Usage Analysis of a Public Website Reconstruction Tool

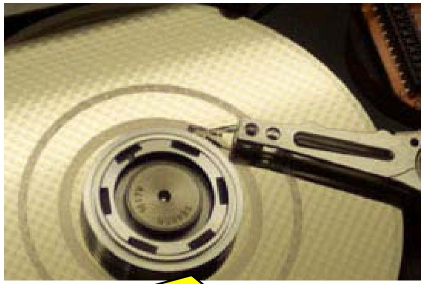
Frank McCown and Michael L. Nelson

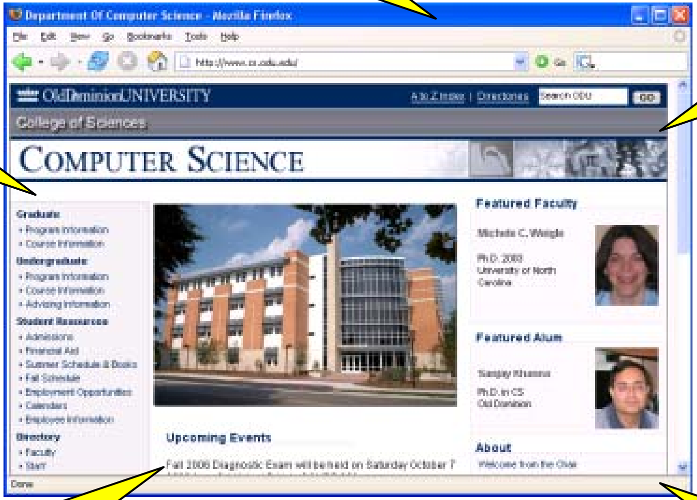
Harding University
Computer Science
Department
Searcy, Arkansas, USA

Old Dominion University
Computer Science
Department
Norfolk, Virginia, USA

JCDL 2008

Pittsburgh, PA
June 19, 2008




Department Of Computer Science - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.cs.odu.edu/
Go
A to Z Index | Directories
Search ODU
GO
Done

Old Dominion UNIVERSITY

College of Sciences

COMPUTER SCIENCE

Graduate
- Program Information
- Course Information

Undergraduate
- Program Information
- Course Information
- Advising Information

Student Resources
- Admissions
- Financial Aid
- Summer Schedule & Books
- Fall Schedule
- Employment Opportunities
- Calendars
- Employee Information

