Client-Side Preservation Techniques for ORE Aggregations

Michael L. Nelson & Sudhir Koneru

Old Dominion University, Norfolk VA

OAI-ORE Specification Roll-Out

Baltimore MD, March 3, 2008

Outline

- Background: Let the "Web Infrastructure" preserve your information
- Premise: ReMs are critical for preservation purposes
- Client-side vs. Server-side approaches to preservation
- Sketch of a possible framework for client-side preservation techniques

Web Infrastructure

Google Scholar BETA

ml nelson ja kaplan — Search

Scholar — Results 1 - 10 of about 9,270 for ml nelson ja kaplan. (0.03 seconds)

12 versions found

A comparison of queueing, cluster and distributed computing systems
JA Kaplan, ML Nelson - View as HTML - Cited by 71 - Web Search
... Joseph A. **Kaplan** (ja.kaplan@larc.nasa.gov) Michael L. **Nelson** (ml.nelson@larc.nasa.gov) NASA Langley Research Center June, 1994 Abstract ...
NASA Technical Memorandum, 1993 - techreports.larc.nasa.gov - fworoads.net - cmpharm.ucsf.edu - phi.sinica.edu.tw - all 12 versions »

[CITATION] A Comparison of Queueing
JA Kaplan, ML Nelson - Cited by 6 - Web Search
Cluster and Distributed Computing Systems, NASA Langley ..., 1994

A Comparison of Queueing, Cluster and Distributed Computing Systems (1994) (Make Corrections) (8 citations)
Joseph A. Kaplan, Michael L. Nelson

View or download:
fsu.edu/pub/drago/dep/kaplan.ps.Z
kari.re.kr/NASA/larc/94 ...tm109025.ps.Z
nasa.gov/pub/techreport.. tm109025.ps.Z
Cached: PS.gz PS PDF Image Update Help

CiteSeer Home/Search
Context Related

From: fsu.edu (more)
From: kari.re.kr/ltrs/1994cit
(Enter author homepages)

3 remote & 4 cached versions

(Enter summary) — Comment on this article

Abstract: Using workstations clusters for distributed computing has become popular with the proliferation of inexpensive, powerful workstations. Workstation clusters offer both a cost effective alternative to batch processing and an easy entry into parallel computing. However, a number of workstations on a network does not constitute a cluster. Cluster management software is necessary to harness the collective computing power. In this paper, we compare a variety of cluster management and queueing... (Update)

Internet Archive Wayback Machine

http://web.archive.org/web/*/http: — Google

INTERNET ARCHIVE
WayBackMachine

2 cached PDF versions

Enter Web Address: http:// — All — Take Me Back — Adv

Searched for http://techreports.larc.nasa.gov/ltrs/PDF/tm109025.pdf — 2 Results

* denotes when site was updated.

Search Results for Jan 01, 1996 - May 03, 2005

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	1 pages	0 pages	1 pages	0 pages	0 pages	0 pages	0 pages	0 pages

