Everyone is a Curator: Human-Assisted Preservation for ORE Aggregations

Frank McCown
Harding University
Computer Science Dept.
Searcy, Arkansas, USA

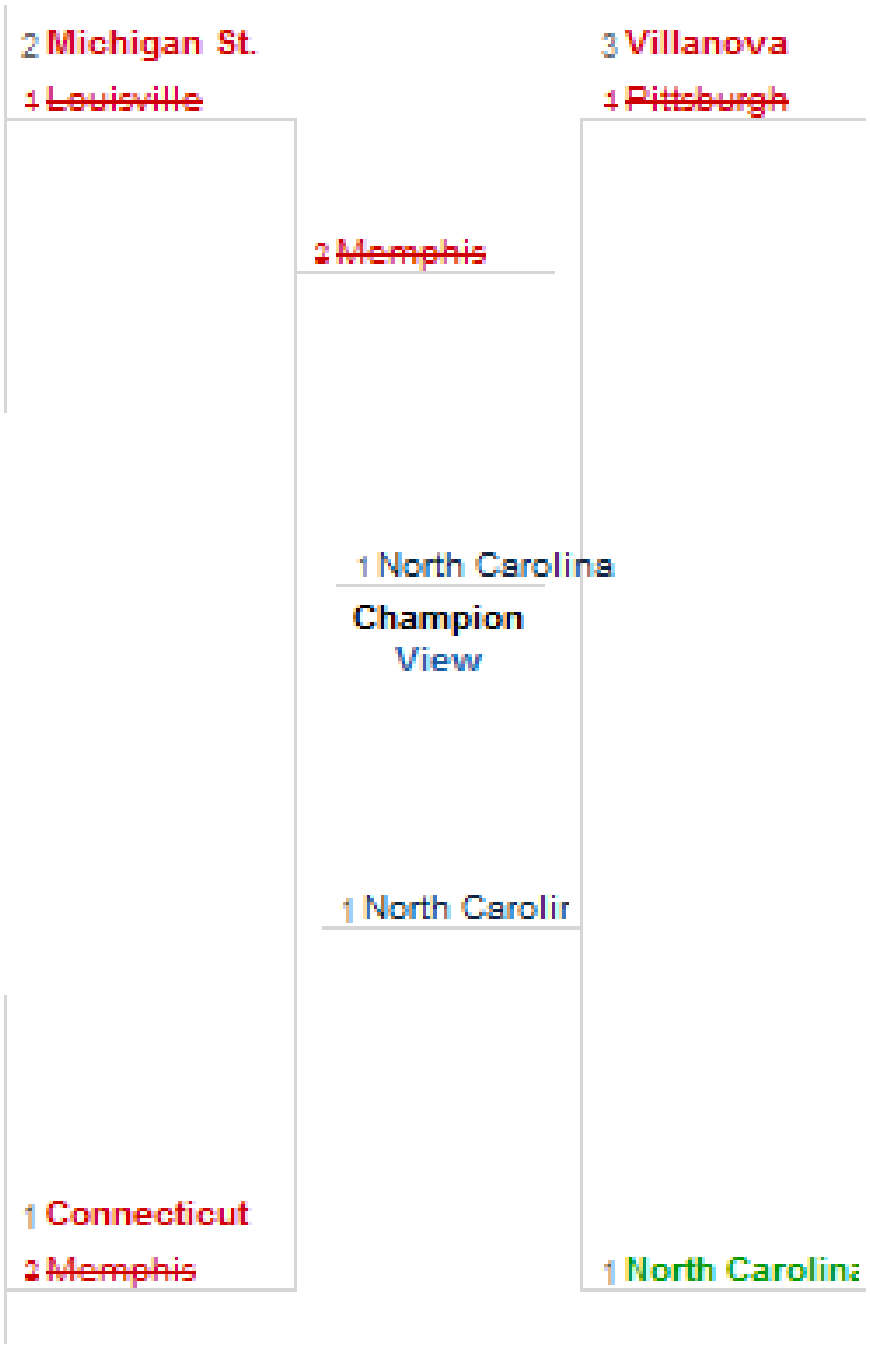
Michael L. Nelson
Old Dominion University
Computer Science Dept.
Norfolk, Virginia, USA

Herbert Van de Sompel
Digital Library Research &
Prototyping Team
Los Alamos National Lab
Los Alamos, NM, USA

DigCCurr 2009

Chapel Hill, NC
April 3, 2009

OAI-ORE work is supported by the Andrew W. Mellon Foundation, Microsoft Corporation, and the National Science Foundation (IIS-0430906).


2 Michigan St.
1 Louisville
3 Villanova
1 Pittsburgh
2 Memphis
1 North Carolina
Champion
View
1 North Carolir
1 Connecticut
2 Memphis
1 North Carolina

Outline

- Aggregations on the Web
- Object Reuse and Exchange (ORE)
- Web Infrastructure (WI)
- Prototype system: ReMember
- Future work

Aggregations

flickr

Signed in as fmccown (1 new) Help Sign Out

Home You Organize Contacts Groups Explore Search your photostream Search

fmccown · Sets

Slideshow Share This

Arrange Edit Order prints Delete this set

Stanford University

Thumbnails Detail Comments

click here to add a description

32 photos | 3 views | Add a comment?

items are from between 29 Feb 2008 & 01 Mar 2008.

D-Lib® Magazine

Search
Go
Search Guidelines

• Continuing publication of *D-Lib Magazine* is made possible by the D-Lib Alliance •

About D-Lib Magazine
About the D-Lib Alliance
Current Issue
Subscriptions
Author Guidelines
Contact D-Lib
•••
RSS D-Lib via RSS
•••

Ads by Google

Academic software
Digital library for research papers Free Software Download
www.Mendeley.com

Digitization Solutions
Customized for libraries, education, research and archives
content-conversion.com

Digital Archive Software

March/April 2009

Vol. 15 No. 3/4
doi:10.1045/dlib.magazine
ISSN: 1082-9873

•••

Table of Contents

•••

EDITORIAL
Some Results from the 2008 D-Lib Magazine Survey
by Bonita Wilson, *CNRI*

•••

LETTERS
To the Editor

•••

OPINION
What's Wrong with Citation Counts?
José H. Canós Cerdá and Manuel Llavador Campos, *Technical University of Valencia, Spain*; and Eduardo Mena Nieto, *University of Zaragoza, Spain*
doi:10.1045/march2009-canos

•••

ARTICLES
Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images
Tracy Powell and Gordon Paynter, *National Library of New Zealand*
doi:10.1045/march2009-powell

How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs
by Rose Holley, *National Library of Australia*
doi:10.1045/march2009-holley

Also This Month
Digital Collections

FEATURED COLLECTION

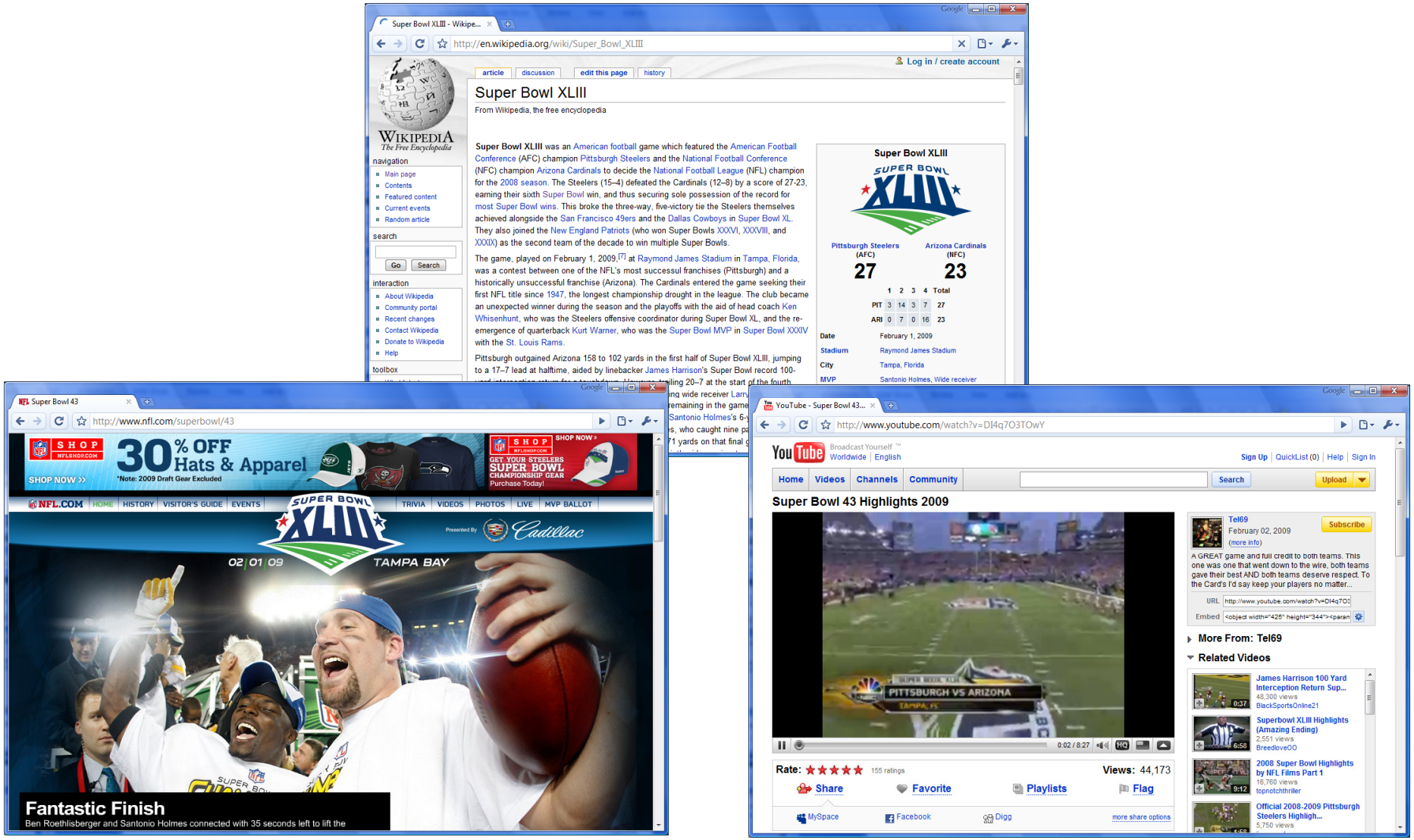
The Washington College of Law Historical Collection

The Washington College of Law Historical Collection from the Pence Law Library contains class schedules, correspondence, newspapers, programs, scrapbooks, and yearbooks documenting its history