Directory
- Faculty
- Staff

Upcoming Events

Fall 2006 Diagnostic Exam will be held on Saturday October 7

Featured Faculty

Michele C. Weigle

Ph.D. 2003
University of North Carolina

Featured Alum

Sanjay Khanna

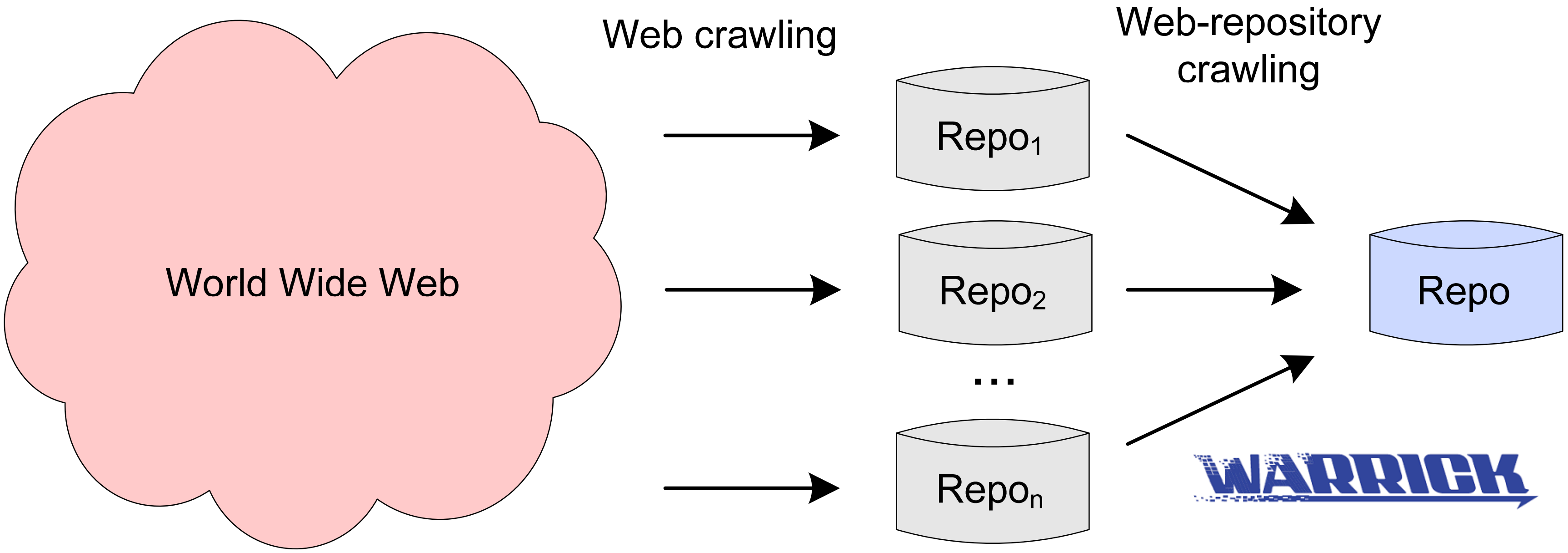
Ph.D. in CS
Old Dominion

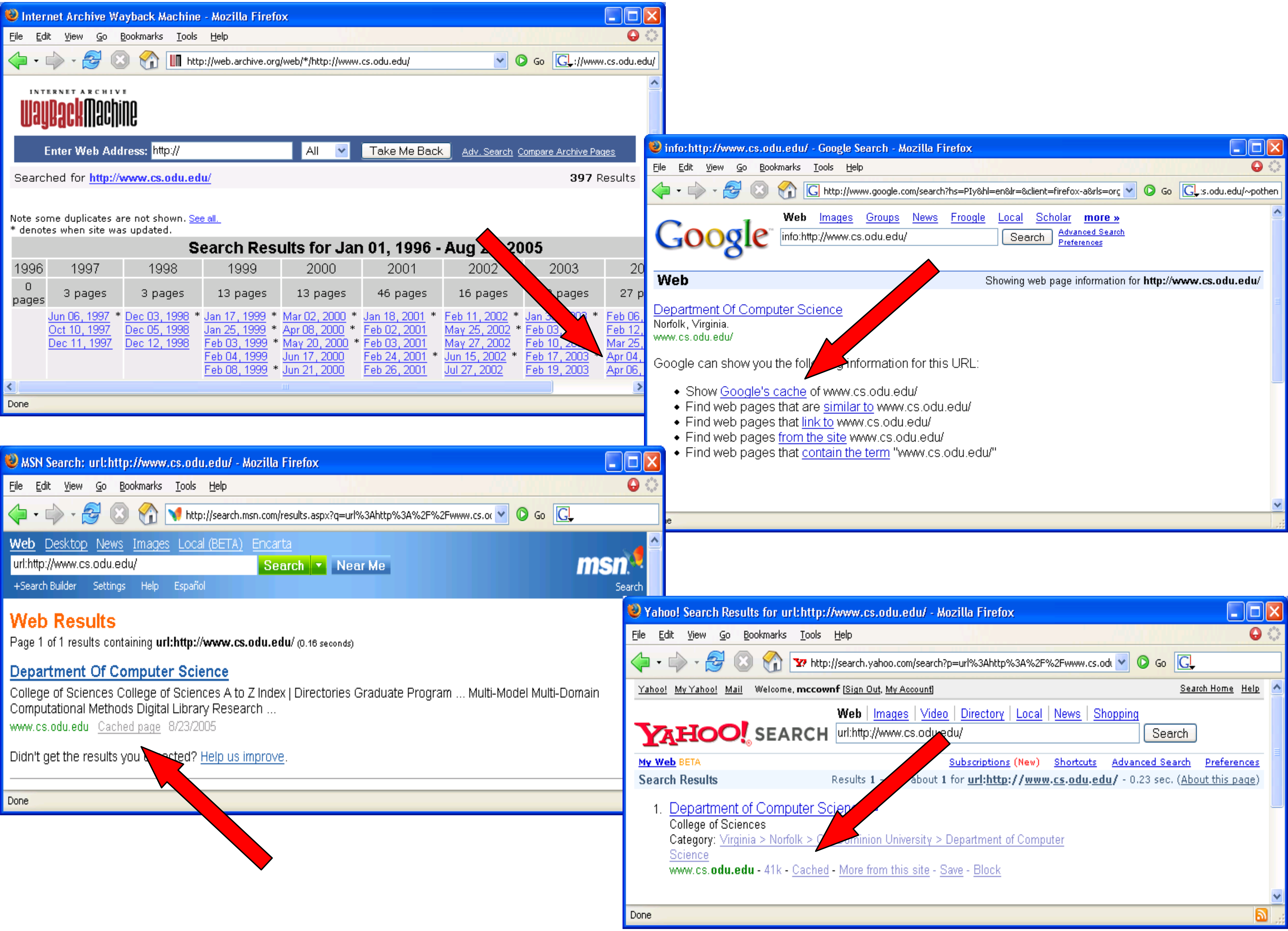
About

Welcome from the Chair

Black hat: http://img.webpronews.com/securitypronews/110705blackhat.jpg
Virus image: http://polarboing.com/images/topics/misc/story.computer.virus_1137794805.jpg
Hard drive: http://www.datarecoveryspecialist.com/images/head-crash-2.jpg

Crawling the Crawlers





Internet Archive Wayback Machine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://web.archive.org/web/*/http://www.cs.odu.edu/
Go
INTERNET ARCHIVE WaybackMachine
Enter Web Address: http://
All
Take Me Back
Adv. Search
Compare Archive Pages
Done

Searched for http://www.cs.odu.edu/ 397 Results

Note some duplicates are not shown. See all.
* denotes when site was updated.