May 03, 1998 *   Sep 25, 2000 *

Home | Help

Copyright © 2001, Internet Archive | Terms of Use | Privacy Policy

---

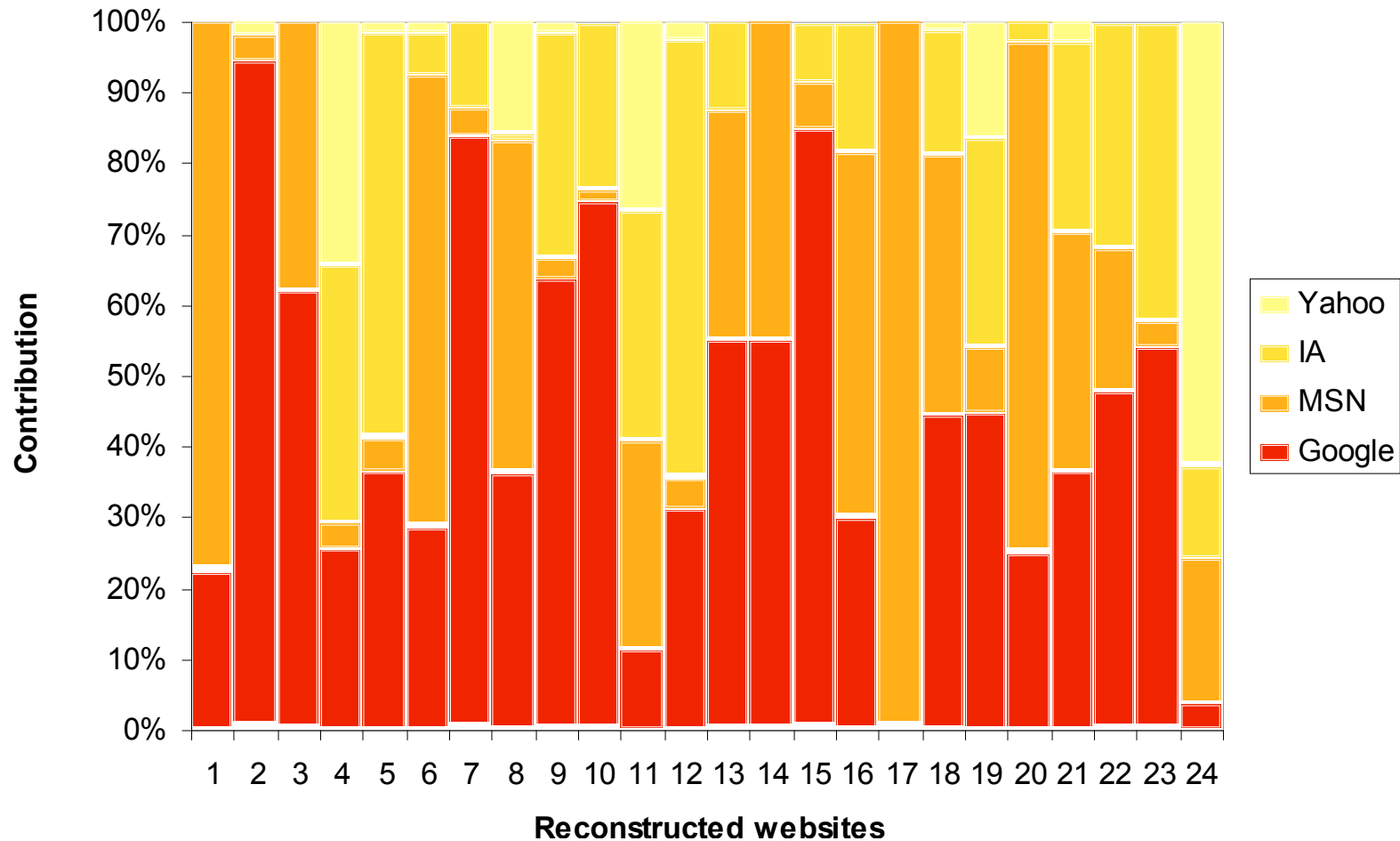Web | Images | Video | Directory | Local | News | Produ

YAHOO! SEARCH — A Comparison of Queueing, Cluster and Distributed

Shortcuts
Computing Systems -citesee

3 versions (2 nasa.gov & 1 mpg.de)

1. A comparison of queueing, cluster and distributed computing systems
A comparison of queueing, cluster and distributed computing systems A comparison of queueing, cluster and distributed computing systems Using workstation clusters for distributed computing has become popular with the proliferation of ...
ntrs.nasa.gov/archive/nasa/.../19940032425_1994032425.pdf - More from this site

2. A Comparison of Queueing, Cluster and Distributed Computing Systems, NASA TM-109025 (Revision 1), June 1994.
Joseph A. Kaplan and Micchael L. Nelson, A Comparison of Queueing, Cluster and Distributed Computing Systems , NASA TM-109025 (Revision 1), June 1994, pp. 50 ,, (95KB PS, 143KB PDF). ... of cluster management and queueing systems: COmputing in DIstributed Networked
techreports.larc.nasa.gov/ltrs/94/tm109025.refer.html - 3k - Cached - More from this site

