At Risk: Capturing and Preserving Web Resources

Special Libraries Association

June 12, 2006

Cathy Nelson Hartman, University of North Texas

Kathleen Murray, University of North Texas

The Web Environment

- Brings significant changes and challenges for Collection Development
- Struggle to meet service demands of students/faculty/staff
- Realign resources to meet new demands and challenges

Traditional Collection Development Processes

- Select
- Acquire
- Describe
- Organize
- Present for access
- Maintain
- Deselect
- Preserve

Web Environment

- Same processes are followed
- New challenges encountered in every process

University of North Texas Libraries

Print on Paper

Selection of Materials

- Establish Authority
  - Peer Review Process
  - Reputation of Publisher
  - Reputation of Author

WEBatRISK

Print on Paper

Selection (cont.)

- Established processes for selection
  - Approval plans
  - Depository Library item selections
  - Vendors / Distributors

University of North Texas Libraries

Web Environment

Selection of Materials

- What will users/researchers value?
- How will authority of Web published materials be established?
- How will we locate/identify Web published materials?

WEBatRISK

Print on Paper

Acquisition of Materials

- Publication/item begins and ends with cover of book or other container
- Established practices for acquiring material facilitated by vendors
- Authenticity/stability

UNIVERSITY of NORTH TEXAS LIBRARIES

Web Environment

Acquisition of Materials

- Where does a publication/item begin and end – Unit of Selection/Acquisition?
  - Links?
  - Databases?
  - Multi-media presentations?
- Can we ensure accuracy of captured material?

WEBatRISK

University of North Texas Libraries

Print on Paper

Describe and Organize

MARC Records – Cooperative Cataloging
(subject/name authority)

Version control

- Changes/updates made
- Item republished as new edition
- Tracked through cataloging/metadata
- Shelved side-by-side on book shelves

Web Environment

Describe and Organize

Metadata – automatic generation?

Less precise Subject control, name authority, etc.

Version control

- Frequent changes made
- Discovery of changed publications/records
- Define a new edition – how many changes required?
- Track changes in metadata
- Serendipitous discovery by browsing less likely

Print on Paper

Presentation of Materials to Users

Usability established

- Centuries of perfecting print on paper

Access rights management

- User community defined
- Presence required in the library
- One person at a time could use an item
- Copyright laws restricted photocopying

Web Environment

Presentation of Materials to Users

Access rights management

- Licensed products with restrictions
  - Specific user group
  - Many simultaneous users
- Determine access rights for captured material

Usability issues

Print on Paper

Maintain and Deselect

- Shelve, repair, provide reference services
- Weed collections

UNIVERSITY of NORTH TEXAS LIBRARIES

Web Environment

Maintain and Deselect

- Storage
- Refresh, migrate
- Provide reference services
- Weed collections

WEBatRISK

Print on Paper

Preservation

- Well developed strategies in place
- Circulating Collections
  - Repairs as needed
  - Acid paper issues
  - Shelf life 75 + years
- Curation of rare materials

Web Environment

Preservation

- Should begin with creation of publication
- Constantly changing technology environments, both hardware and software
- How many copies are needed?
- Standards development – metadata, text mark-up
- Shelf life - ??

Web Environment Issue

Privacy

- Access greatly increased with full-text searching
- Access now from any computer with Internet connection
- Safety issues
- Hackers
- User tracking

Looking for Solutions

National Digital Information Infrastructure and Preservation Program (NDIIPP)

- Sponsored research project –

  The Web-at-Risk

- Partners include:

  California Digital Library, University of North Texas, and New York University

Web-at-Risk Project

Four Research Tracks:

- Development
- Experimental
- Assessment (UNT Lead)
- Partnership Building

Data Collection

- Focus Groups
  - ALA Chicago, Depository Library Council
  - Librarians - UC System, UNT, NYU
- Interviews
  - Data owners
  - Researchers
- Curator Surveys