Search Results for Jan 01, 1996 - Aug 2 2005

1996	1997	1998	1999	2000	2001	2002	2003	20
0 pages	3 pages	3 pages	13 pages	13 pages	46 pages	16 pages	pages	27 p
	Jun 06, 1997 *	Dec 03, 1998 *	Jan 17, 1999 *	Mar 02, 2000 *	Jan 18, 2001 *	Feb 11, 2002 *	Jan 3 *	Feb 06,
	Oct 10, 1997	Dec 05, 1998	Jan 25, 1999 *	Apr 08, 2000 *	Feb 02, 2001	May 25, 2002 *	Feb 03,	Feb 12,
	Dec 11, 1997	Dec 12, 1998	Feb 03, 1999 *	May 20, 2000	Feb 03, 2001	May 27, 2002	Feb 10, 20	Mar 25,
			Feb 04, 1999 *	Jun 17, 2000	Feb 24, 2001	Jun 15, 2002 *	Feb 17, 2003	Apr 04,
			Feb 08, 1999 *	Jun 21, 2000	Feb 26, 2001	Jul 27, 2002	Feb 19, 2003	Apr 06,

info:http://www.cs.odu.edu/ - Google Search - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.google.com/search?hs=PIy&hl=en&lr=&client=firefox-a&rls=org
Go
Web Images Groups News Froogle Local Scholar more »
info:http://www.cs.odu.edu/
Search
Advanced Search
Preferences

Web — Showing web page information for **http://www.cs.odu.edu/**

Department Of Computer Science
Norfolk, Virginia.
www.cs.odu.edu/

Google can show you the foll g information for this URL:

- Show Google's cache of www.cs.odu.edu/
- Find web pages that are similar to www.cs.odu.edu/
- Find web pages that link to www.cs.odu.edu/
- Find web pages from the site www.cs.odu.edu/
- Find web pages that contain the term "www.cs.odu.edu/"

MSN Search: url:http://www.cs.odu.edu/ - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://search.msn.com/results.aspx?q=url%3Ahttp%3A%2F%2Fwww.cs.o
Go
Web Desktop News Images Local (BETA) Encarta
url:http://www.cs.odu.edu/
Search
Near Me
msn Search
+Search Builder Settings Help Español
Done

Web Results

Page 1 of 1 results containing **url:http://www.cs.odu.edu/** (0.16 seconds)

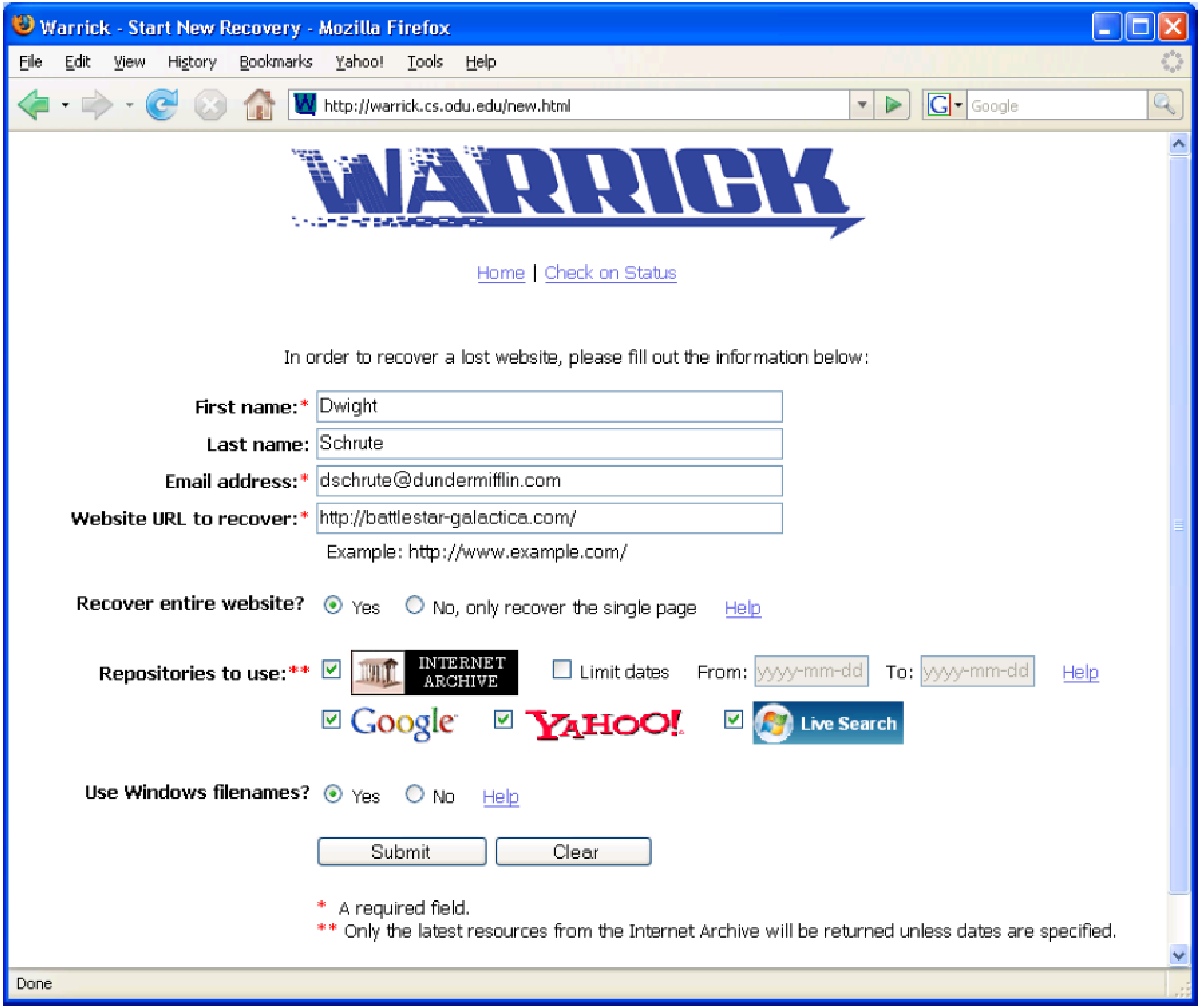
Department Of Computer Science
College of Sciences College of Sciences A to Z Index | Directories Graduate Program ... Multi-Model Multi-Domain Computational Methods Digital Library Research ...
www.cs.odu.edu Cached page 8/23/2005

Didn't get the results you cted? Help us improve.