3. A Comparison of Queueing, Cluster and Distributed Computing Systems (PDF)
... A Comparison of Queueing, Cluster and Distributed Computing Systems ... a variety of cluster management, and queueing systems: COmputing in DIstributed Networked Environments (CODINE ...
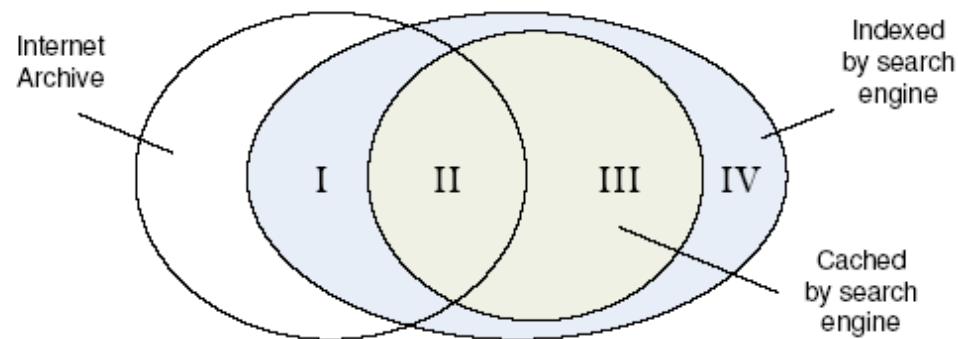www.mpip-mainz.mpg.de/theory/general/tm109025.pdf - 143k - View as html - More from this site

---

preservation = refreshing + migration

Web Repository Contributions

Contribution (y-axis, 0% to 100%)

Reconstructed websites (x-axis, 1–24)

Legend: Yahoo, IA, MSN, Google

Frank McCown, Joan A. Smith, Michael L. Nelson, Johan Bollen, Lazy Preservation: Reconstructing Websites by Crawling the Crawlers, Proceedings of WIDM 2006,pp. 67-74. http://www.cs.odu.edu/~mln/pubs/widm-2006/lazyp-widm06.pdf

Overlap with Internet Archive

	In IA		Not in IA	
SE	Cached (II)	No cache (I)	Cached (III)	No cache (IV)
Ask	9.2%	36.0%	0.3%	54.5%
Google	40.7%	3.7%	50.3%	5.3%
MSN	51.1%	1.1%	43.7%	4.1%
Yahoo	39.3%	1.8%	47.7%	11.2%

Frank McCown, Michael L. Nelson, Characterization of Search Engine Caches, Proceedings of IS&T Archiving 2007, pp. 48-52. http://arxiv.org/abs/cs.DL/0703083

Warrick -- A Service to Recover Lost Websites
warrick.cs.odu.edu

Warrick – View All Jobs

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

PROCESSING (7)

Key	Info	Node	Progress(%)	Age	Actions
fda41ae76a2fb4a8144f6b0dc0713fad [Log File]	Duncan [dmalloch1@gmail.com] http://www.mazdausa.com	c21.seven.research.odu.edu PID: 19678	Proc: 12574 | Rec: 11663 | inQ: 31138	15 days, 6 hrs, 19 min	S | T
0f1831968e4fa6a773fc26f71286d9b8 [Log File]	S Matsuura [smatsu@sa2.so-net.ne.jp] http://www.kangaeroo.net/D-maebashi.html	c23.seven.research.odu.edu PID: 26630	Proc: 11561 | Rec: 11124 | inQ: 14429	14 days, 6 hrs, 2 min	S | T
66b46e6316f9e00e73cebc6d84d9b299 [Log File]	Jamie Jones [jamie_jones@55dsl.tv] http://www.55dsl.com	c25.seven.research.odu.edu PID: 25115	Proc: 4410 | Rec: 3317 | inQ: 7980	11 days, 14 hrs, 4 min	S | T
a241b3d110ad6b7644bfb643072c74fc [Log File]	Igor Bass [webmaster@animeglobe.com] http://www.animewallpapers.com	c26.seven.research.odu.edu PID: 11985	Proc: 3649 | Rec: 3107 | inQ: 25384	8 days, 20 hrs, 4 min	S | T
e73345f7aebc3f1fea9da497cfa9a229 [Log File]	Krzysztof Lewandowski [krzysztof_lew@op.pl] http://www.forbes.com	c22.seven.research.odu.edu PID: 32259	Proc: 5186 | Rec: 4959 | inQ: 28395	4 days, 4 hrs, 5 min	S | T
968035899934549fec453bfd2b3558d1 [Log File]	Bryan [enismirdal@gmail.com] http://www.subreality.com	c24.seven.research.odu.edu PID: 15181	Proc: 5690 | Rec: 5622 | inQ: 4325	3 days, 15 hrs, 7 min	S | T
0dbaafdac3d17642d4ef585d88c7590e [Log File]	emoc [mail@emoc.org] http://www.furoncle.net	c27.seven.research.odu.edu PID: 22124	Proc: 3843 | Rec: 3683 | inQ: 57310	0 days, 12 hrs, 14 min	S | T

