Old Dominion University
ODU Digital Commons

Computer Science Presentations | Computer Science

3-13-2009

Can't Find Your 404s?

Martin Klein
Old Dominion University

Frank McCown
Old Dominion University

Joan Smith
Old Dominion University

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_presentations

Part of the Archival Science Commons, and the Computer Sciences Commons

Recommended Citation

Klein, Martin; McCown, Frank; Smith, Joan; and Nelson, Michael L., "Can't Find Your 404s?" (2009). *Computer Science Presentations*. 24.
https://digitalcommons.odu.edu/computerscience_presentations/24

This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

Can’t Find Your 404s?

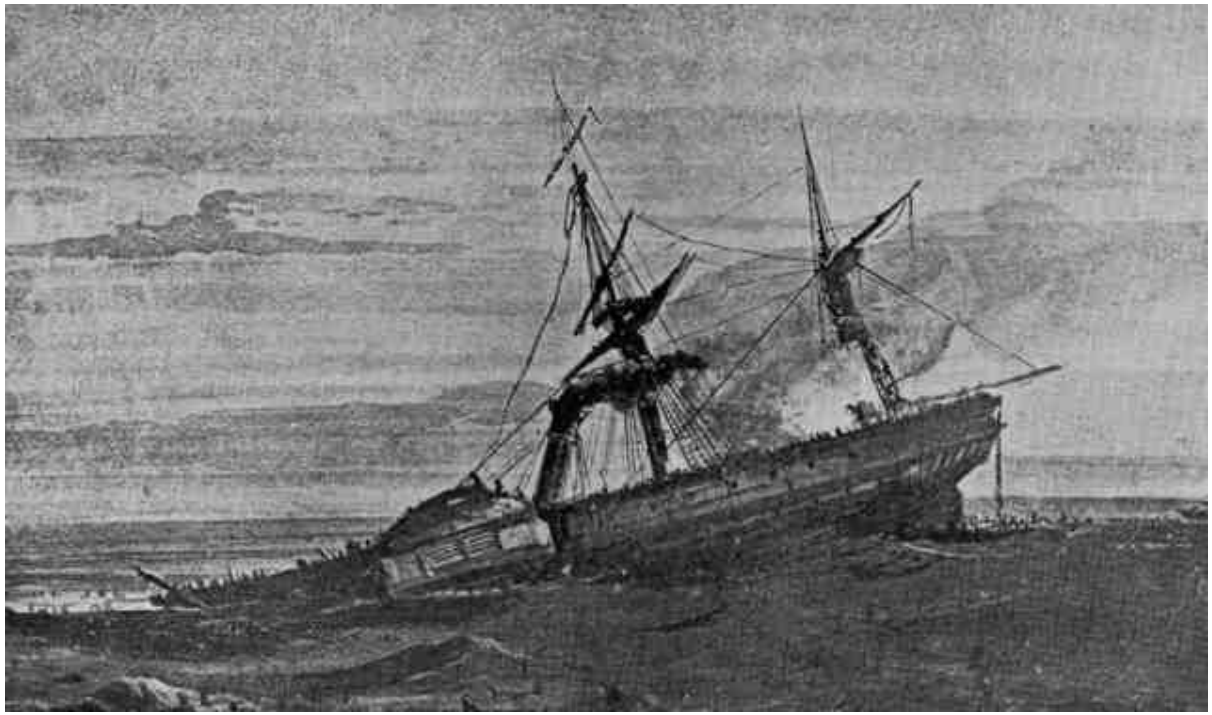
Santa Fe Complex
March 13, 2009

Martin Klein, Frank McCown,
Joan Smith, Michael L. Nelson
Department of Computer Science
Old Dominion University
Norfolk VA

Background

- “We can’t save everything!”
  - if not “everything”, then how much?
  - what does “save” mean?

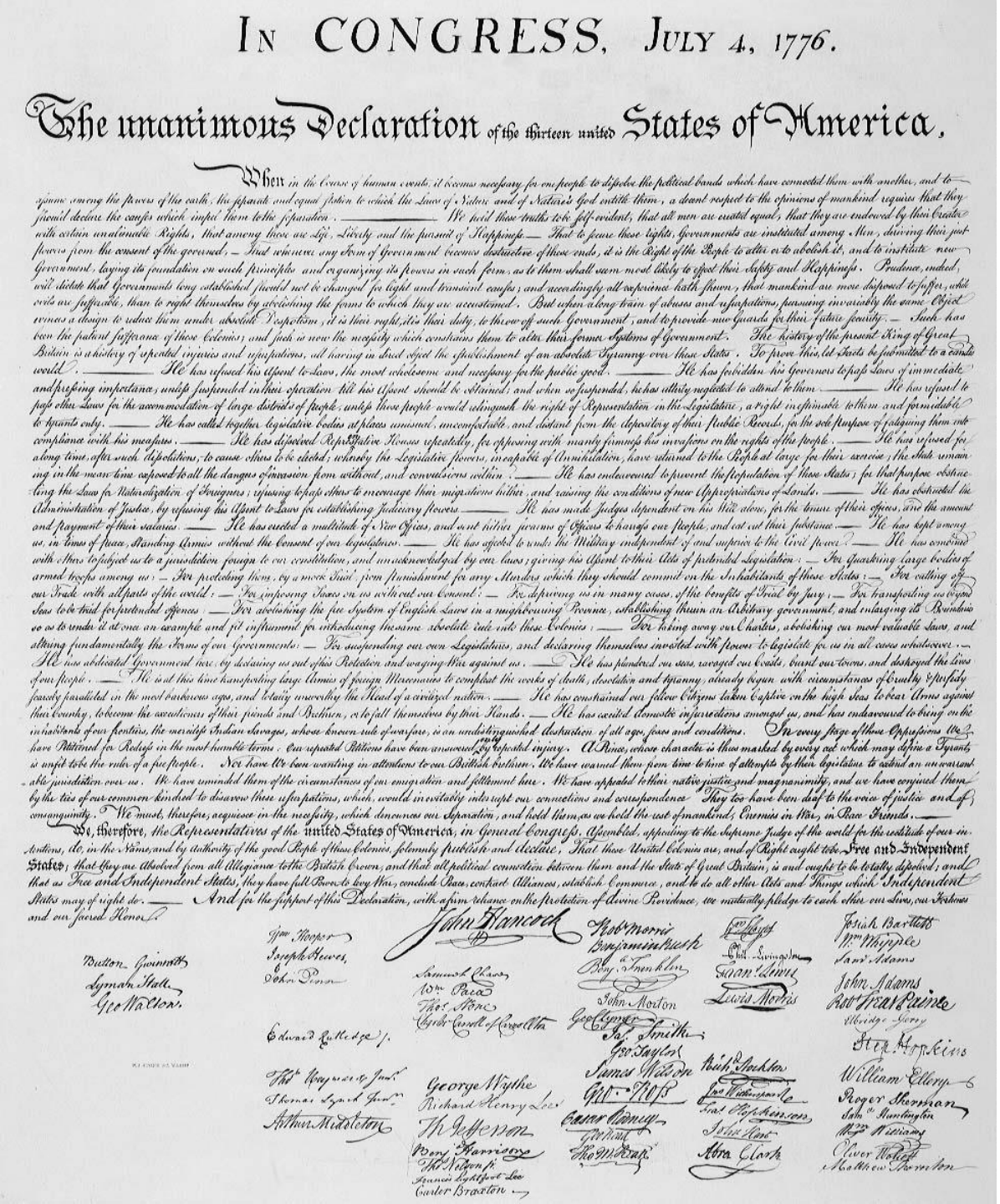
“Women and Children First”



HMS Birkenhead, Cape Danger, 1852

638 passengers 193 survivors all 7 women & 13 children survived

image from: http://www.btinternet.com/~palmiped/Birkenhead.htm



IN CONGRESS, JULY 4, 1776.

The unanimous Declaration of the thirteen united States of America,

When in the Course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to assume among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation. — We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. — That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, — That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness. Prudence, indeed, will dictate that Governments long established should not be changed for light and transient causes; and accordingly all experience hath shewn, that mankind are more disposed to suffer, while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, pursuing invariably the same Object evinces a design to reduce them under absolute Despotism, it is their right, it is their duty, to throw off such Government, and to provide new Guards for their future security. — Such has been the patient sufferance of these Colonies; and such is now the necessity which constrains them to alter their former Systems of Government. The history of the present King of Great Britain is a history of repeated injuries and usurpations, all having in direct object the establishment of an absolute Tyranny over these States. To prove this, let Facts be submitted to a candid world. — He has refused his Assent to Laws, the most wholesome and necessary for the public good. — He has forbidden his Governors to pass Laws of immediate and pressing importance, unless suspended in their operation till his Assent should be obtained; and when so suspended, he has utterly neglected to attend to them. — He has refused to pass other Laws for the accommodation of large districts of people, unless those people would relinquish the right of Representation in the Legislature, a right inestimable to them and formidable to tyrants only. — He has called together legislative bodies at places unusual, uncomfortable, and distant from the depository of their public Records, for the sole purpose of fatiguing them into compliance with his measures. — He has dissolved Representative Houses repeatedly, for opposing with manly firmness his invasions on the rights of the people. — He has refused for a long time, after such dissolutions, to cause others to be elected; whereby the Legislative powers, incapable of Annihilation, have returned to the People at large for their exercise; the State remaining in the mean time exposed to all the dangers of invasion from without, and convulsions within. — He has endeavoured to prevent the population of these States; for that purpose obstructing the Laws for Naturalization of Foreigners; refusing to pass others to encourage their migrations hither, and raising the conditions of new Appropriations of Lands. — He has obstructed the Administration of Justice, by refusing his Assent to Laws for establishing Judiciary powers. — He has made Judges dependent on his Will alone, for the tenure of their offices, and the amount and payment of their salaries. — He has erected a multitude of New Offices, and sent hither swarms of Officers to harass our People, and eat out their substance. — He has kept among us, in times of peace, Standing Armies without the Consent of our legislatures. — He has affected to render the Military independent of and superior to the Civil power. — He has combined with others to subject us to a jurisdiction foreign to our constitution, and unacknowledged by our laws; giving his Assent to their Acts of pretended legislation: — For Quartering large bodies of armed troops among us: — For protecting them, by a mock Trial, from punishment for any Murders which they should commit on the Inhabitants of these States: — For cutting off our Trade with all parts of the world: — For imposing Taxes on us without our Consent: — For depriving us in many cases, of the benefits of Trial by jury: — For transporting us beyond Seas to be tried for pretended offences: — For abolishing the free System of English Laws in a neighbouring Province, establishing therein an Arbitrary government, and enlarging its Boundaries so as to render it at once an example and fit instrument for introducing the same absolute rule into these Colonies: — For taking away our Charters, abolishing our most valuable Laws and altering fundamentally the Forms of our Governments: — For suspending our own Legislatures, and declaring themselves invested with power to legislate for us in all cases whatsoever. — He has abdicated Government here, by declaring us out of his Protection and waging War against us. — He has plundered our seas, ravaged our Coasts, burnt our towns, and destroyed the lives of our people. — He is at this time transporting large Armies of foreign Mercenaries to compleat the works of death, desolation and tyranny, already begun with circumstances of Cruelty & Perfidy scarcely paralleled in the most barbarous ages, and totally unworthy the Head of a civilized nation. — He has constrained our fellow Citizens taken Captive on the high Seas to bear Arms against their Country, to become the executioners of their friends and Brethren, or to fall themselves by their Hands. — He has excited domestic insurrections amongst us, and has endeavoured to bring on the inhabitants of our frontiers, the merciless Indian Savages, whose known rule of warfare, is an undistinguished destruction of all ages, sexes and conditions. In every stage of these Oppressions We have Petitioned for Redress in the most humble terms: Our repeated Petitions have been answered only by repeated injury. A Prince, whose character is thus marked by every act which may define a Tyrant, is unfit to be the ruler of a free people. Nor have We been wanting in attentions to our British brethren. We have warned them from time to time of attempts by their legislature to extend an unwarrantable jurisdiction over us. We have reminded them of the circumstances of our emigration and settlement here. We have appealed to their native justice and magnanimity, and we have conjured them by the ties of our common kindred to disavow these usurpations, which, would inevitably interrupt our connections and correspondence. They too have been deaf to the voice of justice and of consanguinity. We must, therefore, acquiesce in the necessity, which denounces our Separation, and hold them, as we hold the rest of mankind, Enemies in War, in Peace Friends.

We, therefore, the Representatives of the united States of America, in General Congress, Assembled, appealing to the Supreme Judge of the world for the rectitude of our intentions, do, in the Name, and by Authority of the good People of these Colonies, solemnly publish and declare, That these United Colonies are, and of Right ought to be Free and Independent States; that they are Absolved from all Allegiance to the British Crown, and that all political connexion between them and the State of Great Britain, is and ought to be totally dissolved; and that as Free and Independent States, they have full Power to levy War, conclude Peace, contract Alliances, establish Commerce, and to do all other Acts and Things which Independent States may of right do. — And for the support of this Declaration, with a firm reliance on the protection of divine Providence, we mutually pledge to each other our Lives, our Fortunes and our sacred Honor.