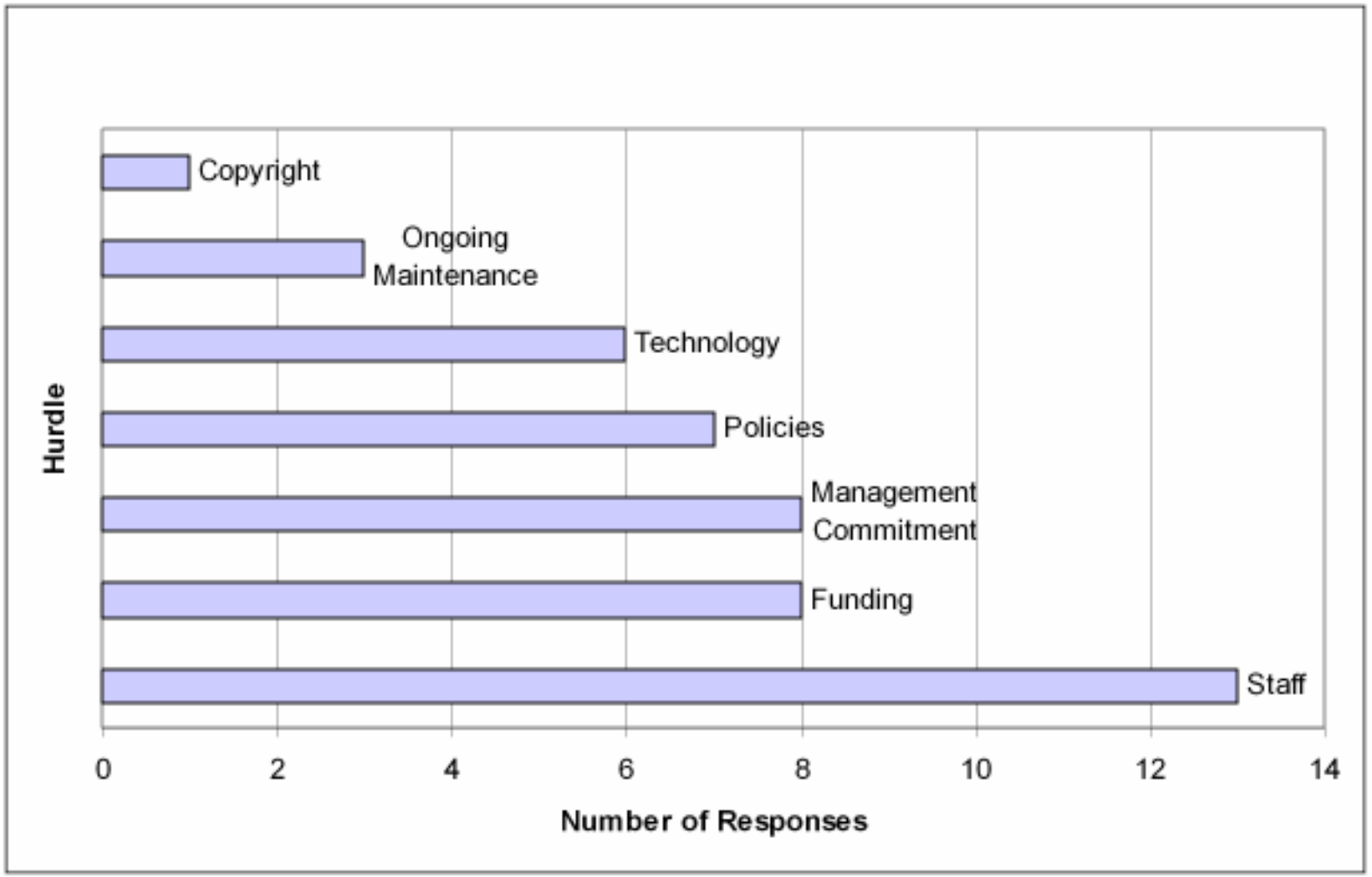
What we learned……

Findings Supporting the Need for a Web Archiving Service

"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

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

"I think that everybody thinks that somebody else is collecting them." - Librarian

Publishers

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

Stewardship in State Government

State Libraries

"Our statute still reads that if it exists in print, we get it. But if print doesn't exist then there's a whole other set of . . ." – Librarian at a State Library

"I called the State Library and asked, 'Are you guys archiving this digital version?' And they said, 'Well, we're planning to but we haven't got to it yet." - 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

"Our State Library is 'at-risk'!" – Librarian

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

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

At "this university, the role of the library is very much under pressure. We're trying to prove it in because the campus plan doesn't see a need for a library in ten years." - Librarian