Top

COMPLETE (64)

Key	Info	Age	URLs Recovered	Picked-up	Actions
b39850d9e33041ddd3a90991e28c774b [Log File]	Mikio [mythalley@hotmail.com] http://www.nicovideo.jp/watch/sm1439373	22 days, 8 hrs, 32 min	22153 of 22805		C | D | S
290e384c00e494fd1dfcbc330642fe91 [Log File]	isao [isao.hatanaka@gmail.com] http://ap.e-direct.ne.jp/	30 days, 5 hrs, 46 min	10311 of 36957		C | D | S
5fbde0415b0cf7960c904b0fdbc774ca [Log File]	Pennye Willcutt [pwillcutt@gmail.com] http://willcuttguitars.com/	17 days, 7 hrs, 58 min	17748 of 29931	2008-02-18 09:55:08	C | D | S

Find:

Next   Previous   Highlight all

Done

How Much Did We Reconstruct?

"Lost" web site
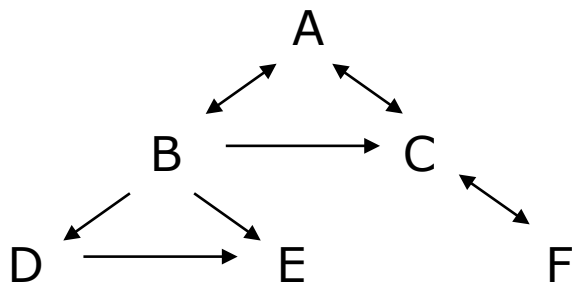
Reconstructed web site
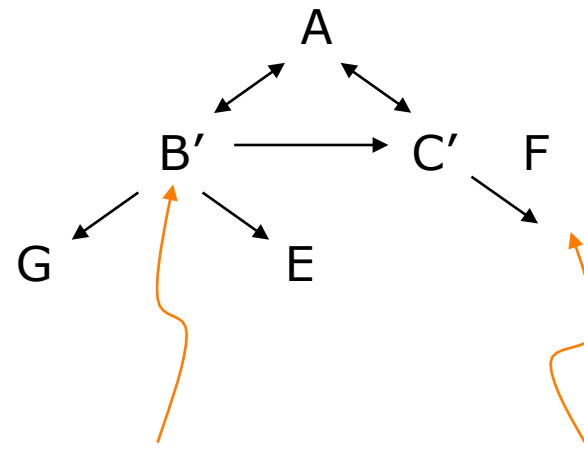
Four categories of
recovered resources:

1) Identical: A, E
2) Changed: B, C
3) Missing: D, F
4) Added: G

Missing link to D;
points to old
resource G

F can't
be found

slide from Frank McCown

Resource Maps Unambiguously Define an Aggregation

- The "manifest" nature of ReMs allow us to know "if we got it all"
  - "known knowns"
  - "known unknowns"
  - "unknown unknowns"

- Assuming the ReM is recovered, the implications for preservation are clear:
  - defines members of the aggregations
  - defines relationships between them

Server-Side Techniques

- Repository A uses ReMs for their aggregations.
- Repository B harvests ReMs to ensure total coverage of Repository A.
- Repository A can use its ReMs to validate transfer to Repository B.
- New ReMs created to reflect migration, refreshing of aggregations.

