



## Archiving Web-published Materials

### Issues & Challenges

Texas Library Association Annual Conference  
San Antonio, Texas  
April 13, 2007  
Kathleen Murray

## Topics



1. Web-at-Risk Project
2. Needs Assessment
3. Key Findings
4. Web Collection Development
5. Web Archiving Service (WAS)

2

## Web-at-Risk Project

- Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) grant
  - Focus: To capture, curate and preserve government and political information from the web
- Partners
  - California Digital Library (CDL)
  - New York University (NYU)
  - University of North Texas (UNT)
- CDL: Development of a Web Archiving Service (WAS)
  - Enable curators to build collections of web-published, at-risk government & political information
- UNT: Needs Assessment and WAS Evaluation






3

## Project Curators

- Project Curators
  - 22 librarians from 13 institutions
  - Government information specialists
  - Range of web archiving experience
- Needs Assessment
  - Survey of curators
  - Focus groups with librarians
  - Interviews with content providers and researchers
- Web Collection Plans
- Evaluation of WAS Releases
  - Survey evaluation by curators
  - Usability testing with curators


Arizona State Library  
New York University  
Stanford  
University of California  
UC Berkeley  
UC Davis  
UC Irvine  
UCLA  
UC Riverside  
UC Santa Barbara  
UC San Diego  
UC San Francisco  
UC Santa Cruz  
University of North Texas

4

## Needs Assessment Activities


Survey: 22 Participants  
6 Collaborations



**Curators**


National: 2 APA - FDLC  
17 Participants

Partners: 3 UNT - CDL - NYU  
26 Participants




**Librarians & Archivists**



Content Producers: 3 Labor Unions  
4 State Gov't. Agencies



**Content Producers & End Users**

End Users: 7 Academic Researchers








5

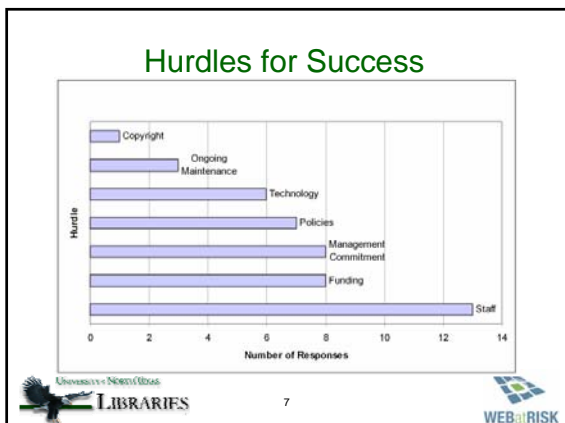
## Findings of the Needs Assessment

*"The things we're talking about are basically the things we've always done with the print collection. But I think they're just much harder with web-archived material."* - Librarian

- Web Archiving Hurdles for Success
- Organizational Roles & Responsibilities
- Transitional Times for Librarians
- Preservation Stewardship and Publishing Anarchy

6



### Roles & Responsibilities

*"IT needs to understand that archiving is not the same as a backup and that preservation goes beyond the 3 months backup copies are retained."* - Librarian

- The necessity of working together
  - Curatorial expertise
  - IT expertise
- The uncertainty of stewardship
  - Publishers
    - Large publishers ought to preserve their publications
    - Small publishers are unable to preserve their publications
  - Government Agencies
    - Regional and local government entities need help: Leadership, direction, & expertise
    - State government agencies
      - Stewardship unclear or non-existent

LIBRARIES 8 WEB@RISK

### Stewardship in State Government

Agencies

"It's frequently true that you call the department and no one seems to sort of be in charge of a publication." - Librarian

One state agency published its annual county-level statistical report on the web in 1998 for the first time. The next year, the agency replaced the 1998 report with the 1999 report. "That has pretty much become our standard bad example." - Librarian at a State Library

Publishers

"I remember asking the publishers, 'Are you printing all the versions?' They replied, 'What versions?'" - Librarian

LIBRARIES 9 WEB@RISK

### Stewardship in State Government

Content Providers

"Our main concern is that the integrity of what is archived be maintained and kept as current as possible and that there is communication between archive and source group to ensure integrity. [We are also concerned] that no one would have access to the back side of the data and possibly change it. Integrity of access is a major concern to ensure that the average person as well as the scholar would have access." - Agency

