

DigCCurr 2007 – April 18-20 UNC
Building Capabilities for Digital Curation Repositories

Web At Risk: Extending the Digital Curation Mission to the Web

Patricia Cruse, Director, Digital Preservation Program
Kirsten Neilsen, Digital Preservation Services Manager
California Digital Library

The Digital Preservation Program

- Established in 2002
- UC-wide program
- Goal: ensure long-term availability and accessibility to materials that are important to the research, teaching, and learning on the UC campuses.
- Centrally managed
- Central and external funds
- A partnership



[Home](#)
[Search](#)
[Browse](#)
[Print View](#)

[Go](#)

[About Us](#)
[Help](#)

Speaking with Vampires

[Acknowledgments](#)
[A Note on Currencies and Talk](#)


[Part One](#)
[1. Blood and Wounds](#)
[2. Histericizing Rumor and Gossip](#)

[Part Two](#)
[3. Translators on Your Mouth](#)
[4. "Why Is Detroit Red?"](#)

[Part Three](#)
[5. "A Special Danger"](#)
[6. "Boast Master Captaincy"](#)
[7. Blood, Buds, and Archives](#)
[8. Citizenship and Censorship](#)
[9. Class Struggle and Cannibalism](#)
[10. Conclusions](#)

[Bibliography](#)
[Credits](#)

[Collapse All](#) | [Expand All](#)



Speaking with Vampires
Rumor and History in Colonial Africa
Luise White
 UNIVERSITY OF CALIFORNIA PRESS
Berkeley · Los Angeles · London
 © 2000 The Regents of the University of California

[Preferred Citation:](#) White, Luise. *Speaking with Vampires: Rumor and History in Colonial Africa*. Berkeley: University of California Press, c2000 2000.
<http://ark.cdlib.org/ark:/13030/tfr29p2x/>



Sylvester Stallone

April 28, 1983

Mr. Bob Kowloff
ASSOCIATED FILM PROMOTION
10100 Santa Monica Blvd.
LOS Angeles, CA 90067


DEAR BOB:

As discussed, I guarantee that I will use Brown & Williamson tobacco products in no less than five feature films.

It is my understanding that Brown & Williamson will pay a fee of \$200,000.00.

Hoping to hear from you soon;

Sincerely,



Sylvester Stallone

ss/sp

UCLA LIBRARY | Digital Collections

Questions

musica de la

FRONTERA

*composiciones de la
tradicin
Mexicana
Americana*

*musica de la
tradicin
Mexicana
Americana*

*musica de la
tradicin
Mexicana
Americana*

Search in English

Búsqueda en español

About the project
Sobre el proyecto

© 2004
© The Berkman Foundation

Cornerstone of the Program: Digital Preservation Repository (DPR)

- Suite of tools & services:
 - Digital Preservation Repository
 - Documentation, guidelines, policies
- Intern'l Standards & Open Source
- Service oriented architecture: flexible, adaptable, simple
- Preservation Partnership
 - Curate
 - Preserve

Digital Preservation Repository core services

- A set of services that support the long-term retention of digital objects:
 - Submit (deposit) digital objects
 - Manage digital objects: add versions, replace, update, delete
 - Request dissemination
 - Request administrative reports (forthcoming)
- What the service is not...



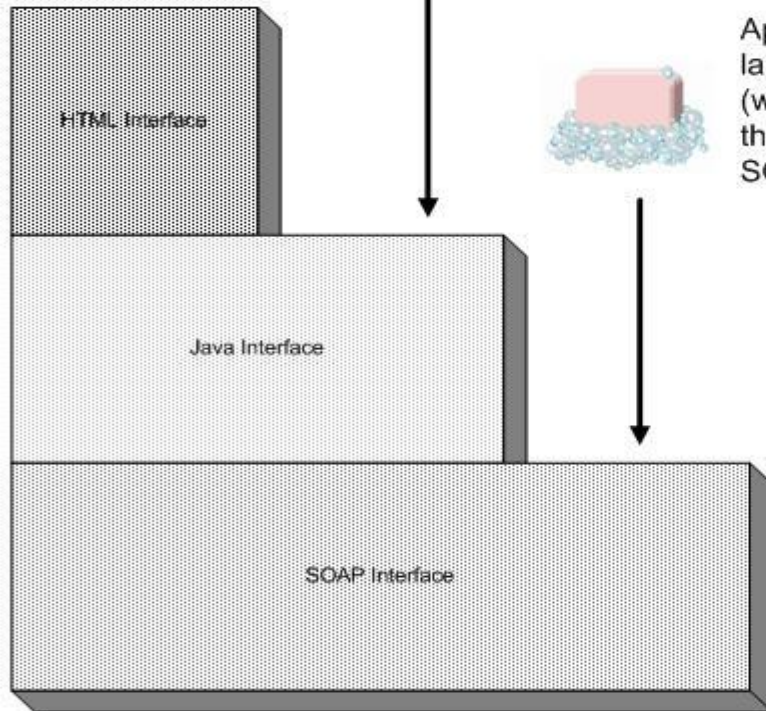
Humans using browsers can interact with the DPR through the CDL-supplied HTML interface.



