

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

Research Supported by the Andrew Mellon Foundation

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



SPURL.NET



LookSmart
FURL | Your Personal Web File

Web Infrastructure



CiteSeer.IST
Scientific Literature Digital Library

Google Scholar

ml nelson ja kaplan

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 - wormads.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](#)

From: [fsu.edu](#) (more)
 From: [kari.re.kr/ltrs/1994cjt](#)
 (Enter author homepages)

(Enter summary) 3 remote & 4 cached versions

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

Enter Web Address: <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					

2 cached PDF versions

Home | Help

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

YAHOO! SEARCH

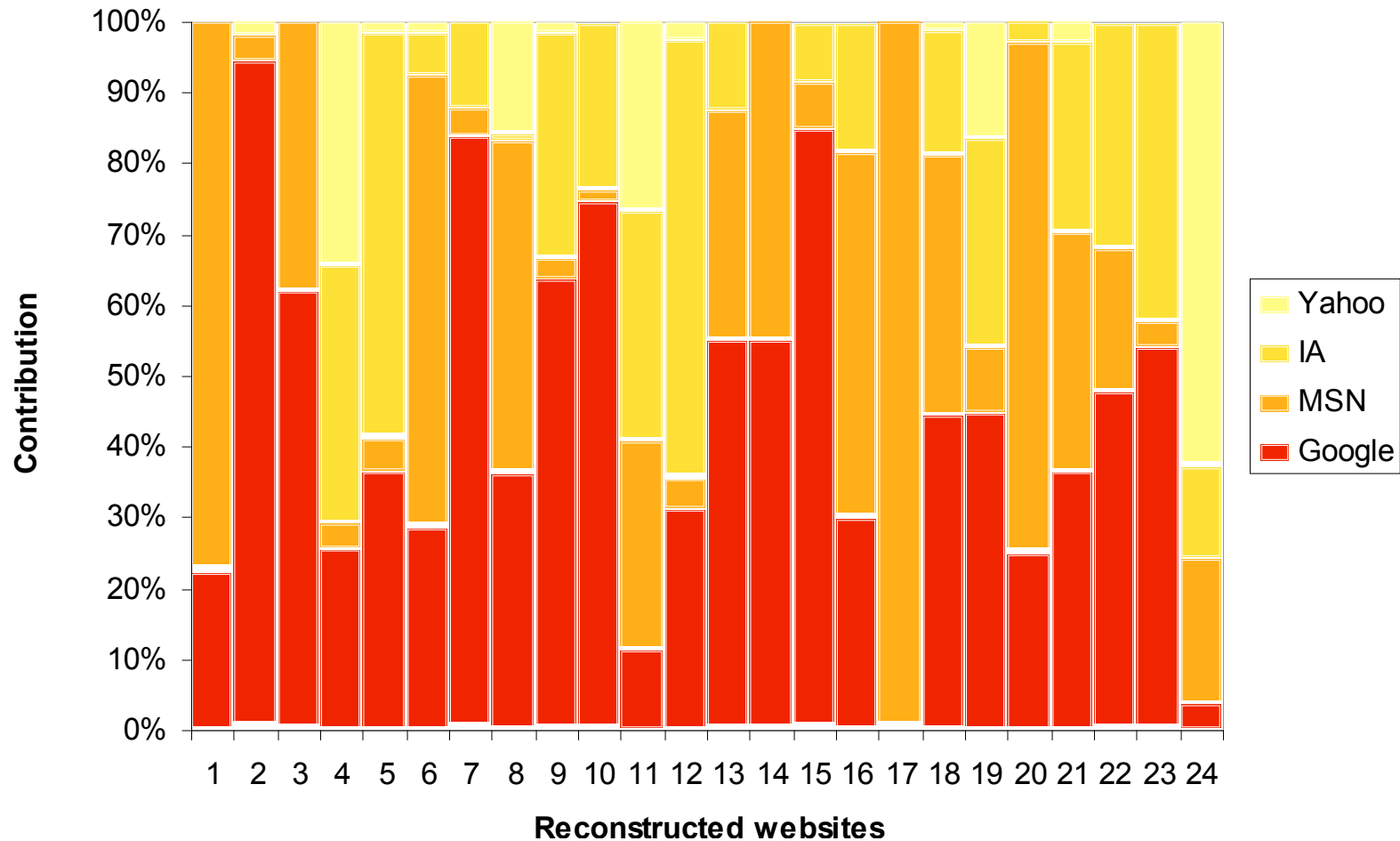
A Comparison of Queueing, Cluster and Distributed Computing Systems -citeseeer

