Digging into Google....
....with Google Diggity

Joel Anderson
OIT Security and Assurance
The Problem of Private Data

What happens when it isn't private anymore?
The University may be **REQUIRED** to notify

http://www.revisor.leg.state.mn.us/bin/getpub.php?pubtype=STAT_CHAP_SEC&%20year=current&section=13.055

**Minnesota Statutes 2005, 13.055**

Copyright 2005 by the Office of Revisor of Statutes, State of Minnesota.

**Minnesota Statutes 2005, Table of Chapters**

**Table of contents for Chapter 13**

13.055 State agencies: disclosure of breach in security.

Subdivision 1. Definitions. For purposes of this section, the following terms have the meanings given to them:

(a) “Breach of the security of the data” means unauthorized acquisition of data maintained by a state agency that compromises the security and classification of the data. Good faith acquisition of government data by an employee, contractor, or agent of a state agency for the purposes of the state agency is not a breach of the security of the data, if the government data is not provided to an unauthorized person.

(b) “Contact information” means either name and mailing...
Consider legally protected data:

http://www.policy.umn.edu/Policies/it/Use/SECUREDATA.html

Non-directory Student Information may not be released except under certain prescribed conditions. Non-releasable information includes:

* Grades
* Courses taken
* Schedule
* Test scores
* Advising records
* Educational services received
* Disciplinary actions
* Student ID
Everyone should **already** know this.... *right?*

http://policy.umn.edu/Policies/it/Use/SECURITYBREACH.html
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: https://moodle.umn.edu/ exists to manage courseware and gradebook functions for classes.

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

*(There are also meta tags define that control reading and caching of data.)*
How do you KNOW it's a NICE Robot?

You don't. You cannot rely on robots.txt to protect information that shouldn't be online in the first place.

It will help guide the well-behaved search engines, but it CAN point the way for less socially responsible robots.

There are other methods (eg. htaccess) that can be used to more assertively restrict access – but the first principle should be:

DOES THIS NEED TO BE ONLINE AT ALL?
The robots.txt file contains "rules" about where web spiders are allowed (and NOT allowed) to look in a website's directory structure. Without over-complicating things, this means that the robots.txt file gives a mini-roadmap of what's somewhat public and what's considered more private on a web site.

Have a look at the robots.txt file itself, it contains interesting stuff.

However, don't forget to check out the other files in these directories since they are usually at the top directory level of the web server!

Here's an interesting search for just that....

http://www.google.com/search?q=intitle:Index.of+robots.txt
Find it first.....

- Many “breach” issues happen when someone “googles” their own name
- You can build searches that will look for this data BEFORE someone else does

http://www.hackersforcharity.org/ghdb/
Okay – proof that Google isn't evil

You can use it to help you find the stuff you DON'T want to be public.

You have a chance to get it off line if you can ....
Google power search

- **numrange** – search for x..y for a range of values.

  111111..9999999 is a good match for student IDs.

  *(emplid's too, though technically they aren't private).*

- **filetype:xls** – finds text IN Excel spreadsheets –
  likewise filetype:pdf

- **intext:(a|b)** finds 'a' OR 'b'

- **site:domain.tld** restricts search to

  site:domain.tld

  *for example.....*
Refining a search for Student data:

- 5,380,000 for filetype:xls
- 6,800 from umn.edu for filetype:xls
- 120 from umn.edu for filetype:xls intext:id
- 153 from umn.edu for filetype:xls intext:id grade (no I don't know why it went

Try adding a number range that covers Student ID's:
1111111..9999999
We're sorry...

... but it appears your computer is sending automated requests.

These days, there's a limit to the fishing Google allows!
Sometimes advanced searches get around this......
But, not always........

Our systems have detected unusual traffic from your computer network. Please try your request again later. Why did this happen?

IP address: 192.168.1.231
Time: 2010-11-11T14:35:02Z
URL: http://www.google.com/search?q=%22student+id%22+filetype%3Axls+site%3Aumn.edu+11111111...999999999&hl=en&num=10&lr=&ft=i&cr=&safe=images
Ouch!

So … how can you use Google to keep searching for stuff?

http://www.stachliu.com/index.php/resources/tools/google-hacking-diggity-project/
GoogleDiggity - Version 0.2
Google Hacking Utility - Using Google AJAX API
By: Fran Brown and Rob Ragan
Stach & Liu, LLC - http://www.stachliu.com
Copyright 2010 Stach & Liu LLC

---

** Note: Google AJAX API won't block you, but does limit you to 64 results per query.
** Note2: Output also written to tab delimited file: output.txt. Easily open with MS Excel.