Applications written in Java can interact through the CDL-supplied Java client library.



Applications written in any language that supports SOAP (with Attachments) can interact through the CDL-supplied SOAP interface.



Main Menu

Submit Objects

Create new objects; or **add** or **replace** versions of objects.

Submit a single object

You may wish to **validate** the object first.

Submit multiple objects

You may wish to **validate** the objects first.

Browse

Browse by **inventory**

- [UCD - Eastman Photographs](#)
- [UCD - Yolo Aerial Photographs](#)
- [UCD - Faculty Entomology Images](#)

Search

Locate an object by its **identifier**

The **locate** functionality performs exact matches on identifiers.

Object ID (ARK):

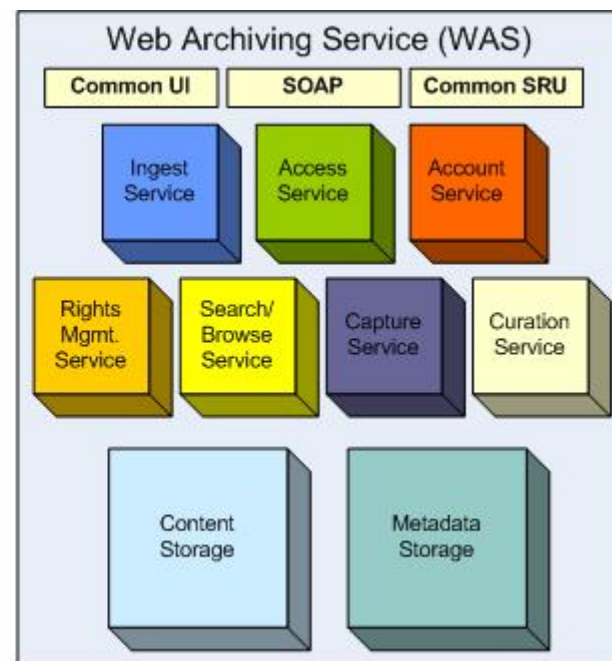
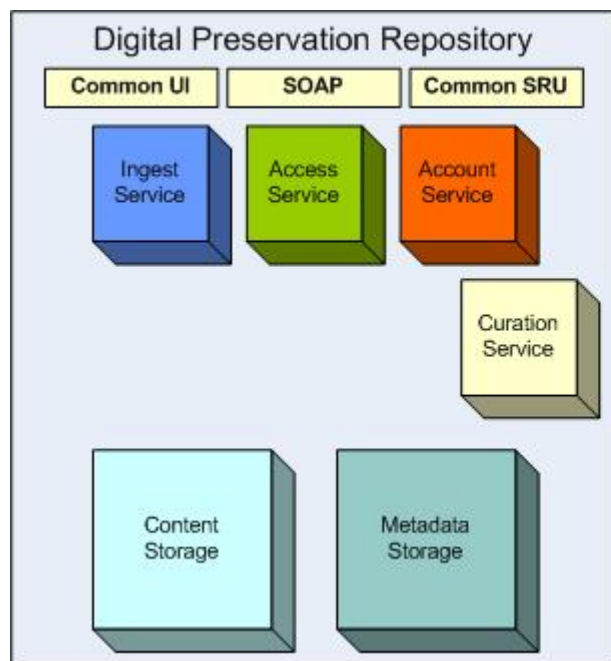
ARK syntax: `ark:/nnnnn/ccccccccc`

[Locate](#)

To locate an object by its **Local Object ID**, **choose the inventory** to which that object belongs.