John Hancock

Button Gwinnett
Lyman Hall
Geo Walton

Wm Hooper
Joseph Hewes
John Penn

Edward Rutledge
Thos Heyward Junr
Thomas Lynch Junr
Arthur Middleton

Samuel Chase
Wm Paca
Thos Stone
Charles Carroll of Carrollton

George Wythe
Richard Henry Lee
Th Jefferson
Benj Harrison
Thos Nelson jr
Francis Lightfoot Lee
Carter Braxton

Robt Morris
Benjamin Rush
Benj Franklin
John Morton

Geo Clymer
Jas Smith
Geo Taylor
James Wilson
Geo Ross

Caesar Rodney
Geo Read
Tho M:Kean

Wm Floyd
Phil Livingston
Frans Lewis
Lewis Morris

Richd Stockton
Jno Witherspoon
Fras Hopkinson
John Hart
Abra Clark

Josiah Bartlett
Wm Whipple

Saml Adams
John Adams
Robt Treat Paine
Elbridge Gerry

Step Hopkins
William Ellery

Roger Sherman
Sam Huntington
Wm Williams
Oliver Wolcott

Matthew Thornton

We should probably save a copy of this…

Or maybe we don't have to…

the Wikipedia link is in the top 10, so we're ok, right?

NACA TN 2434

NATIONAL ADVISORY COMMITTEE FOR AERONAUTICS

TECHNICAL NOTE 2434

A BIHARMONIC RELAXATION METHOD FOR CALCULATING THERMAL STRESS IN COOLED IRREGULAR CYLINDERS

By Arthur G. Holms

Lewis Flight Propulsion Laboratory
Cleveland, Ohio

LIBRARY COPY
MAY 12 1953
LANGLEY RESEARCH CENTER
LIBRARY NASA
HAMPTON, VIRGINIA

NACA

Washington

August 1951

FOR REFERENCE

NOT TO BE TAKEN FROM THIS ROOM

Surely we're saving copies of this…

2 copies in the UK

2 Dublin Core records

That's probably good enough…

What about the things that we know we don't need to keep?

You *DO* support recycling, right?

Yugoslav phone books may be the last record of a people

http://www.umich.edu/~urecord/9900/Jan24_(

The University Record, January 24, 2000

Yugoslav phone books may be the last record of a people

By Joanne Nesbit
News and Information Services

Janet Crayne with one of the Yugoslav phone books that have been collected to help those who fled Kosovo establish residency. The project involves transferring more than 16 million entries to microfilm. Photo by Paul Jaronski, U-M Photo Services

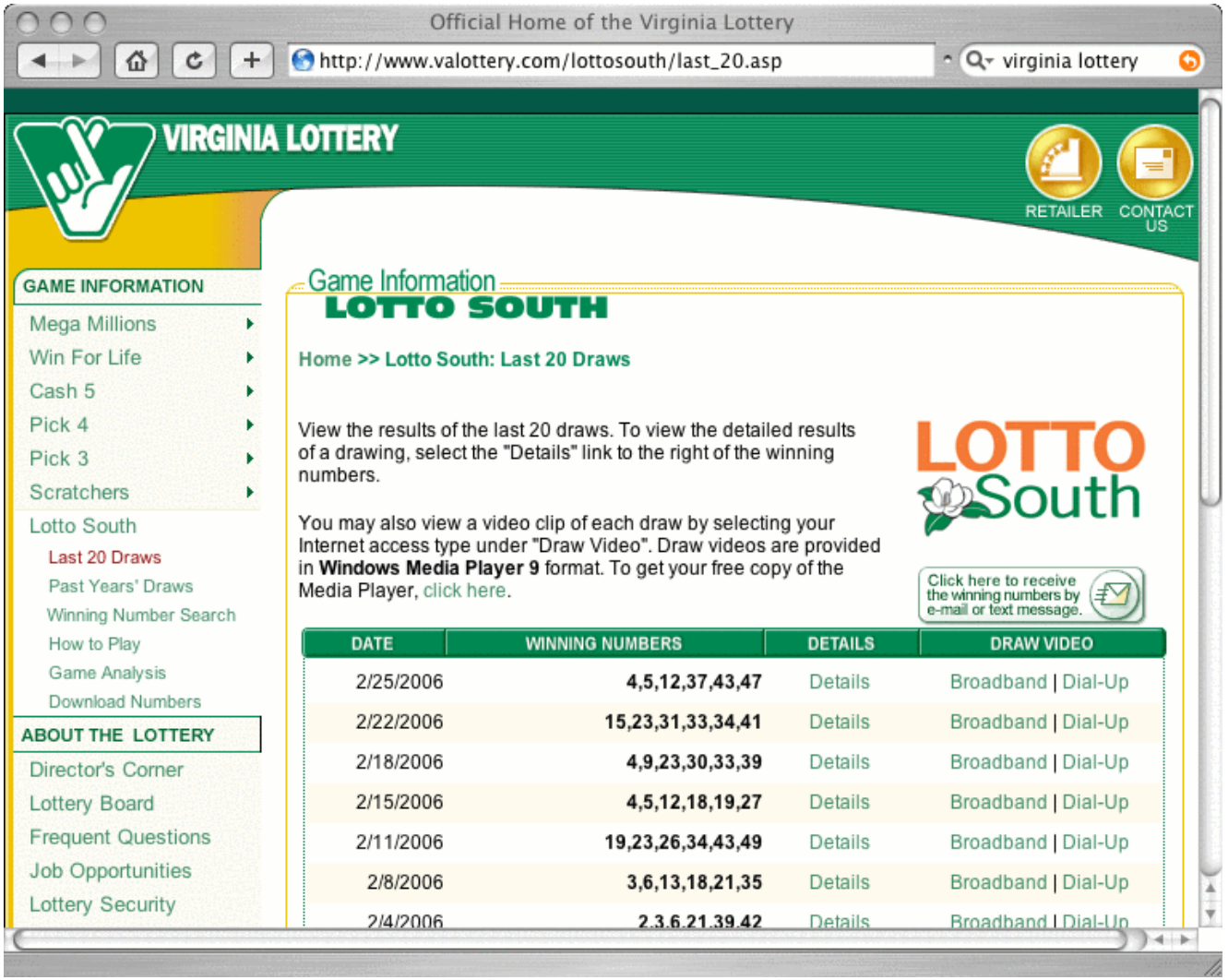
When the call went out last summer, Yugoslav phone books began arriving at the University from around the world. The telephone books from 1968 to 1995 contained what might be the only records available to document residency for the close to one million people forced to evacuate Kosovo.

Most of the government archives in Kosovo have been burned, and personal documentation was stripped from refugees prior to their crossing the Kosovo borders to other countries. The major library holdings in Bosnia, including the collections in the National and University Library and the Oriental Institute, were destroyed by Serb Nationalist artillery during the siege of Sarajevo. The Sarajevans were forced to burn the few surviving books for fuel during the winters. And the last Yugoslav census did not provide reliable and/or consistent demographic information, says Janet Crayne, project coordinator at the Library's Slavic Division.

An advisory board representing the U-M, Harvard University, the University of Montreal and University of Washington, in collaboration with microfilmer Norman Ross Inc., coordinated the collection and filming of phone books from former Yugoslav republics and their successor states Bosnia and Hercegovina, Croatia, Macedonia, Yugoslavia (Serbia, Montenegro, Vojvodina and Kosovo) and Slovenia. Donations were international in scope

A higher moral calling for pack rats?

Just Keep the Important Stuff!



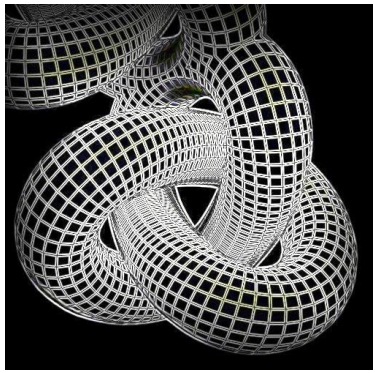
Preservation: Fortress Model

Five Easy Steps for Preservation:

1. Get a lot of $
2. Buy a lot of disks, machines, tapes, etc.
3. Hire an army of staff
4. Load a small amount of data
5. “Look upon my archive ye Mighty, and despair!”

image from: http://www.itunisie.com/tourisme/excursion/tabarka/images/fort.jpg

Alternate Models of Preservation



- Lazy Preservation
  - "How much preservation do I get if I do absolutely nothing?"
- Just-In-Time Preservation
  - Wait for it to disappear first, then a "good enough" version
- Shared Infrastructure Preservation
  - Push your content to sites that *might* preserve it
- Web Server Enhanced Preservation
  - Use Apache modules to create archival-ready resources

image from: http://www.proex.ufes.br/arsm/knots_interlaced.htm

The Problem

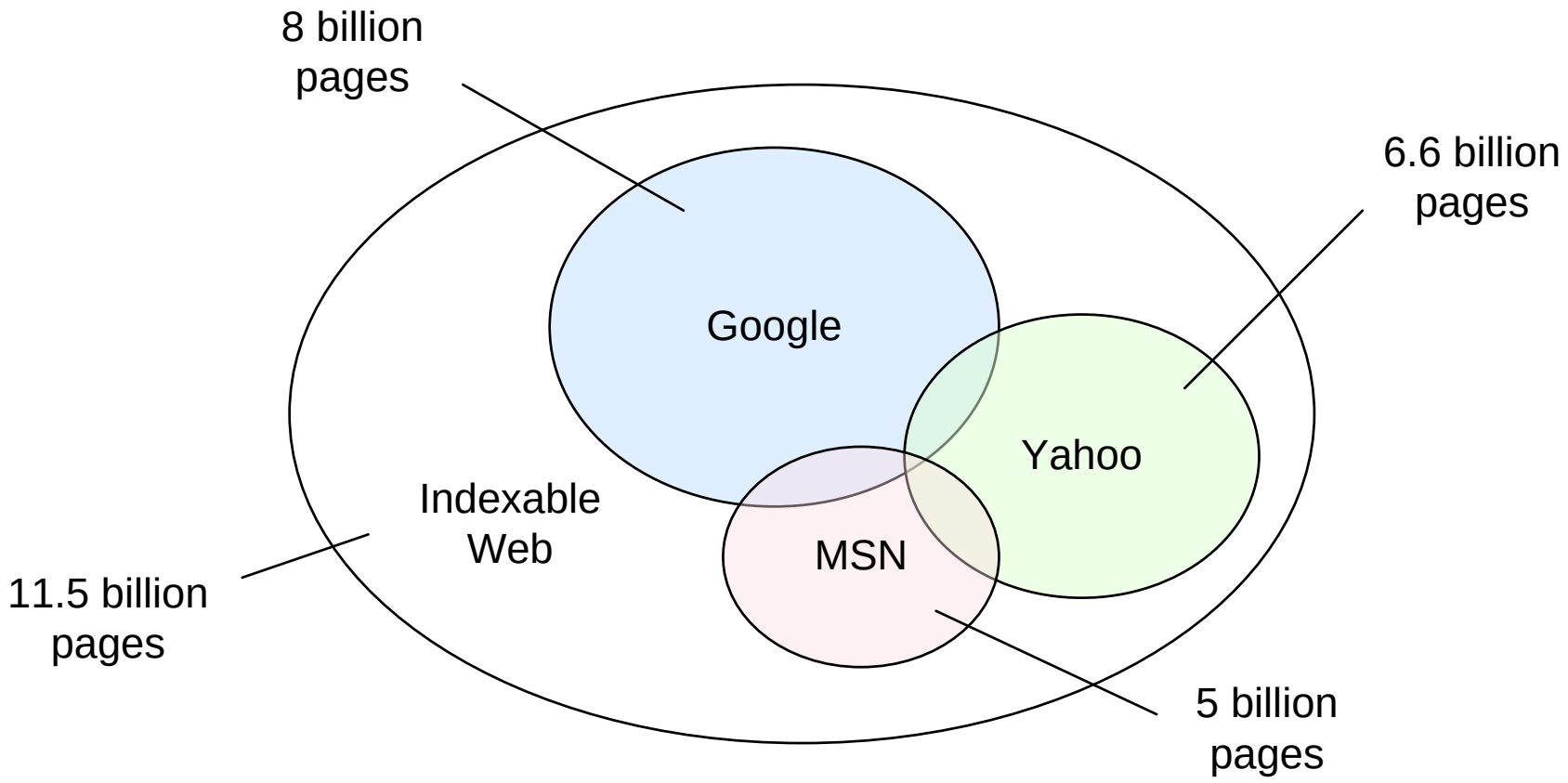
- Web users experience 404 errors
  - Expected lifetime of a web page is 44 days [Kahle97]
  - 2% of web disappears every week [Fetterly03]
- Are they really gone? Or just relocated?
  - Has anybody crawled and indexed them?
  - Do Google, Yahoo, MSN or the Internet Archive have a copy of the page?
- To find new or different copies, we need to combine *information retrieval* techniques with *digital preservation*