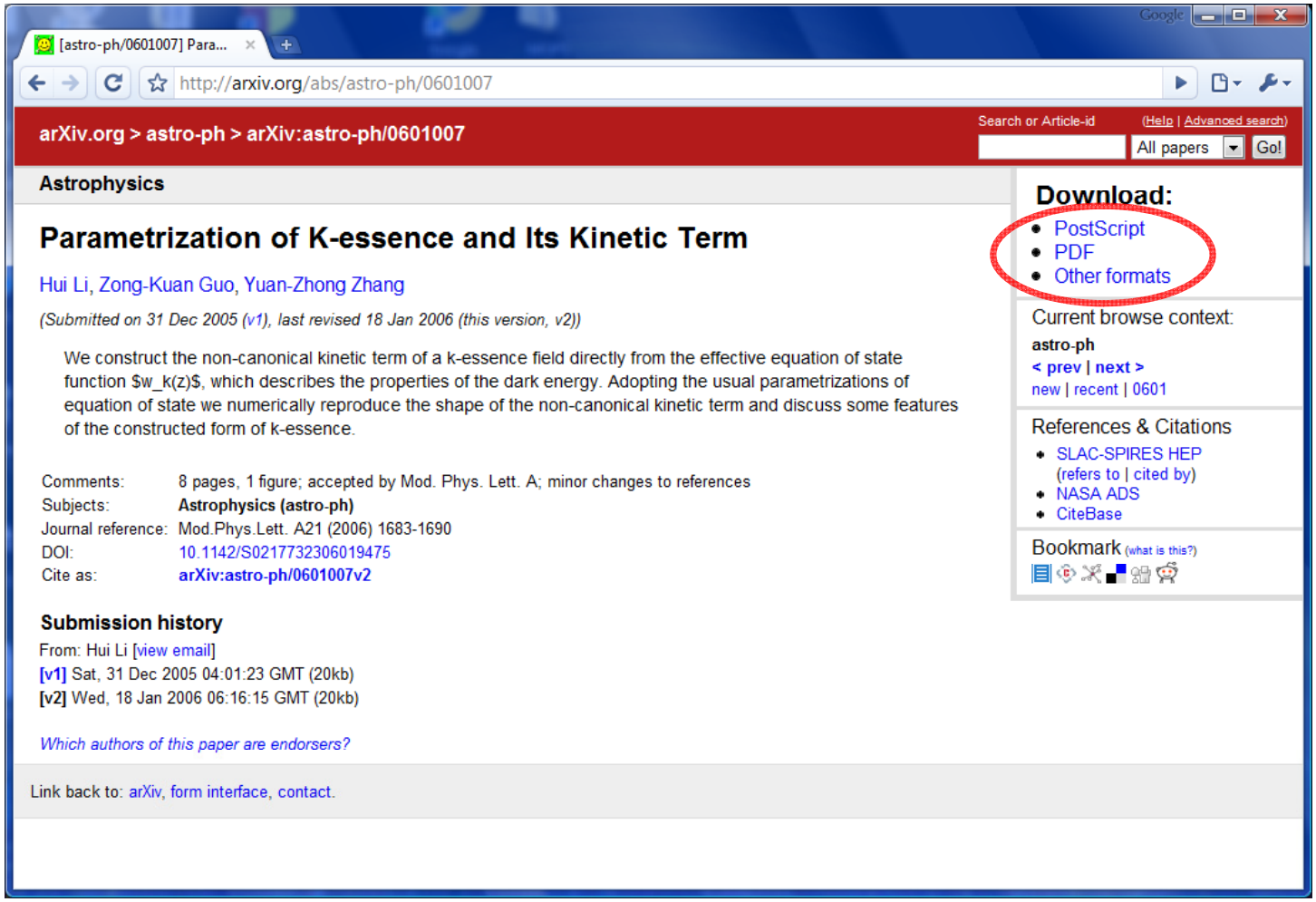
Aggregation spans multiple pages

http://www.sciam.com/article.cfm?id=preventing-next-pandemic

Aggregation spans multiple web sites

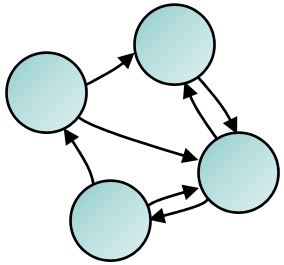


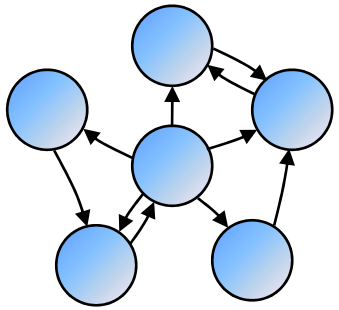
Aggregation composed of multiple representations

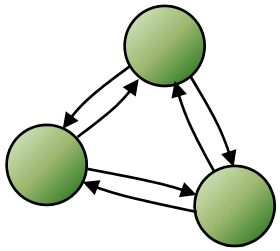


Robot's view of web aggregations

Web sites

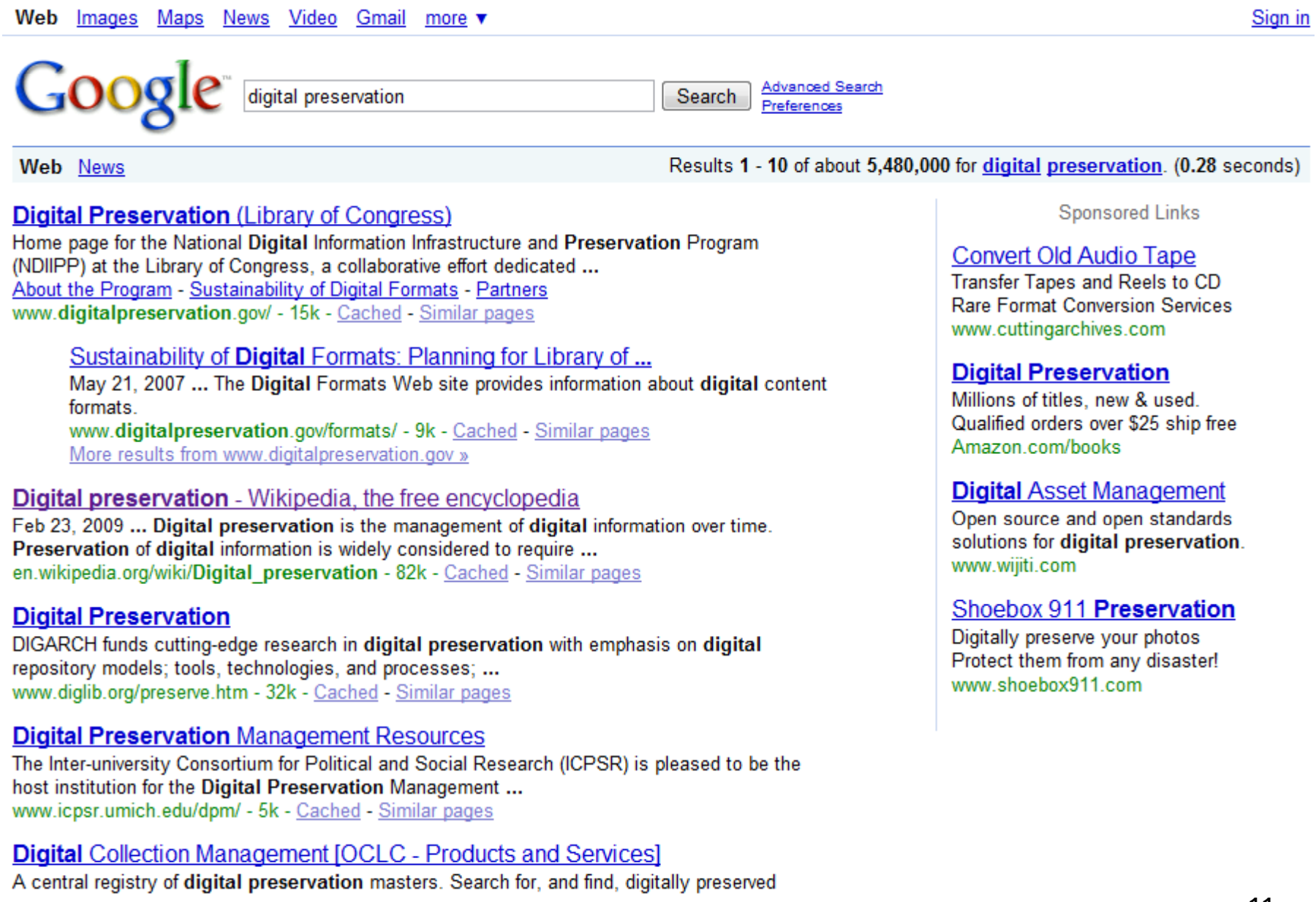






Web pages

Web resources


Web Images Maps News Video Gmail more ▼
Sign in
Google
digital preservation
Search
Advanced Search
Preferences
Web News
Results 1 - 10 of about 5,480,000 for digital preservation. (0.28 seconds)
Digital Preservation (Library of Congress)
Home page for the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress, a collaborative effort dedicated ...
About the Program - Sustainability of Digital Formats - Partners
www.digitalpreservation.gov/ - 15k - Cached - Similar pages
Sustainability of Digital Formats: Planning for Library of ...
May 21, 2007 ... The Digital Formats Web site provides information about digital content formats.
www.digitalpreservation.gov/formats/ - 9k - Cached - Similar pages
More results from www.digitalpreservation.gov »
Digital preservation - Wikipedia, the free encyclopedia
Feb 23, 2009 ... Digital preservation is the management of digital information over time. Preservation of digital information is widely considered to require ...
en.wikipedia.org/wiki/Digital_preservation - 82k - Cached - Similar pages
Digital Preservation
DIGARCH funds cutting-edge research in digital preservation with emphasis on digital repository models; tools, technologies, and processes; ...
www.diglib.org/preserve.htm - 32k - Cached - Similar pages
Digital Preservation Management Resources
The Inter-university Consortium for Political and Social Research (ICPSR) is pleased to be the host institution for the Digital Preservation Management ...
www.icpsr.umich.edu/dpm/ - 5k - Cached - Similar pages
Digital Collection Management [OCLC - Products and Services]
A central registry of digital preservation masters. Search for, and find, digitally preserved
Sponsored Links
Convert Old Audio Tape
Transfer Tapes and Reels to CD
Rare Format Conversion Services
www.cuttingarchives.com
Digital Preservation
Millions of titles, new & used.
Qualified orders over $25 ship free
Amazon.com/books
Digital Asset Management
Open source and open standards solutions for digital preservation.
www.wijiti.com
Shoebox 911 Preservation
Digitally preserve your photos
Protect them from any disaster!
www.shoebox911.com

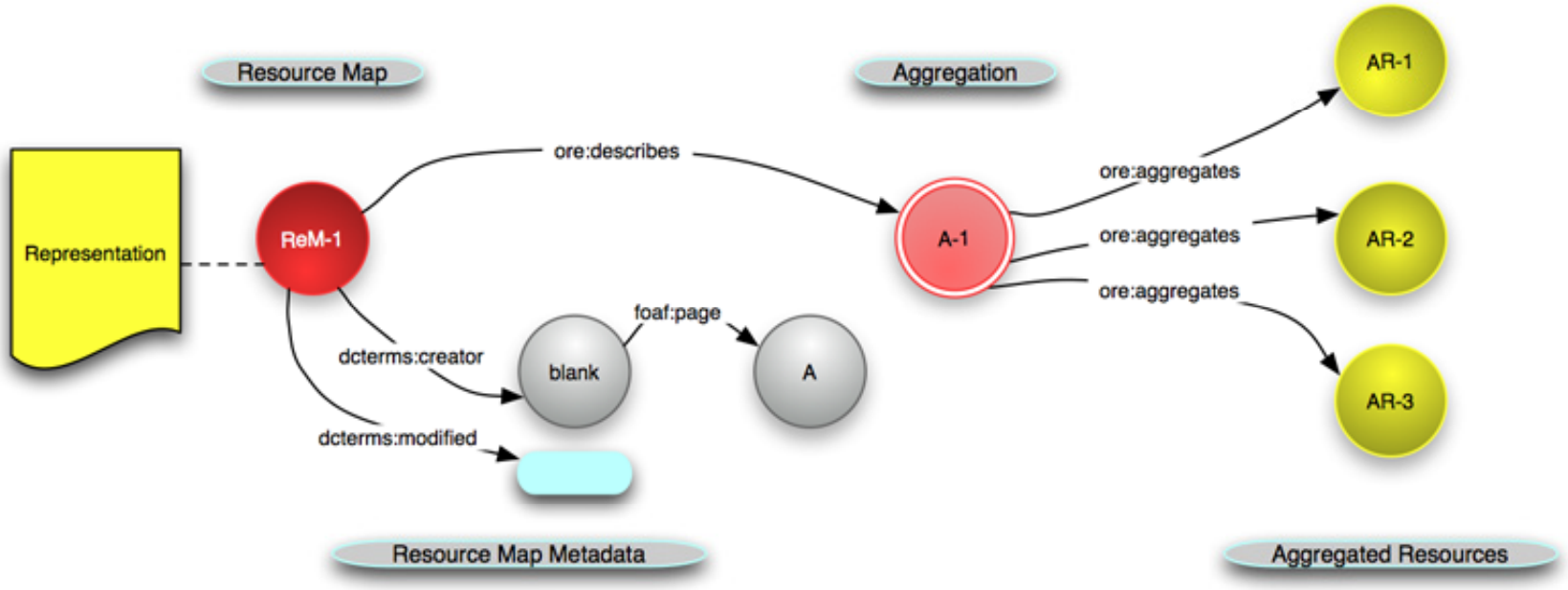
OAI-ORE solution

- OAI-ORE provides standards to
  - Describe an aggregation of web resources
  - Provide boundaries for an aggregation
  - State the relationships between aggregated resources
- OAI-ORE uses existing web standards and technologies
- Example application: improve web search

