Emptying the Cache!
I found something – NOW WHAT?

1) Report it to OIT Security: abuse@umn.edu. We'll determine if this requires breach notification.

2) Remove the data from public sites.

3) Contact search engines that contain the data for removal from indexes AND caches.
Getting the data offline

Most major search engines have information about removal policies:

- **MSN**: [http://search.msn.com/docs/siteowner.aspx](http://search.msn.com/docs/siteowner.aspx)

Bear in mind: these services are not nimble. They have to be cautious (after all, they don't want just ANYONE messing up your webpage). Some use Robots.txt to process requests (after all, if you can modify the page to insert meta tags, your request is probably valid).
Specifically – how do we deal with cleaning up a Google cache? Instructions can be found:


You can:

* Remove your entire website
* Remove part of your website
* Remove snippets
* Remove cached pages
* Remove an outdated ("dead") link
What it takes to get action.....

1. Change / Remove content (in the case of protected content REMOVE.)

2. Wait for GOOGLE to re-index – OR contact them to actively update the index (in the case of protected content CONTACT GOOGLE.)
Contact Google's "Automatic URL Removal System"

http://services.google.com:8882/urlconsole/controller

Your account has been created!

A message has been sent to your email account. Follow the instructions in that message. Please note that you must activate your account within 24 hours or it will be deleted.

Return to information on URL removal.

 Removing your URL

Greetings from Google!

This email was automatically generated when your account was created. Visit the following link to activate your account:


You must activate your account within 24 hours or it will be removed.

If you did not request this account, or do not wish to proceed with the removal of a URL, please ignore this message.

Regards,
The Google Team
Options

You may process your URL for removal from Google's search results. URLs will be removed after we've verified your request. Bear in mind that verification can take several days or longer and all pages submitted via the automatic URL removal system will be removed from the Google index temporarily for six months. You may review the status of submitted requests in the column to the right. To remove your Usenet posts from Google Groups, please click here.

Select only one:

- **Remove pages, subdirectories or images using a robots.txt file.**
  Your robots.txt file need not be in the root directory.

- **Remove a single page using meta tags.**

- **Remove an outdated link.**

Please visit our Contact Us page if you have further questions or problems.
Remove a single page using meta tags.

You must insert the appropriate meta tags before proceeding. If you have not, please return to our Remove URL page for further information.

Note: Your page will remain out of Google's index as long as it contains the appropriate meta tags.

URL to remove:  

E.g. http://www.google.com/page.html
robots.txt

http://www.robotstxt.org

this is the way to control what the search engines SEE and what they retain when they discover and visit your pages:

**In a nutshell**, when a Robot visits a Web site, say http://www.foobar.com/, it firsts checks for http://www.foobar.com/robots.txt. If it can find this document, it will analyse its contents for records like:

User-agent: *
Disallow: /
OTHER things to worry about...
You MAY have cleaned up Google/Yahoo/MSN/ASK.com..... BUT your data may STILL Be out there!

Archive.org is an archive of web pages that contains historical archives of web pages.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

* denotes when site was updated.

Search Results for Jan 01, 1996 - Apr 13, 2006

<table>
<thead>
<tr>
<th>Date</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 02, 2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dec 06, 2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jul 26, 2002</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep 21, 2002</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep 24, 2002</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov 24, 2002</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov 30, 2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>May 04, 2003</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jun 07, 2003</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jun 23, 2003</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jul 23, 2003</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct 27, 2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dec 30, 2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 30, 2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jul 12, 2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep 28, 2004</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb 12, 2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fortunately – archive.org is 1) not text searchable and 2) easy to contact for removal

http://www.archive.org/about/contact.php
The Google Search Appliance – crawls UMN.EDU space from WITHIN *.umn.edu. This means that content is exposed (and searchable) beyond the campus. SEARCH.UMN.EDU is available to off-campus locations.
http://umn.edu/google
for assistance on removing content from the search. (This CACHE is independent of Google.com archives/indexes.)
Reminder - You can **KEEP** this data **OFFLINE**

- Question *everything* – why does the data need to be online?

- There is no need to publicly provide student communications: [http://webct.umn.edu](http://webct.umn.edu) exists for this purpose.

- The “robots.txt” standard exists to define ways to wall off data from the search engines
references

- **ROBOTS.TXT standard**: http://www.robotstxt.org/
- **Google URL Removal System**  http://services.google.com:8882/urlconsole/controller
- **UMN Google Search appliance** http://www.umn.edu/google
- **Archive.org Contact** http://www.archive.org/about/contact.php
- **Multi-Search Engine Tool**:  http://www.dogpile.com/
- **Report Data Exposure**: abuse@umn.edu
- **MSN**:  http://search.msn.com/docs/siteowner.aspx
- **ASK.COM**:  http://about.ask.com/en/docs/about/webmasters.shtml