IMLS Award Number LG-06-07-0040-07

**Classification of the End-of-Term Archive:**
**Extending Collection Development Practices**
**to Web Archives**

# FOCUS GROUP REPORT:
# METRICS for WEB ARCHIVES

JULY 2010

Kathleen Murray
kathleen.murray@unt.edu

University of North Texas
UNT Libraries
1155 Union Circle #305190
Denton, TX 76203-5017

# Contents

Introduction ..... 1
Project Overview ..... 1
Summary of Findings ..... 1
Key Findings ..... 2
Methodology ..... 4
Objective ..... 4
Participants ..... 4
Data Collection ..... 5
Data Analysis ..... 5
Findings ..... 5
Digital Content ..... 5
Acquisition ..... 8
Metrics ..... 11
Closing ..... 15
Appendix A Participant Questionnaire ..... 16
Appendix B. Discussion Guide ..... 17

# Introduction

## Project Overview

The Classification of the End-of-Term Archive (EOT Archive) project builds on a previous project conducted collaboratively by the Library of Congress, the US Government Printing Office, the Internet Archive, the University of North Texas (UNT) Libraries, and the California Digital Library. That project[1] captured the entirety of the federal government's public Web presence before and after the 2009 change in presidential administrations. The result is an approximately 25 terabyte web archive of government information that is replicated in repositories at the collaborating organizations, including UNT.

As web archives become more available and accessible, many libraries will be collecting materials from these important information repositories. Librarians will need the capability to identify and select materials in accord with collection development policies. Additionally, libraries will need to characterize these materials using common metrics; however, such metrics are not established for web archives, making it difficult for librarians to communicate the scope and value of these materials to administrators.

The *Classification of the End-of-Term Web Archive* project will utilize the EOT Archive to investigate innovative solutions to address these needs. Research will be conducted concurrently in two work areas:

1. *EOT Archive Classification* The materials in the EOT Archive will be classified according to the Superintendent of Documents (SuDocs) Classification Numbering System. Classifying government information in accordance with SuDocs will allow government information librarians to use existing collection development policies to select materials from the Archive.

2. *Web Archive Metrics* A set of metrics for materials in Web archives will be identified. These will enable characterization of materials in Web archives in units of measurement more familiar to libraries and their administrations.

## Summary of Findings

Providers of web archives are an emerging class of content suppliers for which the business cases and service models have yet to be developed. Web archives are essentially repositories of born-digital and digitized resources. In terms of current collection management practices in libraries, web archives are most akin to the broad category of electronic resources and to the specific material types of electronic journals and databases. It is largely current collection management practices and librarian experiences in the electronic resources arena, particularly selection and retention practices, that may best inform the development of models and usage statistics for Web archives.

It is important, however, to consider the characteristics of web archives that resist conformance to current library practices. For example, web archives include resource types of recognized value for future research that are beyond the current scope of electronic resource providers, for example, web

[1] *Library Partnership Preserves End-of-Term Government Web Sites* (2008, August 14); [http://www.loc.gov/today/pr/2008/08-139.html]

sites, blogs, and wikis. While both current selection practices and service provider models can extend to these material types, current statistical reporting requirements fail to do so.

Fundamentally we found that current service and acquisition models for electronic resources provide guidance for the extension of these models to include web archives. However, the demand for ownership by libraries of web archive content, versus networked access to shared content, is not certain and needs to be further understood so that appropriate service models and business models can be developed for web archive providers.

Additionally, there are differential statistics for "ownership" versus "networked access" acquisition models that must be collected and reported by libraries. Once again, current practices provide some guidance for web archive content. However, some content types in web archives lack direct and measurable corollaries to the material types for which library statistics must be reported. The end result is that these materials, which will only increase in importance in library collections, especially research library collections, are generally not represented in the annual survey data reported by libraries.

## Key Findings

### *Collection Management: Selection and Retention*

ACQUISITION MODELS FOR ELECTRONIC RESOURCES

1. Access.
   A library provides discovery services to users, via its catalog or web pages, and network services for resources that are served by another entity. This model applies both to licensed and freely-available materials.
2. Purchase.
   A library acquires materials and typically provides the following services: storage, maintenance, discovery, and access. These services may be provided to users who are associated with the library (e.g., students, faculty, and staff), to other libraries via consortia and consortia-like arrangements, or to web users of any ilk.
3. Production.
   A library digitizes materials and collections or acquires born-digital materials and collections. The same services identified in the purchasing model apply to this model.
4. User-driven.
   For particular materials or collections, such as eBooks, a library follows an access model. Based on usage, a library purchases specific materials.