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

- [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
- [A Comparison of Queueing, Cluster and Distributed Computing Systems, NASA TM-109025 \(Revision 1\), June 1994](#)
 Joseph A. Kaplan and Michael 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: CComputing in Distributed Networked
[techreports.larc.nasa.gov/ltrs/94/tm109025.refer.html](#) - 3k - Cached - More from this site
- [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: CComputing in Distributed Networked Environments (CODINE ...
[www.mpip-mainz.mpg.de/theory/general/tm109025.pdf](#) - 143k - View as html - More from this site

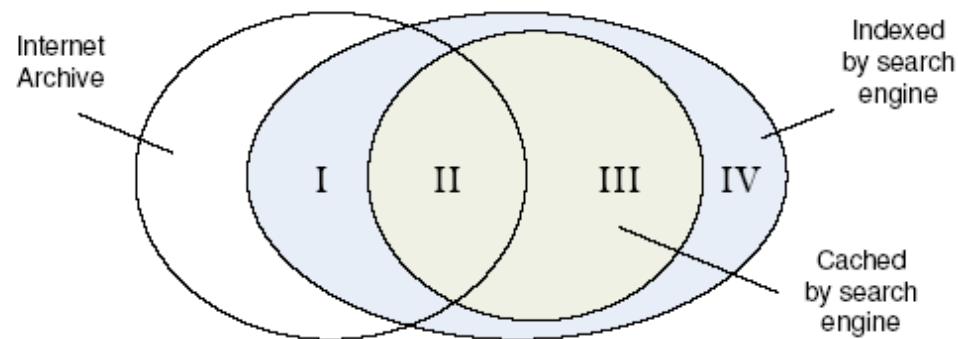
preservation = refreshing + migration

Web Repository Contributions



Overlap with Internet Archive

SE	In IA		Not in IA	
	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%



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

Google

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 I I
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 I I
66b46e6316f9e00e73cebc5d84d9b299 [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 I I
a241b3d110ad6b7644fb643072c74fc [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 I I
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 I I
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 I I
0dbaafdac3d17642d4ef585d88c7590e [Log File]	emoc [mail@emoc.org] http://www.furuncle.net	c27.seven.research.odu.edu PID: 22124	Proc: 3843 Rec: 3683 inQ: 57310 	0 days, 12 hrs, 14 min	S I I

[Top](#)

COMPLETE (64)

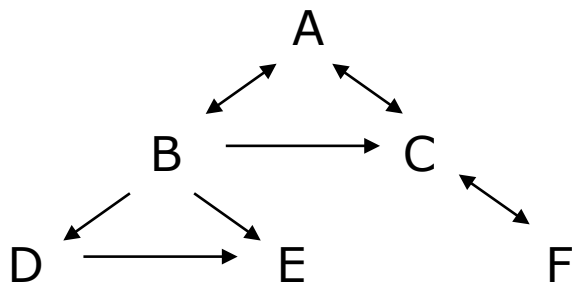
Key	Info	Age	URLs Recovered	Picked-up	Actions
b39850d9e33041ddd3a90991e28c774b [Log File]	Mikio [mythalley@hotmail.com] http://www.nicovideo.jp/watch/sm1439373	22 days, 8 hrs, 32 min	22153 of 22805		C D I S
290e384c00e494fd1dfcb330642fe91 [Log File]	isao [isao.hatanaka@gmail.com] http://ap.e-direct.ne.jp/	30 days, 5 hrs, 46 min	10311 of 36957		C D I 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 I S

Find: [Next](#) [Previous](#) [Highlight all](#)