Locate an object to either **add** or **replace** versions, or to **view** or **download** a version.

DPR to Web Archiving Service



Web-at-Risk: NDIIPP Funds

Jan 2005 – Jan 2008

- Build tools to allow librarians to capture, curate and preserve web-based government and political information.
 - Create topical and event-based archives
 - Capture individual sites and documents
- Assess the impact of these tools on traditional collection development practices.
- Explore web archiving service sustainability.

Project Partners

Technical Partners

San Diego
Supercomputer Center

Computer Science Dept.,
Stanford University

Sun Microsystems

Library of Congress

Martha Anderson - Agreement Officer
Technical Representative

California Digital Library

Daniel Greenstein – P.I.

New York University

James Bullen

University of North

Texas

Cathy Hartman

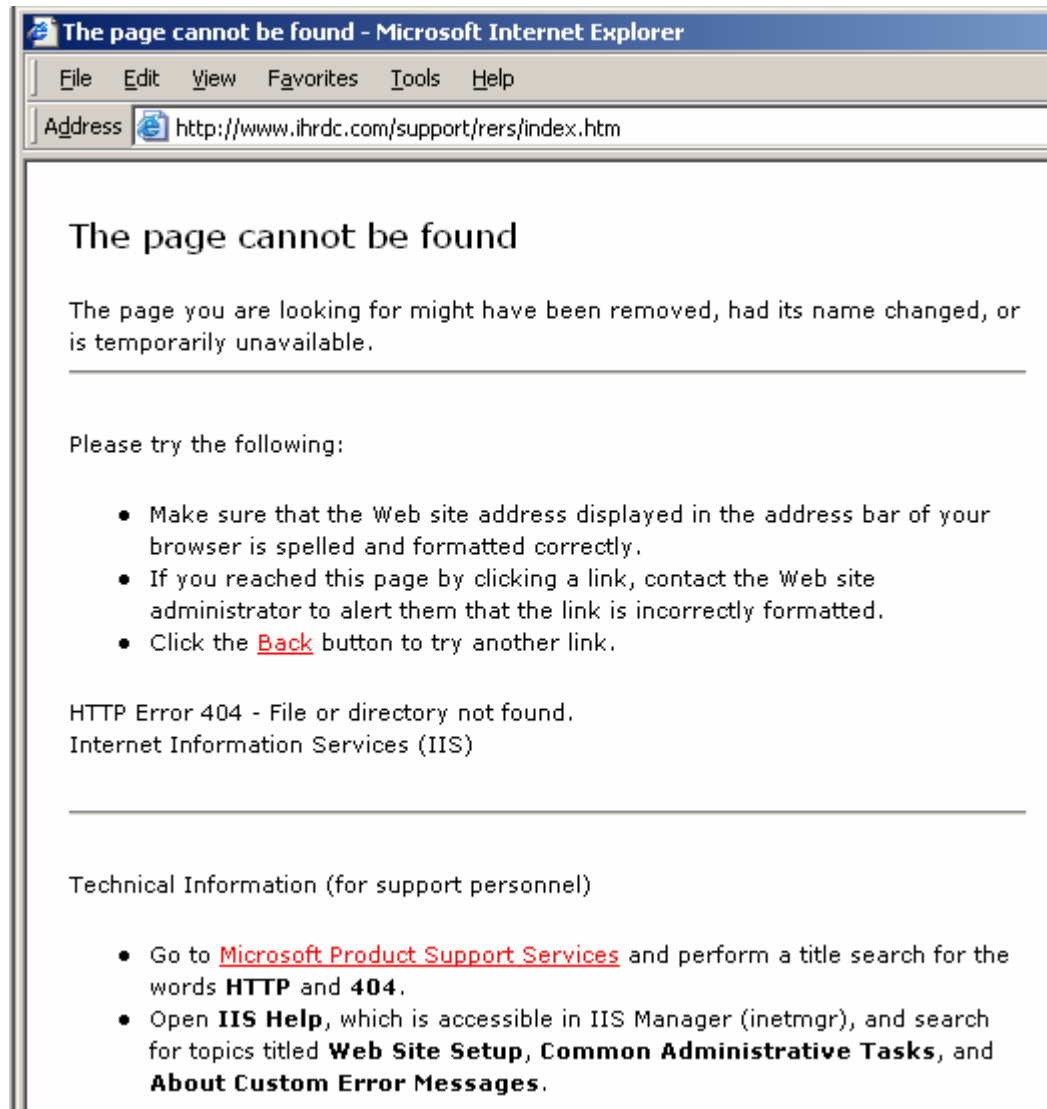
Curatorial Partners

Arizona State Library
Stanford
UC Davis
UCLA
UC San Francisco

New York University
UC Berkeley
UC Irvine
UC San Diego
UC Santa Cruz

Preserving the Web

- Why all the fuss?
- What is “Web Archiving?”
- Web Archiving Service (WAS)
 - Collecting content
 - Curating content
- Current status & future plans





U.S. DEPARTMENT of STATE

U.S. Department of State

Home

Issues & Press

Travel & Business

Youth & Education

About State Department

404 - Document Not Found