The Environment

Web Infrastructure (WI) [McCown07]

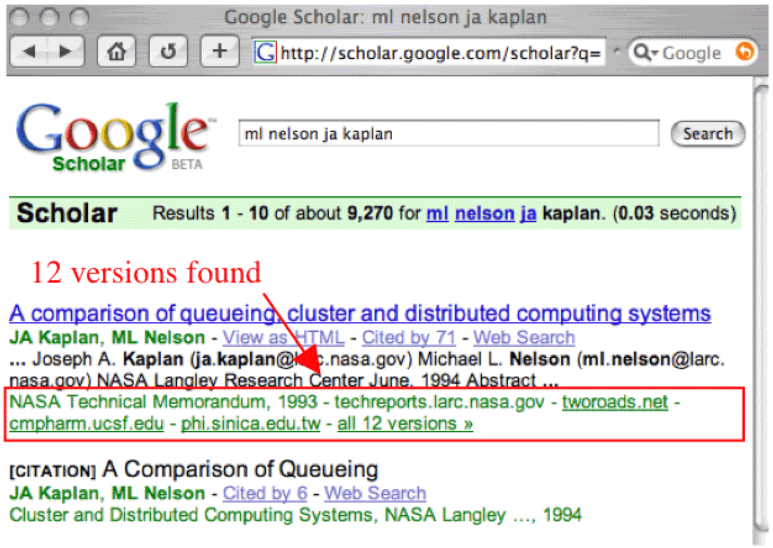
- Web search engines (Google, Yahoo, MSN Live) and their caches
- Research Projects (CiteSeer, NSDL)
- Web archives (Internet Archive)

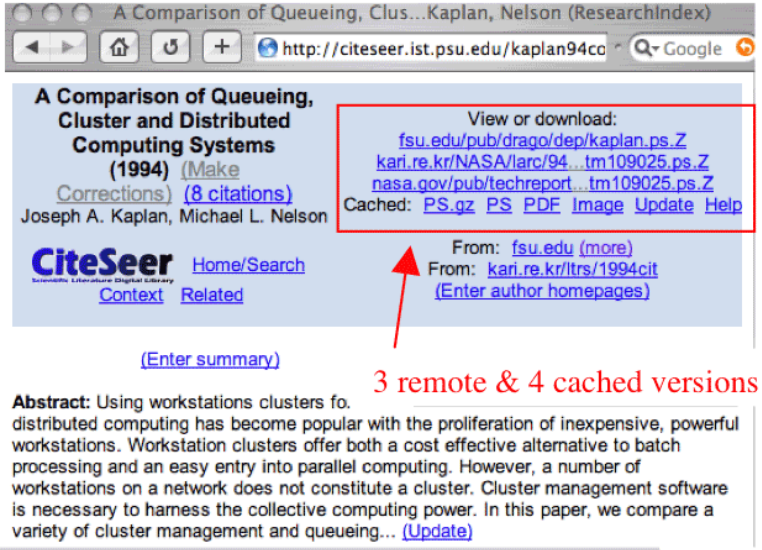
How much of the Web is indexed?

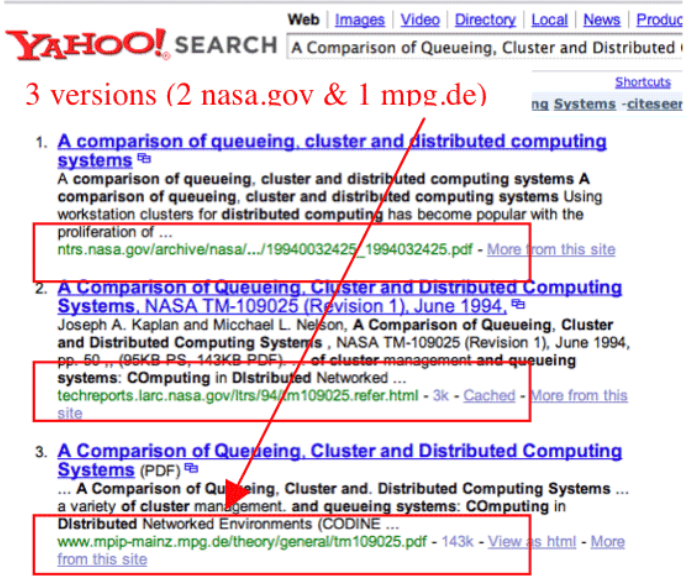


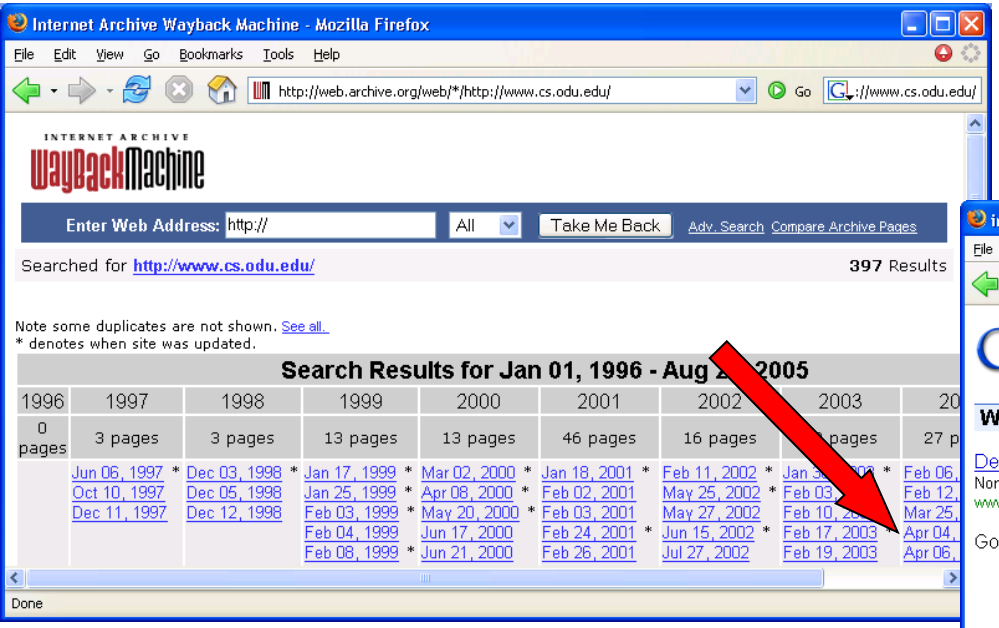
Estimates from “The Indexable Web is More than 11.5 billion pages” by Gulli and Signorini (WWW’05)

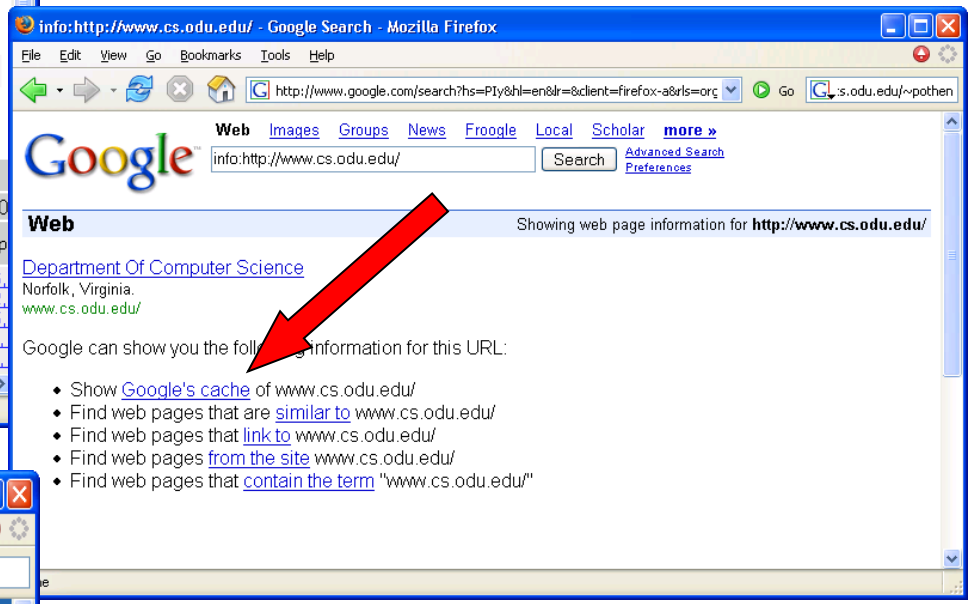
Web Infrastructure: Refreshing & Migrating

