline copy' of the website you are interested in, you can tell the browser to work offline. You can then navigate the archived version of the website. Should you come across a portion of the site that has not been archived, the browser gives you the option of going back online to retrieve the information.

To create an 'offline copy' of a valuable web resource using IE, follow these instructions.

- Navigate to the first page of the site you wish to archive.
- From the Favorites Menu, Choose Add to Favorites
- Check the small box that says Make available offline



- Note that when you check the box, the **Customize** button becomes active. Clicking this button starts an automated wizard system that will step you through the archiving process. *You'll be able to specify how many links deep you wish to make your archive.* You can also choose an automated update plan that will keep the resource current. It is often difficult to automatically update password-protected pages. The IE wizard offers you a way to store your passwords and automate this process as well.
- If you skip the customization wizard, all you need to do is click **OK** or **Create In>>** to make your

archive. You'll notice an animated file transfer icon appears on the screen as your browser saves a copy of the pages you visit to your hard drive.

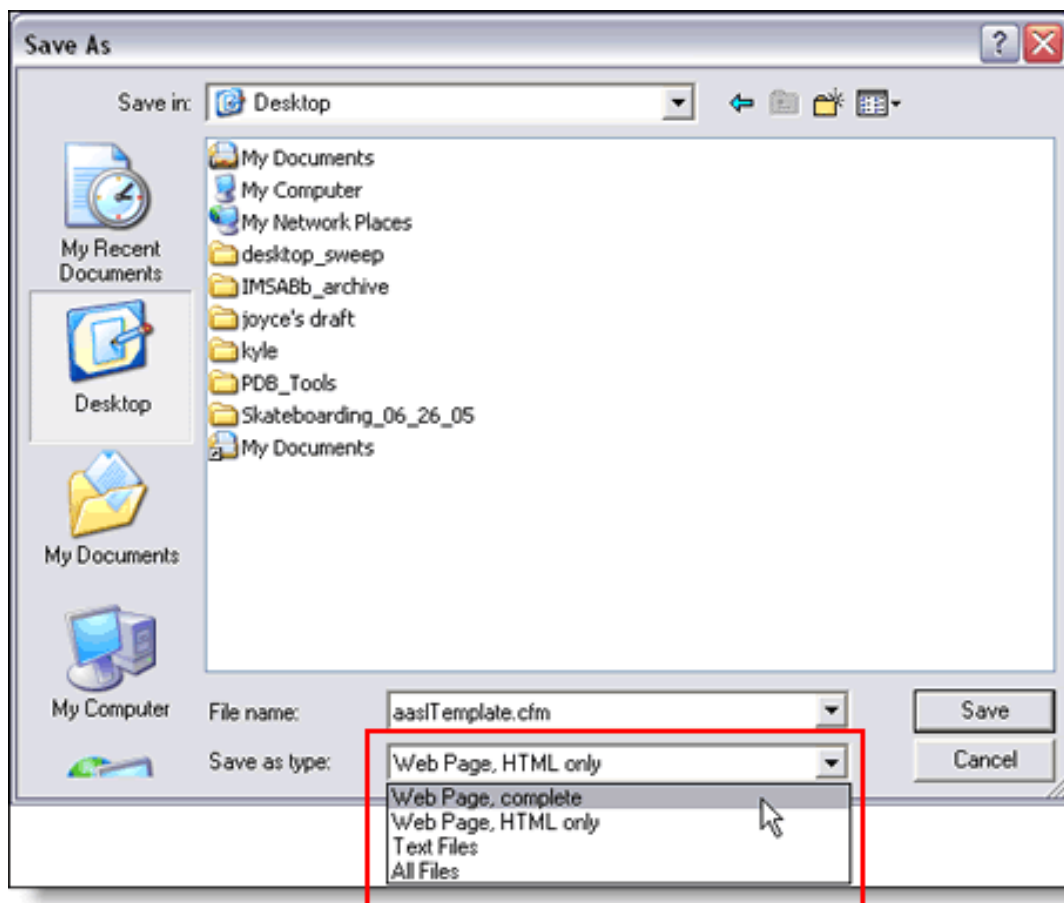
- Next time you want to use the archived materials, choose **Work Offline** from the file menu, and then click the link on your Favorites list. You'll be working with the resources 'offline' in no time!

## Netscape

**Netscape:** To create an 'offline' copy using Netscape **Click File** and choose **Save Page As**:

**At bottom of the Save As panel you'll see two sets of options:**

- File Name (here you enter a descriptive name).
- Save As Type (Here you choose from three types by clicking on the inverted triangle)
- Web Page Complete (This option creates an archive of the entire page, including graphics.)
- Web Page HTML only (This option saves space by excluding graphics)
- Text file (This option translates the page into text, excluding HTML formatting, and graphics).



\*It should be noted that this procedure saves just the page you have open in your browser, and not

the other linked pages on the site.

## Popular commercial archiving packages

There are also third party software products that help you create your own Internet archive. These products are powerful and provide more automation than the simple functions of the most popular browsers. These products can also be used to maintain local 'mirrors' of popular websites typically used by students. Typically a mirror would be updated daily. Students could then access it without actually going online. This is a strategy to consider when you have limited Internet access or wish to limit student access to a few selected sites.

- WebWacker <http://bluesquirrel.com/products/whacker>
- WebZIP <http://www.spidersoft.com>
- HTTrack: Website Copier <http://www.httrack.com>

Armed with the knowledge to use Google's Cache, research the Internet Archives, and make personal backups of treasured web pages, you are better equipped to deal with the Vanishing Web. Luckily the resources and materials available on the Internet continue to grow; still it is useful to be able to snatch back an important page from the verge of extinction!

## What about Copyright?

An archival copy of a website that is strictly for personal reference does not violate copyright for fair use requirements. However, Copyright protections apply. It is best to have permission of the website author before creating an archive.

Authored by Dennis O'Connor 2003



[1](#) [2](#) [3](#) [4](#) [\[5\]](#)

*End of Micromodule - vanishing.*

*Return to Micromodule [List](#)*