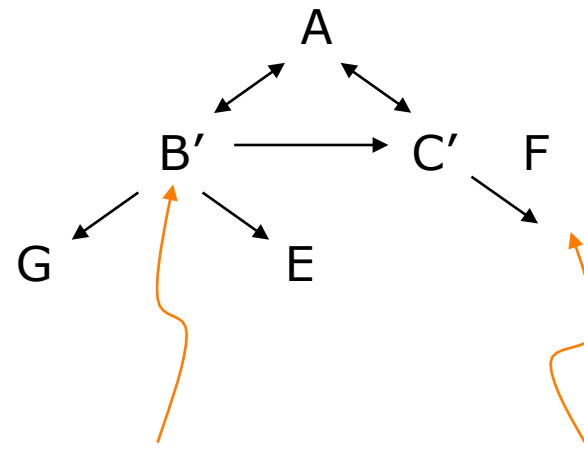
Done

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

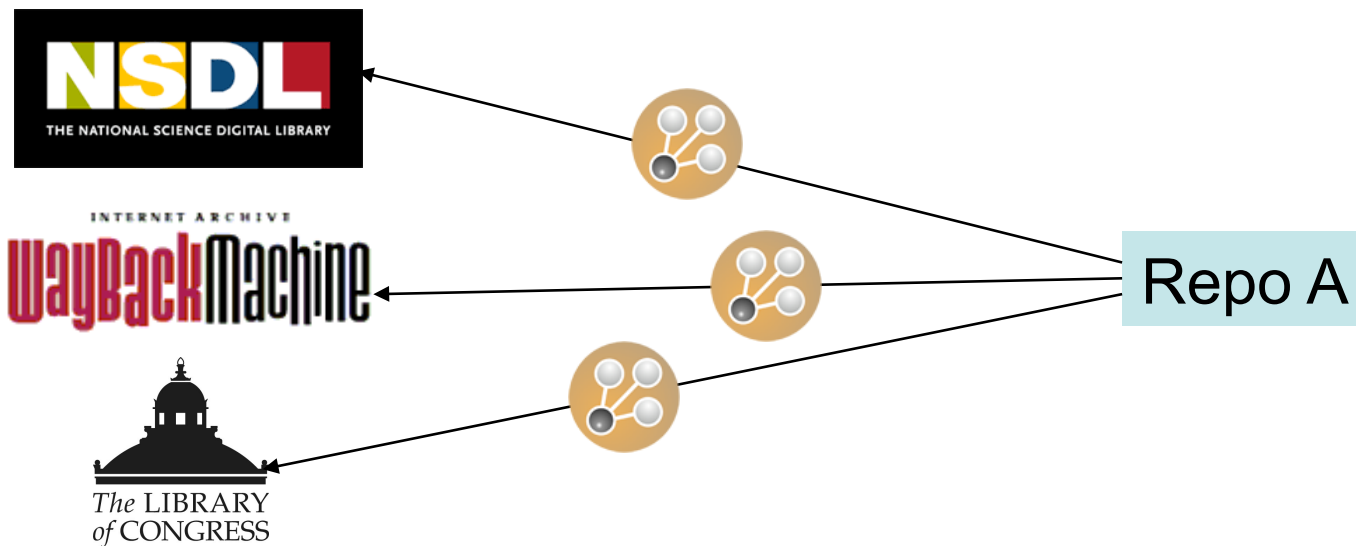
Resource Maps Unambiguously Define an Aggregation

- 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.



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>
```

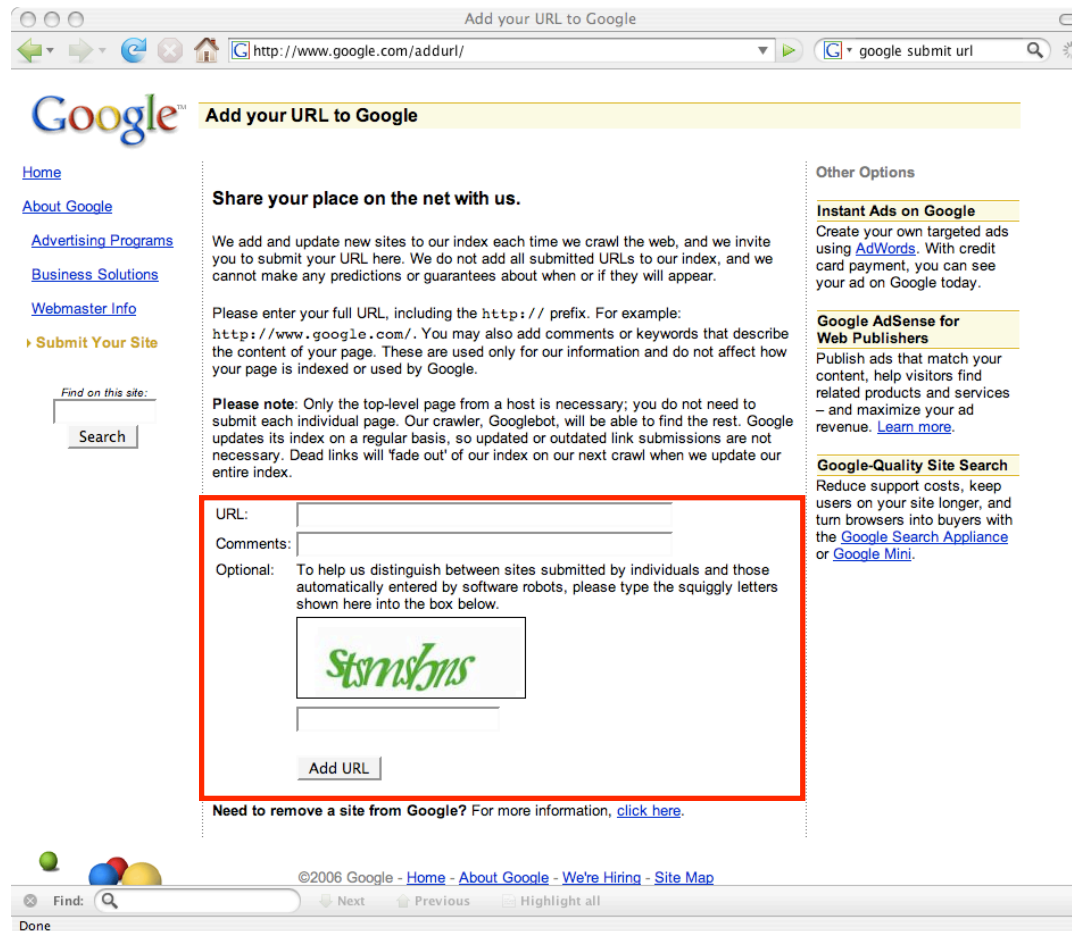
```
<script type="text/javascript"
```