OAI-ORE solution

- Resource Map (ReM) describes the aggregation
  - Has a URI (new resource)
  - May express relationships and properties pertaining to all aggregated resources
  - Contains metadata about the ReM (e.g., who created it, when it was last modified, etc.)
  - Machine readable (serialized as Atom XML, RDF/XML, RDFa, n3, turtle, and other RDF serialization formats)

Example ReM



http://www.openarchives.org/ore/1.0/primer.html

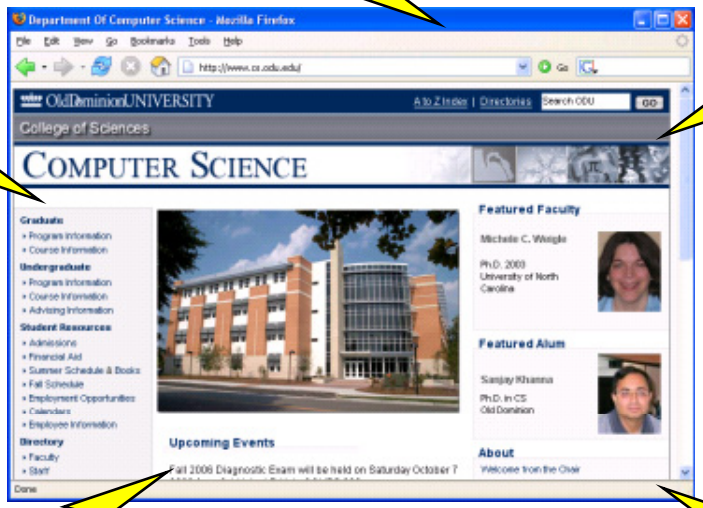
Atom XML ReM Serialization Example

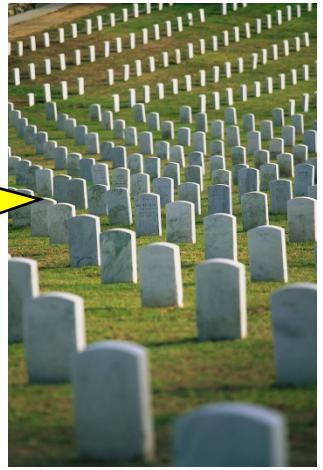
```
<?xml version="1.0" encoding="UTF-8"?>
<atom:entry
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:atom="http://www.w3.org/2005/Atom"
    xmlns:ore="http://www.openarchives.org/ore/terms/"
    xmlns:foaf="http://xmlns.com/foaf/0.1/">
    <atom:id>tag:arxiv.org,2008:astro-ph:0601007</atom:id>

    <!-- About the Aggregation for the ArXiv document -->

    <!-- This Atom entry describes a specific ORE Aggregation -->
    <atom:link href="http://arxiv.org/aggregation/astro-ph/0601007"
        rel="http://www.openarchives.org/ore/terms/describes"/>
    <atom:category term="http://www.openarchives.org/ore/terms/Aggregation"
        scheme="http://www.openarchives.org/ore/terms/" label="Aggregation"/>
    <!-- The Aggregation aggregates ... -->
    <atom:link href="http://arxiv.org/abs/astro-ph/0601007"
        rel="http://www.openarchives.org/ore/terms/aggregates"
        title="[astro-ph/0601007] Parametrization of K-essence and Its Kinetic Term"
        type="text/html" />
    <atom:link href="http://arxiv.org/ps/astro-ph/0601007"
        rel="http://www.openarchives.org/ore/terms/aggregates"
        title="Parametrization of K-essence and Its Kinetic Term" type="application/postscript"
        hreflang="en"/>
    <atom:link href="http://arxiv.org/pdf/astro-ph/0601007"
        rel="http://www.openarchives.org/ore/terms/aggregates"
        title="Parametrization of K-essence and Its Kinetic Term" type="application/pdf"
        hreflang="en"/>
    <!-- Metadata about the Aggregation: title and authors -->
    <atom:title>Parametrization of K-essence and Its Kinetic Term</atom:title>
    <atom:author>
        <atom:name>Hui Li</atom:name>
        <atom:email>lihui@somewhere.cn</atom:email>
    </atom:author>
```

http://www.openarchives.org/ore/1.0/primer.html#Serialization

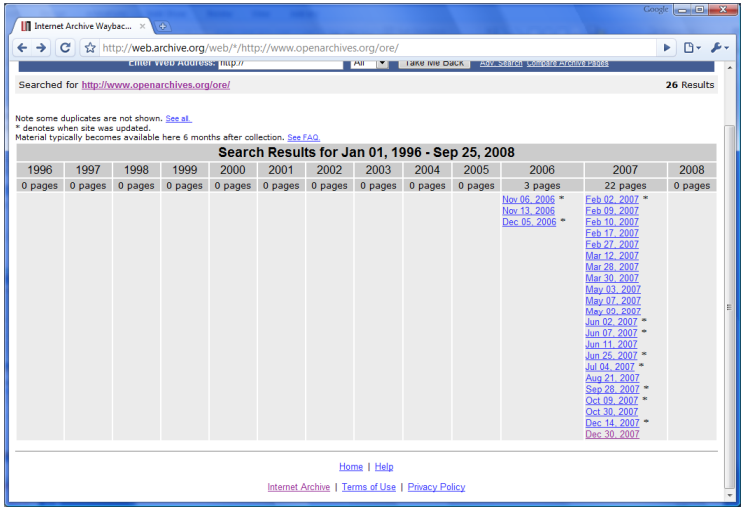




Black hat: http://img.webpronews.com/securitypronews/110705blackhat.jpg
Virus image: http://polarboing.com/images/topics/misc/story.computer.virus_1137794805.jpg
Hard drive: http://www.datarecoveryspecialist.com/images/head-crash-2.jpg