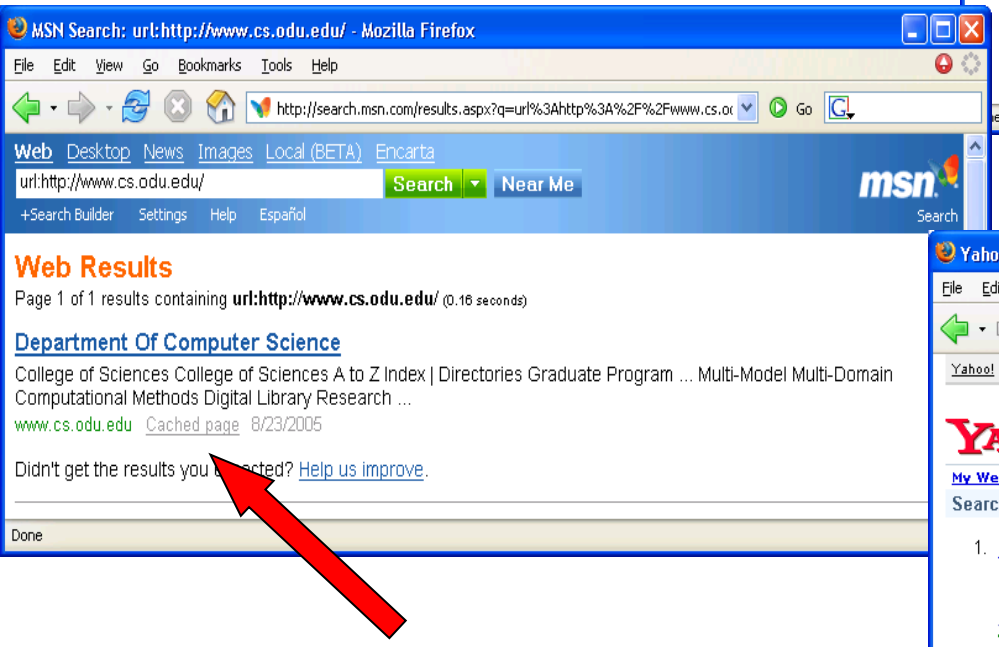


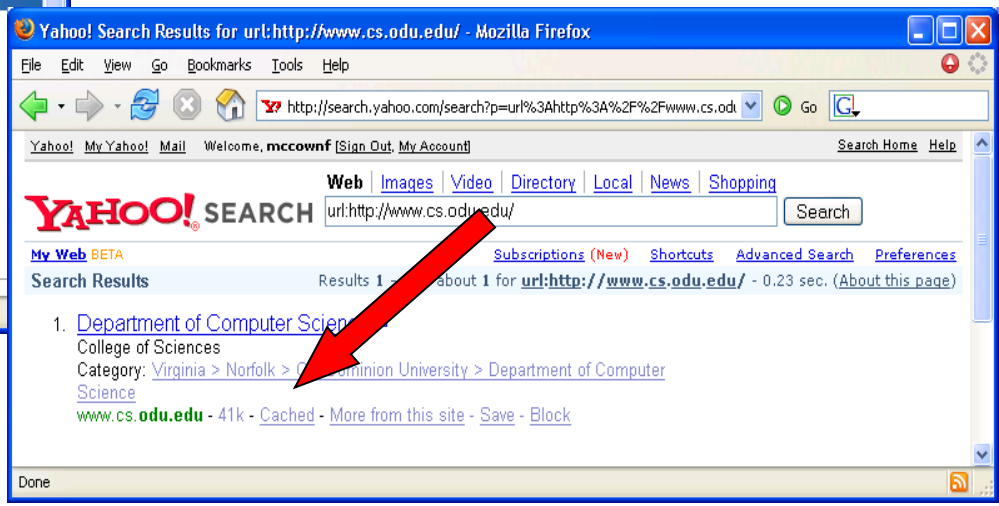




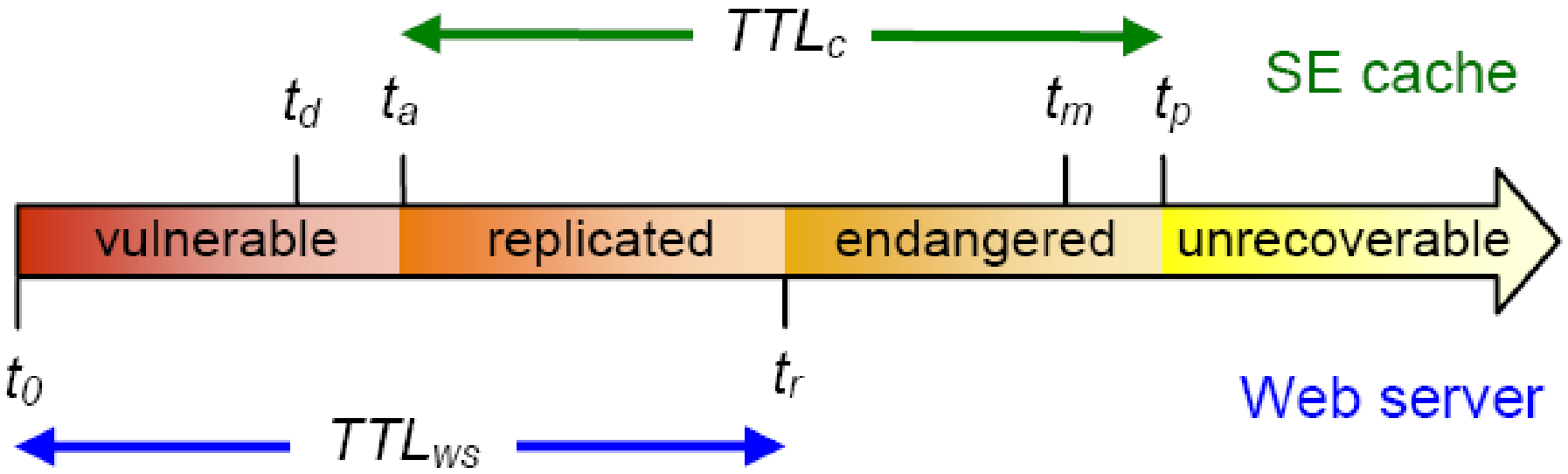

Internet Archive Wayback Machine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://web.archive.org/web/*/http://www.cs.odu.edu/
INTERNET ARCHIVE
WaybackMachine
Enter Web Address: http://
All
Take Me Back
Adv. Search
Compare Archive Pages
Searched for http://www.cs.odu.edu/
397 Results
Note some duplicates are not shown. See all.
* denotes when site was updated.
Search Results for Jan 01, 1996 - Aug 2005
1996
1997
1998
1999
2000
2001
2002
2003
0 pages
3 pages
3 pages
13 pages
13 pages
46 pages
16 pages
27 p
Jun 06, 1997
Oct 10, 1997
Dec 11, 1997
Dec 03, 1998
Dec 05, 1998
Dec 12, 1998
Jan 17, 1999
Jan 25, 1999
Feb 03, 1999
Feb 04, 1999
Feb 08, 1999
Mar 02, 2000
Apr 08, 2000
May 20, 2000
Jun 17, 2000
Jun 21, 2000
Jan 18, 2001
Feb 02, 2001
Feb 03, 2001
Feb 24, 2001
Feb 26, 2001
Feb 11, 2002
May 25, 2002
May 27, 2002
Jun 15, 2002
Jul 27, 2002
Feb 17, 2003
Feb 19, 2003
Apr 04
Apr 06
Done

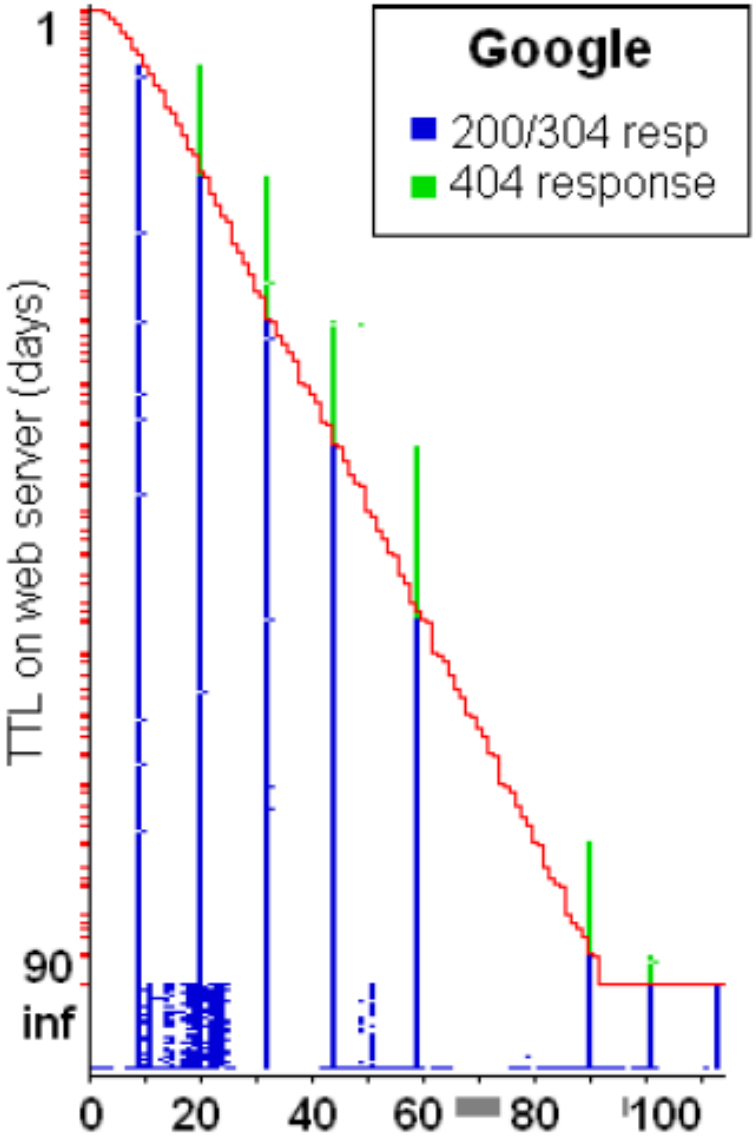

info:http://www.cs.odu.edu/ - Google Search - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Google
Web Images Groups News Froogle Local Scholar more »
info:http://www.cs.odu.edu/
Search
Advanced Search
Preferences
Web
Showing web page information for http://www.cs.odu.edu/
Department Of Computer Science
Norfolk, Virginia.
www.cs.odu.edu/
Google can show you the following 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/"

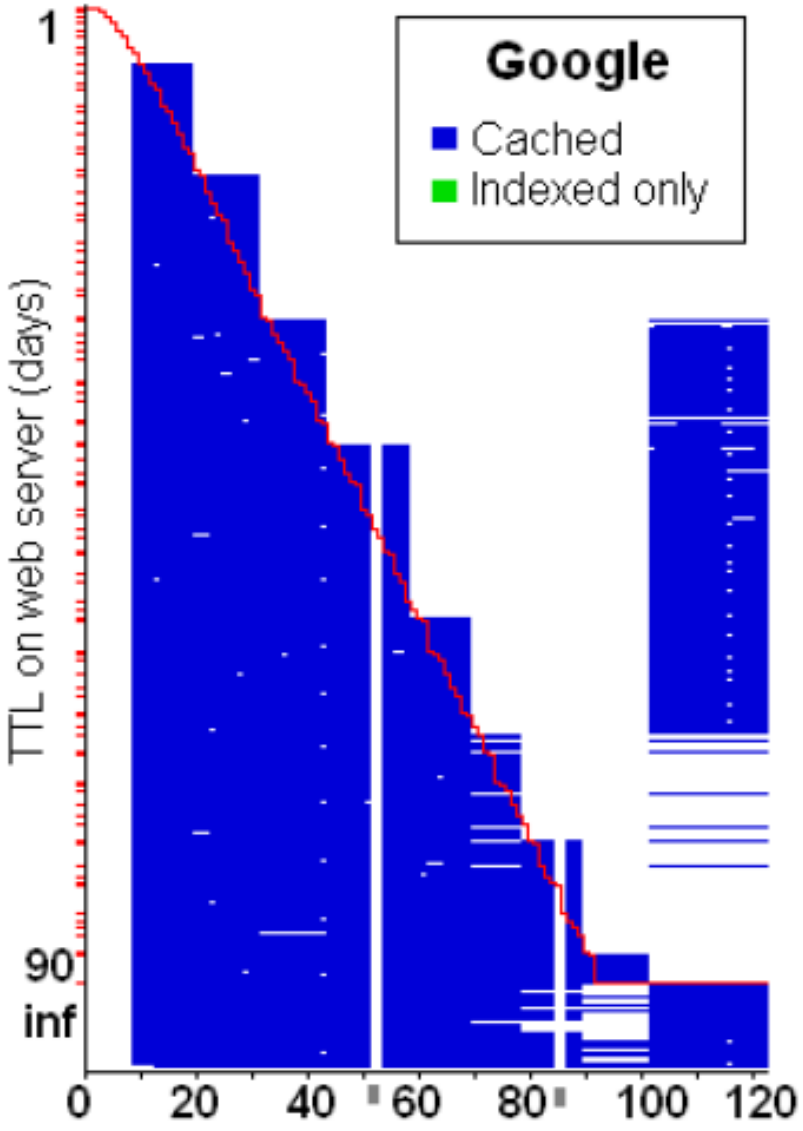

MSN Search: url:http://www.cs.odu.edu/ - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Web Desktop News Images Local (BETA) Encarta
url:http://www.cs.odu.edu/
Search
Near Me
msn
+Search Builder Settings Help Español
Search
Web Results
Page 1 of 1 results containing url:http://www.cs.odu.edu/ (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 expected? Help us improve.
Done

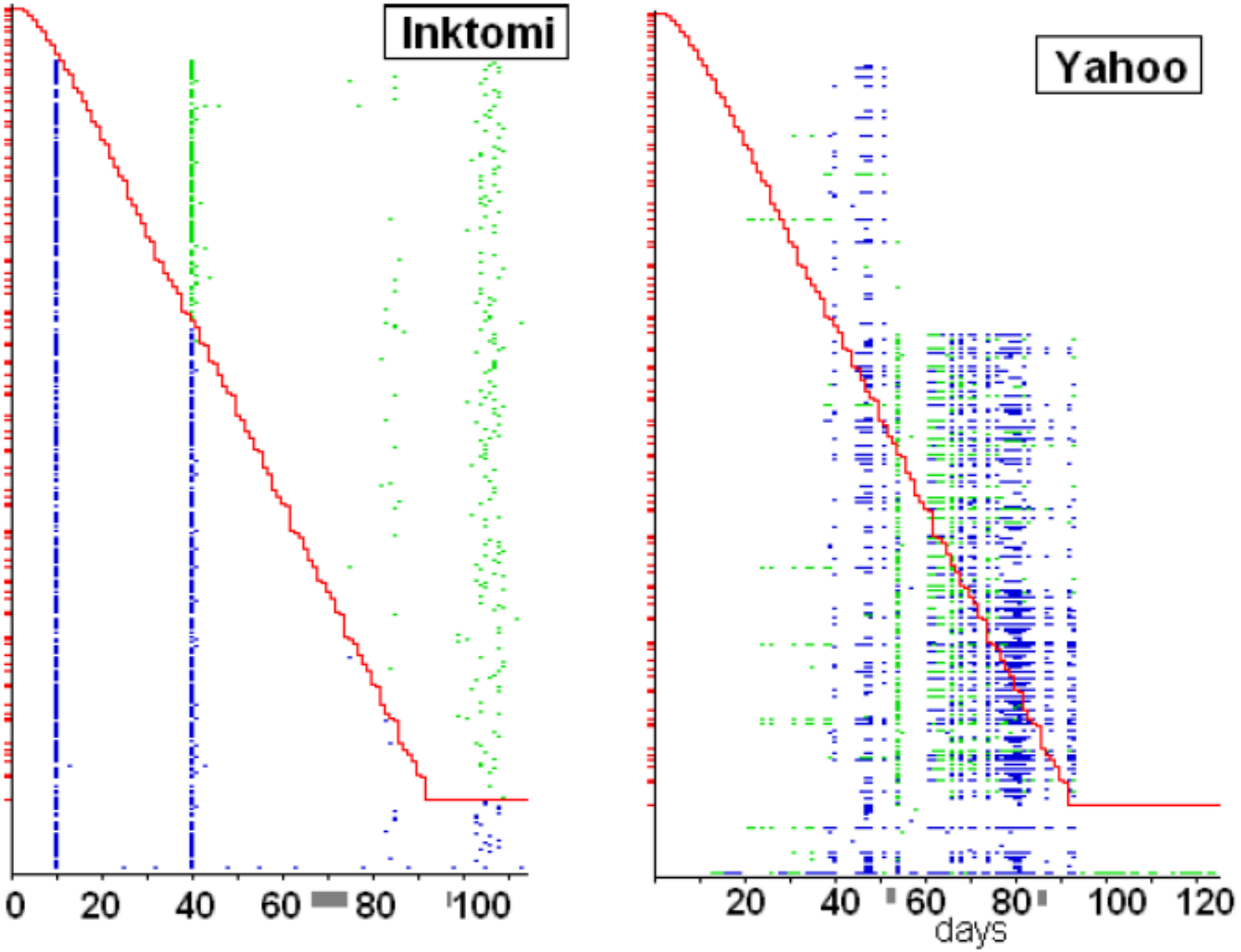

Yahoo! Search Results for url:http://www.cs.odu.edu/ - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Yahoo! My Yahoo! Mail Welcome, mccownf [Sign Out, My Account]
Search Home Help
Web Images Video Directory Local News Shopping
YAHOO! SEARCH
url:http://www.cs.odu.edu/
Search
My Web BETA
Subscriptions (New) Shortcuts Advanced Search Preferences
Search Results
Results 1 - about 1 for url:http://www.cs.odu.edu/ - 0.23 sec. (About this page)
1. Department of Computer Science
College of Sciences
Category: Virginia > Norfolk > Old Dominion University > Department of Computer Science
www.cs.odu.edu - 41k - Cached - More from this site - Save - Block
Done

