USING WEB SEARCH ENGINES TO FIND INFORMATION LEAKAGE AND SECURITY HOLES WITHIN YOUR ORGANIZATION

Hans Pongratz[1]

[1] Technische Universität München, Arcisstr. 21, 80333 München, Germany, pongratz@tum.de.

Keywords
information leakage, information disclosure, data leakage prevention, google hacking, web search engines

1. EXECUTIVE SUMMARY

The world wide web got ubiquitous and pervasive in our lives, organizations and higher education institutions (HEIs). Search engines try to locate, sort and catalogue the web and help us find the wanted information. Due to configuration mistakes, bugs or other human failure there are many cases of unwanted publication of information through the web.

Information leakage is one of the next great challenges for the information society. “Google Hacking” has been established as acronym for the use of search engines to gain attack vectors or to find confidential information or privacy threats via search queries. The spectrum reaches from traitorous error messages to login credentials up to special file types and browseable directories.

1.1. Background

Particularly in organizations with a widely distributed and divided IT structure, as common within HEIs, usually a wide variety of web systems with decentralized responsibility are in use. The maintenance and security level of each single system is often very different. Due to the decentralized structure a central content filtering of sensible data is not possible or in the case of HEIs is often neither wanted nor enforceable. An alleged intrusion, information theft or loss is often noticed too late.

A famous example for search engine based information leakage was the publication of a database with 43,861 student records of the University of Magdeburg, Germany, through the web in May 2008. The search engine Google indexed the file for about ten days. Affected data were e.g. name, address, phone, gender, faculty, and enrolment date.

1.2. Conclusion

To face the challenge of information leakage through search engines, a classification model for sensible data at HEIs and a software prototype were developed and implemented. This could help CIOs and their staff to keep track of their web systems.

Based on a summary of the technique of “Google Hacking”, the author presents a brief description of simple and complex search queries, the classification model and the software prototype. Concrete examples of information leakage and security holes found with the help of search engines in the field of HEIs are shown and possible countermeasures are discussed.

2. PROBLEM STATEMENT

The landscape of information technology (IT) within organizations gets more and more complex and meshed up. Especially so called self service functionalities via web are rolled-out. Particularly at HEIs, with often widely distributed and divided IT structures, a wide variety of web systems with decentralized responsibilities are in use (Bode et al 2007) (Stratmann and Kerres 2008).

In February 2009 a TYPO3 security issue was published (typo3-sa-2009-002[1]), the following days the typo3-based website of the Federal Minister of the Interior of the Federal Republic of Germany Schäuble (www.wolfgang-schaeuble.de) was hacked. The defacement stated: "Typo3 Please update it ;) And change passwords" including a hyperlink to the typo3 security advisories. Hackers could use web search engines to find very easily and unnoticed attackable web sites by combing search operators in a skillful way. This is just one recent and prominent example for the many different ways in which a security hole can be used to change, receive or delete published data.

3. GOOGLE HACKING

The term “Google Hacking” is an acronym for using web search engines to find information leakage and attack vectors. Search engines crawl billions of web pages by following hyperlinks and analyze and evaluate their contents by special (in most cases confidential) criteria. Most search engines mine the www for content of web pages, like text, images or files. Data which is added to the database of a search engine is called indexed and can be found via search queries. Some search engines even mine databases, social networks, news or directories (Kraft and Weyert 2007) (Lancor and Workmann 2007).

Google, Yahoo and MSN Live Search (now Bing) are the market leader for web search in Germany (Schultz 2009).

Table 1: Common web search engine operators

Operator	Description	Google	Bing	Yahoo
+	Adds words to search query, even those which are normally ignores, like „the“	yes	yes	yes
-	Exclude words or arguments from search	yes	yes	yes
"EUNIS 2009"	Searches the exact phrase	yes	yes	yes
|	or-Operator	yes	yes	yes
site:	site:url will only search within specific website or domain.	yes	yes	yes
intext:[2]	Query term must be in the text of result(s).	yes	yes	yes
intitle:	Query term must be in the title of result(s).	yes	yes	yes
inurl:	Given word must be in the URL of result(s).	yes	yes	no
filetype:	Results must be particular file format, e.g. .pdf, .xls or .doc.	yes	yes	yes
daterange:	Searches for sites, which have been created or modified in specified timeslot.	yes	no	yes
numrange:	Results will be restricted to those containing numbers in specified range	yes	no	no

[1] http://typo3.org/teams/security/security-bulletins/typo3-sa-2009-002/

[2] At www.bing.com: „inbody:“

cache:	cache:url will not display the current version of the page, but Google's cached version.	yes	no	no
~car	Looks for car and its synonyms	yes	no	no
ip:	Searches for sites, which are hosted by the given IP address, e.g. 193.144.75.244.	no	yes	no
contains:	Only sites with links to the specified file types are shown.	no	yes	no

Table 1 gives a brief overview of the complexity of search options and compares a selection of featured search operators of Google, Bing and Yahoo. The operators are very powerful and can help e.g. narrowing search results. For example the search query "inurl:index.htm site:xyz.org" finds web pages within the domain "xyz.org", where the url contains "index.htm". The operator "ip:" at Bing helps to search for all indexed web pages of a specified web server by its IP-address.