Yahoo! Search Results for url:http://www.cs.odu.edu/ - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://search.yahoo.com/search?p=url%3Ahttp%3A%2F%2Fwww.cs.od
Go
Yahoo! My Yahoo! Mail Welcome, mccownf [Sign Out, My Account]
Search Home Help
Web Images Video Directory Local News Shopping
YAHOO! SEARCH
url:http://www.cs.od edu/
Search
My Web BETA
Subscriptions (New) Shortcuts Advanced Search Preferences
Done

Search Results — Results 1 about 1 for **url:http://www.cs.odu.edu/** - 0.23 sec. (About this page)

1. Department of Computer Scien
College of Sciences
Category: Virginia > Norfolk > O minion University > Department of Computer Science
www.cs.**odu.edu** - 41k - Cached - More from this site - Save - Block



Home | Check on Status

In order to recover a lost website, please fill out the information below:

First name:*	Dwight
Last name:	Schrute
Email address:*	dschrute@dundermifflin.com
Website URL to recover:*	http://battlestar-galactica.com/

Example: http://www.example.com/

Recover entire website? ● Yes ○ No, only recover the single page Help

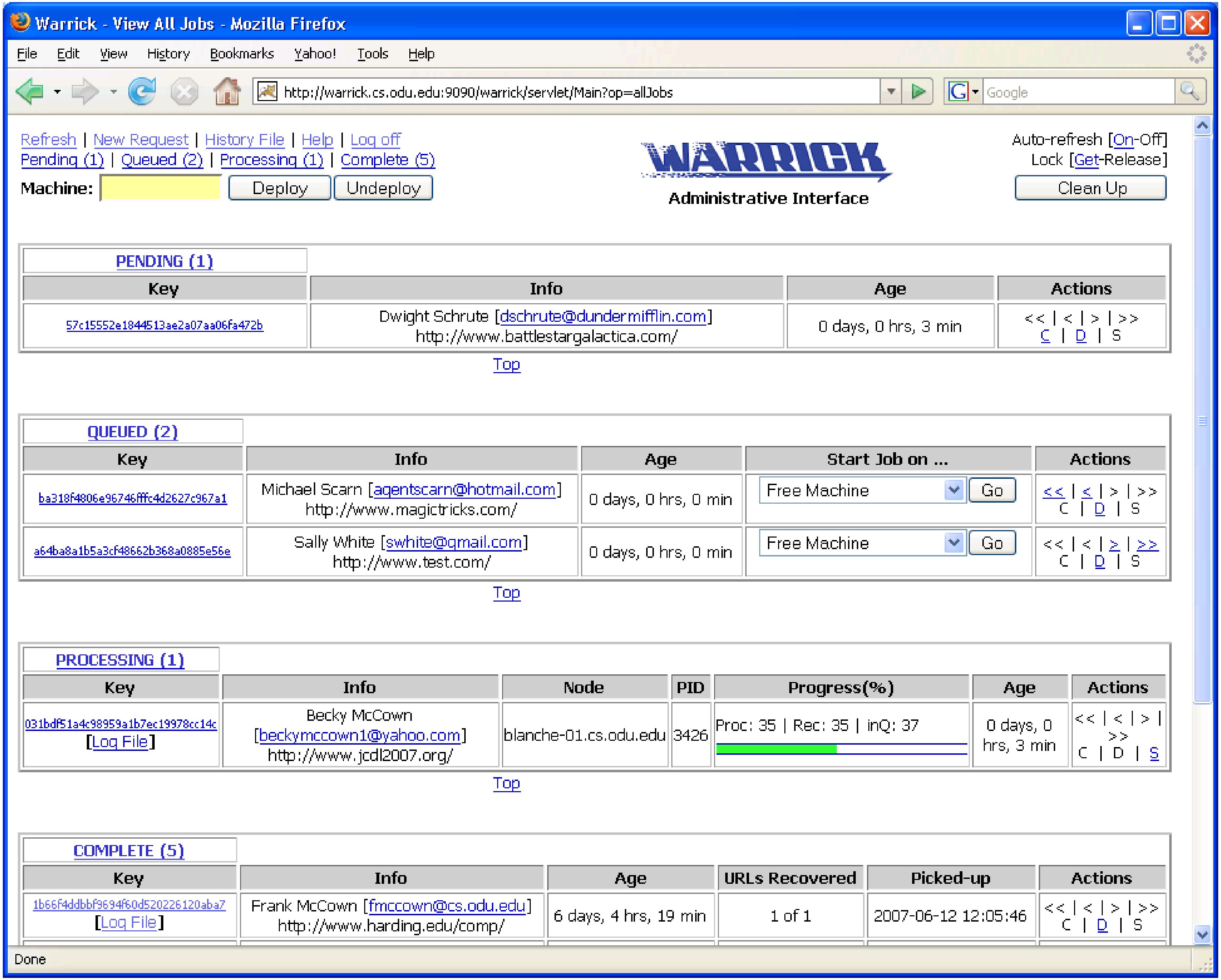
Repositories to use:** ☑ Internet Archive □ Limit dates From: yyyy-mm-dd To: yyyy-mm-dd Help

☑ Google ☑ Yahoo! ☑ Live Search

Use Windows filenames? ● Yes ○ No Help

* A required field.

