### Topics

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

### 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**

### 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

### Needs Assessment Activities

- **Survey:** 22 Participants
- **Crawls:** 18 Participants
- **National:** APA - FDLC
  - 17 Participants
- **Partners:** UNT - CDL - NYU
  - 26 Participants
- **Content Producers:** 3 Labor Unions
  - 4 State Govt. Agencies
- **End Users:** 7 Academic Researchers

### 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**
Hurdles for Success

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

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

Stewardship in State Government

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

“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

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

Stewardship in State Government

“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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**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

**Perspectives on Selection**

- **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

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

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
• 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

Needs Assessment Activities

Survey: 22 Participants
6 Collaborations
Crawls: 18 Participants
3 Collaborations
National: 2 APA - FDLC
17 Participants
Partners: 3 UNT - CDL - NYU
26 Participants
Content Producers: 3 Labor Unions
4 State Gov’t. Agencies
End Users: 7 Academic Researchers
Curators
Survey & Test Case Findings
Librarians & Archivists
Focus Group Findings
Interview Findings
Requirements
Content Producers & End Users
Web Collection Development

<table>
<thead>
<tr>
<th>PHASES</th>
<th>SELECTION</th>
<th>CURATION</th>
<th>PRESERVATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selection</td>
<td>Description</td>
<td>Preservation</td>
</tr>
<tr>
<td></td>
<td>Acquisition</td>
<td>Organization</td>
<td>Presentation</td>
</tr>
<tr>
<td></td>
<td>Maintenance</td>
<td>Deselection</td>
<td></td>
</tr>
</tbody>
</table>

Web Collection Plans

Arizona State Agencies Web Publications
- National Parks Service
- Arizona State Library, Archives and Public Records
- The Tamiment Library at New York University

Islamic and Middle Eastern Political Web
- John A. King, University of California, Irvine
- International Government Organizations and Developing Countries

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 Bilbao; University of California, Irvine
- UCLA NGO & Local Government Information
- Rini Basavanhody; University of California, Los Angeles
- UCR Inland Empire (CA) Web Archive
- Lyne-Raeenar and Kenneth-Funde; University of California, Riverside
- UCSB Santa Barbara, Ventura, and San Luis Obispo Counties Local Planning Documents and Water Archive
- Sherry DeCoskey and Janet Maravilla; 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 O’Reilly; University of California, San Diego

Web Archiving Service

8 Stage Development Schedule

- Requirements
  - 1: Basic Capture
  - 2: Search & Display
  - 3: Analysis & Reports
  - 4: Collection Building
  - 5: Rights Management
  - 6: Event Capture
  - 7: Preservation
  - 8: Enhancements

Summary

Collection Development Framework
- Needs Assessment Toolkit
- Data Collection
  - Survey
  - Focus Groups
  - Interviews
- Data Analysis & Reports
- WAS Requirements

Web Collection Planning Guidelines & Templates
- WAS RT Plan
- WAS RT Trial & Evaluation

More Information

Web Site: http://web3.unt.edu/webatrisk/
Wiki: http://wiki.cdlib.org/WebAtRisk/

Kathleen Murray krmurray@unt.edu

Thank You
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: kmurray@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

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

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

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

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