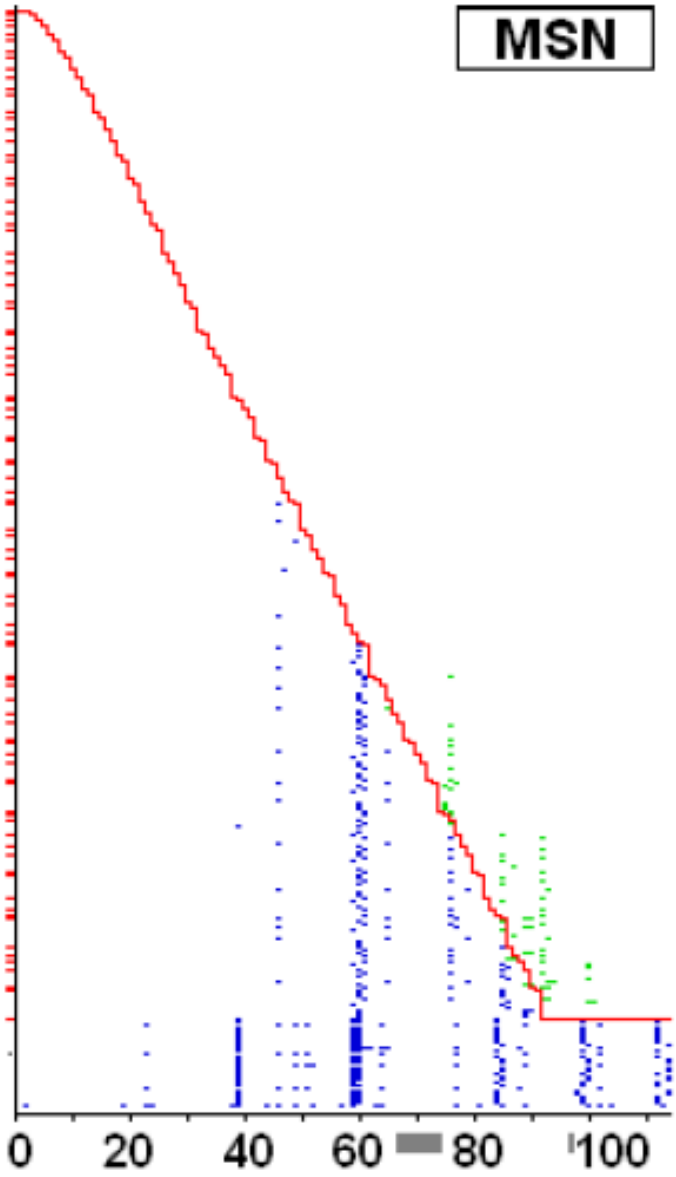
Timeline of Web Resource

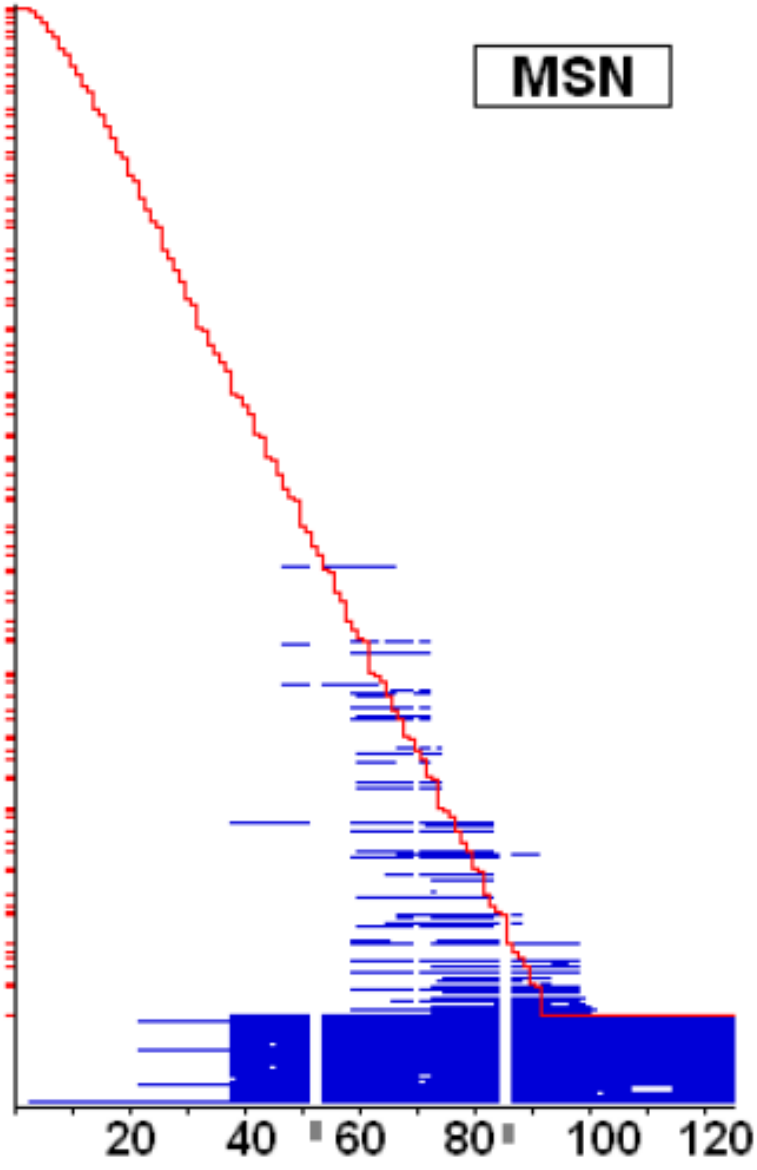



Google
200/304 resp
404 response
TTL on web server (days)
1
90
inf
0
20
40
60
80
100


Google
Cached
Indexed only
TTL on web server (days)
1
90
inf
0
20
40
60
80
100
120