** Only the latest resources from the Internet Archive will be returned unless dates are specified.


Warrick - View All Jobs - Mozilla Firefox
File
Edit
View
History
Bookmarks
Yahoo!
Tools
Help
Google
Refresh | New Request | History File | Help | Log off
Pending (1) | Queued (2) | Processing (1) | Complete (5)
Deploy
Undeploy
WARRICK
Auto-refresh [On-Off]
Lock [Get-Release]
Clean Up
Top
Done

http://warrick.cs.odu.edu:9090/warrick/servlet/Main?op=allJobs

Machine:

Administrative Interface

PENDING (1)

Key	Info	Age	Actions
57c15552e1844513ae2a07aa06fa472b	Dwight Schrute [dschrute@dundermifflin.com] http://www.battlestargalactica.com/	0 days, 0 hrs, 3 min	<< | < | > | >> C | D | S

QUEUED (2)

Key	Info	Age	Start Job on ...	Actions
ba318f4806e96746fffc4d2627c967a1	Michael Scarn [agentscarn@hotmail.com] http://www.magictricks.com/	0 days, 0 hrs, 0 min	Free Machine Go	<< | < | > | >> C | D | S
a64ba8a1b5a3cf48662b368a0885e56e	Sally White [swhite@gmail.com] http://www.test.com/	0 days, 0 hrs, 0 min	Free Machine Go	<< | < | > | >> C | D | S

PROCESSING (1)

Key	Info	Node	PID	Progress(%)	Age	Actions
031bdf51a4c98959a1b7ec19978cc14c [Log File]	Becky McCown [beckymccown1@yahoo.com] http://www.jcdl2007.org/	blanche-01.cs.odu.edu	3426	Proc: 35 | Rec: 35 | inQ: 37	0 days, 0 hrs, 3 min	<< | < | > | >> C | D | S

COMPLETE (5)

Key	Info	Age	URLs Recovered	Picked-up	Actions
1b66f4ddbbf9694f60d520226120aba7 [Log File]	Frank McCown [fmccown@cs.odu.edu] http://www.harding.edu/comp/	6 days, 4 hrs, 19 min	1 of 1	2007-06-12 12:05:46	<< | < | > | >> C | D | S

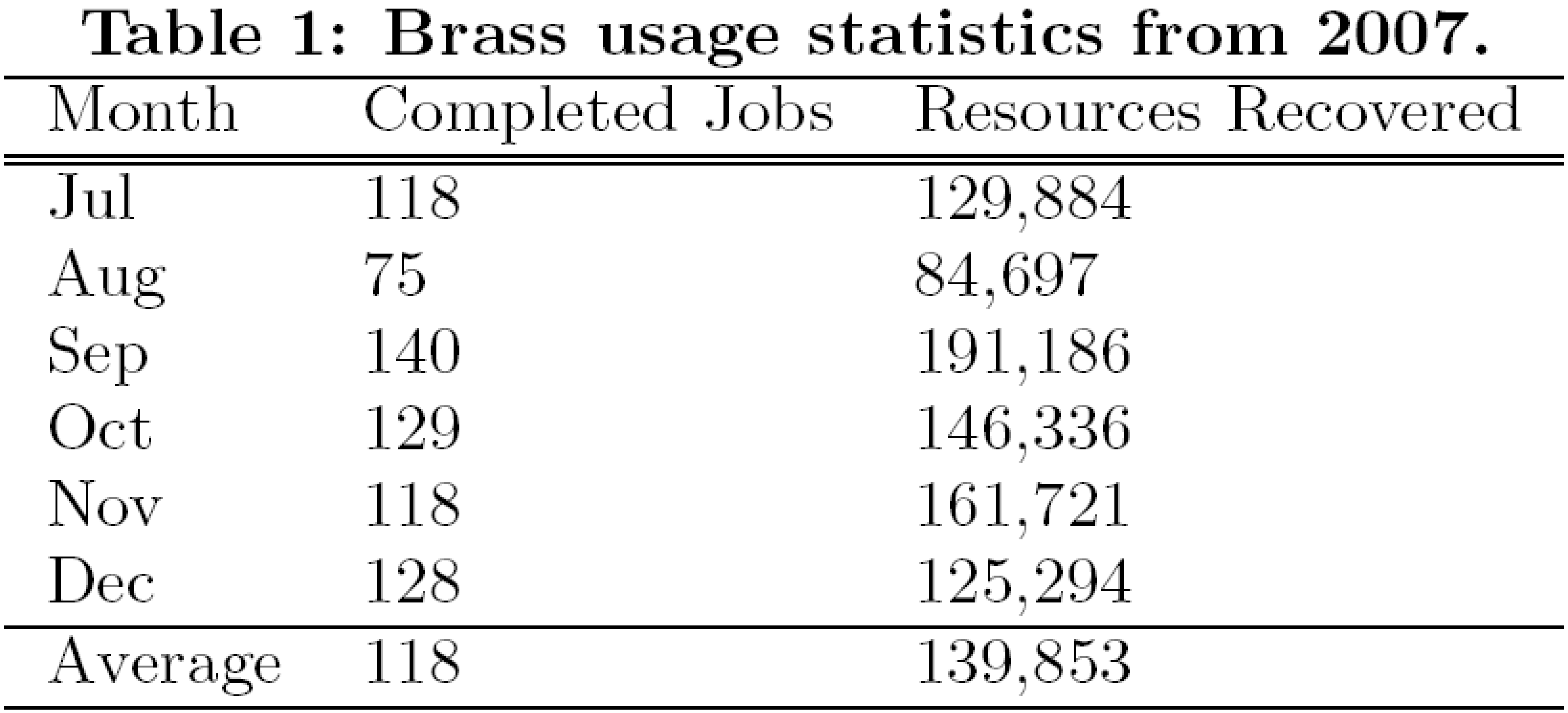


Table 1: Brass usage statistics from 2007.