Researcher

"State legislatures don't usually archive their own materials from the Web. They just replace last session's materials in favor of this session's. You can't get at committee assignments from 1999 to 2004." - Researcher

LIBRARIES 10 WEB@RISK

### Transitional Times

- Stewardship unclear or non-existent
- Collection development models transfer at great expense in resources
  - Expensive to select
  - Expensive to harvest
  - Expensive to create metadata
- Preservation practices are not readily available
- Consortial efforts are not yet established
- Existing policies & practices lack scope

LIBRARIES 11 WEB@RISK

### Adapting in Transitional Times

*"I have been known to archive web publications by printing them out and having them bound in buckram and then cataloged."* - Librarian

- "Doing what we can do"
  - Print archives
  - CD-ROM archives
  - Preservation archives
- Collaborative efforts have begun
  - State libraries
  - Universities
- Policies & practices for web collections are being formulated

LIBRARIES 12 WEB@RISK

### Collaboration

**Question**  
Is some organization already archiving these materials in a manner that meets the needs of my user groups?

**Benefits**

- Expand access to materials
- Eliminate redundancy of effort
- Control preservation costs

### External Partnerships

**Motivation**  
Web materials are disappearing and universities have both a self-serving and an altruistic interest in preserving them.

*"A Community of Creators"*

**Benefits**

- Preserve historical record in areas of interest
- Fulfill mission to serve community
- Foster a sustainable business model

### Internal Partnerships

**Motivation**  
The history and intellectual products of the institution are being lost and the library cannot preserve it alone.

**Benefits**

- Preserve historical record in areas of interest
- Fulfill mission to serve community
- Foster a sustainable business model

### Web Collection Development Considerations

- Selection
- Intellectual Property
- Capture
- Content v. Content
- Authenticity

### Perspectives on Selection

The Value of Content is in the Eye of the User

- Librarians: Discipline-Related Web Content
  - Relative newness of a discipline
  - Demand for current information
  - Cultural & political studies
- Researchers: Key Content Genres
  - Journals, periodicals, databases
  - Government records or documents
  - Newspapers
- Content Providers: Related Web Content
  - National labor union & local affiliates
  - State government agency & federal counterpart

### Intellectual Property

- Concerns about federal government publications
  - GPO repositioning itself as a vendor or supplier
  - Licensing agreements with distribution strings attached will become more common
- State government agencies
  - General commitment to open access
  - Exception: Copyrighted materials need permission
- Content providers not amenable to ceding intellectual property rights to an archive provider
 

One-half of surveyed curators were unsure if permission would need to be obtained to collect their targeted materials.

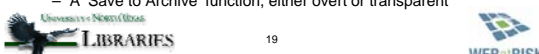
## Selection & Capture

Preservation Begins at Creation: Extending the Deposit Model

**Problem**  
Selection of web sites consumes an inordinate amount of time and *push button preservation* is not on the horizon.

**Solution**  
Deposit Models for Authors, Creators, and Publishers

- Newspapers: "It's the only way!"
  - Publisher provides the content and the metadata
  - Archive preserves the newspapers
- University Members:
  - Mandatory deposit policy for faculty
  - Mandatory project or research funding requirement that a preservation process is documented and executed
  - A 'Save to Archive' function, either overt or transparent




## Selection

*"When you're close to a subject, like 'progressive social movements', you realize how much variety there is and if we're collecting 100 websites, do we collect 50 about terrorism or do we collect a representative sample of the variety of the whole?" - Archivist*

- Materials representing range of topics in an area
- Materials limited to one or more topics
- What is important to preserve?
  - The inertia that follows asking this question


*"If I were making an archive I'd put all those association websites in and the data websites, but I might pick up a blog here and a rant there and put them in. I assume that even if nobody's going to use it today, somebody might want to use it in the future." - Librarian*



## Content v. Context

Related Decisions: Unit of Selection & Unit of Description

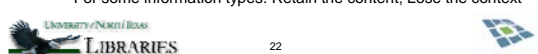
- Depends on the discipline
  - Social Science v. History v. Anthropology
- Depends on the research purpose
  - Comparison of images within ads over time
  - Comparison of role of images in publications over time
- The website becomes 'evidence' for those who study it