Homeland Security

Contact Us | Site Map

Search:

Go

Home

Information Sharing
and Analysis

Prevention &
Protection

Preparedness &
Response

Research

Commerce &
Trade

Travel Security &
Procedures

Immigration

About the Department

Open for Business

Press Room

About the Department

404 : Page can not be found

Most Requested



UNITED STATES DEPARTMENT OF VETERANS AFFAIRS

Veterans Affairs banner with U.S. Flag

VA Home

About VA

Organizations

Apply Online



Find a Facility

Contact VA

Search

Health Care



Benefits



Burial & Memorials



VA MID-ATLANTIC HEALTH CARE NETWORK - VISN 6

HTTP 404: Page Not Found



**Web-Based Government Information: Evaluating Solutions for Capture,
Curation, and Preservation**

An Andrew W. Mellon Funded Initiative of the California Digital Library

- 2003 survey of the .gov domain:
 - as much as 65 percent of all government publications that are distributed to libraries through the federal depository library program are currently produced exclusively in electronic form and distributed via the web.

What is a “Web Archive?”

- Automated method to gather web content
- Collections composed of multiple sites
- Captured content preserved
- Meaningful access to content provided
 - Public or end-user access may not be available

Web

55 billion pages



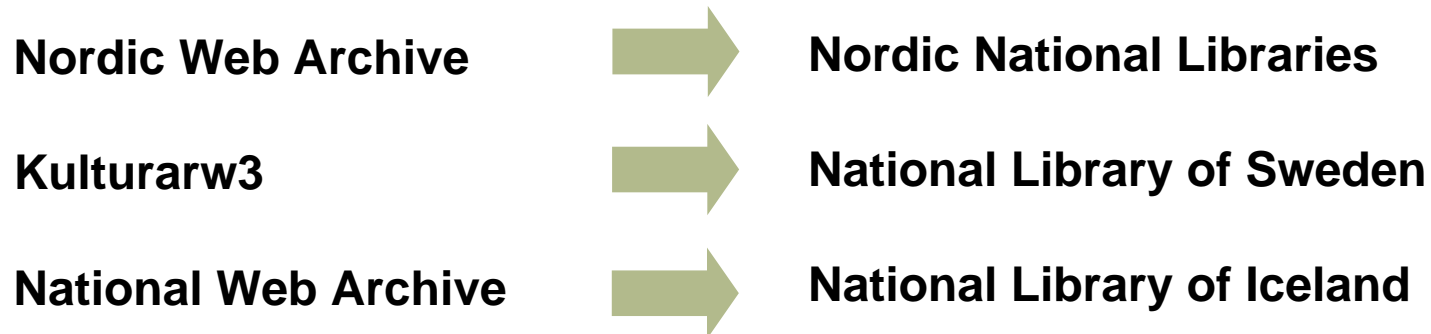
Take Me Back

[Advanced Search](#)