Some search features are only available via the web interface, like the "daterange"-option at Yahoo or the feature to show cached web sites at Bing and Yahoo. For a complete list of search operators of a search engine and to see which operators are combinable, please check with the respective search tips.

By combining search operators in a skillful way, someone could find unintentionally accessible documents, back doors and other information for attack vectors, like traitorous error messages. The following paragraph gives some examples on simple and complex search queries and how to find TYPO3 instances within your web domain.

3.1. Search Queries

Name	Login dialog
Search query	intitle:login site:eunis.org
Description	Looks for potential login pages of a web presence. The query searches for web pages, where the page title contains "login" within the domain xyz.org.
Hits[3]	Google: 52 Bing: 51 Yahoo: 32
Comment	There a lots of other search queries to find login dialogs, e.g. via intext: login or login.asp.

Name	Directory browsing
Search query	intitle:"index of" eunis
Description	Searches for browseable directories, where the term "eunis" is used.
Hits[3]	Google: 201 Bing: 6 Yahoo: 275
Comment	It should be checked, if anonymous users are allowed to see the directory structure. Directory-listing could be deactivated at the web server.

Name	Confidential documents
Search query	(confidential | secret | internal use| restriced) filetype:pdf site:eunis.org
Description	Searches for pdf documents within the site eunis.org, where one of the terms "confidential", "secret", "internal use" or "restriced" appears.

[3] Search queries where done on May 29th 2009 at www.google.de, www.yahoo.de and www.bing.de. Hit count doesn't imply anything on the quality of search result(s).

Hits[3]	Google: 22 Bing: 1 Yahoo: 0
Comment	Further file types and terms should be checked for a concrete analysis.

Name	TikiWiki pages
Search query	intext:tikiwiki site:eunis.org
Description	Searches for web pages within the web appearance of EUNIS which contain the phrase “tikiwiki”.
Hits[3]	Google: 173 Bing: 284 Yahoo: 33
Comment	There are lots of other search queries to find specific TikiWiki pages, like login screens or forgot your password pages.

3.2. TYPO3 Search Queries

TYPO3[4] is a PHP-based, open source content management system, which is quite frequently used in Germany. There are large repositories of includable extensions, which add special features. Security bulletins inform about recent security issues within the main application or extensions. The following search queries should give the reader an idea of possible ways to find instances of TYPO3 within the own organization. This could help finding relevant TYPO3 Instances to patch.

Name	TYPO3
Search query	typo3 site:yoursite.xyz
Description	This query searches for typo3 terms with your web presence. To narrow the result the “inurl:”-operator could be used.

Name	TYPO3 temp directory
Search query	typo3temp site:yoursite.xyz
Description	The typo3temp directory is used by TYPO3 to provide temporary files and pictures. Filenames within this directory are md5-coded. This query searches for typo3temp terms with your web presence. To narrow the result the “inurl:”-operator could be used.

Name	Typo3 fileadmin directory
Search query	fileadmin site:yoursite.xyz
Description	The fileadmin directory stores uploaded files and templates within a TYPO3 instance. This query searches for fileadmin terms with your web presence. To narrow the result the “inurl:”-operator could be used.

Name	Typo3 URL
Search query	inurl:/index.php?id site:yoursite.xyz

[4] For further information see www.typo3.org.

Description	Lots of TYPO3 instances don't use URL encoding, so "index.php?id" is part of the URL. To narrow the result the "-"-operator could be used, e.g. "-joomla", because other content management systems have similar URLs.

4. SENSIBLE DATA AT HIGHER EDUCATION INSTITUTIONS

Particularly in organizations with a widely distributed and divided IT structure, as common within HEIs, usually a wide variety of web systems with decentralized responsibilities are in use. The maintenance and security level of each single system is often very different. Due to the decentralized structure a central content filtering of sensible data is not possible or in the case of HEIs is often neither wanted nor enforceable. An alleged intrusion, information theft or loss is often noticed too late. HEIs are great targets for attackers, because of a high density of sensitive data, good IT-infrastructures and often no central security precautions.

So which data should be protected? Due to the great variety in the German HEI landscape and the resulting different business processes within HEIs no complete listing of all sensitive data can be provided. Figure 1 shows a classification model, where a distinction is drawn between identity groups (subjects), organizational units (objects) and the IT-systems of a HEI with examples to sensitive information (Pongratz 2009). These segments are called domains and group data sets, like personal data or credentials. Information within data sets are called attributes.