## Capture

*"If there's something on the Internet that's critically important to my research, I capture it." - Researcher*

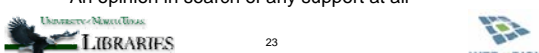
- Authenticity
  - How modifications to websites within the archive are handled
  - Researchers want statements of provenance and lots of contextual tagging for any alteration.
- Frequency of capture
  - Highly dependent on web site and will vary considerably
  - Very critical variable
- Versions and formats
  - In general capture it all: You never know what someone will need in the future
  - For some information types: Retain the content; Lose the context






## Authenticity

*"I would want the archive from an institution that I have faith and confidence in; if it's done in the university or the federal government, that would satisfy me." - Researcher*


- Trusted sources
- Where is the original?
- Certification of authenticity
- Variance by user groups & by discipline
  - Law publications: Print source citations
  - History-Social Science: Researcher age factor
  - An opinion in search of any support at all



## Needs Assessment Activities

Survey: 22 Participants 6 Collaborations	 <b>Curators</b>	Survey & Test Crawl Findings
Crawls: 18 Participants 3 Collaborations		
National: 2 APA - FDLC 17 Participants	 <b>Librarians &amp; Archivists</b>	Focus Group Findings
Partners: 3 UNT - CDL - NYU 26 Participants		
Content Producers: 3 Labor Unions 4 State Gov't. Agencies	 <b>Content Producers &amp; End Users</b>	Interview Findings
End Users: 7 Academic Researchers		

R  
E  
Q  
U  
I  
R  
E  
M  
E  
N  
T  
S



## Web Collection Development

PHASES		
SELECTION	CURATION	PRESERVATION
Selection	Description	Preservation
Acquisition	Organization	
	Presentation	
	Maintenance	
	Deselection	

25

## Web Collection Plans

- Arizona State Agencies Web Publications**  
Richard Pearce-Moses; Arizona State Library, Archives and Public Records
- The Tamiment Library at New York University**  
Michael Nash and Peter Filardo; New York University
- Islamic and Middle Eastern Political Web**  
John A. Ellis; Stanford University
- International Government Organizations and Developing Countries**  
James A. Church; University of California, Berkeley
- AFL-CIO/Change to Win Web Archive**  
Terence K. Howe; University of California, Berkeley
- California Political Blogs and Interest Group Websites**  
Nick Robinson; University of California, Berkeley
- UCLA Online Campaign Literature Archive**  
Gabriella Gray and Scott Martin; University of California, Los Angeles
- CyberCemetery**  
Valerie Glenn and Starr Hoffman; University of North Texas

26

## Web Collection Plans: Collaboration

- Local Government and Local Area Flood Control Collection**  
Marcia Meister and Juri Stratford; University of California, Davis
- UCI Orange County Government Information Web Collection**  
Yvonne Wilson; University of California, Irvine
- UCLA NGO & Local Government Information**  
Kris Kasianovitz; University of California, Los Angeles
- UCR Inland Empire (CA) Web Archive**  
Lynne Reasoner and Kenneth Furuta; University of California, Riverside
- UCSB Santa Barbara, Ventura, and San Luis Obispo Counties Local Planning Documents and Water Archive**  
Sherry DeDecker and Janet Martorana; University of California, Santa Barbara
- Monterey Bay Area Local and Regional Government Websites**  
Lucia Orlando; University of California, Santa Cruz
- UCSD Local Government Information Archive**  
Megan Dregler; University of California, San Diego

27

## Web Archiving Service

**8 Stage Development Schedule**

Requirements

1

2

3

4

5

6

7

8

**Release 1: Basic Capture**

**Release 2: Search & Display**

**Release 3: Analysis & Reports**

**Release 4: Collection Building**

**Release 5: Rights Management**

**Release 6: Event Capture**

**Release 7: Preservation**

**Release 8: Enhancements**

28

## Summary

**Collection Development Framework** (1/05 - 4/06)

- Needs Assessment Toolkit (2/05 - 6/05)
- Data Collection: Survey, Focus Groups, Interviews (6/05 - 12/05)
- Data Analysis & Reports (12/05 - 4/06)

↓

WAS Requirements

**Web Collection Planning Guidelines & Template** (5/06 - 8/06)

**Web Collection Plans** (9/06 - 2/07)

↓

WAS R1 Trial & Evaluation (5/06 - 8/06)