Repo A
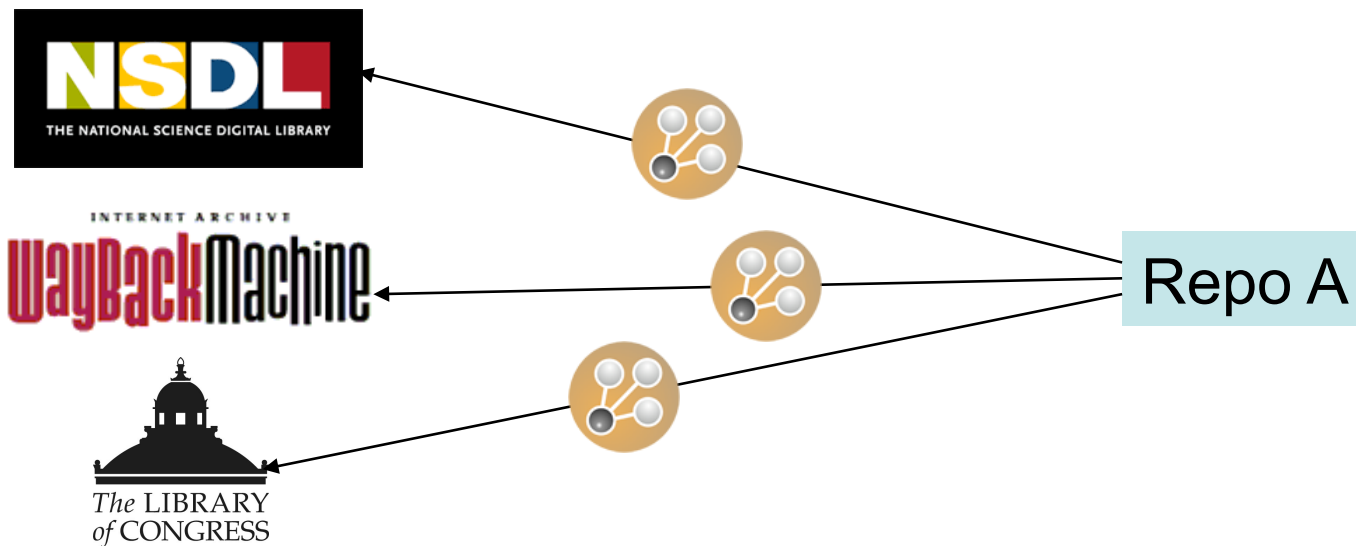
Can We Involve End-Users in the Preservation Process?