Month	Completed Jobs	Resources Recovered
Jul	118	129,884
Aug	75	84,697
Sep	140	191,186
Oct	129	146,336
Nov	118	161,721
Dec	128	125,294
Average	118	139,853

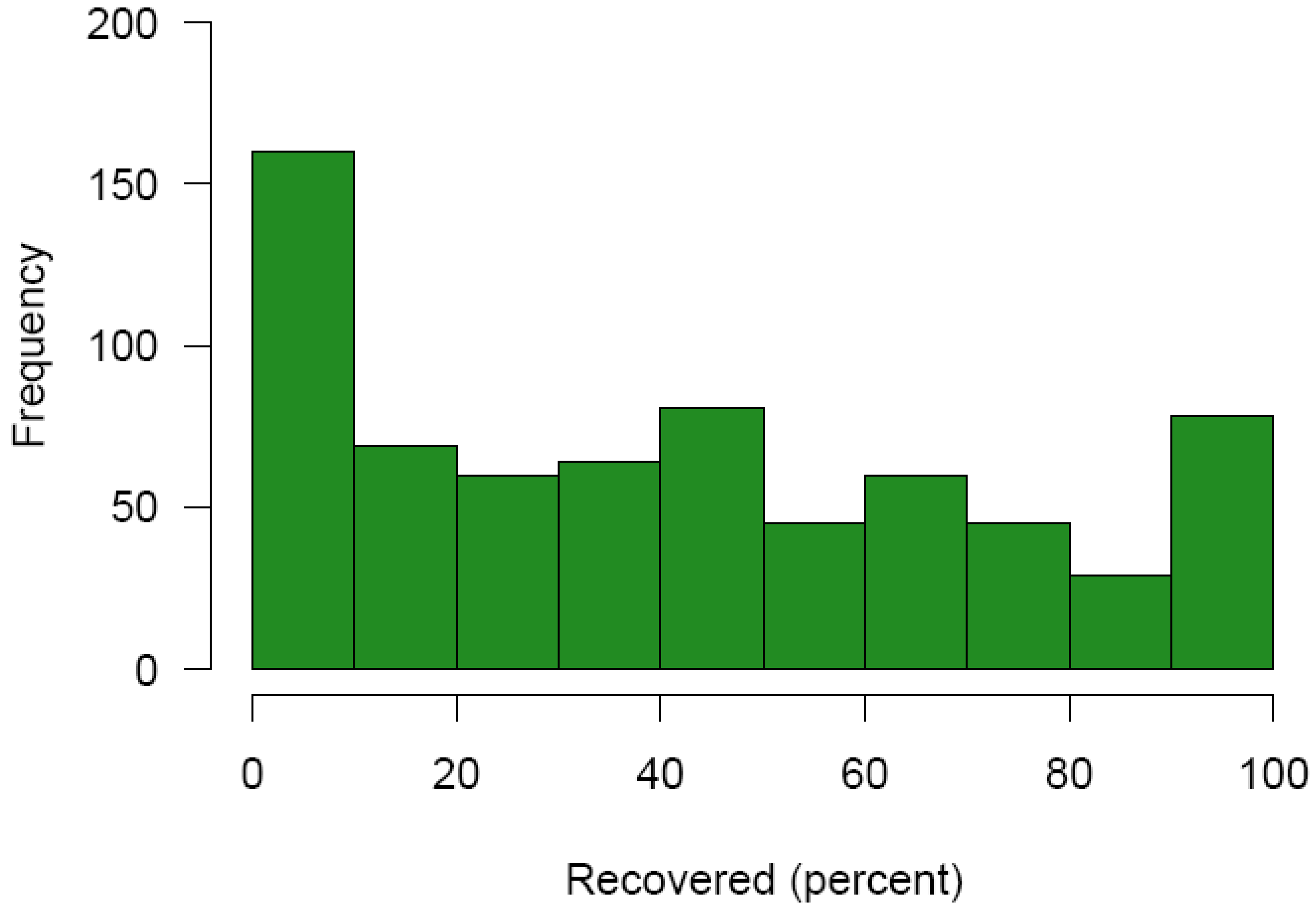


Figure 2: Distribution of websites by percentage of recovered resources.