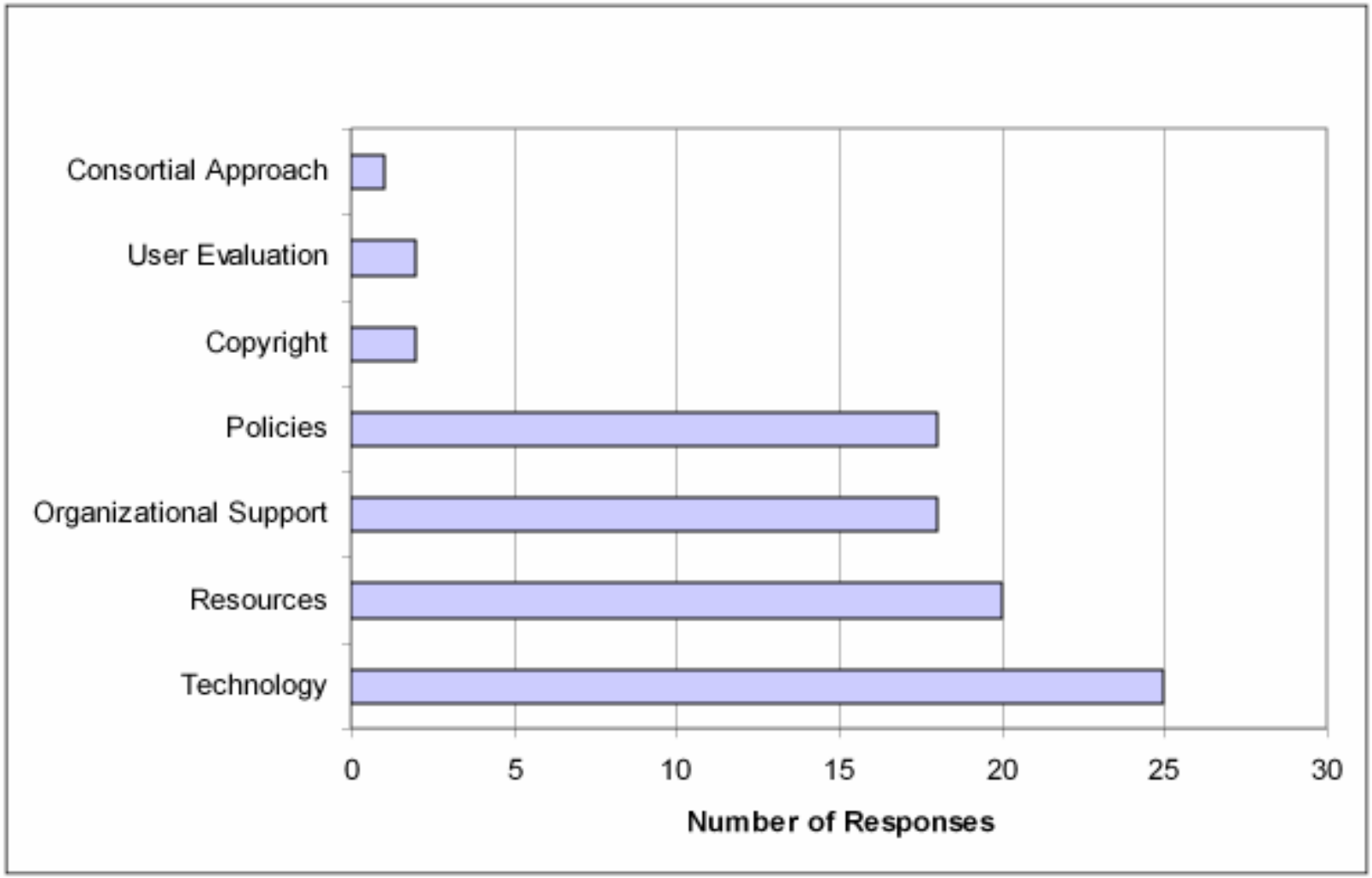
Adapting in Transitional Times

"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

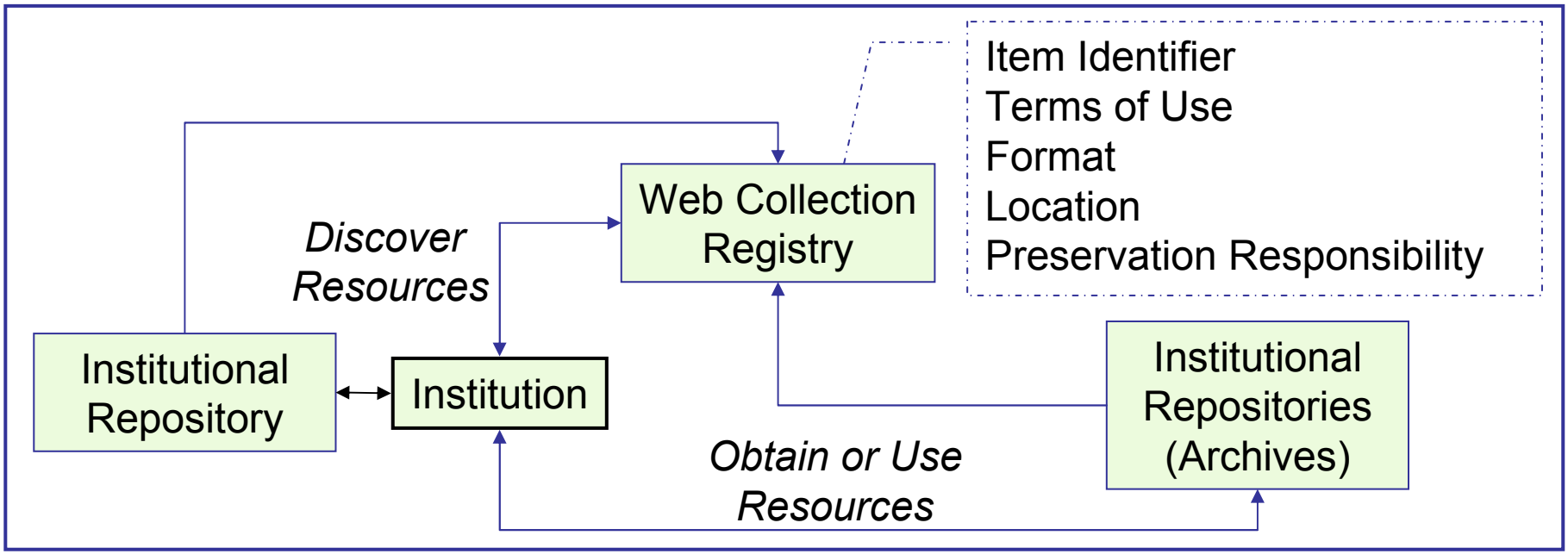
"When we're just going out and harvesting websites we are gong to be less familiar with the individual pieces of information than we were when were able to actually handle each document. Losing that control doesn't necessarily have to be a bad thing. It's just really difficult to make that transition." - Librarian