Search Results for Jan 01, 1996 - Oct 22, 2006

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
0 pages	2 pages	5 pages	13 pages	15 pages	19 pages	18 pages	33 pages	51 pages	13 pages	0 pages
	Oct 14, 1997 * Dec 11, 1997 *	Jan 09, 1998 * Feb 11, 1998 * Apr 25, 1998 * Dec 12, 1998 * Dec 12, 1998 *	Jan 25, 1999 * Jan 28, 1999 * Feb 02, 1999 * Feb 08, 1999 * Feb 10, 1999 * Feb 18, 1999 * Apr 17, 1999 * Apr 28, 1999 * Apr 29, 1999 * May 08, 1999 * Oct 08, 1999 * Oct 13, 1999 * Nov 04, 1999 *	Feb 29, 2000 * May 11, 2000 * May 20, 2000 * Jun 06, 2000 * Jun 11, 2000 * Jun 20, 2000 * Jun 22, 2000 * Aug 15, 2000 * Aug 16, 2000 * Oct 02, 2000 * Oct 19, 2000 * Oct 22, 2000 * Oct 25, 2000 * Nov 21, 2000 * Dec 07, 2000 *	Jan 21, 2001 * Feb 01, 2001 * Feb 03, 2001 * Feb 04, 2001 * Feb 24, 2001 * Feb 26, 2001 * Mar 01, 2001 * Mar 02, 2001 * Apr 01, 2001 * Apr 04, 2001 * Apr 09, 2001 * Apr 18, 2001 * Jun 04, 2001 * Jun 08, 2001 * Aug 03, 2001 * Oct 18, 2001 * Oct 18, 2001 * Nov 18, 2001 * Dec 13, 2001 *	Feb 08, 2002 * May 23, 2002 * Jun 02, 2002 * Jun 05, 2002 * Aug 02, 2002 * Aug 03, 2002 * Sep 13, 2002 * Sep 22, 2002 * Sep 25, 2002 * Sep 27, 2002 * Sep 29, 2002 * Oct 01, 2002 * Oct 10, 2002 * Oct 16, 2002 * Nov 13, 2002 * Nov 24, 2002 * Nov 29, 2002 * Nov 30, 2002 *	Feb 03, 2003 * Feb 04, 2003 * Feb 10, 2003 * Feb 13, 2003 * Feb 16, 2003 * Mar 22, 2003 * Mar 26, 2003 * Apr 04, 2003 * Apr 09, 2003 * Apr 19, 2003 * May 01, 2003 * Jun 05, 2003 * Jun 05, 2003 * Jun 10, 2003 * Jun 12, 2003 * Jun 20, 2003 * Jun 24, 2003 * Jul 29, 2003 * Jul 30, 2003 * Aug 01, 2003 * Aug 07, 2003 * Aug 08, 2003 * Sep 26, 2003 *	Jan 29, 2004 * Feb 16, 2004 * Mar 04, 2004 * Mar 17, 2004 * Mar 28, 2004 * May 19, 2004 * May 26, 2004 * Jun 06, 2004 * Jun 10, 2004 * Jun 11, 2004 * Jun 16, 2004 * Jun 19, 2004 * Jun 27, 2004 * Jun 29, 2004 * Jul 04, 2004 * Jul 10, 2004 * Jul 18, 2004 * Jul 21, 2004 * Jul 25, 2004 * Aug 11, 2004 * Aug 15, 2004 * Sep 23, 2004 * Sep 25, 2004 *	Jan 24, 2005 * Feb 05, 2005 * Feb 06, 2005 * Feb 06, 2005 * Feb 13, 2005 * Feb 21, 2005 * Feb 25, 2005 * Mar 02, 2005 * Mar 03, 2005 * Mar 06, 2005 * Mar 06, 2005 * Mar 09, 2005 * Mar 21, 2005 *	

Domain-Based Web Archives



Topical Web Archives

CyberCemetery
University of North Texas Libraries

[CyberCemetery Home](#)
[Search the CyberCemetery](#)
[Browse the CyberCemetery](#)
[Related Resources](#)
[Contact Us](#)
[Digital Collections of the Government Documents Department](#)

Browse the CyberCemetery: Agencies by Name
[A](#) | [B](#) | [C](#) | [D](#) | [E](#) | [F](#) | [G](#) | [H](#) | [I](#) | [J](#) | [K](#) | [L](#) | [M](#) | [N](#) | [O](#) | [P](#) | [Q](#) | [R](#) | [S](#) | [T](#) | [U](#) | [V](#) | [W](#) | [X](#) | [Y](#) | [Z](#)

UCLA LIBRARY

Digital Collections

Questions

A

[Acc](#)

[Adv](#)

[Adv](#)

[Adv](#)

[Adv](#)

[Bro](#)

[Ant](#)

UCLA LIBRARY | Digital Collections
Online Campaign Literature Archive

[Search the Archive](#)

[Browse the Archive](#)

[About the Archive](#)

UCLA Online Campaign Literature Archive
a century of Los Angeles elections


Every American election produces thousands of campaign flyers, pamphlets, posters, and bumper stickers, generally called "campaign literature." These documents provide an important record of the campaign, its participants, issues, and tactics. Despite this value, the small size, short production period, and irregular distribution of the documents, all outside the bounds of the traditional publishing industry, put most campaign literature beyond the scope of standard library collections. These materials are seldom saved for posterity.

