University of North Texas Libraries

New York University Libraries

Web-at-Risk Project

A Distributed Approach to Preserving our Nation's Political Cultural Heritage

Kathleen Murray – University of North Texas

Tracy Seneca – California Digital Library

Fall Federal Depository Library Conference &
Depository Library Council Meeting

October 2005

California Digital Library

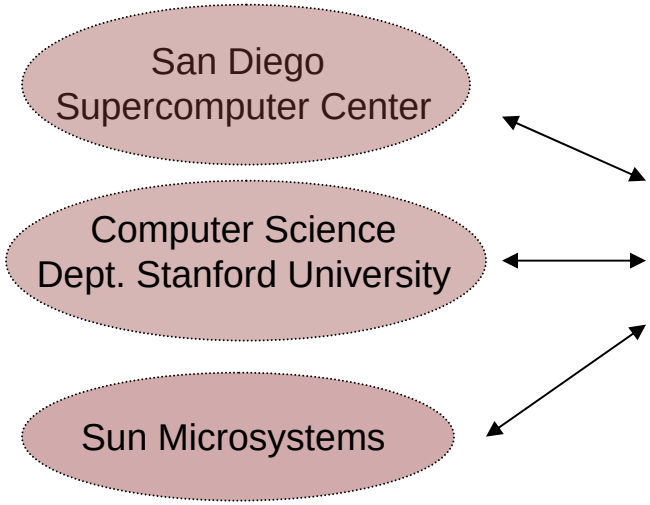
CDL

Contents

Web-at-Risk Goal

Web Archiving Service (WAS):

Develop web archiving tools that will be used by libraries to capture, curate, and preserve collections of web-based government and political information



The Library of Congress

Martha Anderson, Agreement Officer
Technical Representative

California Digital Library

(Lead Partner)

Dan Greenstein, P.I.
Peter Brantley, Director of Technology
Lynne Cameron, Manager, Ingest Group
Patricia Cruse, Director
John Kunze, Technical Lead
Shifra Raffel, Web Archive Programmer
Tracy Seneca, Web Archive Coordinator
Stu Sugarman, Arch. and Insfrast. Manager

University of North Texas

Cathy Hartman, Fellow, Texas Center for Digital Knowledge
Kathleen Murray, Assessment Analyst

New York University

David Ackerman, Executive Director of eServices
Rasan Rasch, Web Archive Programmer

Selection and Curator Agents

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

Four overlapping paths of activity:

- Content Identification and Selection CISA)
- Content Ingest, Retention and Transfer (CIRT)
- Content Harvest and Analysis (CHA)
- Partnership Building (PB)

Content Identification, Selection, & Acquisition

- Needs assessment: surveys, focus groups and interviews
  - Curator (librarian) needs
  - End user needs
- Analysis of test crawls based on curators' content
- Case studies, best practices, and collection development guidelines

Content Ingest, Retention, and Transfer

- Modify CDL's Digital Preservation Repository
- Data model for Web Archive Digital Objects (WADO)
- Development and testing of remote replication strategies

Content Harvest and Analysis

- Tools to conduct and support curators' web crawling:
  Curator User Interface
  - Crawler
  - Collection Management
  - Reports
  - Export/Import Handler
- File formats for web crawling: WARC