```
  src="http://ore.cs.odu.edu/preserve.js"> </script>
```

```
...
```

```
</html>
```

One Reason Why We Need Humans in the Loop



The screenshot shows a web browser window titled "Add your URL to Google" with the address bar containing "http://www.google.com/addurl/". The page content includes the Google logo, navigation links (Home, About Google, Advertising Programs, Business Solutions, Webmaster Info, Submit Your Site), and a search box. The main heading is "Add your URL to Google" with the sub-heading "Share your place on the net with us." Below this, there is explanatory text and a "Please note" section. A red rectangular box highlights the form fields: "URL:", "Comments:", "Optional:" (with a text box containing "Stinsons"), and an "Add URL" button. To the right, there are sections for "Other Options" including "Instant Ads on Google", "Google AdSense for Web Publishers", and "Google-Quality Site Search". The footer contains copyright information and navigation links.

Add your URL to Google

[Home](#)
[About Google](#)
[Advertising Programs](#)
[Business Solutions](#)
[Webmaster Info](#)
[Submit Your Site](#)

Find on this site:

Share your place on the net with us.

We add and update new sites to our index each time we crawl the web, and we invite you to submit your URL here. We do not add all submitted URLs to our index, and we cannot make any predictions or guarantees about when or if they will appear.

Please enter your full URL, including the `http://` prefix. For example: `http://www.google.com/`. You may also add comments or keywords that describe the content of your page. These are used only for our information and do not affect how your page is indexed or used by Google.

Please note: Only the top-level page from a host is necessary; you do not need to submit each individual page. Our crawler, Googlebot, will be able to find the rest. Google updates its index on a regular basis, so updated or outdated link submissions are not necessary. Dead links will 'fade out' of our index on our next crawl when we update our entire index.

URL:

Comments:

Optional: To help us distinguish between sites submitted by individuals and those automatically entered by software robots, please type the squiggly letters shown here into the box below.

Other Options

Instant Ads on Google
Create your own targeted ads using [AdWords](#). With credit card payment, you can see your ad on Google today.

Google AdSense for Web Publishers
Publish ads that match your content, help visitors find related products and services – and maximize your ad revenue. [Learn more.](#)

Google-Quality Site Search