Web Infrastructure

WI capture of web page

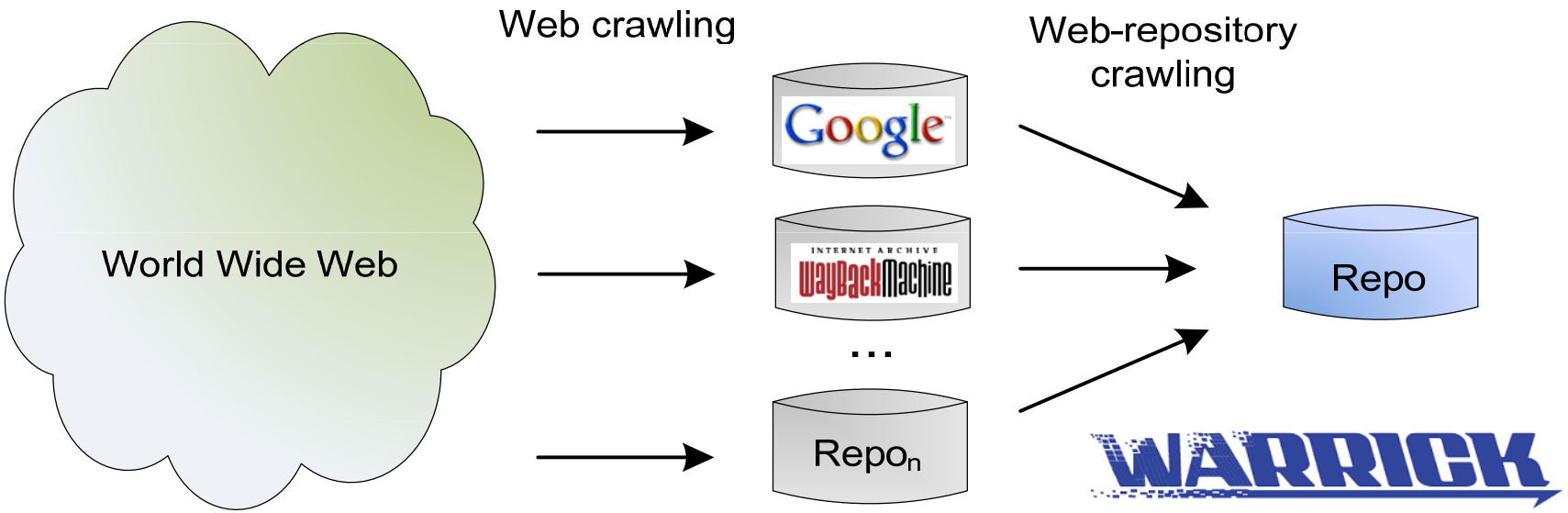


Google

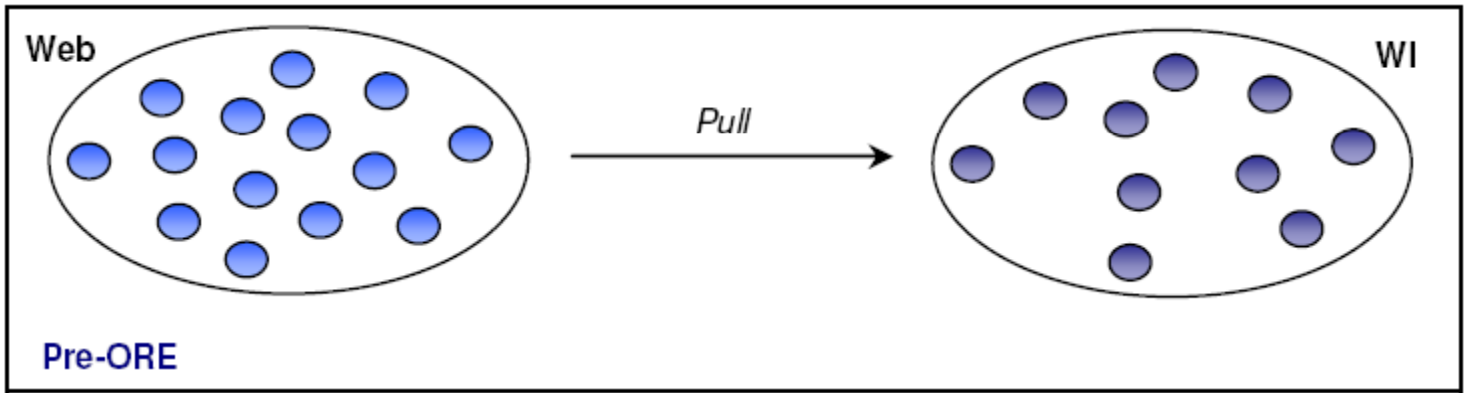
INTERNET ARCHIVE
WayBackMachine

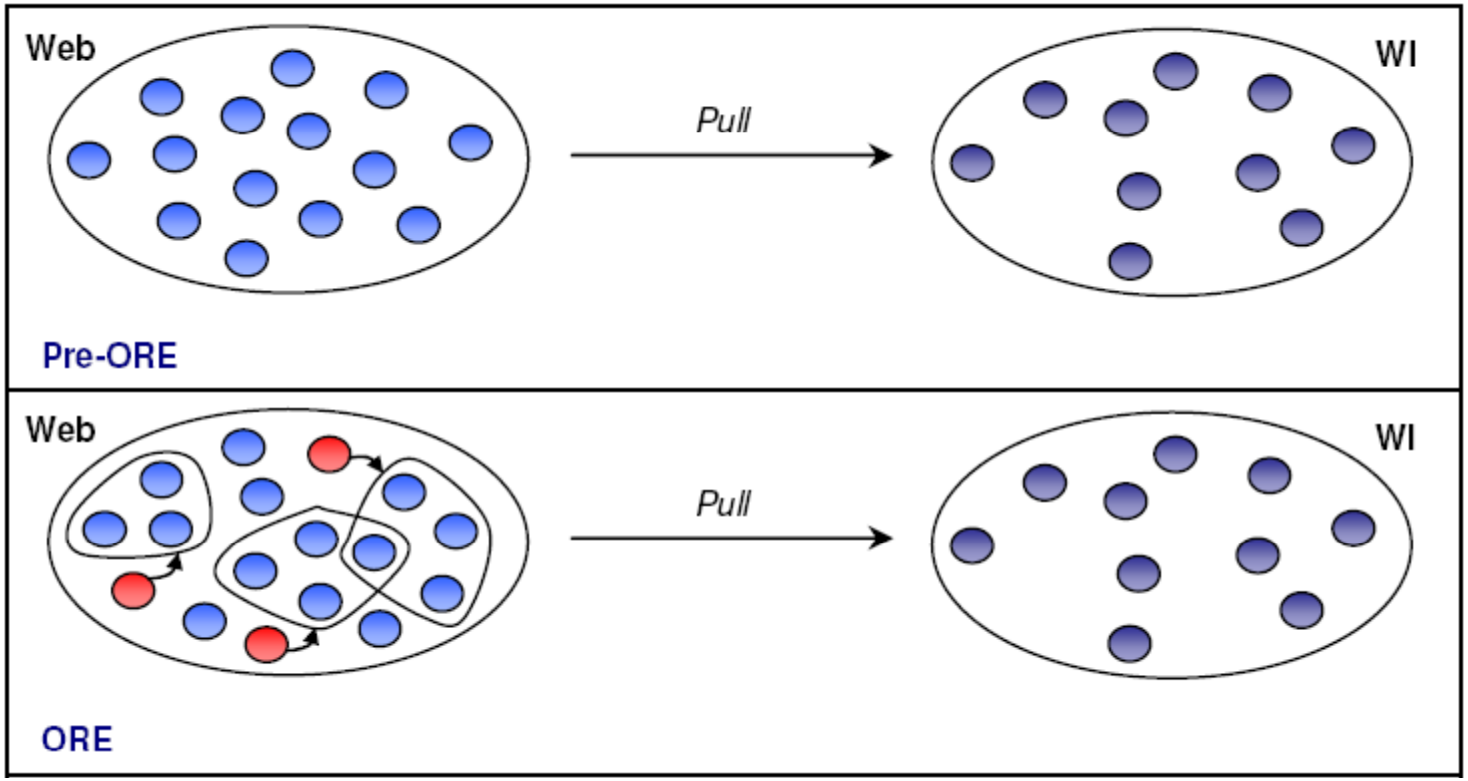
WebCite

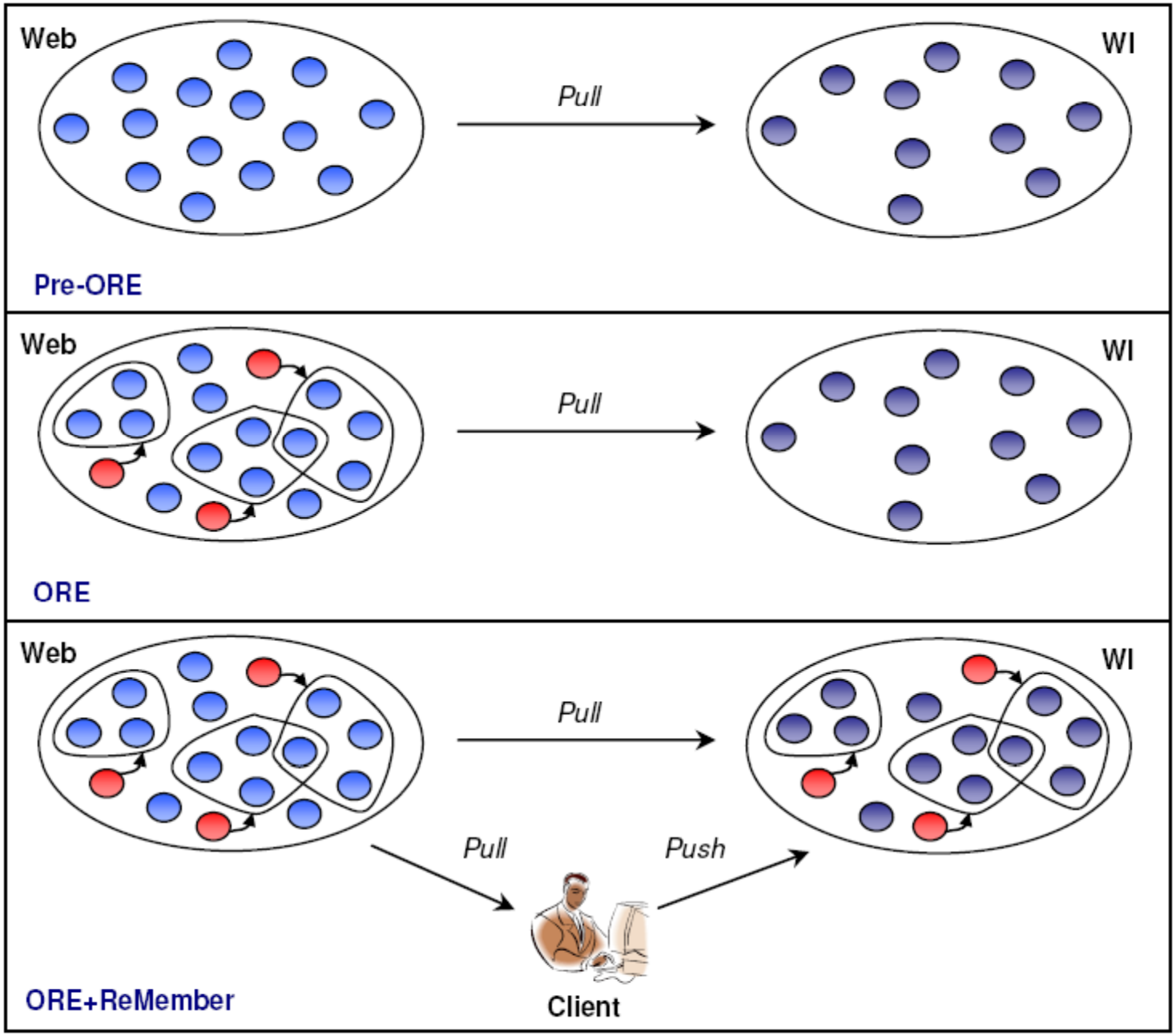
Warrick: Crawling the crawlers to reconstruct entire web sites



http://warrick.cs.odu.edu/


Web
Pull
WI
Pre-ORE


Web
WI
Pull
Pre-ORE
Web
WI
Pull
ORE

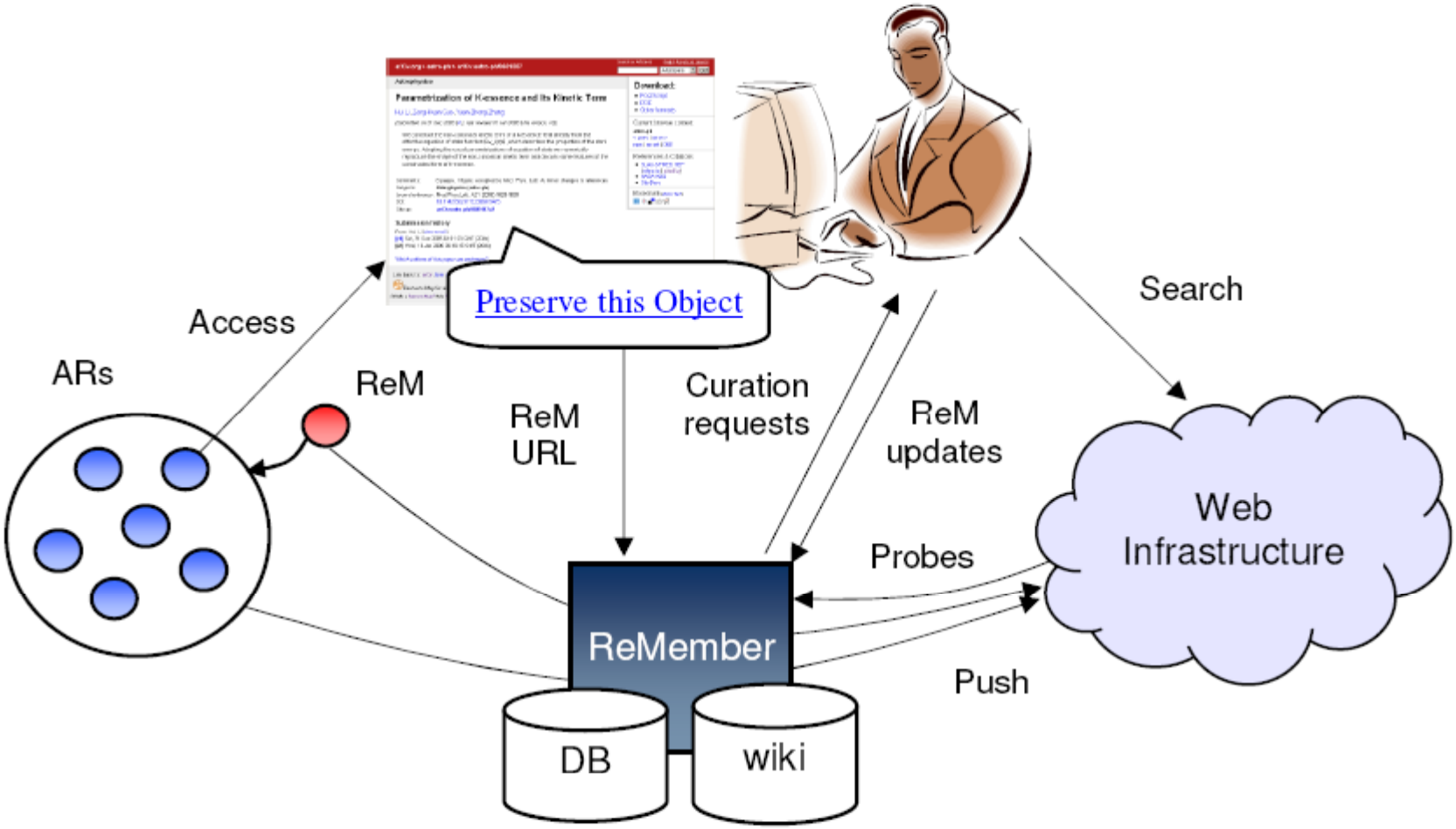

Web
Pull
WI
Pre-ORE
Web
Pull
WI
ORE
Web
Pull
WI
Pull
Push
Client
ORE+ReMember

Can we harness the web community to curate ReMs?