Inktomi
Yahoo
0
20
40
60
80
100
120
days


MSN
0
20
40
60
80
100

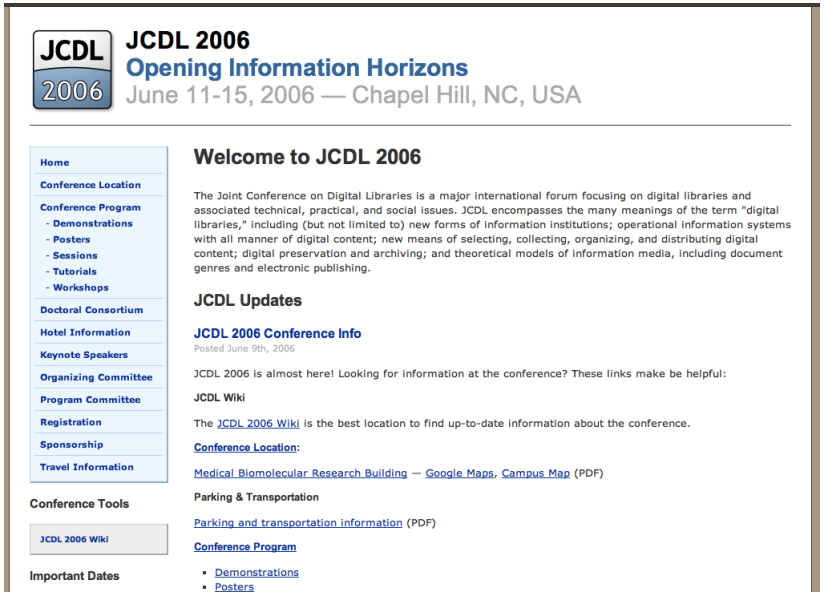

MSN
20
40
60
80
100
120

Lapsed Website

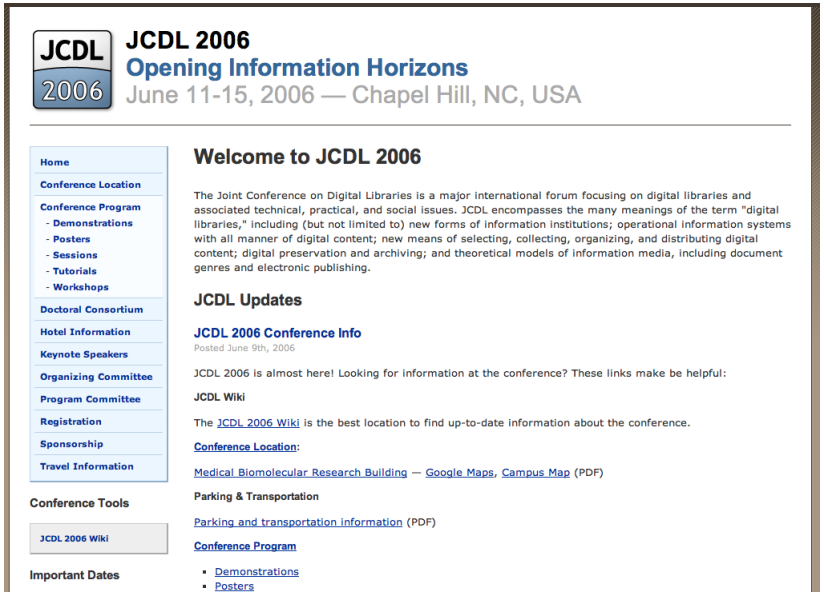
Scenario 1: Same URI, Same Content

JCDL 2006

http://www.jcdl2006.org/
July 2006



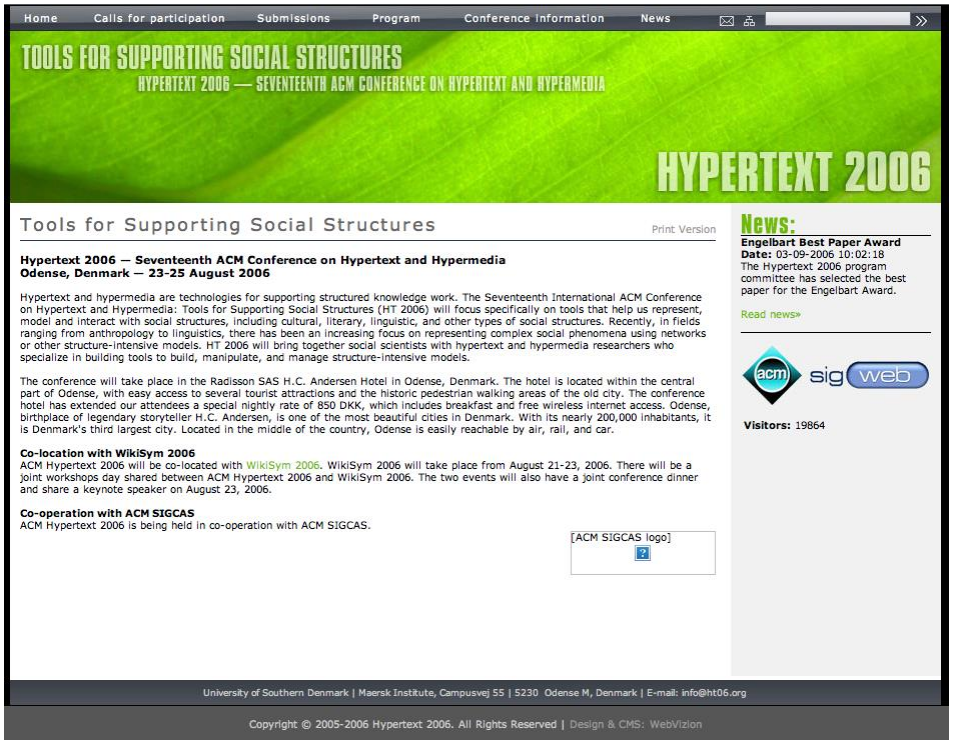
http://www.jcdl2006.org/
Today



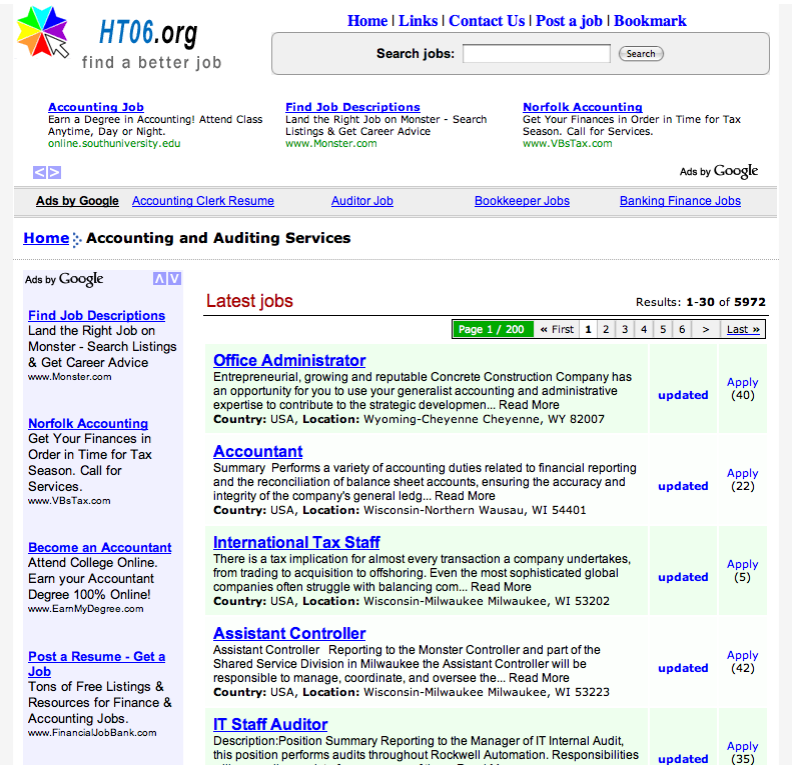
Scenario 2: Same URI, Different Content

Hypertext 2006

http://www.ht06.org/
August 2006



http://www.ht06.org/
Today

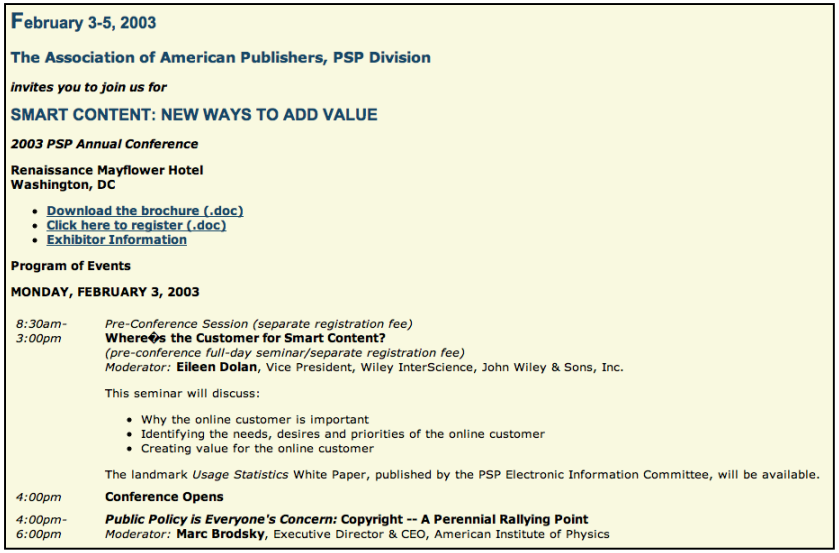


Scenario 3a: Same Content, Different URI

PSP 2003

http://www.pspcentral.org/events/annual_meeting_2003.html

August 2003



February 3-5, 2003

The Association of American Publishers, PSP Division

invites you to join us for

SMART CONTENT: NEW WAYS TO ADD VALUE

2003 PSP Annual Conference

Renaissance Mayflower Hotel
Washington, DC

- Download the brochure (.doc)
- Click here to register (.doc)
- Exhibitor Information

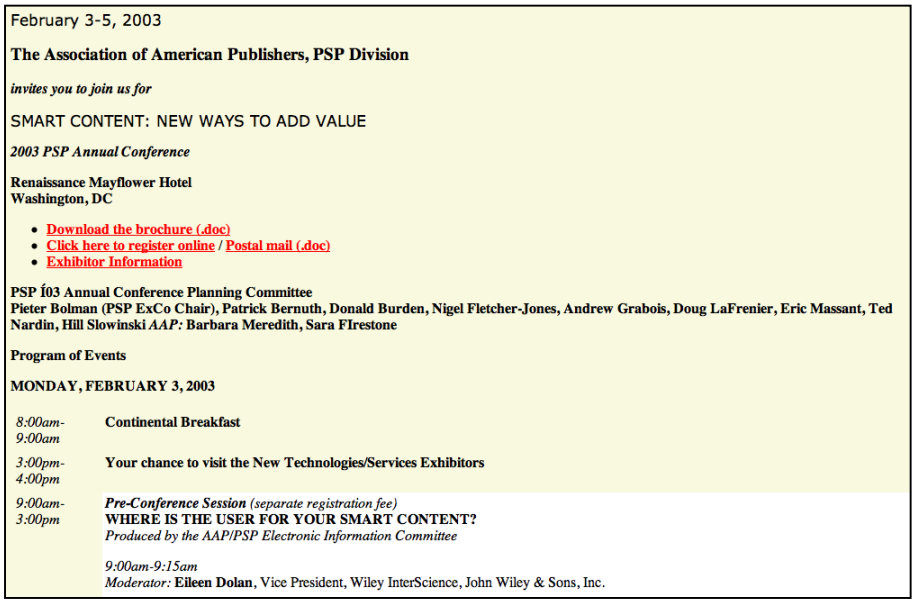
Program of Events

MONDAY, FEBRUARY 3, 2003

8:30am-3:00pm	*Pre-Conference Session (separate registration fee)* **Where's the Customer for Smart Content?** *(pre-conference full-day seminar/separate registration fee)* *Moderator:* **Eileen Dolan**, Vice President, Wiley InterScience, John Wiley & Sons, Inc. This seminar will discuss: • Why the online customer is important • Identifying the needs, desires and priorities of the online customer • Creating value for the online customer The landmark *Usage Statistics* White Paper, published by the PSP Electronic Information Committee, will be available.
4:00pm	**Conference Opens**
4:00pm-6:00pm	***Public Policy is Everyone's Concern:* Copyright -- A Perennial Rallying Point** *Moderator:* **Marc Brodsky**, Executive Director & CEO, American Institute of Physics

http://www.pspcentral.org/events/archive/annual_meeting_2003.html

Today



February 3-5, 2003

The Association of American Publishers, PSP Division

invites you to join us for

SMART CONTENT: NEW WAYS TO ADD VALUE

2003 PSP Annual Conference

Renaissance Mayflower Hotel
Washington, DC

- Download the brochure (.doc)
- Click here to register online / Postal mail (.doc)
- Exhibitor Information

PSP Í03 Annual Conference Planning Committee
Pieter Bolman (PSP ExCo Chair), Patrick Bernuth, Donald Burden, Nigel Fletcher-Jones, Andrew Grabois, Doug LaFrenier, Eric Massant, Ted Nardin, Hill Slowinski *AAP:* Barbara Meredith, Sara FIrestone

Program of Events

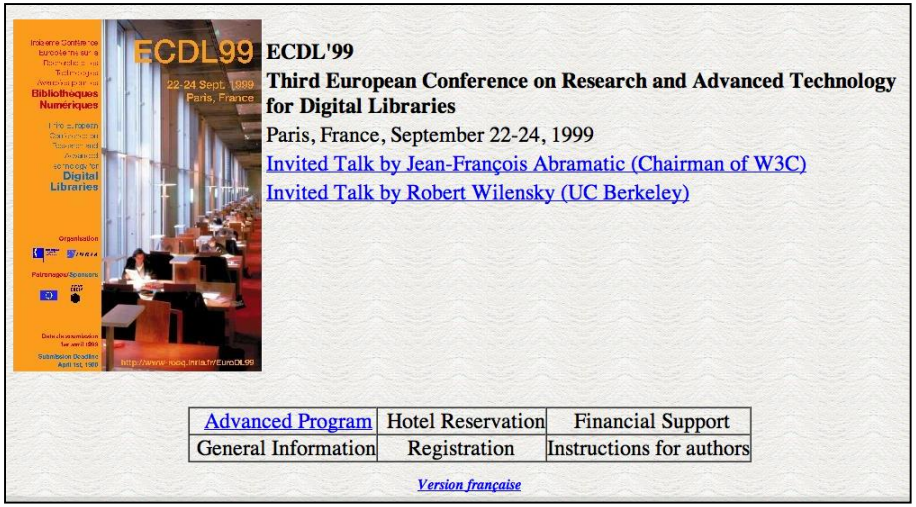
MONDAY, FEBRUARY 3, 2003

8:00am-9:00am	**Continental Breakfast**
3:00pm-4:00pm	**Your chance to visit the New Technologies/Services Exhibitors**
9:00am-3:00pm	***Pre-Conference Session*** *(separate registration fee)* **WHERE IS THE USER FOR YOUR SMART CONTENT?** *Produced by the AAP/PSP Electronic Information Committee* *9:00am-9:15am* *Moderator:* **Eileen Dolan**, Vice President, Wiley InterScience, John Wiley & Sons, Inc.

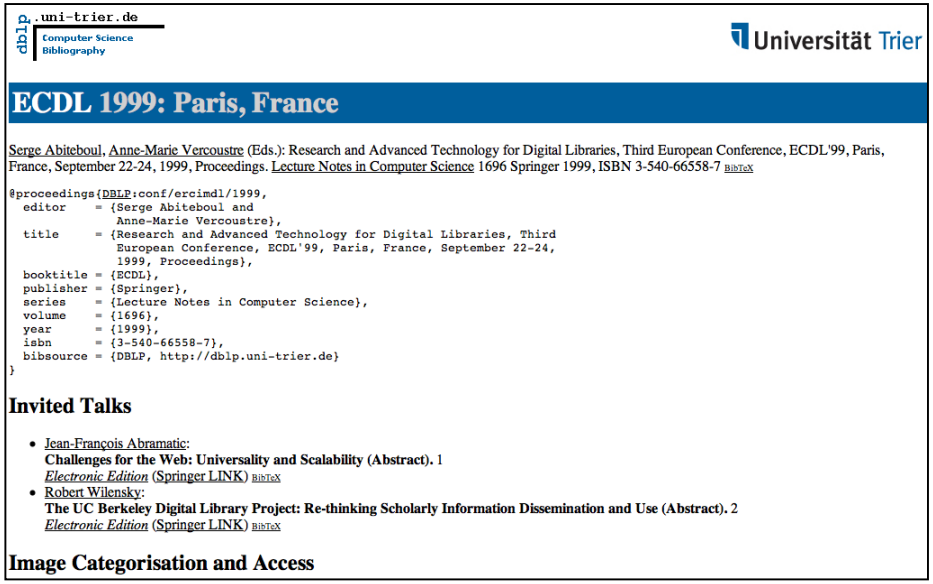
Scenario 3b: Similar Content, Different URI

ECDL 1999

http://www-rocq.inria.fr/EuroDL99/
October 1999



http://www.informatik.uni-trier.de/~ley/db/conf/ercimdl/ercimdl99.html
Today



Scenario 4: Content Not Findable At Any URI

Greynet 1999

http://www.konbib.nl/infolev/greynet/2.5.htm

1999

Today

?

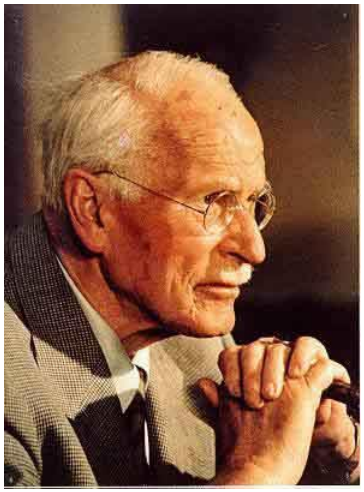
?

Miller: A lot o' people don't realize what's really going on. They view life as a bunch o' unconnected incidents 'n things. They don't realize that there's this, like, lattice o' coincidence that lays on top o' everything. Give you an example; show you what I mean: suppose you're thinkin' about a plate o' shrimp. Suddenly someone'll say, like, *plate, or shrimp, or plate o' shrimp* out of the blue, no explanation. No point in lookin' for one, either. It's all part of a cosmic unconsciousness.