↓

WAS R2 Trial & Evaluation (9/06 - 2/07)

29

## More Information

Web Site: <http://web3.unt.edu/webatrisk/>

Wiki: <http://wiki.cdlib.org/WebAtRisk/>

Kathleen Murray    krmurray@unt.edu

Thank You

30

## Archiving Web-published Materials: Issues and Challenges

Kathleen Murray  
 Post Doctoral Research Associate  
 University of North Texas Libraries  
 P.O. Box 305190  
 Denton, TX 76203-5190  
 Email: krmurray@unt.edu

### Web Collection Development

Every area within traditional collection development is impacted when web-published materials are included in a collection. Table 1 identifies some considerations for each area. Additionally, Table 1 depicts the importance of setting policies to support and guide web collection development.

Table 1.  
*Considerations for Web Collection Development*

Policy Setting	Political mandates, organizational mission, financial parameters, & technical capabilities.	
	Selection	Factors: Focus of the collection, unit of selection, web boundaries, copyright obligations, and authenticity of materials.
	Acquisition	Requirements for crawling tools: Global or selective capture.
	Description	Baseline metadata: Machine-generated Enriched metadata: Specific to an organization; both human-generated & machine-generated metadata.
	Organization	Considerations: Retain or modify the organizational structure of the materials as they existed on the web.
	Presentation	Considerations: Mirror the web at the time of their capture or selectively present (searching & browsing).
	Maintenance	Functions: Training, hardware and software maintenance, performance optimization, backups, upgrades, & duplicate detection.
	Deselection	Reasons: Duplication, errors, legal or social considerations.
	Preservation	Challenges: Persistent naming, format migration and/or emulation, inventory management, volatility, replication, re-validation, & storage.

### Guidelines for Web Collection Plans

The findings from the needs assessment informed a set of guidelines for web collection development plans. The Web-at-Risk project curators used these guidelines and a companion template to create web collection plans. Web collection plans provide guidance for managing collections of web-published materials created for specific groups of users. A web collection typically consists of a group of web-sites related by a common subject, theme, or event.

Some concepts and practices from collection planning for print materials easily transfer to collection planning for web-published materials while some new concepts and unfamiliar practices are introduced. To effectively manage collections of web-published materials, it is good practice to either create new plans or modify existing plans to address these concepts and practices. Table 2 is the outline for the collection plans created by the project curators.

Table 2.  
*Outline for Web Collection Plans*

<b>Section 1. Mission &amp; Scope</b>	<b>Section 5. Presentation &amp; Access</b>
A. Mission Statement	A. Discovery
B. User Group(s)	B. Access
C. Collection Subject, Theme, or Event	C. Look-and-Feel
D. Curator(s)	D. Dynamic Content
<b>Section 2. Selection</b>	E. Multiple Types/Formats
A. Seed List	F. Authenticity
B. Capture Scope	<b>Section 6. Maintenance &amp; Weeding</b>
C. Rights Management	A. Maintenance Activities
<b>Section 3. Acquisition</b>	B. Deselection Guidelines
A. Frequency of Capture	C. Collection Evaluation
B. Capture Scope	<b>Section 7. Preservation</b>
C. Material Types & Formats	A. Technology Obsolescence
B. Interactive & Dynamic Content	C. Preservation Metadata
<b>Section 4. Descriptive Metadata</b>	<b>Appendices</b>
A. Level of description	A. Submission Agreements
B. Metadata elements	B. Web Archiving Service Agreement
C. Controlled vocabularies	C. Collaboration Agreements

*Note:* The web collection plan guidelines, template, and resulting collection plans are available at: <http://web3.unt.edu/webatrisk/cpg.php>.

### Related Web Sites

Digital Projects Unit at the University of North Libraries  
<http://www.library.unt.edu/digitalprojects/>

National Digital Information Infrastructure and Preservation Program (NDIIPP)  
<http://www.digitalpreservation.gov/>

UNT Web Collection: Cyber Cemetery  
<http://digital.library.unt.edu/browse/department/govdocs/cybercemetery/>

Web-at-Risk Assessment Activities  
<http://web3.unt.edu/webatrisk/>

Web-at-Risk Project Wiki  
<http://wiki.cdlib.org/WebAtRisk/tiki-index.php>