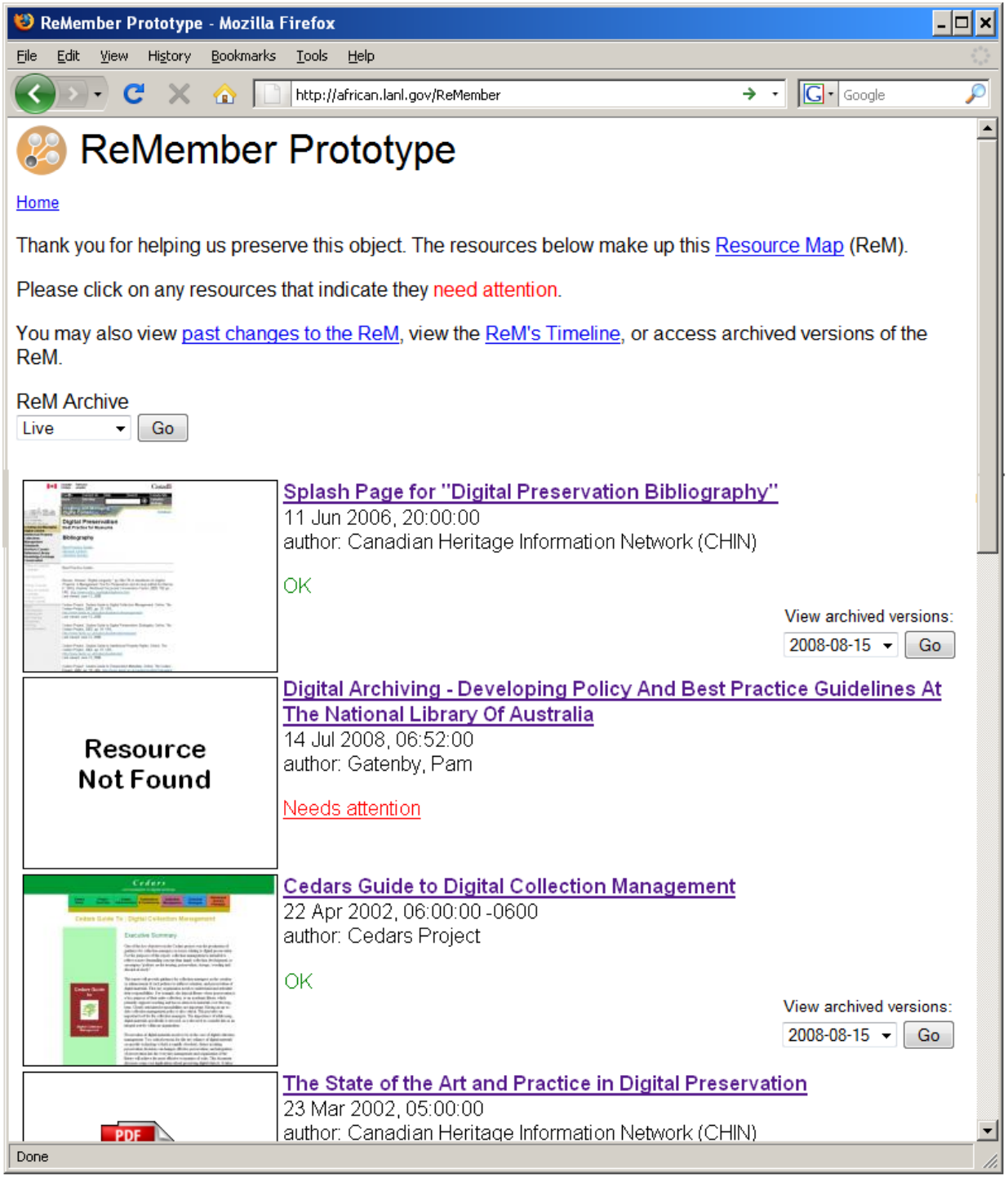
- Anyone can create a ReM
- ReM's aggregated resources may change
- Aggregated resources may go 404
- Aggregated resources may move location
- Humans are better at resolving these problems than computers

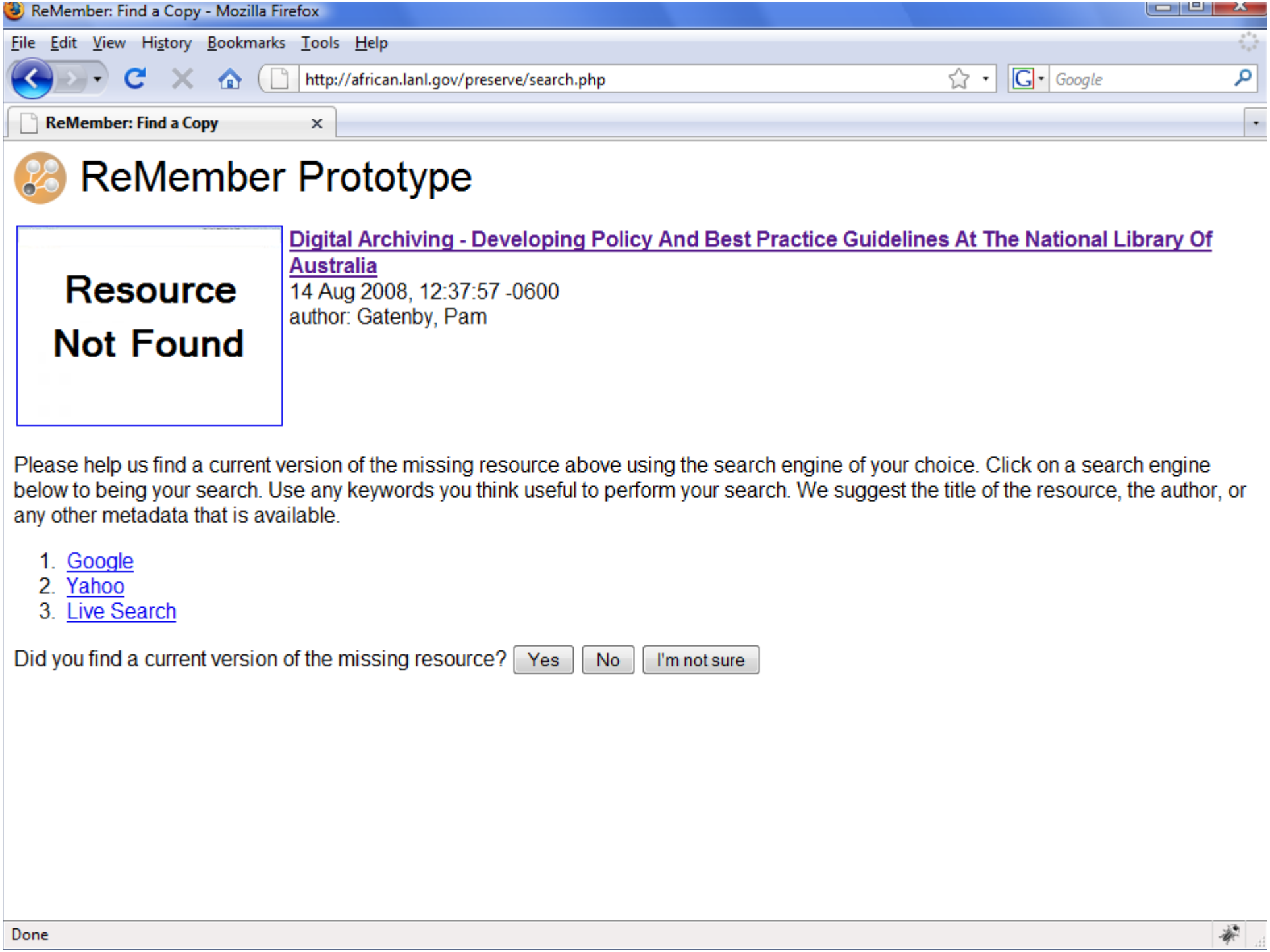
Many people
doing a little work
=
curation on a grand scale

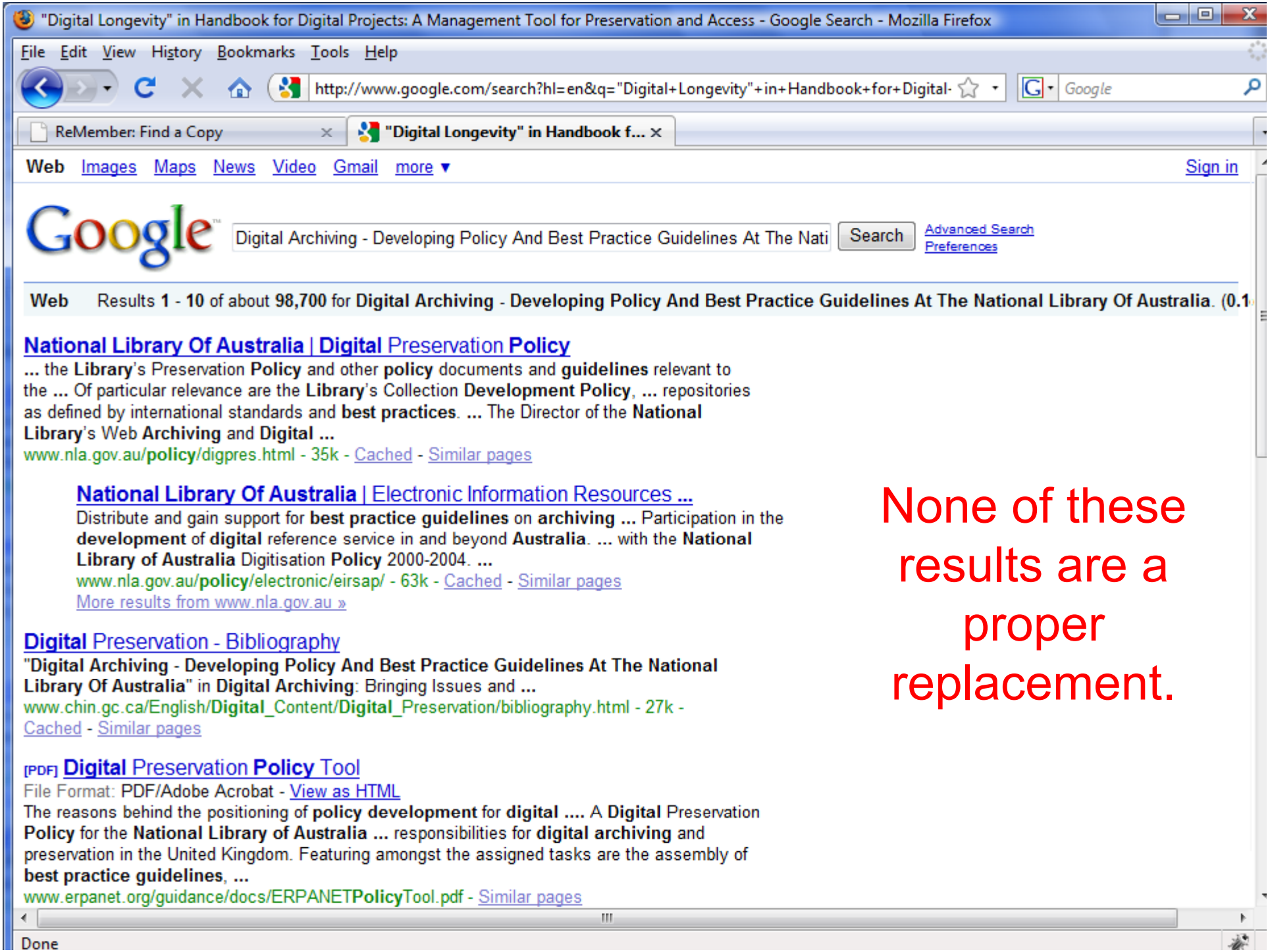
ReMember Framework



Original URL is 404:
http://www.icsti.org/2000
workshop/gatenby.html




ReMember: Find a Copy - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://african.lanl.gov/preserve/search.php
Google
ReMember: Find a Copy
ReMember Prototype
Resource Not Found
Digital Archiving - Developing Policy And Best Practice Guidelines At The National Library Of Australia
14 Aug 2008, 12:37:57 -0600
author: Gatenby, Pam
Please help us find a current version of the missing resource above using the search engine of your choice. Click on a search engine below to being your search. Use any keywords you think useful to perform your search. We suggest the title of the resource, the author, or any other metadata that is available.
1. Google
2. Yahoo
3. Live Search
Did you find a current version of the missing resource? Yes No I'm not sure
Done



None of these results are a proper replacement.