Reduce support costs, keep users on your site longer, and turn browsers into buyers with the [Google Search Appliance](#) or [Google Mini](#).

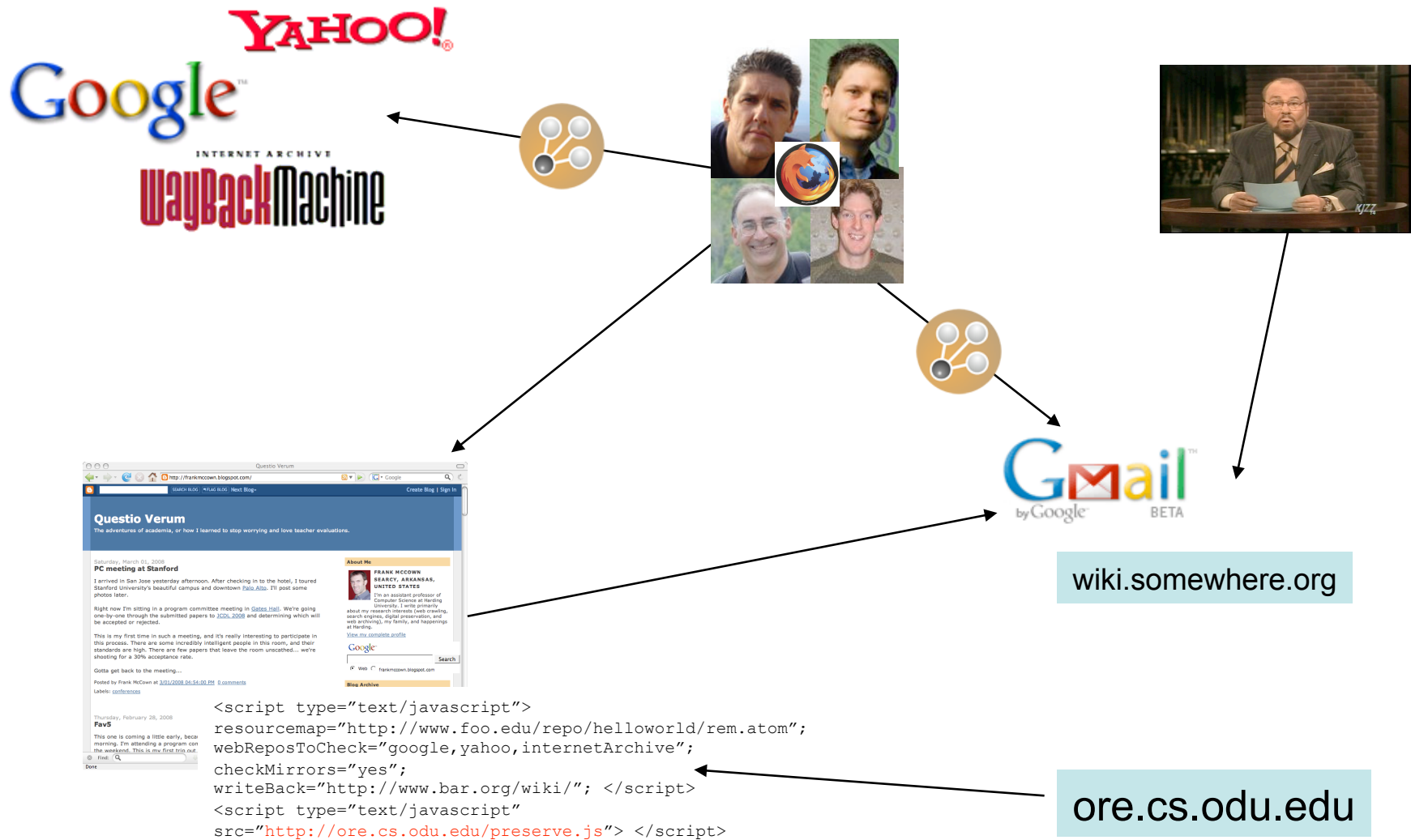
Need to remove a site from Google? For more information, [click here](#).

©2006 Google - [Home](#) - [About Google](#) - [We're Hiring](#) - [Site Map](#)

Find:

Done

A Possible Scenario...



Wikis Would Make a Nice Inter-Client Message Store

The screenshot shows a browser window displaying the revision history of the 'Open Archives Initiative' article on Wikipedia. The page title is 'Revision history of Open Archives Initiative - Wikipedia, the free encyclopedia'. The main content area lists a series of revisions, each with a date, time, and user name. The most recent revision is from 03:18 on 5 December 2007 by user 'Bdewh'. The list includes various minor edits and section edits, with some entries marked as 'removed' or 'added'. The left sidebar contains navigation links such as 'Main page', 'Contents', and 'Featured content', as well as a search box and a toolbox.

Revision history of Open Archives Initiative - Wikipedia, the free encyclopedia
can help wikipedia change the world: [Donate now!](#)

From the fundraising blog - **Preservation of Knowledge, Decades From Now**
"Helping to support one of the greatest uses of modern technology." - Anon.

Revision history of Open Archives Initiative

From Wikipedia, the free encyclopedia
[View logs for this page](#)

(Latest | **Earliest**) View (previous 50) (next 50) (20 | 50 | 100 | 250 | 500)
For any version listed below, click on its date to view it.
For more help, see [Help:Page history](#) and [Help:Edit summary](#).
(cur) = difference from current version, (last) = difference from preceding version,
m = minor edit, → = section edit, ← = automatic edit summary

[Compare selected versions](#)

- (cur) (last) [03:18, 5 December 2007 Bdewh \(Talk | contribs\)](#) (14,007 bytes) *(Picture Australia has changed its branding and the two words are no longer concatenated)* (undo)
- (cur) (last) [19:29, 30 May 2007 Ashenfelder \(Talk | contribs\)](#) **m** (14,001 bytes) (*→See also*) (undo)
- (cur) (last) [19:29, 30 May 2007 Ashenfelder \(Talk | contribs\)](#) **m** (14,001 bytes) (*→See also*) (undo)
- (cur) (last) [00:15, 28 April 2007 Petermr \(Talk | contribs\)](#) **m** (13,877 bytes) *(removed I)* (undo)
- (cur) (last) [00:14, 28 April 2007 Petermr \(Talk | contribs\)](#) (13,878 bytes) *(added section on OAI-ORE (might be worth seperate page later?))* (undo)
- (cur) (last) [20:02, 13 February 2007 Rocinante9x \(Talk | contribs\)](#) (undo)
- (cur) (last) [02:45, 29 January 2007 68.201.244.25 \(Talk\)](#) (*→See also*) (undo)