In order to fill this gap, since the 1920s the government documents section of the UCLA Library has built and maintained a Campaign Literature Collection, containing printed ephemeral election materials distributed by campaigns for local, state, and federal offices and ballot measures affecting the Los Angeles area. In response to changes in election campaigning brought about by the internet, in 1998 the Library also began to capture and preserve copies of campaign websites.

The UCLA Online Campaign Literature Archive presents a subset of the

C

[Col](#)



UC LIBRARIES

Digital Preservation Program

Event-Based Web Archives



Hurricanes Katrina & Rita Web Archive

Keyword Search

About this Web Archive

Internet Archive and many individual contributors created a [comprehensive list](#) of websites documenting the historic devastation and massive relief effort due to Hurricane Katrina. The sites were crawled between the dates of September 4 - October 17th. This collection, containing more than 25 million searchable documents, will be preserved by Internet Archive with access to historians, researchers, scholars and the general public.

Querying works generally as it does in Google with some caveats. See [how querying works](#) in the Katrina search for discussion.

News Sites



[The Times-Picayune](#)

Relief Sites



[USA Freedom Corps](#)

Complete List of Crawled URLs

The Library of Congress >> More Online Collections

MINERVA

Mapping the Internet Electronic Resources Virtual Archive

[home](#) >> [about: collection overview](#)

Election 2000 Web Archive

SEARCH

ABOUT

[Overview](#) | [Selection Criteria](#) | [Metadata](#) | [Technical Architecture](#) | [Copyright](#) | [FAQs](#) | [Partners](#)

Overview

The *Election 2000: an Internet Library* is a selective collection of nearly 800 sites archived daily between August 1, 2000 and January 21, 2001, covering the United States national election period.

The Library of Congress commissioned the [Internet Archive](#) to create the Election 2000: an Internet Library. The Internet Archive provided project coordination and quality assurance. Compaq Computer Corporation captured and stored all the digital materials collected. The Internet Archive currently hosts the collection and provides access to the archive using its Wayback Machine technology. The Election 2000: an Internet Library was made available to the public in June of 2000.





Sites	Collections	Rights	Search	
-----------------------	-----------------------------	------------------------	------------------------	--

See [What's New: WAS Release 2](#) for details about new features. See [Capturing Sites: Overview](#) for a brief guide to the capture process.

Add Site

Create entries for sites you plan to capture.
Provide seed URLs, capture settings and descriptions.

Manage Sites

Capture, edit or delete existing sites.
Review site capture history & site details.

View Results

Search and browse archived content.
Review capture job reports.



Web Archiving Lingo

- Crawler
- Host
- Site
- Seed
- Capture
- Robots.txt



Sites	Collections	Rights	Search	
-----------------------	-----------------------------	------------------------	------------------------	--

New Site:

[Capture Settings](#)[Descriptive Data](#)[Rights Data](#)* Site Name:

* Seed URLs:

Scope: Max Time: [Cancel](#)[Save \(all tabs\)](#)

Tips: Add Site

Site Name: What you name a site is up to you. A site names does not need to match the HTML <title> text of its pages.

URLs: In some cases a web site is delivered from multiple servers. When you create a site entry, you can enter as many server URLs as needed to define the site.

You may learn more about a site's architecture if you choose a 10 minute, host +1 capture setting. When you review this initial test, you will be able to see if there are additional hosts you need to include as seed URLs. After analyzing these results, you can edit the capture settings as needed and capture the site again.