Internet Archive Wayback Machine - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://web.archive.org/web/*/http://www.icsti.org/2000workshop/gatenby.html
Google
ReMember Prototype
Internet Archive Wayback Machine
INTERNET ARCHIVE
WayBackMachine
Enter Web Address: http://
All
Take Me Back
Adv. Search
Compare Archive Pages
Searched for http://www.icsti.org/2000workshop/gatenby.html
31 Results
Note some duplicates are not shown. See all.
* denotes when site was updated.
Material typically becomes available here 6 months after collection. See FAQ.
Search Results for Jan 01, 1996 - Sep 25, 2008
1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
0 pages 0 pages 0 pages 0 pages 0 pages 0 pages 4 pages 5 pages 6 pages 6 pages 5 pages 3 pages 1 pages
Feb 08, 2002 * Jun 21, 2002 Aug 04, 2002 Oct 16, 2002
Feb 10, 2003 * Apr 01, 2003 Aug 04, 2003 Oct 20, 2003 Dec 24, 2003
Mar 01, 2004 Apr 06, 2004 Jun 21, 2004 Aug 19, 2004 * Oct 15, 2004 Dec 23, 2004
Feb 11, 2005 Feb 19, 2005 Apr 25, 2005 Oct 15, 2005 * Dec 11, 2005 Dec 16, 2005
Feb 09, 2006 May 04, 2006 Jun 30, 2006 Oct 03, 2006 * Oct 19, 2006 *
Jan 06, 2007 Aug 31, 2007 Nov 18, 2007
Jan 18, 2008
Home | Help
Internet Archive | Terms of Use | Privacy Policy
See if this is the correct resource
Done

ICSTI, Digital Archiving, P.Gatenby, Jan. 2000 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://web.archive.org/web/20080118212818/http://www.icsti.org/2000workshop/gatenby.h

Google

ReMember Prototype | ICSTI, Digital Archiving, P.Gatenb...

Digital Archiving — Developing Policy And Best Practice Guidelines At The National Library Of Australia

ReM's URL updated to point to archived version.

By Pam Gatenby, National Library of Australia

1. Introduction

The main focus of my presentation is the business model the National Library of Australia has put in place to archive digital on-line publications. I'll explain the strategic framework and rationale for the approach we are currently taking; the processes we follow and then outline areas we are addressing to ensure cost-effective and workable solutions to providing long-term access to digital publications.

Like most national libraries, the National Library of Australia has responsibility for collecting, preserving and making available the published national output. It accepts that this responsibility extends to digital publications in all forms.

We have been archiving Internet Australian publications since 1996 when the PANDORA project was established. PANDORA stands for Preserving and Accessing Networked Documentary Resources of Australia. Through this project we have been working at two levels to develop our strategy for providing long-term access to Australian on-line publications — at the conceptual level to establish the standards, technical, and business requirements for a durable, robust archive; and at the practical level to actually start building an archive of titles. This now numbers 450 titles which includes some very complex ones which have provided us with significant technical challenges, and many which are gathered on an on-going basis.

At the Library we attach great importance and value to the "hands-on" approach and to collaborative effort. After an initial period of intensive discussion and grappling with issues back in 1995, we decided that the only way to make progress was to actually start doing something in order to learn through experience and to develop our ideas further. The progress that the PANDORA project has made since that time is a tribute to the intelligence, perseverance and dedication of a small team of staff who have worked closely together on many difficult issues. PANDORA was launched based on a set of basic software tools developed in-house, and a modest level of staffing funded from recurrent budget. This approach has had some inherent operational frustrations but has allowed us to test and refine our theoretical principles.

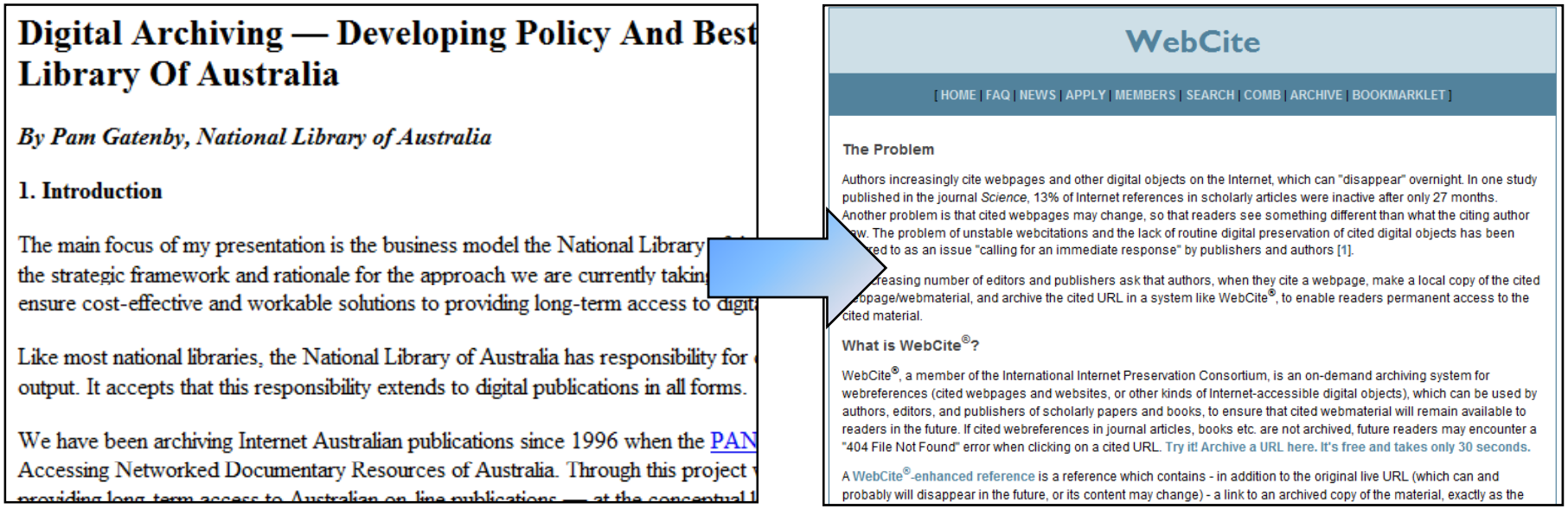
2. Strategic framework for digital archiving implementation

The Library's overall approach to managing the archiving and preservation of digital publications and to providing access to them, takes place within a

Done

What if Resource is found on “live” Web?

1. Update ReM to point to new location
2. Push live version to WebCite



Digital Archiving — Developing Policy And Best
Library Of Australia

By Pam Gatenby, National Library of Australia

1. Introduction

The main focus of my presentation is the business model the National Library
the strategic framework and rationale for the approach we are currently takin
ensure cost-effective and workable solutions to providing long-term access to digit

Like most national libraries, the National Library of Australia has responsibility for
output. It accepts that this responsibility extends to digital publications in all forms.

We have been archiving Internet Australian publications since 1996 when the PAN
Accessing Networked Documentary Resources of Australia. Through this project

WebCite

[HOME | FAQ | NEWS | APPLY | MEMBERS | SEARCH | COMB | ARCHIVE | BOOKMARKLET]

The Problem

Authors increasingly cite webpages and other digital objects on the Internet, which can "disappear" overnight. In one study published in the journal *Science*, 13% of Internet references in scholarly articles were inactive after only 27 months. Another problem is that cited webpages may change, so that readers see something different than what the citing author saw. The problem of unstable webcitations and the lack of routine digital preservation of cited digital objects has been referred to as an issue "calling for an immediate response" by publishers and authors [1].

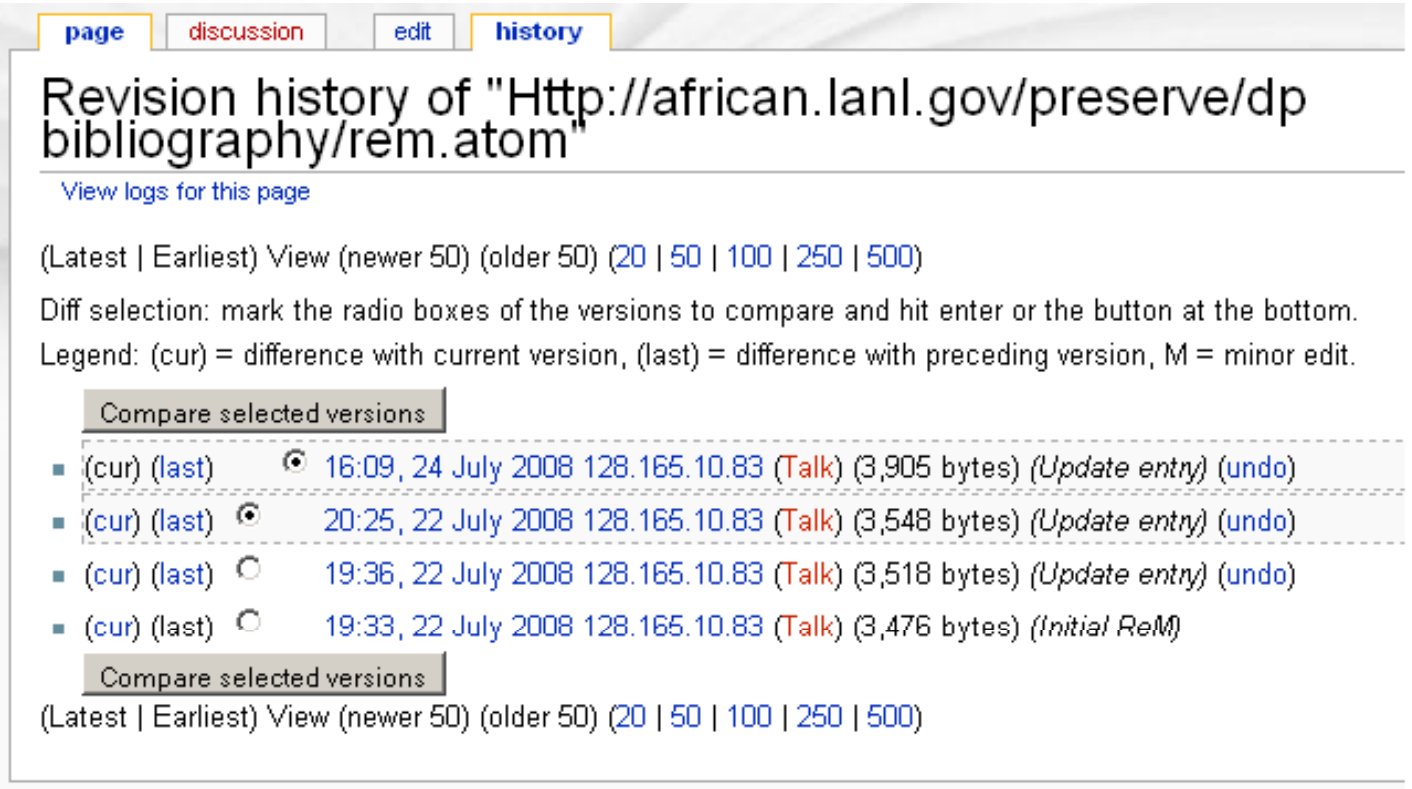
An increasing number of editors and publishers ask that authors, when they cite a webpage, make a local copy of the cited webpage/webmaterial, and archive the cited URL in a system like WebCite®, to enable readers permanent access to the cited material.