Usages:

GoogleDiggity.exe [-f Google_Search_Query_File] [-d Target_Domain] [-k GoogleCustomSearchEngineID] [-a Google AJAX Search API Key]

-q [google search query]
Optional, can alternatively load queries from file using the -f option.

-f [Google Hack queries file]
Optional, text file contain Google search queries, one per line.
GHDB and FSDB txt files provided (e.g. "Error Messages.txt", "Files containing passwords.txt", etc.).

-d [Domain to search]
Optional, example: whatever.com

-t [Target_Domain_or_Site_File]
Optional, text file containing list of sites and domains to target (e.g. "www.stachliu.com").

-s [Search_Safety_Level]
Optional, search safety level: off, moderate, or full. Defaults to Moderate.

-a [Google AJAX API Key]
Optional, Google AJAX Search API Key
For more info, go to http://code.google.com/apis/ajaxsearch/signup.html

-k [Google Custom Search Engine ID]
Optional, this can also be set in the configuration file.
Note, to use the google cse ID in the config file, specify the -k option without any params.
For more info, go to http://www.google.com/cse/

Examples:
GoogleDiggity.exe -q test -d stachliu.com
GoogleDiggity.exe -f "GHDB and FSDB - All.txt" -d stachliu.com
GoogleDiggity.exe -f "GHDB and FSDB - All.txt" -t mycompaniesDomains.txt
GoogleDiggity.exe -f "Error Messages.txt" -k 00611383271342003645:ytae6rpi4i
GoogleDiggity.exe -f "Error Messages.txt" -s off
Here's a search built to look for grade spreadsheets:

UMN-grades;;grade +"student id" filetype:xls numrange:111111..9999999
UMN-grades;;grade +emplid filetype:xls numrange:111111..9999999
UMN-grades;;grade +"student id"  numrange:111111..9999999
UMN-grades;;grade +emplid  numrange:111111..9999999

The first two lines look for Excel spreadsheets with grades, the second without the filetype: finds ANY content with the keywords.
Note: the three results were false positives. Html and pdf's that included student id+grade+a number in the range of student ID.
Google Diggity gives you tools to build automated searches that you can run to discover potential data breaches.

It also includes searches to look for things from the GoogleHack Database (GHDB):

Welcome to the Google Hacking Database (GHDB)!
We call them ‘googledorks’: Inapt or foolish people as revealed by Google. Whatever you call these fools, you've found the center of the Google Hacking Universe! Stop by our forums to see where the magic happens!

Advisories and Vulnerabilities (215 entries)
These searches locate vulnerable servers. These searches are often generated from various security advisory posts, and in many cases are product or version-specific.

Error Messages (66 entries)
Really retarded error messages that say WAY too much!

Files containing juicy info (230 entries)
No usernames or passwords, but interesting stuff none the less.

Files containing passwords (135 entries)
PASSWORDS, for the LOVE OF GOD!!! Google found PASSWORDS!

http://www.hackersforcharity.org/ghdb/
There are a lot of search engines! How do I find ALL the data?

First – it might not be possible. However, once you've found it on one search engine, you should at least try to find it on the major search engines. Dogpile.com is a useful site that will search the major search engines Google, Yahoo, MSN and Ask.com. Knowing WHAT the public text *is* makes it easier to find it on other sites – just search for fragments of text you KNOW are online.

http://www.dogpile.com
I found something – NOW WHAT?

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

2) Notify the local culprit data-owner. (OITSEC is happy to assist in this discussion.)

3) Remove the data from public sites. Retrieve logs that assist in determining if the data was seen.

4) Contact search engines that contain the data for removal from indexes AND caches. Virtually EVERY search engine has a cache – you should assume this is the case – removing the data from your website is not enough.
Getting the data offline

Most major search engines have information about removal policies:

- **GOOGLE**:  
  http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=164734&rd=1
- **YAHOO**:  
- **BING**:  
- **ARCHIVE.ORG** http://www.archive.org/about/terms.php
- **ASK.COM**:  
  http://about.ask.com/en/docs/about/webmasters.shtml

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).
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: [https://moodle.umn.edu/](https://moodle.umn.edu/) exists to manage courseware and gradebook functions for classes.

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