- Leverage the actions of end users?
- Make preservation more accessible?
  - light-weight and easy like Google Analytics and reCAPTCHA?

```
<html>
…
<h1>hello world</h1>
…
<script type="text/javascript">
resourcemap="http://www.foo.edu/repo/helloworld/rem.atom";
webReposToCheck="google,yahoo,internetArchive";
checkMirrors="yes";
writeBack="http://www.bar.org/wiki/"; </script>
<script type="text/javascript"
src="http://ore.cs.odu.edu/preserve.js"> </script>
…
</html>
```
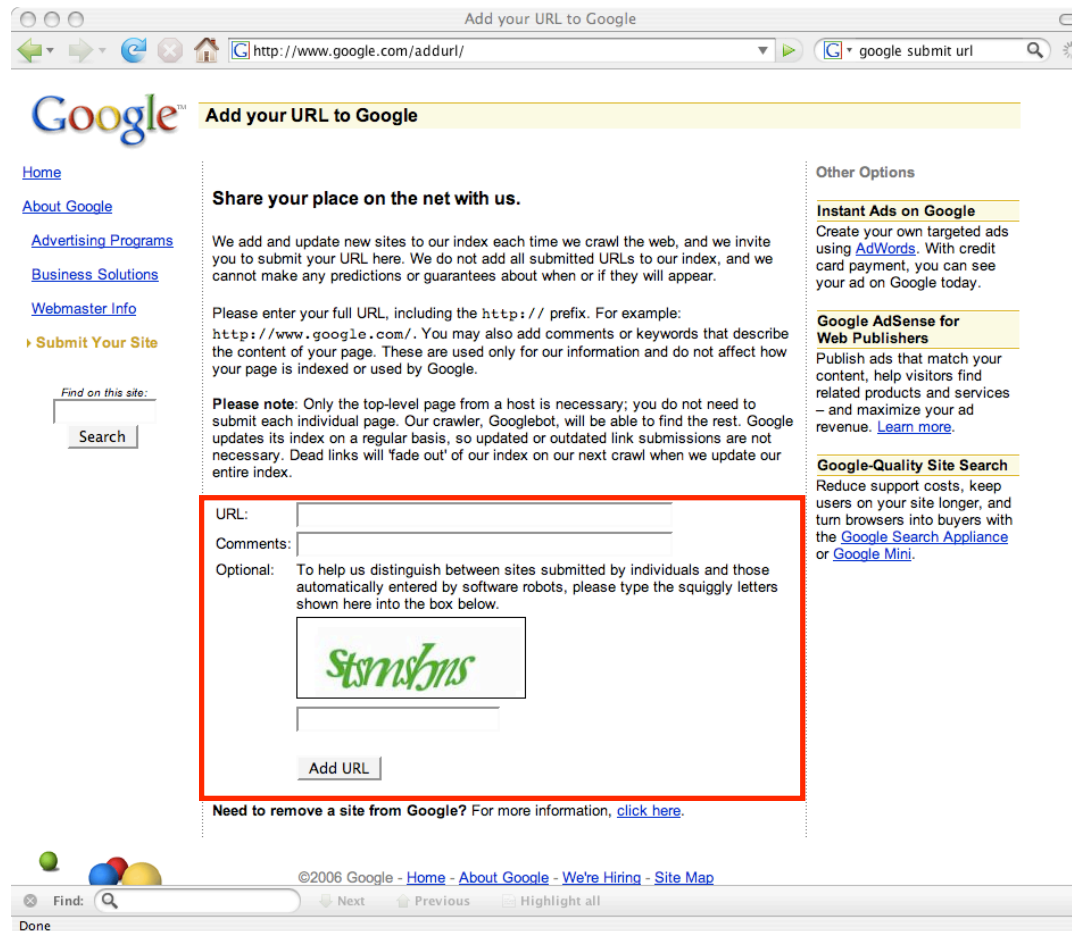
One Reason Why We Need Humans in the Loop

A Possible Scenario…


```
<script type="text/javascript">
resourcemap="http://www.foo.edu/repo/helloworld/rem.atom";
webReposToCheck="google,yahoo,internetArchive";
checkMirrors="yes";
writeBack="http://www.bar.org/wiki/"; </script>
<script type="text/javascript"
src="http://ore.cs.odu.edu/preserve.js"> </script>
```

wiki.somewhere.org

ore.cs.odu.edu

Wikis Would Make a Nice Inter-Client Message Store

Function as a publicly (computers + humans) readable revision control system for ReMs

Client-Side Techniques

- Operations on the ReM and Aggregated Resources (ARs)
  - validation, http status, ReM visualization, etc.
- Interacting with the Web Infrastructure
  - checking for ReM, ARs in Internet Archive, search engine caches, etc.
  - reconstructing aggregation for a given time interval
  - submitting ReM, ARs to WI
- Inter-client communication
  - my client updates/repairs ReM -- how to communicate that to other clients and servers?

"Help Preserve This Object"

http://www.cs.odu.edu/~skoneru/indstd/arxiv.html

Current Status

- Hierarchical view of ReM
- Finds copies of Aggregated Resources in Internet Archive, Google, Yahoo
- Next up:
  - use Simile time line software (http://simile.mit.edu/timeline/) to display ARs in time
  - chose a time interval for reconstruction
  - send edited ReMs to a wiki or public email service
  - write a program to read & vet edited ReMs from public store