Otto: You eat a lot of acid, Miller, back in the hippie days?

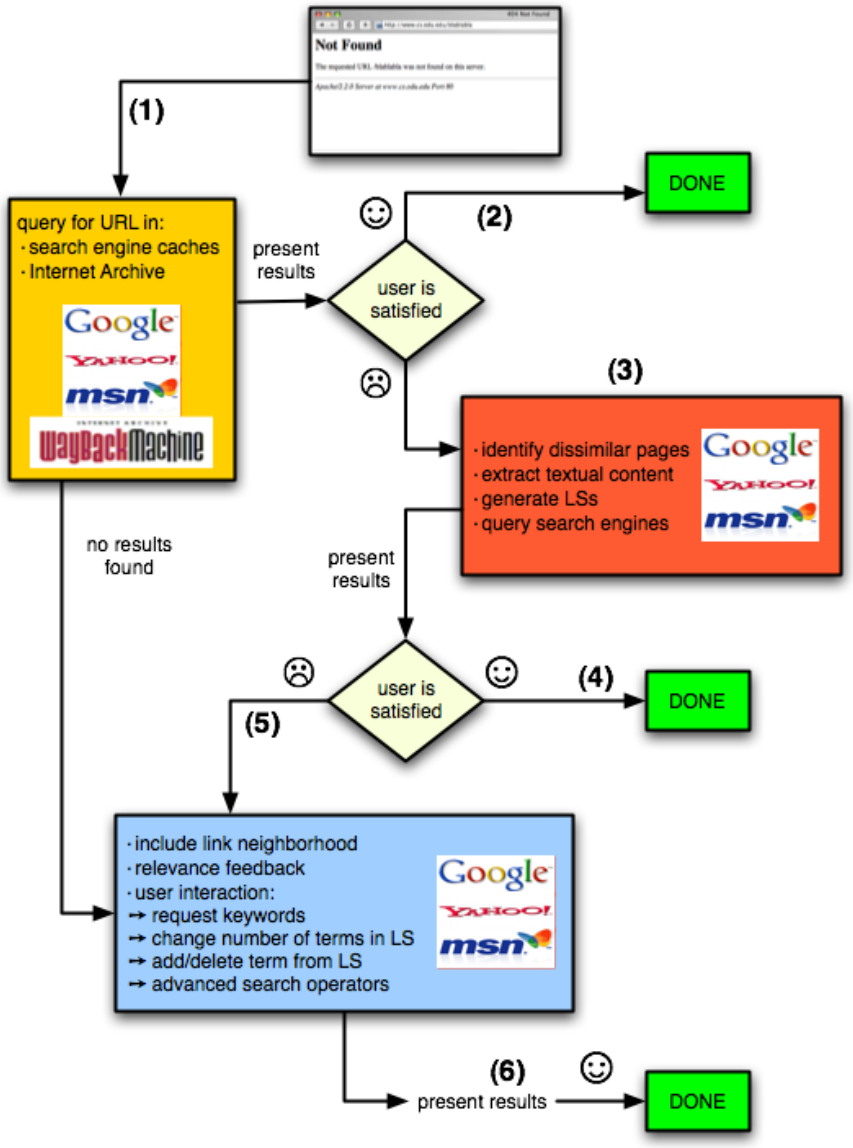
Synchronicity

- Experience of causally unrelated events occurring together in a meaningful manner
- Events reveal underlying pattern, framework bigger than any of the synchronous systems
- Carl Gustav Jung (1875-1961)
  - “meaningful coincidence”



picture from http://www.crystalinks.com/jung.html

Synchronicity Architecture



- Browser plugin catches 404 ``Page not found'' error
- Discovers copy of missing page in WI (1) and provides to user (2)
- Generates a search engine query based on what the missing page is ``about'' (3)
- Provides ``good enough'' alternative page (4)

What is a Signature?

(aka “message digest”, examples include “md5” and “sha-1”)

image from Eddie Kohler http://www.cs.ucla.edu/~kohler/

What is a Lexical Signature?

- First introduced by Phelps and Wilensky [Phelps00]
- Small set of terms capturing the "aboutness" of a document
  - Phelps and Wilensky assumed 5
- "lightweight metadata"

Resource

Removal Policies in Network Caches for World-Wide Web Documents

Stephen Williams, Marc Abrams (Computer Science, Virginia Tech, Blacksburg, VA 24061-0106)

Charles R. Standridge (Industrial Eng., FAMU-FSU College of Eng., Tallahassee, FL 32310)

Ghaleb Abdulla, Edward A. Fox (Computer Science, Virginia Tech, Blacksburg, VA 24061-0106)

{williams,abrams}@cs.vt.edu, standfeng.fsu.edu, {abdulla,fox}@cs.vt.edu
Revised and Corrected version of Original Publication

Abstract

World-Wide Web proxy servers that cache documents can potentially reduce three quantities: the number of requests that reach popular servers, the volume of network traffic resulting from document requests, and the latency that an end-user experiences in retrieving a document. This paper examines the first two using the measures of cache hit rate (or fraction of client-requested URLs returned by the proxy) and weighted cache hit rate (or fraction of client-requested bytes returned by the proxy). A client request for an uncached document may cause the removal of one or more cached documents. Variable document sizes and types allow a rich variety of policies to select a document for removal, in contrast to policies for CPU caches or demand paging, that manage homogeneous objects. We present a taxonomy of removal policies. Through trace-driven simulation, we determine the maximum possible hit rate and weighted hit rate that a cache could ever achieve, and the removal policy that maximizes hit rate and weighted hit rate. The experiments use five traces of 38 to 190 days of client URL requests. Surprisingly, the criteria used by current proxy-server removal policies (LRU and a proposal by Pitkow and Recker) are among the worst performing criteria in our simulation; instead, replacing documents based on size maximizes hit rate in each of the studied workloads.

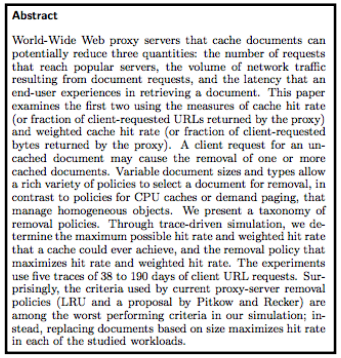
1 Introduction

It is ironic that what is arguably the most popular application of the Internet, namely the World-Wide Web (WWW or "Web"), is inherently unscalable. The protocol underlying the Web, HTTP, is designed so that every time a user accesses a large body of documents, each with a low volume of readership. However the Web today is characterized by high volume of access to popular Web pages. Thus identical copies of many documents pass through the same network links. This has several costs: network administrators are growing utilization that requires bandwidth upgrades, Web site administrators are growing server utilization that requires upgrading or replacing servers, and end users see longer latency for document requests to be satisfied.

These problems can be alleviated by widespread migration of copies of popular documents from the server to places closer to users. Migration can follow a distribution model, in which servers control where document copies are stored, or a cache model, in which copies automatically migrate in response to user requests. Distribution is popular with commercial users that want to protect material with copyrights, and will only trust certain sites to keep copies of such material. However the cache model is the most popular method to date: caching at a server (e.g., [10]), caching at a client (e.g., caches built into Web browsers), and caching in the network itself through so-called proxy servers (e.g., [11]). This paper examines the last of these.

In the following, a *document* is any item retrieved by a Universal Resource Locator (URL), such as a dynamically created page or an audio file. We call a document on the World-Wide Web *cacheable* if it does not change while in a cache. Consider a client configured to use proxy P that requests document D from Web server S. The client sends an HTTP GET request message for URL http://S/D to P. There cases are now possible: (1) P has a copy of D that is consistent with the original, (2) P has a copy that is considered inconsistent, and (3) P does not have a copy. (Various algorithms not considered here are used to estimate consistency.) In case (1), P serves the local copy; this is a hit. In (2), P sends an HTTP conditional GET message to S containing the Last-Modified time of its copy. If the original was modified after that time, S replies to P with the new version (a miss). Otherwise, P serves the local copy; this is also a hit. In (3), P requests D from S (a miss), and may or may not cache it. If D is cacheable, P may be able to satisfy future requests for D more quickly. For example, in one of the five Web traffic workload traces used in this paper, representing remote client requests for documents on a network backbone in the Computer Science Department at Virginia Tech, 88% of the bytes transferred in a 37 day measurement period were audio. This surprising result arose because a student created what has been recognized in the WWW as the authoritative Web site for a popular British recording artist. Unfortunately, every request worldwide to listen to a song on this Web site consumes

Abstract



Abstract

World-Wide Web proxy servers that cache documents can potentially reduce three quantities: the number of requests that reach popular servers, the volume of network traffic resulting from document requests, and the latency that an end-user experiences in retrieving a document. This paper examines the first two using the measures of cache hit rate (or fraction of client-requested URLs returned by the proxy) and weighted cache hit rate (or fraction of client-requested bytes returned by the proxy). A client request for an uncached document may cause the removal of one or more cached documents. Variable document sizes and types allow a rich variety of policies to select a document for removal, in contrast to policies for CPU caches or demand paging, that manage homogeneous objects. We present a taxonomy of removal policies. Through trace-driven simulation, we determine the maximum possible hit rate and weighted hit rate that a cache could ever achieve, and the removal policy that maximizes hit rate and weighted hit rate. The experiments use five traces of 38 to 190 days of client URL requests. Surprisingly, the criteria used by current proxy-server removal policies (LRU and a proposal by Pitkow and Recker) are among the worst performing criteria in our simulation; instead, replacing documents based on size maximizes hit rate in each of the studied workloads.

LS

REMOVAL
HIT
RATE
PROXY
CACHE

Query Google

"Removal Policies in Network Caches for World-Wide Web Documents"

LS as Proposed by Phelps and Wilensky

- "Robust Hyperlink Cost Five Words Each"
- Append LS to URL

  http://www.cs.berkeley.edu/~wilensky/NLP.html?lexical-signature=texttiling+wilensky+disambiguation+subtopic+iago

- **Limitations:**
  1. Applications (browsers) need to be modified to exploit LSs
  2. LSs need to be computed a priori
  3. Works well with most URLs but not with all of them

Lexical Signatures -- Examples

Rank/Results	URL	LS
1/1	**http://www.cs.berkeley.edu/~wilensky/NLP.html**	**texttiling wilensky disambiguation subtopic iago** http://www.google.com/search?q=texttiling+wilensky+disambiguation+subtopic+iago
na/10	**http://www.dli2.nsf.gov**	**nsdl multiagency imls testbeds extramural** **http://www.google.com/search?q=nsdl+multiagency+imls+testbeds+extramural**
1/221,000 (1/174,000 in 01/2008)	**http://www.loc.gov**	**library collections congress thomas american** **http://www.google.com/search?q=library+collections+congress+thomas+american**
1/51 (2/77 in 01/2008)	**http://www.jcdl2008.org**	**libraries jcdl digital conference pst** **http://www.google.com/search?q=libraries+jcdl+digital+conference+pst**

Generating LSs

- Term Frequency (TF)
  - “How often does this term occur in this document?”
- Inverse document frequency
  - “In how many documents does this term appear?”