Success Factors



Collaboration

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

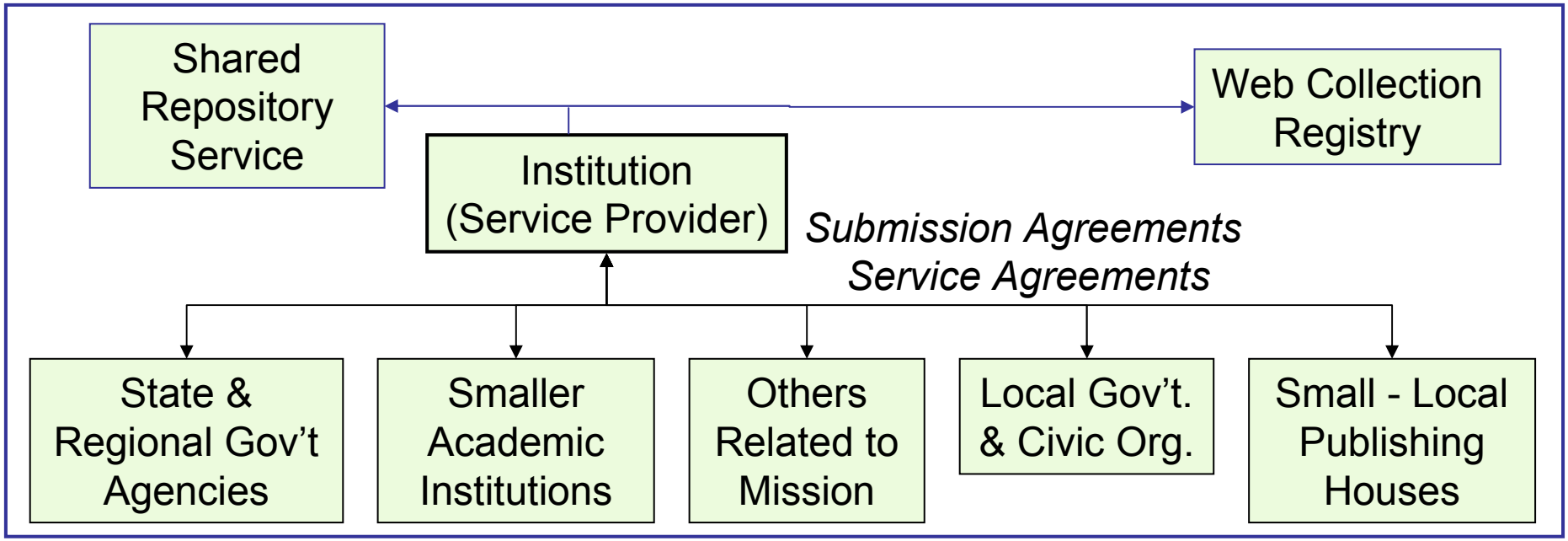


Benefits

- 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.



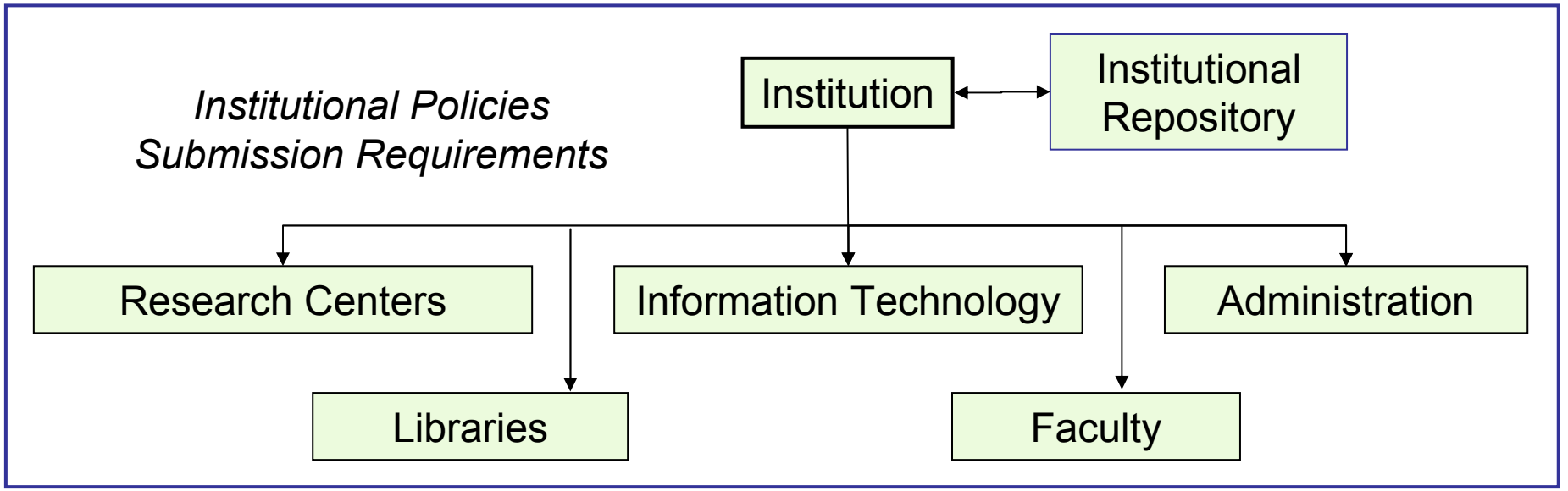
Benefits

"A Community of Creators"

- 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

University of North Texas Libraries

Collection Development Findings

- Selection
- Capture
- Metadata
- Presentation
- Weeding
- Preservation

WEBatRISK

Perspectives on What to Archive

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 the 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 Conflict: Unit of Selection

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

"I think the hardest thing to deal with as an historian with any publication is knowing about the readers, knowing how widely this [publication] was disseminated, who accessed it. So I think any information that you can archive about people who access these web sites is valuable.