SELECTION CRITERIA

- Broadness of applicability
  - Scope or breadth of material coverage to serve the "broadest possible group of users"
  - Promotes buy-in from multiple departments
- Usage Data
  - Generally vendor provided
  - Vendor compliance with standards needed
- Number of "titles"
  - A measure of the volume or amount of materials
- Unique Content
  - Number of unique items in the archive, that is, materials not available elsewhere

- Duplicate Content
  - The "titles" (or materials) in the existing collection that are duplicated
- Appropriateness for Collection
  - Particularly in regard to the degree of "completeness" needed for in a particular subject

ACQUISITION BUDGETS FOR ACADEMIC LIBRARIES

- Tend to be long-established
- Difficult to reallocate among departments
- Expenses for electronic resources (i.e., serials and databases) are allocated across departments
- Endowed collections are handled uniquely

METRICS DRIVING ACQUISITIONS

- Retention: Cost per use
- Selection: Usage when available

## *Library Statistics*

REPORTING AUTHORITIES

Academic libraries report statistics to one or more of three organizations:

- IPEDS
  Integrated Postsecondary Education Data Systems (IPEDS); U.S. Department of Education's Institute of Education Sciences' National Center for Education Statistics (NCES)
- ARL
  Association of Research Libraries (ARL)
- ACRL
  Association of College & Research Libraries (ACRL)

CLASSES OF DATA REPORTED

1. Scope (How much; how many)
2. Expenditures (Cost)
3. Usage (Counts)
4. Quality (Outcomes; Value)

RELATIONSHIP TO COLLECTION MANAGEMENT

- Much of the data reported is of limited value for:
  - Analyzing collections
  - Making rational collection management decisions
- Informed acquisition decisions require:
  - Standard data elements for comparable material types
  - For networked electronic resources, counts based on IP addresses for:
    - Specific pages and collections *accessed*
    - Specific files / materials *retrieved*

## *Web Archives*

POSSIBLE SERVICE PROVIDER MODELS

- Access Model
- Purchase Model

STATISTICS REQUIRED

- Scope (How much; how many)
- Usage (Counts)

The next section describes the study methodology. Following it are the detailed findings from the focus group discussion. The closing section describes the next steps the project will take in determining the metrics needed for web archives.

# Methodology

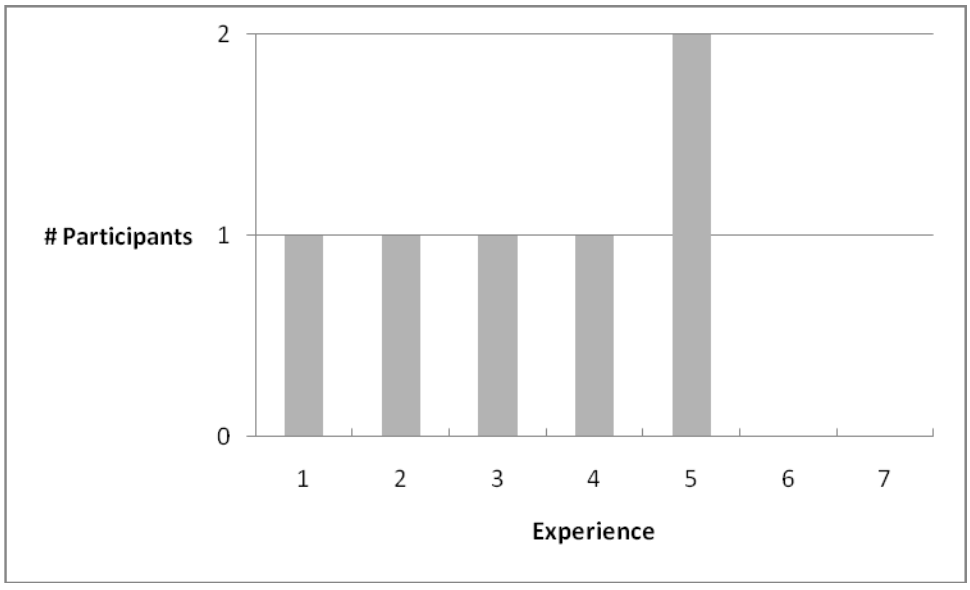
As part of the *Web Archive Metrics* work area, a focus group discussion was conducted at the first meeting of the project's Subject Matter Experts (SMEs) and project staff members. The meeting was held on April 25, 2010 at the Adam's Mark Hotel in Buffalo, New York.[2]

## Objective

The objective of the focus group discussion was to identify the criteria libraries use in making material acquisition decisions, in particular the countable units that play a critical role in these decisions.