What is WebCite®?

WebCite®, a member of the International Internet Preservation Consortium, is an on-demand archiving system for webreferences (cited webpages and websites, or other kinds of Internet-accessible digital objects), which can be used by authors, editors, and publishers of scholarly papers and books, to ensure that cited webmaterial will remain available to readers in the future. If cited webreferences in journal articles, books etc. are not archived, future readers may encounter a "404 File Not Found" error when clicking on a cited URL. Try it! Archive a URL here. It's free and takes only 30 seconds.

A WebCite®-enhanced reference is a reference which contains - in addition to the original live URL (which can and probably will disappear in the future, or its content may change) - a link to an archived copy of the material, exactly as the

Wiki for revision control



page | discussion | edit | history

Revision history of "Http://african.lanl.gov/preserve/dp bibliography/rem.atom"

View logs for this page

(Latest | Earliest) View (newer 50) (older 50) (20 | 50 | 100 | 250 | 500)

Diff selection: mark the radio boxes of the versions to compare and hit enter or the button at the bottom.

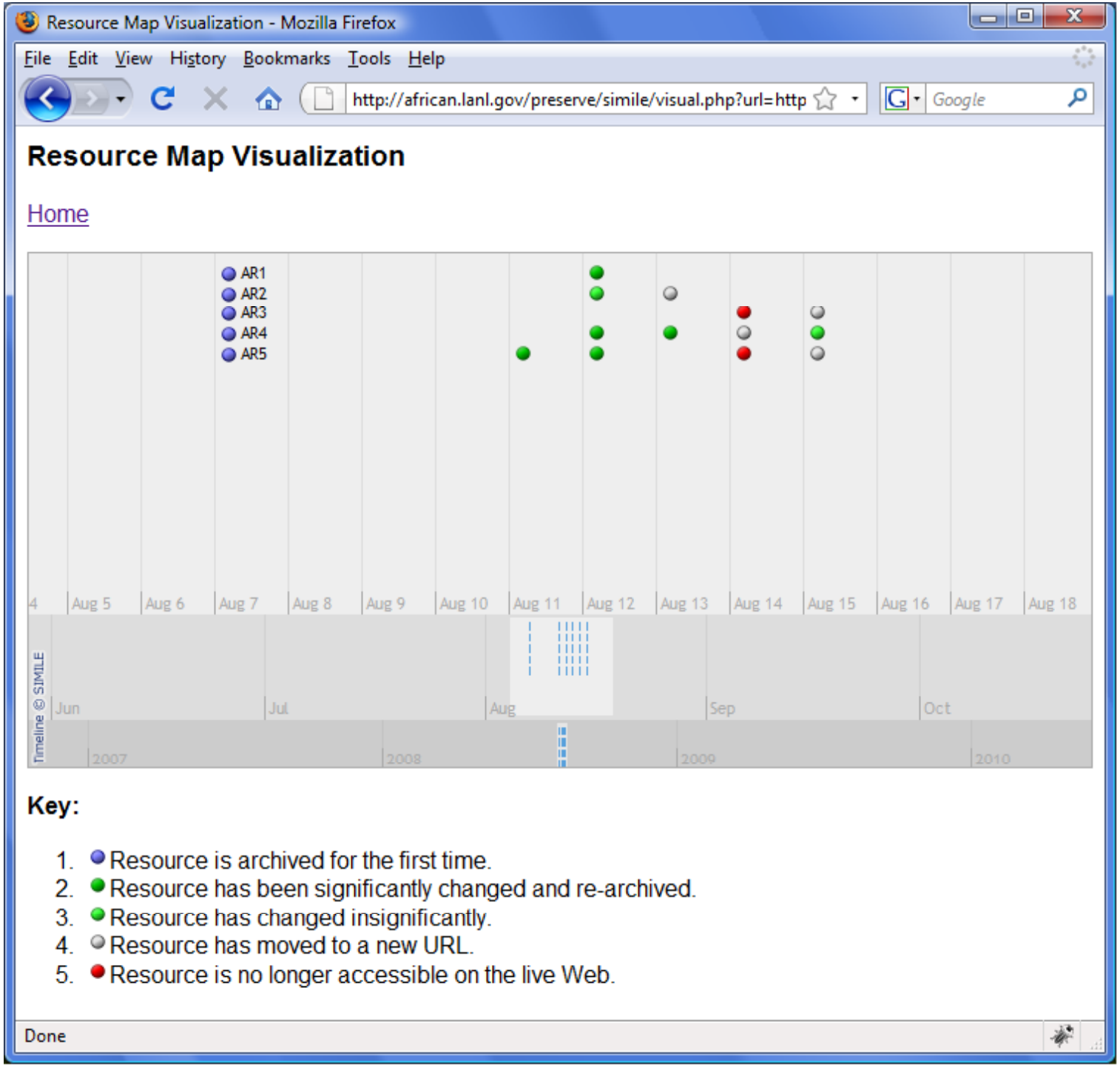
Legend: (cur) = difference with current version, (last) = difference with preceding version, M = minor edit.

Compare selected versions

- (cur) (last) 16:09, 24 July 2008 128.165.10.83 (Talk) (3,905 bytes) *(Update entry)* (undo)
- (cur) (last) 20:25, 22 July 2008 128.165.10.83 (Talk) (3,548 bytes) *(Update entry)* (undo)
- (cur) (last) 19:36, 22 July 2008 128.165.10.83 (Talk) (3,518 bytes) *(Update entry)* (undo)
- (cur) (last) 19:33, 22 July 2008 128.165.10.83 (Talk) (3,476 bytes) *(Initial ReM)*

Compare selected versions

(Latest | Earliest) View (newer 50) (older 50) (20 | 50 | 100 | 250 | 500)


Resource Map Visualization - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://african.lanl.gov/preserve/simile/visual.php?url=http
Google
Resource Map Visualization
Home
AR1
AR2
AR3
AR4
AR5
4 Aug 5 Aug 6 Aug 7 Aug 8 Aug 9 Aug 10 Aug 11 Aug 12 Aug 13 Aug 14 Aug 15 Aug 16 Aug 17 Aug 18
Jun Jul Aug Sep Oct
2007 2008 2009 2010
Timeline © SIMILE
Key:
1. Resource is archived for the first time.
2. Resource has been significantly changed and re-archived.
3. Resource has changed insignificantly.
4. Resource has moved to a new URL.
5. Resource is no longer accessible on the live Web.
Done

Future Work

- Waiting for ORE to catch on (Version 1.0 just released Oct 2008)
- Will user community adopt ReMember?
- How do we prevent or detect malicious alterations?
- Allow public to add and remove Resources from ReMs?