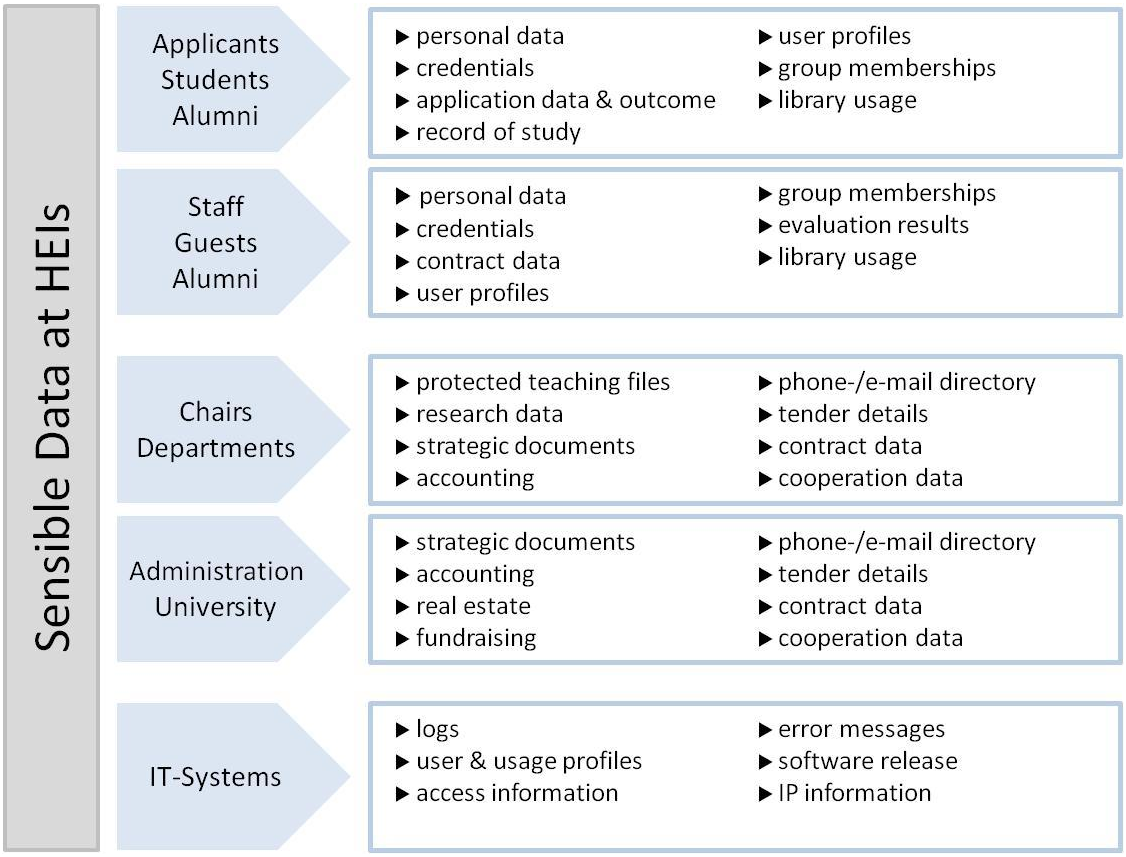

Figure 1: Classification model for sensible data at HEIs

In the following some examples for data sets, their attributes and search terms are listed.

Domain	Applications, Students, Alumni
Data set	personal data
Attribute	matriculation number

Search term(s)	Matrikelnummer, Matrikelnr, Matr, Matnr, Matrikel-nummer, numrange-Matrikelnummer, …
Comment	Numrange depending on HEI, e.g. 6-digits matriculations number google search term: „site:xyz.de 010000..029999“

Domain	IT-Systems
Data set	software release
Attribute	Apache version
Search term(s)	Apache/2.2.11, …
Comment	Searches for Apache Version 2.2.11. “site:yoursite.xyz apache/2.2.” would show you all 2.2.x releases.

Domain	IT-Systems
Data set	Logs
Attribute	web server logs
Search term(s)	filetype:log get, webalizer, awstats, usage, …
Comment	Numrange depending on HEI, e.g. 6-digits matriculations number google search term: „site:xyz.de 010000..029999“

5. SOFTWARE PROTOTYPE

The given examples show the outline of using web search engines to find information leakage and security holes within an organization. A manual inquiry of all conceivable search words for a domain on a regular basis is due to the abundance of possible search patterns not possible. Because of this, a proof of concept prototype for a meta-scanner was developed in the context of a bachelor thesis (Wirtz 2008). The prototype is constantly improved and extended for further research work. The software is PHP-based and offers the possibility to add new search categories, patterns or search engines. Search results can be analyzed by regular expressions.