There's clearly an ethical issue -- clearly a legal issue -- which is not my expertise." - Researcher

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

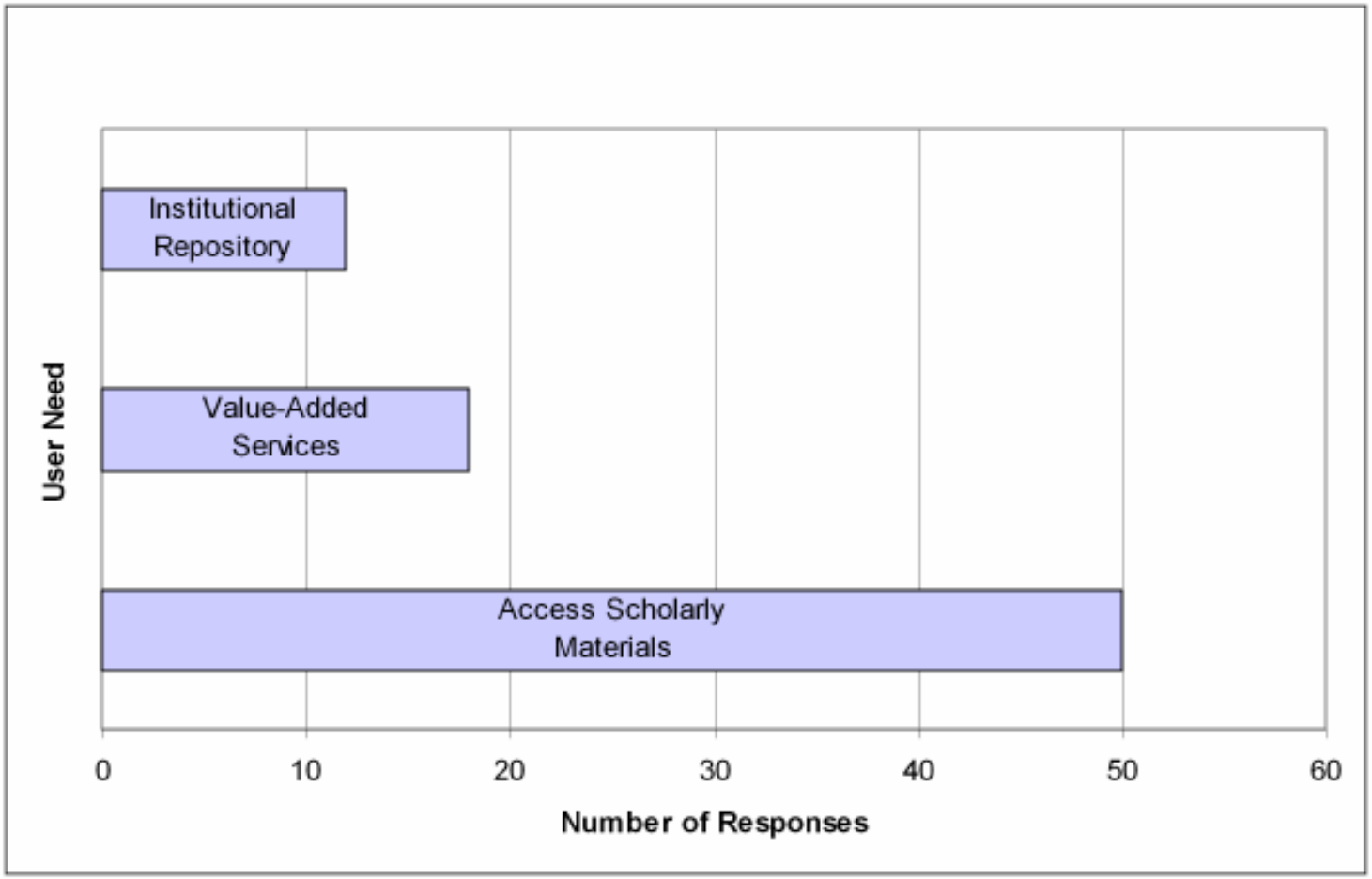
"I would want the archive from an institution that I have faith and confidence in; if it's done in the university or the federal government, that would satisfy me." - Researcher

- 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

"It behooves the person who's looking at it to see: Where are the authors or publishers coming from? What's their bias? Can I trust this or not?" - Librarian

Value of Web Archive to Users

"Talking Points for Discussions with Administrators"



More Information

Web Site: http://web2.unt.edu/webatrisk/

Select “Reports” to see reports of all data collection activities. Collection Planning Guidelines will soon be posted.

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

Kathleen Murray, Assessment Analyst, UNT
krmurray@unt.edu