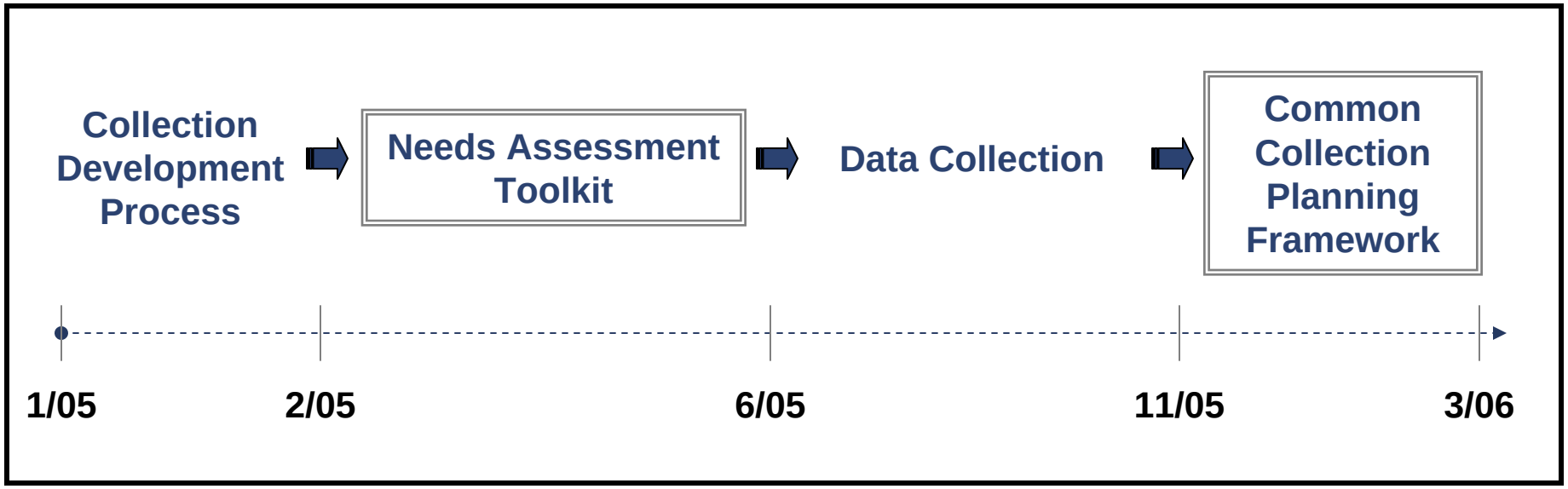
Partnership Building

- Develop model agreements (partner MOUs)
- Assess costs of sustaining a distributed approach to web archiving

CISA Path: Needs Assessment

I. Activities
II. Status
III. Future Activities

Needs Assessment Activities



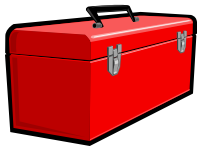
Collection Development

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.

Needs Assessment Toolkit

- Responsibilities
- Timeframes
- Participants
- Guidelines
- Consent Forms
- Instruments
  - Survey Instrument
  - Focus Group Guide
  - End User Interview Guide
  - Content Provider Interview Guide

Available:



http://web2.unt.edu/webatrisk

Data Collection Activities: 2005

		Jun	Jul	Aug	Sep	Oct
1	Survey	X	X			
2	Focus Group: National	X				X
3	Focus Group: Local			X	X	
4	Interviews: End Users				X	X
5	Interviews: Content Providers				X	X

Survey Status

- Purpose
  - End Users' Needs
  - Curators' Needs
  - Functional Requirements
    - Web Crawler Tool
- Online Survey
  - 5 Sections
  - 59 Questions

- Respondents (N = 16)
  - Curatorial Partners
    - University of CA
    - New York University
    - University of North TX
    - Stanford University
    - AZ State Library

Focus Group Status

PARTICIPANTS	LOCATION	N = 44
Special & Academic Libraries	Chicago, IL	8
University of North Texas Libraries	Denton, TX	7
University of California & Stanford University Libraries	Oakland, CA	10
New York University Libraries	New York, NY	8
Public, State, & Academic Libraries	Washington, DC	11

Interview Status

INTERVIEWS	PARTICIPANTS	PURPOSE
Face-to-Face	Researchers (N = 9–15) – Faculty – Students	• End User Access Needs
Telephone	Web Publishers (N = 6-9) – Government Agencies • State, Regional, & Local – Political & Labor Organizations	• Needs • Concerns

CISA: Future Activities

Collection Development Overview ➡

1. Needs Assessment Toolkit
- Survey
- Focus Groups
- Interviews

6/05

➡ **Assessment Data Collection** ➡

Data Analysis and Demographic Analysis ➡

2. Common Collection Planning Framework
- Guidelines
- Tools

3/06

➡

3. Collection Plans
- User needs
- Policies
- Goals
- Practices, etc.

7/06

➡

Case Study Methodology and Template ➡

5. Case Studies
- Processes
- Practices
- Goals

6/07

➡

6. Final Report

Assessment:
- Goal achievement
- Tools
- Costs & benefits

Identification:
- Best practices

9/07

4. Survey of Archive Implementation Experience
- Cost & benefits
- Needs assessment tools
- Technical tools
- Best practices

6/07

➡ **Survey Data Collection** ➡

Project Status

- Test crawls being conducted
  - Government sites provided by curators
  - Hurricane Katrina
- Rights management
  - Web-at-Risk rights management protocol

Rights Management

- 3 rights management schemes
  - A: consent assumed
  - B: consent sought
  - C: consent required
- Tools for recording rights contact activity

24 Sample Sites for Test Crawling

- 10 are .gov
  - 3 federal
  - 3 local govt
  - 4 state agencies
- 3 are .com
  - 1 of these is for a county agency
- 11 are .org
  - One .org site contains warning that material may not be reproduced in any format without permission.

Rights Oddities

- 3 different curators selected sites that are associations of local governments. Two of those (San Diego and Sacramento) are .org sites, and one (Southern California) is a .gov site.

- http://www.ocsd.com City agency with a .com address that states that all of its material is in the public domain.

- http://www.city.davis.ca.us: A city site that aggressively asserts copyright and demands permission for any reproduction.

Rights Classification

- In 5 cases, I changed the original classification of the site after I read the copyright warnings/info on the site itself.

- I determined that if I were curating these sites, I would classify their rights as follows:
  - Scheme A: Consent implied: 6 sites
  - Scheme B: Consent sought: 14 sites
  - Scheme C: Consent required: 4 sites

Crawling Tools

- Heritrix
  - Contributing code back
- WERA for display

More info:

http://wiki.cdlib.org/WebAtRisk