6. COUNTERMEASURES

This paragraph will point out possible countermeasures to prevent information leakage through web search engines. First of all - don’t put sensitive data on your web server. This is not always possible, e.g. due to human failure. In an investigated case of unwanted publication of sensible data, a secretary stored a database within a personal web folder, which was indexed by search engines. She had thought it was her personal backup folder.

Web servers should be checked on a regular basis for threats, confidential information and needed updates. Hacker search e.g. for unpatched apache web server or TYPO3 instances. You should consider using a robots.txt file with meta-tags, like noarchive. This would prevent web crawlers of search engines to add your site or folder to the index of the search engine. But hackers could analyze just your robots.txt file to gain more information about your web site. For further information about robots.txt see (NoArchive website 2009) and (Web Robots website 2009).

TYPO3 instances should publish as little information about directories, URLs, patch level and so on as possible. TYPO3 directory information can be masked by using URL encoding. TYPO3 html code contains information about the CMS, like “<meta name="generator" content="TYPO3 4.2 CMS" />”, but it’s configurable. To change this meta tag, edit typo3/sysext/cms/tslib/class.tslib_pagegen.php and manipulate the variable $GLOBALS[‘TSFE’]. Standard setting: “$GLOBALS['TSFE']->content.='<meta name="generator" content="TYPO3 '.TYPO3_branch.' CMS" />'; “.

In case of unwanted information leakage through web search engines, you should:

- Remove the information from your web server
- Check your web server log files, which IP addresses access the information
- Contact the search engine for removal, e.g.:
  - Google: www.google.com/remove.html
  - Yahoo: www.help.yahoo.com/l/us/yahoo/search/siteexplorer/delete/index.html
  - Bing: https://support.discoverbing.com/eform.aspx?productKey=bingcontentremoval&ct=eformts&scrx=1
- Check other search engines and their caches, too! For example the Internet Archive (2009).

To be aware of new, critical security issues subscribe to the respective mailing lists and patch your instances!

7. CONCLUSION AND OUTLOOK

The reasons for information leakage are various, e.g. lacking security awareness, operating errors, incorrect configuration, software bugs, inner company job-relocation, forgetfulness, negligence and/or the underestimation of the consequences. Apart from a general web content filtering, which is neither always possible nor easy to maintain, there are other ways to face the challenge, like workflow & approval processes for web content and/or automated search tools for vulnerabilities.

In the context of the research work, domain-specific sensible data at HEIs were defined and search heuristics for e.g. matriculation numbers, personal data, credentials or marks were generated and integrated into a search query client (meta-scanner). Thanks to the meta-scanner some sensible web content could be removed, e.g. defacements and a remote shell. Further research focuses on automatic filtering and rating of search results.

8. REFERENCES

Bode, A., Borgeest, R., Pongratz, H. (2007). The ICT Strategy of the Technische Universität München. Desnos Epelboin (Hrsg.). Proceedings EUNIS 2007, Grenoble, Frankreich.

Internet Archive website (2009). The Wayback Machine. Retrieved May 30, 2009, from http://web.archive.org.

Kraft, P. B., Weyert, A. (2007). Network Hacking. Franzis.

Lancor, L., Workman, R. (2007). Using google hacking to enhance defense strategies. SIGCSE Bull. 39, 1, pp. 491-495.

NoArchive website (2009). The NoArchive Initiative. Retrieved May 30, 2009, from http://noarchive.net/meta/.

Pongratz, H. (2009). Suchmaschinen-basierte Informationsausspähung. In Christian Paulsen (Hrsg.), Sicherheit in vernetzten Systemen, 16. DFN Workshop, DFN CERT.

Schultz, C. D. (2009). Suchmaschinenmarketing. In Dirk Lewandowski (Hrsg.), Handbuch Internet-Suchmaschinen, AKA Verlag Heidelberg.

Stratmann, J., Kerres, M. (2008). E-Strategy – Strategisches Informationsmanagement für Forschung und Lehre. Waxmann.

Studierendenrat Universität Magdeburg website (2009). Studierendenrat Universität Magdeburg: Datenschutz an der Otto-von-Guericke-Universität Magdeburg. Retrieved May 30, 2009, from http://www.studentenrat.org/?p=129

Web Robots website (2009). The Web Robots Pages. Retrieved May 30, 2009, from http://www.robotstxt.org/robotstxt.html.

Wirtz, A. (2008). IT-Sicherheit: Informationsausspähung im Web durch Suchmaschinen. Analyse, Klassifizierung und Gegenmaßnahmen im universitären Umfeld. Bachelor Thesis, Technische Universität München, Fakultät für Informatik, 2008.