Detailed Guides

- [Selecting Sites](#)
- [Adding New Sites](#)
- [Default Settings](#)



Overview Search Files Reports

General Info

Description:
Capture Job Status: Finished
Start Time: Apr 05, 2007 - 10:02:04 AM
Finish Time: Apr 05, 2007 - 12:05:19 PM
Duration: 2h 54s
Size: 41.1 MB
Documents: 387

Add this capture job to collection:

California Water Sites

Robot Exclusions

Robots.txt files

The following robots.txt files were discovered during your capture. The robots.txt files are archived in order to document the host server policies on the date this capture job was run. Click the link to view the archived version of the robots.txt file.

- <http://www.lacity.org/robots.txt>
- <http://download.macromedia.com/robots.txt>

See [Robots Help](#) for more information.

Mimetypes

Below are the most common mimetypes for this capture job. See the [Mimetype Report](#) for complete statistics.

Mimetype	URLs
application/pdf	282
text/html	66

Results: United Nations Population Fund (Apr 09, 2007 - 10:10:23 AM)

Overview

Search

Files

Reports

You can search the files retrieved in this capture job by keyword. To browse by seed URL or file type, use the **Files** tab.

When reviewing search results, click on a title, URL or thumbnail to display the item. Click on **Show Detailed Record** to review all item details. The title and abstract are derived automatically by the search engine.

Search By Keyword: video

Search

501 search result(s) found for jpg < 1-10 of 501 >



Title: <http://video.unfpa.org/adver/53616883881755063252006.jpg>

24167 image/jpeg

Captured: Mon Apr 09 10:14:21 -0700 2007

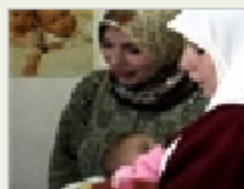
URL: <http://video.unfpa.org/adver/53616883881755063252006.jpg>

Overview

Search

Files

Reports



Title: <http://video.unfpa.org/adver/53616883881755063252006.jpg>

6179 image/jpeg

Captured: Mon Apr 09 10:14:21 -0700 2007

URL: <http://video.unfpa.org/adver/53616883881755063252006.jpg>



Title: <http://video.unfpa.org/adver/53616883881755063252006.jpg>

5324 image/jpeg

Captured: Mon Apr 09 10:14:21 -0700 2007

URL: <http://video.unfpa.org/adver/53616883881755063252006.jpg>

Below are the files retrieved during this capture job. You can limit the list by mimetype and URL.

application/pdf



Limit

21 file(s) found < 1-20 of 21 >

URL	Mimetype	Size
http://www.ci.san-luis-obispo.ca.us/finance/download/pers...	application/pdf	
http://www.ci.san-luis-obispo.ca.us/publicworks/download/...	application/pdf	
http://www.ci.san-luis-obispo.ca.us/publicworks/transport...	application/pdf	
http://www.ci.san-luis-obispo.ca.us/publicworks/documents...	application/pdf	
http://www.ci.san-luis-obispo.ca.us/publicworks/documents...	application/pdf	
http://www.ci.san-luis-obispo.ca.us/publicworks/download/...	application/pdf	



Sample Collection Plan

- **Section 1.** **Mission & Scope**
- **Section 2.** **Selection**
- **Section 3.** **Acquisition**
- **Section 4.** **Descriptive Metadata**
- **Section 5.** **Rights and Access**
- **Section 6.** **Maintenance and Weeding**
- **Section 7.** **Preservation**

- **Appendix A.** **Letter of Agreement**
- **Appendix B.** **Seed List**
- **Appendix C.** **Metadata**

Flexibility in the face of uncertainty

What metadata will you need?

Title:
Subject:
Author:
Date:

Rights Management Approaches

- **Library of Congress**

- Extensive rights management efforts
- Permission secured for any site not clearly in the public domain
 - If no response, the site is not captured

- **Internet Archive**

- Opt-out policy
- Obey robots.txt

- **WAS**

- Flexibility

Preservation

- Content preserved in the DPR
 - Bit preservation (fixity, integrity)
 - Replication
 - Desiccation
- Massive storage requirements
 - Multiple projects investigating mass storage environments

WAS: Now & into the Future

- Current Status
 - in development
 - 12/07 roll out to current curators
- Beyond 2007
 - Extending service to additional curators
 - Developing end user access
 - Exploring release of open access tools

Acknowledgements

- Tracy Seneca, Web Archiving Coordinator
 - CDL WAS development team
- Kathleen Murray
 - UNT Partners
- NDIIPP