- (cur) (last) [08:18, 1 December 2006 164.67.233.84 \(Talk\)](#) (undo)
- (cur) (last) [15:44, 29 November 2006 83.103.62.218 \(Talk\)](#) (undo)
- (cur) (last) [17:14, 17 November 2006 217.162.95.218 \(Talk\)](#) (undo)
- (cur) (last) [14:56, 20 August 2006 12.149.241.66 \(Talk\)](#) (*→References*) (undo)
- (cur) (last) [20:55, 19 August 2006 89.150.64.144 \(Talk\)](#) (*→See also*) (undo)
- (cur) (last) [02:22, 15 July 2006 Ghewgill \(Talk | contribs\)](#) **m** *(it's -> its)* (undo)
- (cur) (last) [21:07, 26 April 2006 132.230.108.236 \(Talk\)](#) *(OAIster)* (undo)

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”

The screenshot shows the arXiv.org interface for the article 'Parametrization of K-essence and Its Kinetic Term' (astro-ph/0601007). The page includes a search bar, a navigation menu, and a main content area with a red border. The main content area contains the title, authors (Hui Li, Zong-Kuan Guo, Yuan-Zhong Zhang), submission information, abstract, and submission history. A red box highlights the 'Link back to' section, which includes a Resource Map icon and a text input field containing the URL 'http://arxiv.org/rem/astro-ph/060100C'. Below the input field is a small note: '(What's a Resource Map? Help Preserve This Object.)'.

[astro-ph/0601007] Parametrization of K-essence and Its Kinetic Term

arXiv.org > astro-ph > arXiv:astro-ph/0601007

Search or Article-id (Help | Advanced search)

All papers Go!

Astrophysics

Parametrization of K-essence and Its Kinetic Term

Hui Li, Zong-Kuan Guo, Yuan-Zhong Zhang

(Submitted on 31 Dec 2005 (v1), last revised 18 Jan 2006 (this version, v2))

We construct the non-canonical kinetic term of a k-essence field directly from the effective equation of state function $w_k(z)$, which describes the properties of the dark energy. Adopting the usual parametrizations of equation of state we numerically reproduce the shape of the non-canonical kinetic term and discuss some features of the constructed form of k-essence.

Comments: 8 pages, 1 figure; accepted by Mod. Phys. Lett. A; minor changes to references
Subjects: **Astrophysics (astro-ph)**
Journal reference: Mod.Phys.Lett. A21 (2006) 1683-1690
DOI: [10.1142/S0217732306019475](https://doi.org/10.1142/S0217732306019475)
Cite as: [arXiv:astro-ph/0601007v2](https://arxiv.org/abs/astro-ph/0601007v2)

Submission history

From: Hui Li [[view email](#)]
[v1] Sat, 31 Dec 2005 04:01:23 GMT (20kb)
[v2] Wed, 18 Jan 2006 06:16:15 GMT (20kb)

Link back to: [arXiv](#), [form interface](#), [contact](#).

Resource Map for arXiv:astro-ph/0601007
(What's a Resource Map? Help Preserve This Object.)

[previous](#) | [next](#)

<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