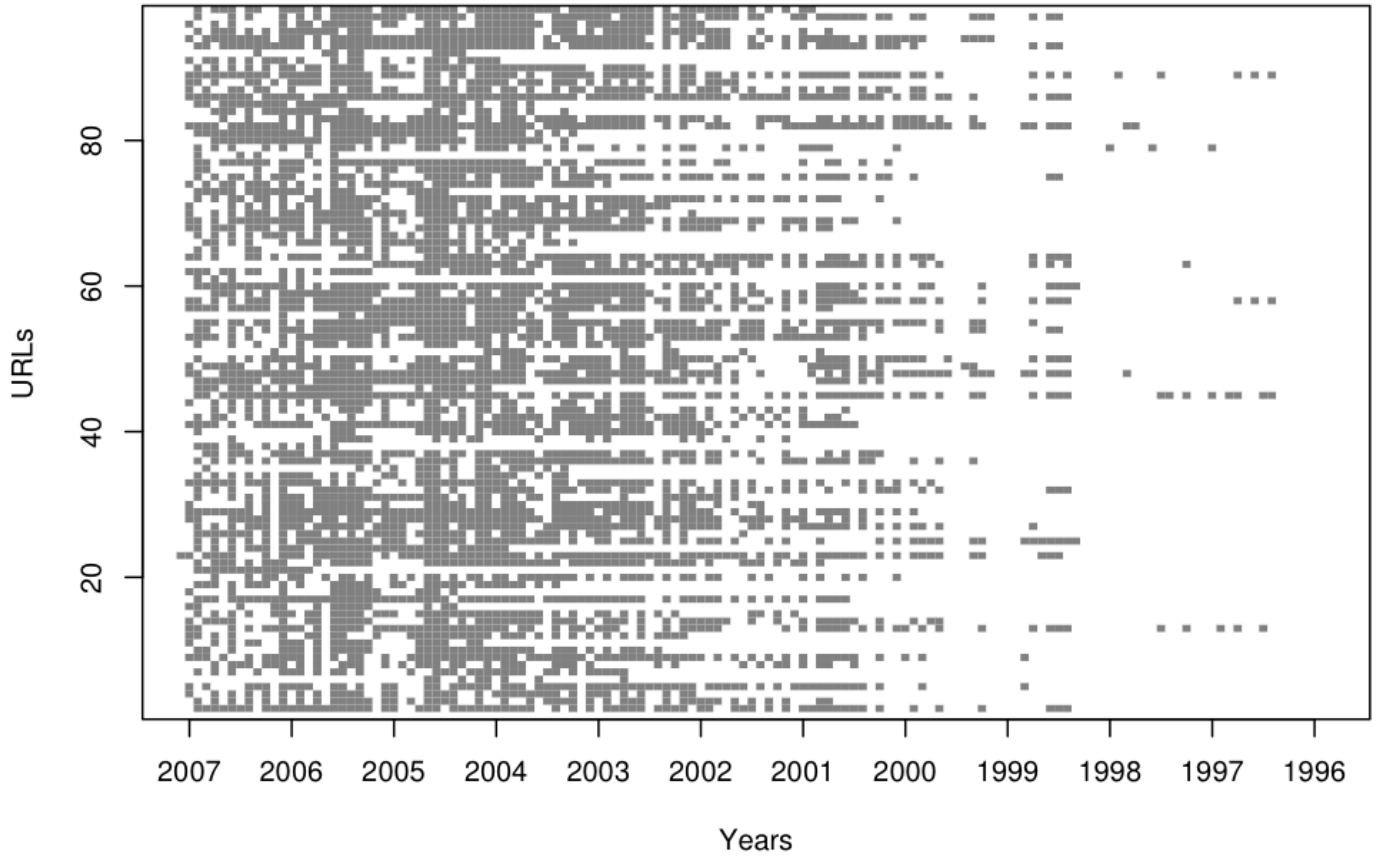
Generating LSs

- Park et al. [Park03] investigated performance of various LS generation algorithms
- Evaluated “tunability” of TF and IDF
  - Weight on TF increases recall (completeness)
  - Weight on IDF improves precision (exactness)
- Also assumed “5” to be a good number
- Compared results after 6 months, but did not do an in-depth analysis of LSs over time

Evolution Over Time

Data

- Copies of web pages from the IA (1996-2007)



300 Random URLs, winnowed to 98, 10493 observations over 12 years

Evolution Over Time -- Example

10-term LSs generated for http://www.perfect10wines.com

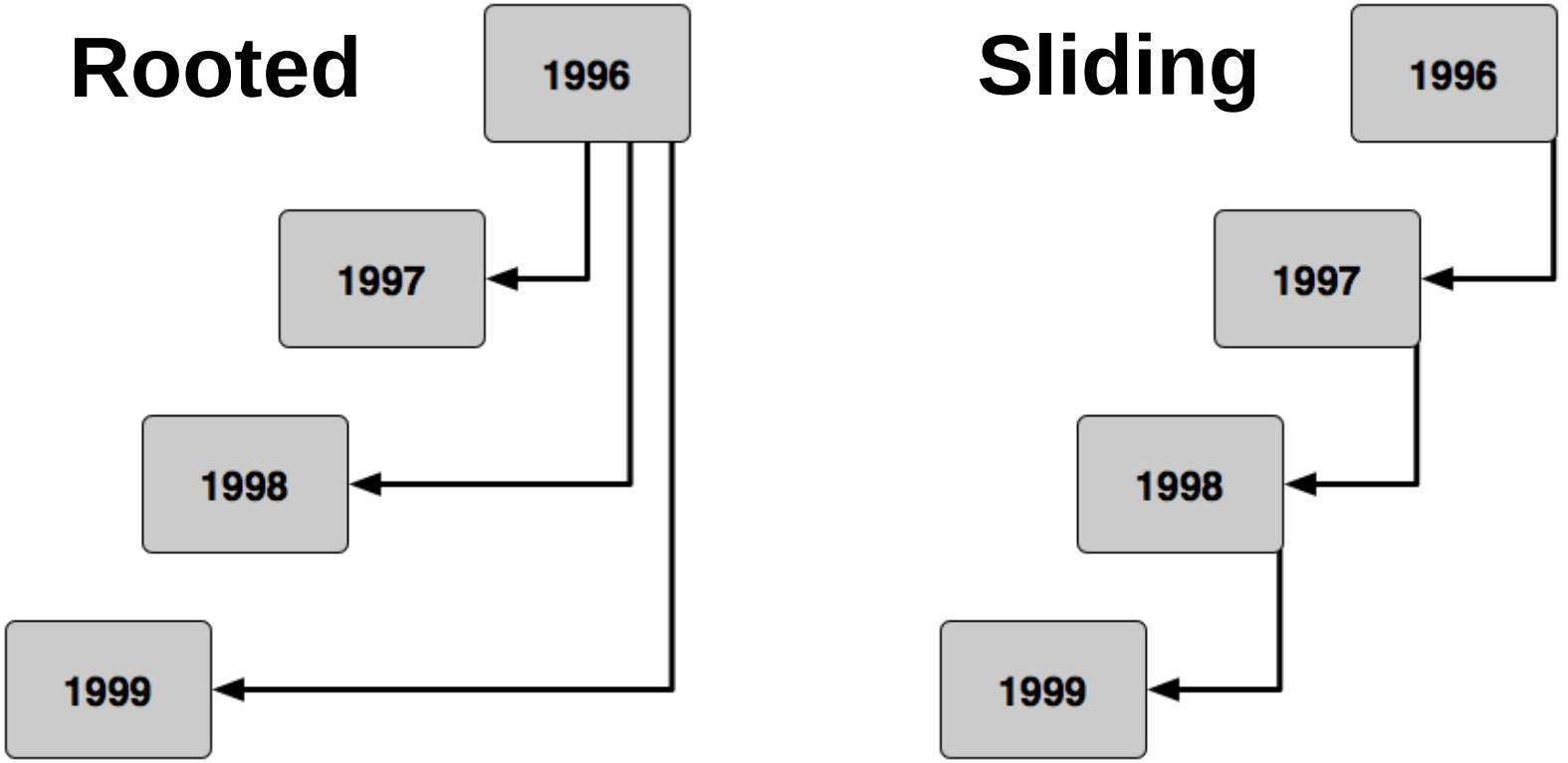
	2005		2006		2007	
	Term	**Score**	**Term**	**Score**	**Term**	**Score**
1	wines	8.56	wines	6.52	wines	5.25
2	perfect	5.00	wine	4.80	wine	4.50
3	wine	3.03	perfect	4.70	paso	4.50
4	10	2.60	10	3.45	perfect	4.10
5	monterey	2.24	paso	3.01	robles	3.75
6	chardonnay	2.24	robles	2.89	10	3.40
7	merlot	2.20	monterey	2.79	monterey	2.25
8	robles	1.99	chardonnay	2.79	cabernet	2.25
9	paso	1.99	ripe	1.86	chardonnay	2.25
10	blonde	1.38	vanilla	1.86	sauvignon	2.25

$$for\ all\ terms: \left|\cup\right| = 14\ and\ \left|\cap\right| = 8$$

Evolution Over Time

Idea

- Generate LSs from copies of URLs
- Conduct overlap analysis



Evolution Over Time

Rooted

compare to	Year of First Observation										
	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
1997	0.33										
1998	0.13	0.33									
1999	0.13	0.20	0.56								
2000	0.13	0.33	0.49	0.51							
2001	0.20	0.27	0.31	0.46	0.58						
2002	0.13	0.33	0.33	0.32	0.48	0.64					
2003	0.13	0.13	0.40	0.40	0.47	0.54	0.66				
2004	0.13	0.13	0.36	0.35	0.40	0.53	0.60	0.66			
2005	0.13	0.07	0.38	0.37	0.37	0.42	0.50	0.63	0.58		
2006	0.13	0.20	0.31	0.35	0.38	0.48	0.51	0.46	0.62	0.80	
2007	0.20	0.20	0.27	0.29	0.37	0.44	0.50	0.37	0.52	0.60	0.90

- Little overlap between the early years and more recent ones
- Highest overlap in the first 1-2 years after creation of the LSs
- Rarely peaks after that – once terms are gone they do not return

Evolution Over Time

Sliding

comparison	Year of First Observation										
	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
1996-1997	0.33										
1997-1998	0.40	0.33									
1998-1999	0.73	0.27	0.56								
1999-2000	0.53	0.40	0.49	0.51							
2000-2001	0.47	0.87	0.56	0.62	0.58						
2001-2002	0.53	0.73	0.51	0.52	0.63	0.64					
2002-2003	0.60	0.73	0.67	0.55	0.67	0.64	0.66				
2003-2004	0.93	0.80	0.76	0.69	0.80	0.83	0.73	0.66			
2004-2005	0.87	0.80	0.73	0.66	0.82	0.68	0.83	0.74	0.58		
2005-2006	0.93	0.47	0.71	0.72	0.77	0.72	0.84	0.51	0.76	0.80	
2006-2007	0.87	0.53	0.80	0.68	0.83	0.76	0.81	0.49	0.68	0.80	0.90

- Overlap increases over time
- Seem to reach steady state around 2003

Performance of LSs

Idea

• Measure performance in respect to age of LS and number of terms it contains
• Query Google search API with LSs
• Identify URL in result set:

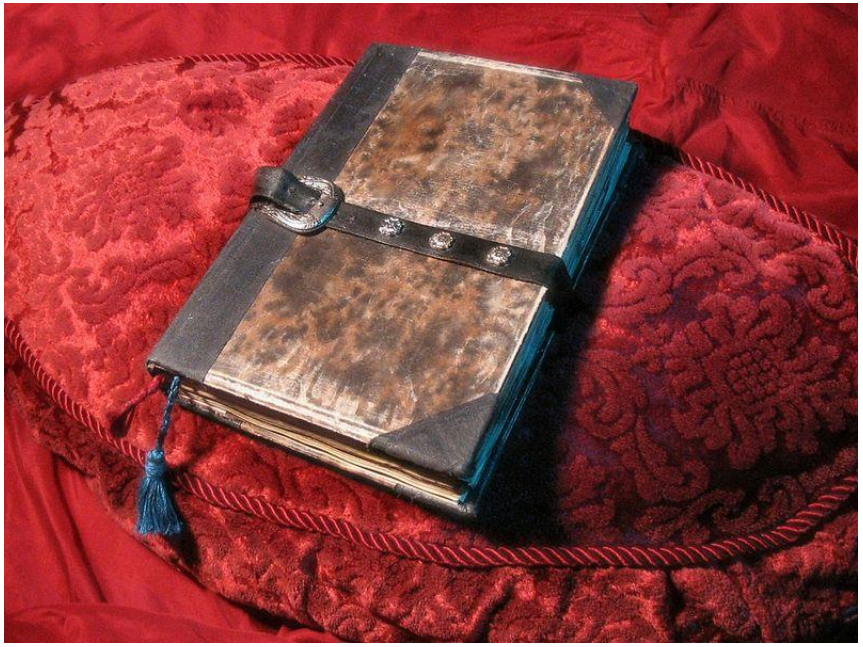
1. Top ranked
2. Ranked between 2-10
3. Ranked between 11-100
4. Ranked beyond 100 (considered undiscovered)

Performance – Number of Terms

	1	2-10	11-100	≥101	MR
2-term	24.3	14.9	13.2	47.6	53.1
3-term	40.2	15.0	15.0	29.8	36.5
4-term	43.9	15.7	11.4	29.0	33.8
5-term	47.0	19.4	3.4	30.2	32.7
6-term	51.2	11.4	3.4	34.1	36.0
7-term	54.9	9.4	1.5	34.2	35.5
8-term	49.8	7.7	2.2	40.4	41.9
9-term	47.0	6.6	0.9	45.5	46.4
10-term	46.1	4.0	0.9	49.0	49.8
15-term	39.8	0.8	0.6	58.9	59.5

- 2-, 3- and 4-term LSs perform poorly
- 5-, 6- and 7-term LSs seem best
  - Top mean rank (MR) value with 5 terms
  - Most top ranked with 7 terms
  - Binary pattern: either top 10 or undiscovered
- 8 terms and beyond do not show improvement

Conclusion & Future Work

- LSs decay over time
  - Rooted: quickly after generation
  - Sliding: seem to stabilize
- 5-, 6- and 7-term LSs seem to perform best
- 8 terms and beyond hurt performance
- Our other results have shown:
  - Google can be “screen scraped” to get the necessary information to builds LSs
  - HTML titles, when available, work slightly better than LS (titles + LSs can be combined for even better performance)
- Build the synchronicity Mozilla plug-in
- Construct a list of 404 web pages…

Necronomicon