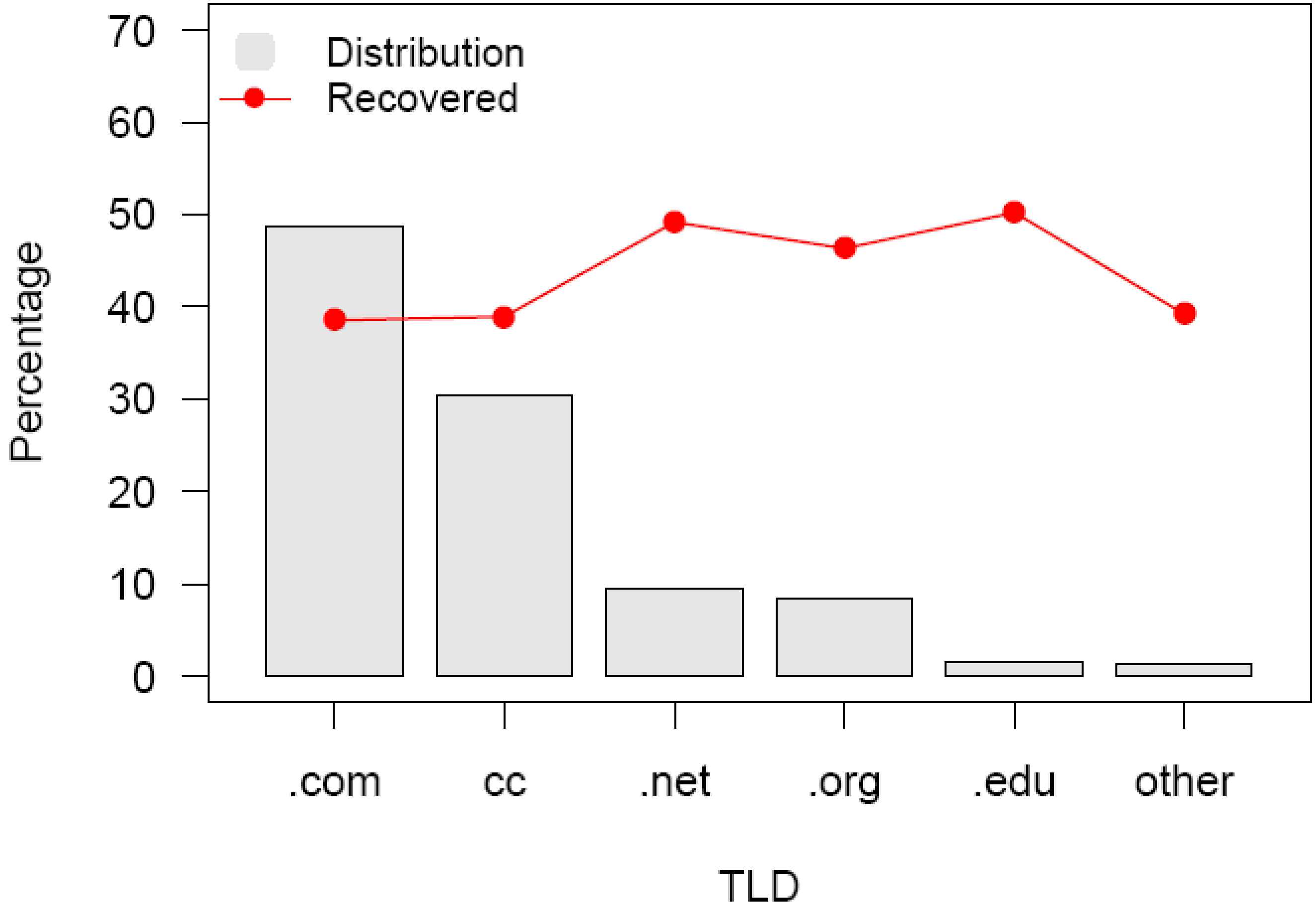


Figure 3: Brass recovered websites by TLD.