Thank You

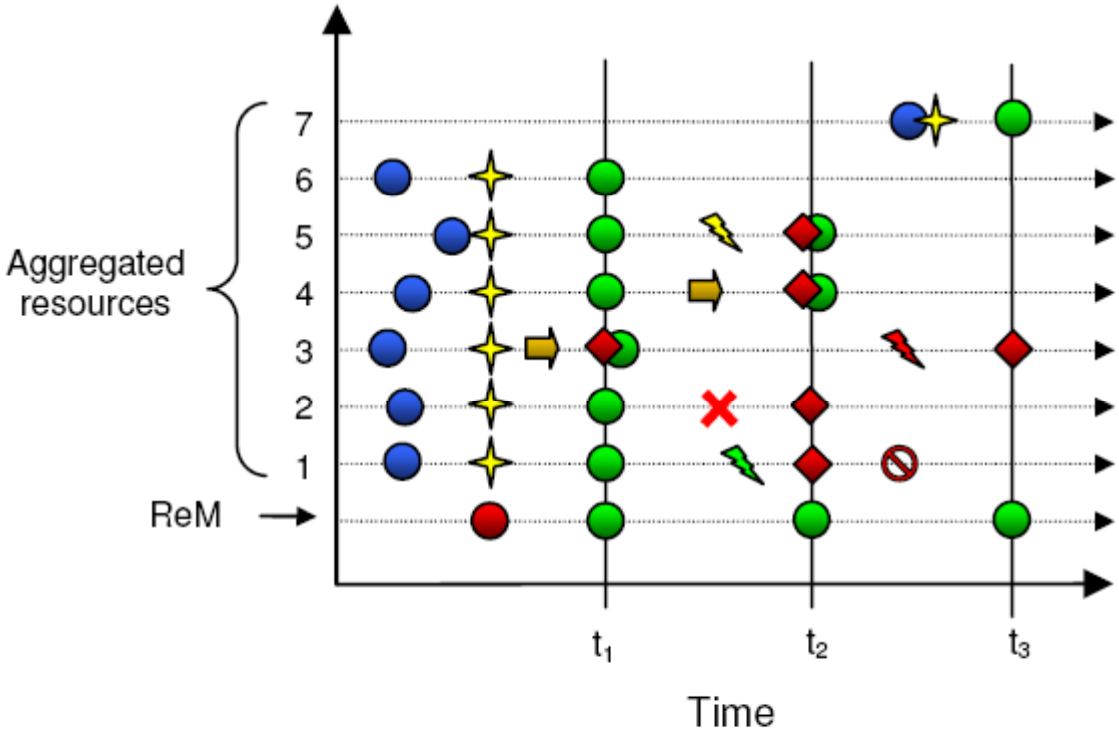
Frank McCown

fmccown@harding.edu

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

HARDING

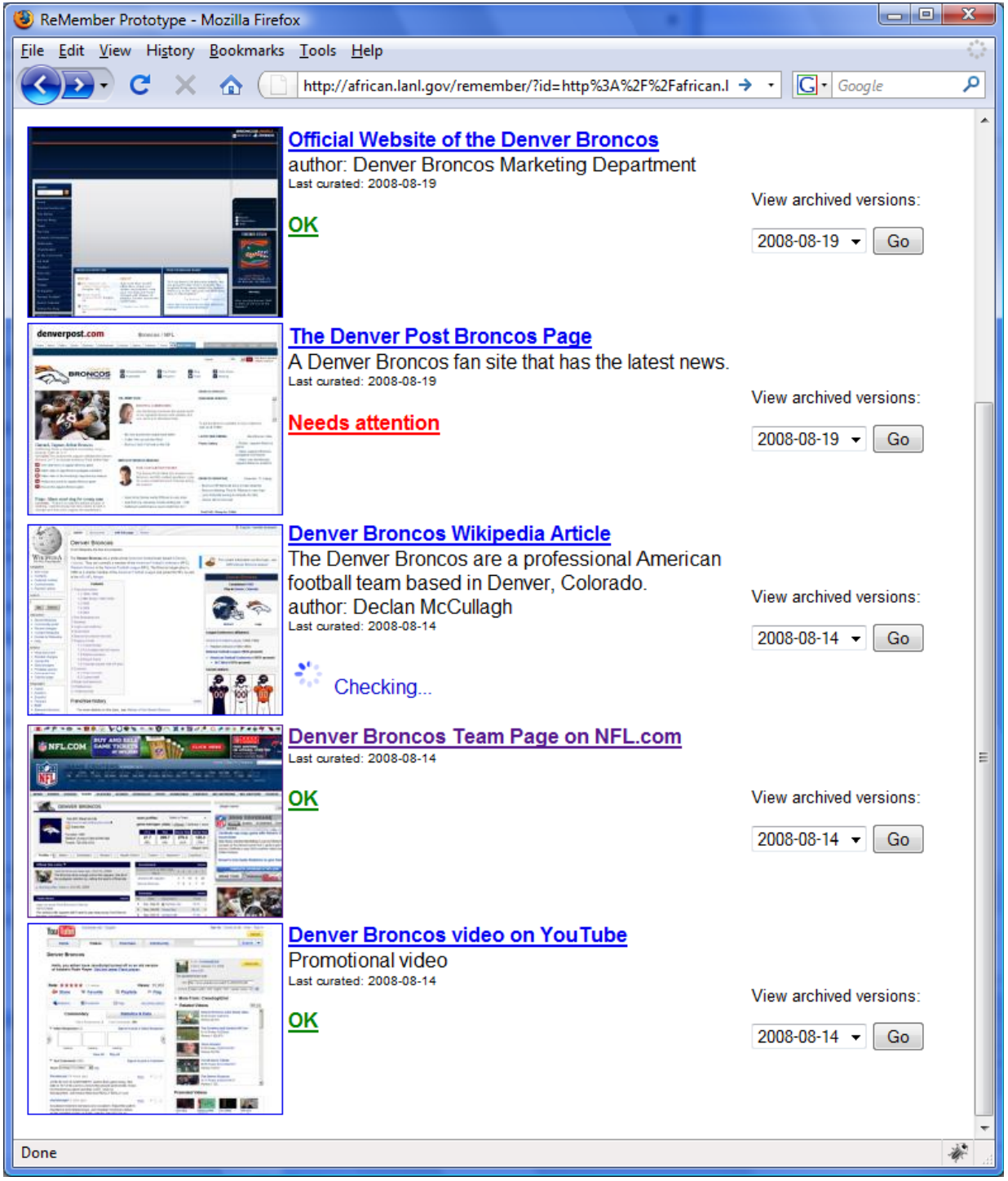
OAI-ORE work is supported by the Andrew W. Mellon Foundation, Microsoft Corporation, and the National Science Foundation (IIS-0430906).



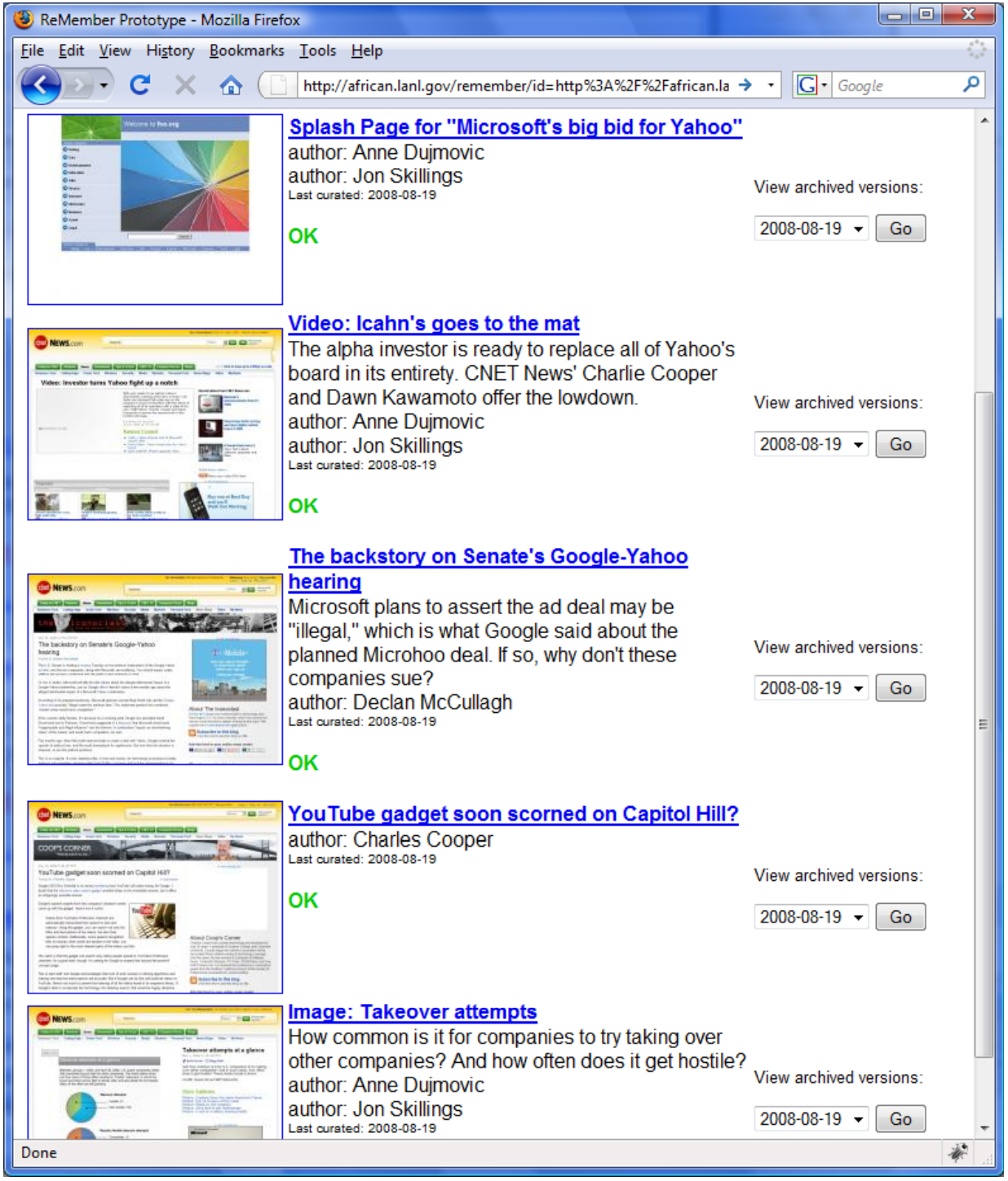
Event Key

- AR is created
- ReM is created
- Resource is archived in WI
- AR is removed from ReM
- AR is added to ReM
- AR needs human attention
- AR changes URI (previous URI is 404)
- AR is removed from the Web (404, no new URI)
- AR is changed in an insignificant way
- AR is changed in a significant way
- AR's URL is pointing to different content (no new URI)

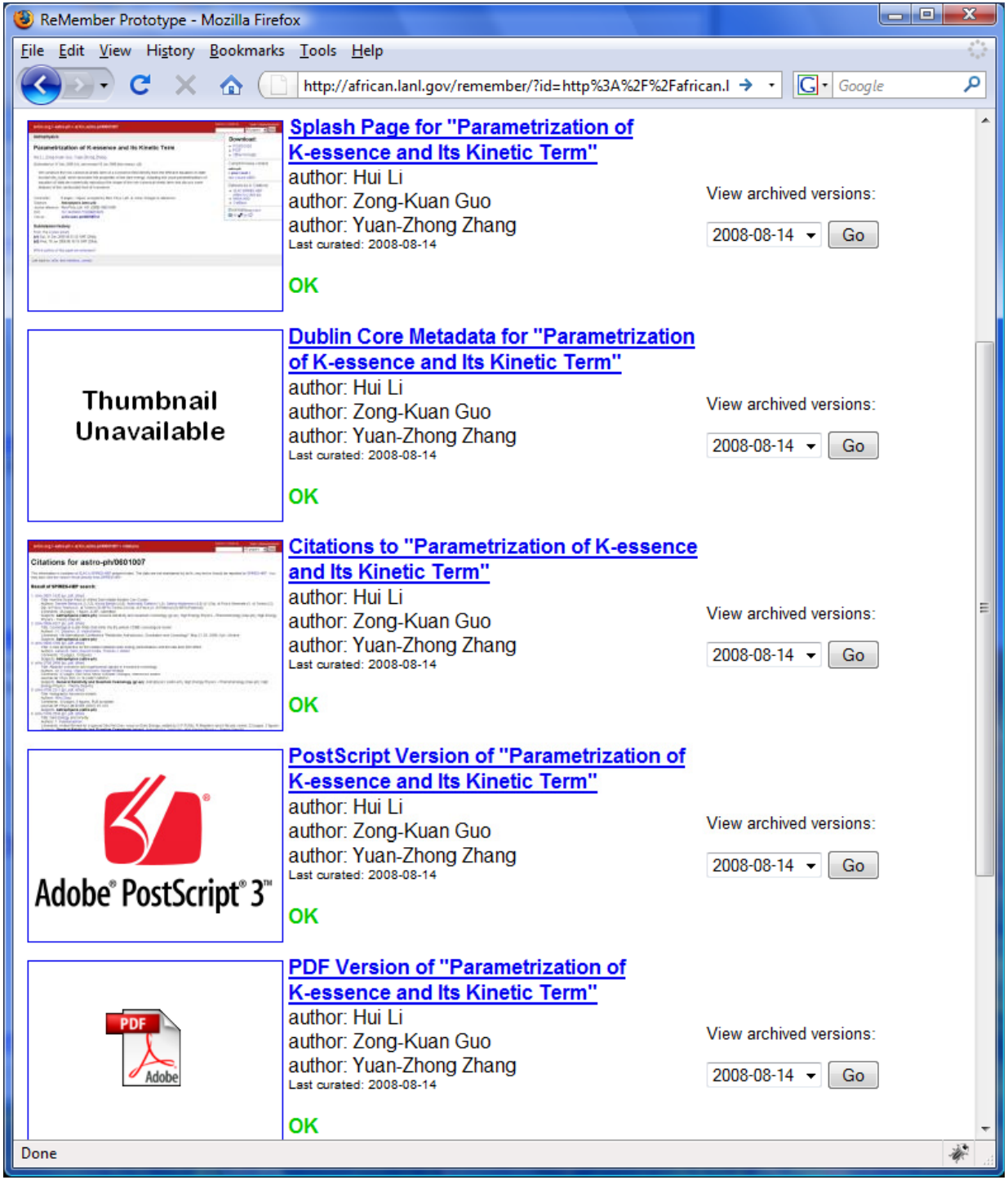
Denver Broncos ReM



Microsoft-Yahoo Merger ReM



arXiv ReM


ReMember Prototype - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://african.lanl.gov/remember/?id=http%3A%2F%2Fafrican.l
Google
Splash Page for "Parametrization of K-essence and Its Kinetic Term"
author: Hui Li
author: Zong-Kuan Guo
author: Yuan-Zhong Zhang
Last curated: 2008-08-14
OK
View archived versions:
2008-08-14
Go
Thumbnail Unavailable
Dublin Core Metadata for "Parametrization of K-essence and Its Kinetic Term"
author: Hui Li
author: Zong-Kuan Guo
author: Yuan-Zhong Zhang
Last curated: 2008-08-14
OK
View archived versions:
2008-08-14
Go
Citations to "Parametrization of K-essence and Its Kinetic Term"
author: Hui Li
author: Zong-Kuan Guo
author: Yuan-Zhong Zhang
Last curated: 2008-08-14
OK
View archived versions:
2008-08-14
Go
PostScript Version of "Parametrization of K-essence and Its Kinetic Term"
Adobe PostScript 3
author: Hui Li
author: Zong-Kuan Guo
author: Yuan-Zhong Zhang
Last curated: 2008-08-14
OK
View archived versions:
2008-08-14
Go
PDF Version of "Parametrization of K-essence and Its Kinetic Term"
PDF
Adobe
author: Hui Li
author: Zong-Kuan Guo
author: Yuan-Zhong Zhang
Last curated: 2008-08-14
OK
View archived versions:
2008-08-14
Go
Done