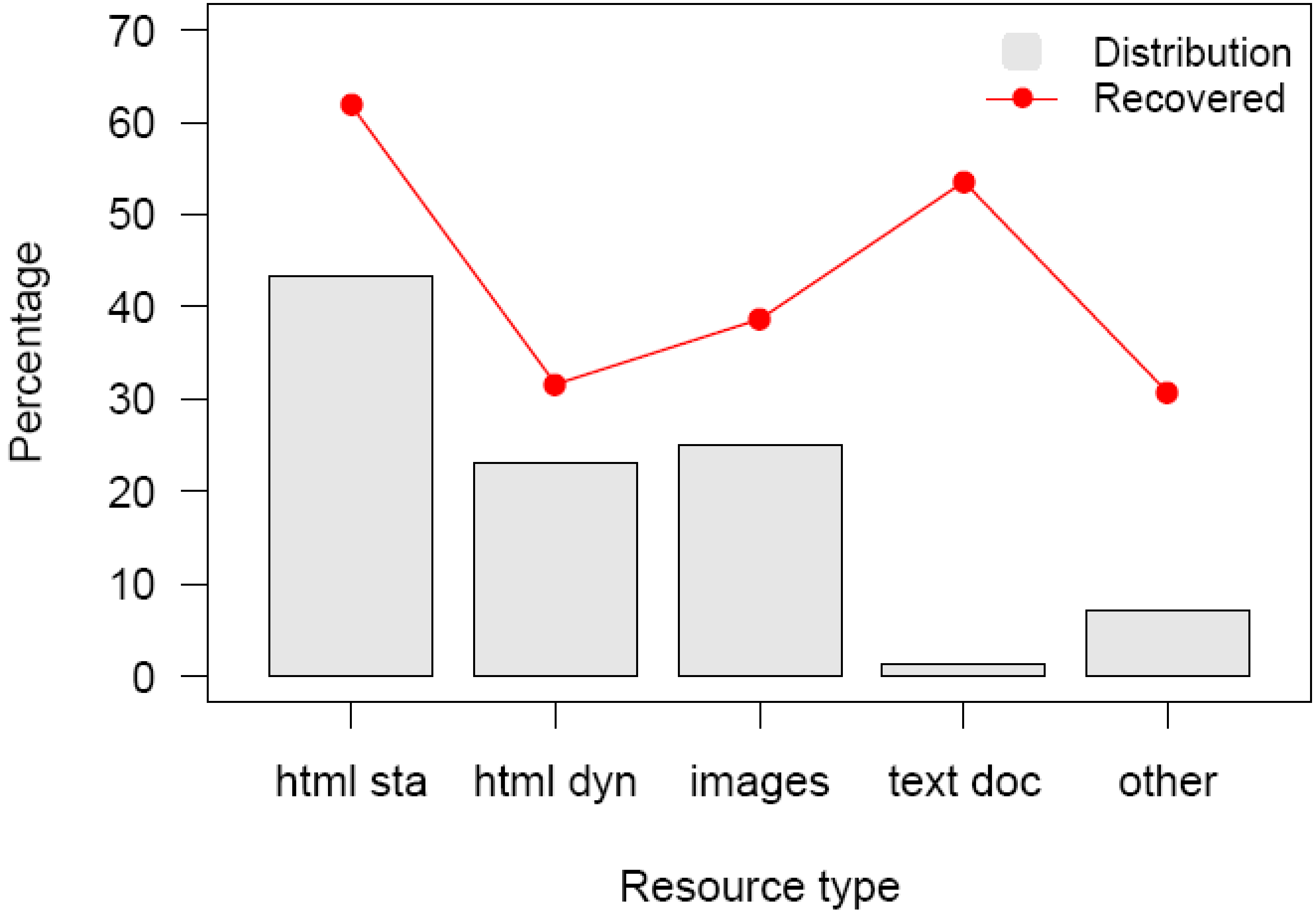


Figure 4: Brass recovered resources by type.

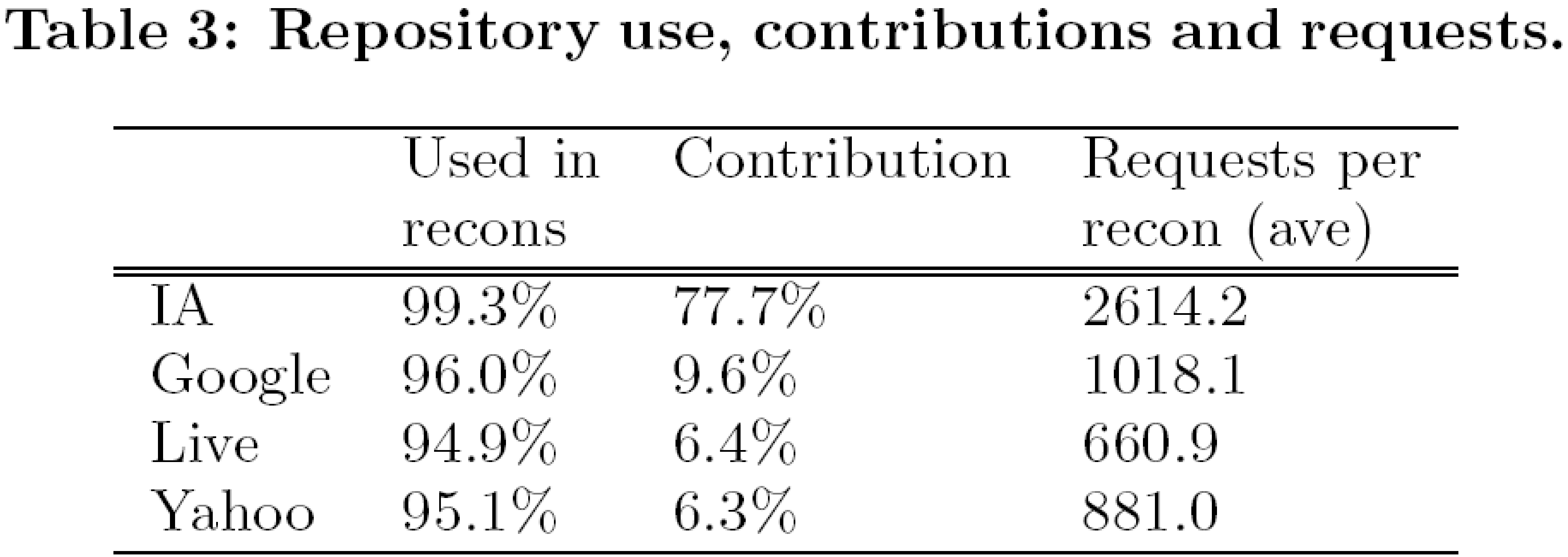


Table 3: Repository use, contributions and requests.

	Used in recons	Contribution	Requests per recon (ave)
IA	99.3%	77.7%	2614.2
Google	96.0%	9.6%	1018.1
Live	94.9%	6.4%	660.9
Yahoo	95.1%	6.3%	881.0

Summary

- 118 websites are recovered each month on average
- 84% of jobs have at least 1 recovered resource
- 41% of resources are recovered on average
- 62% of static HTML resources are recovered on average
- 78% of resources are recovered from Internet Archive
- Many .jp URLs are submitted, but few jobs are picked up – spam?

Thank You

One recovered website coming up…

Frank McCown

fmccown@harding.edu

http://www.harding.edu/fmccown/