## Participants

Participants (*N*=6) were government information librarians from depository libraries who were recruited by the principal investigator to serve as SMEs for the project. Within their academic libraries, these individuals have responsibility for collection development in the area of government information and have an average of 21 years (*range*: 6-37 years) experience doing so. Five participants are responsible for reporting statistics within their libraries for the materials in their collection(s).




Note. Experience Range: 1 = Novice; 7 = Expert

Figure 1. Experience with Web Archives (*N*=6)

[2] This was the site of the spring 2010 Federal Depository Library Meeting, held April 26-28, 2010.

When asked to describe their experience with Web archives, participants reported a range of experience, from novice to just above average (Figure 1). None of the participants characterized themselves as experts in this area.

## Data Collection

The meeting opened with a presentation by project staff members.[3] The presentation included an overview of the End-of-Term Web Archive project and descriptive statistics of the Archive itself. Additionally, the project's objectives, work areas, and key activities were discussed. Following a lunch break the focus group discussion was conducted.

The discussion was facilitated by the project's assessment manager, who explained the purpose of the group discussion and encouraged participants to engage in dialog with one another and to react to the views expressed by other participants. After participants signed forms indicating their consent to participate in the discussion group, the facilitator led the semi-structured discussion. The discussion guide in Appendix B informed the discussion areas. An audio recording was made of the discussion. Additionally, two staff members took notes during the discussion and one project SME, who was unable to attend the meeting, contributed written comments regarding the topics discussed.

At the end of the discussion, participants completed the questionnaire in Appendix A. The questionnaire identified participants' demographic characteristics, the types of digital content they select for their collections, and their experience with web archives.

## Data Analysis

Descriptive statistics characterized the responses to the participant questionnaire and the results are included in the Introduction and Finding sections of this document. The audio recording was transcribed and content analysis of the transcription, as well as the written comments from one SME, identified the key findings. In order to best represent the group discussion, the findings were refined and augmented by the summary notes recorded by staff members.

# Findings

The findings are grouped into three major areas. These areas are: Digital Content, Acquisition, and Metrics.

## Digital Content

### *Collection management*

TREND TOWARDS ALL DIGITAL COLLECTIONS

*"We're moving towards managing all digital collections."*

*"We've started to trust the vendors - that they're not going to drop that title."*

[3] Presentation available on the project website:
http://research.library.unt.edu/eotcd/w/images/1/12/SME_Meeting_20100425_Buffalo_NY.pdf

For many types of digital content in their collections, libraries offer networked access through catalog entries and library web pages, in lieu of owning and hosting the content locally. There is a strong, if not almost exclusive, preference for providing networked access to digital serials versus providing physical access to print formats. Over the last few decades, the practice of collection management for serials has progressed from only acquiring print-formatted materials, to a mixture of acquiring print formats and digital formats, to licensing digital formats and providing networked access to them. To a lesser extent, this trend extends to monographs.

INCREASED TRUST

*"We trust you. You hold it for us."*

Trust is a critical component between libraries and digital content providers. Digital content providers include trusted vendors of journals and databases, trusted digital archives (e.g., JSTOR[4]), and trusted consortia (e.g., CIC[5]). All of these content providers manage and provide access to their digital content on behalf of their customers and members.

HOSTING AND ACCESS

*"We're digitizing anything we can get our hands on."*

Some libraries either acquire and/or create born-digital and digitized content. Some preserve this content and provide access to others. This content includes:

- Born-digital topographical maps
- Born-digital state government information
- Data sets in support of faculty research
- Digitized materials from the library's collections:
    - map collections
    - photographic collections
    - poster collections
    - government information from the legacy collections
- Born-digital university publications (e.g., blogs)

PERSPECTIVES ON PRINT MATERIALS

*"We also hear increasingly from our students that they don't like paper."*

_____

*". . . the more digital the better . . ."*

Some researchers require access to older works that are not available in digital formats, whereas for some faculty the accessible digital materials are sufficient for certain subjects. Some senior faculty members have a preference for printed materials, both serials and primary source materials, and some

[4] JSTOR is a service of ITHAKA, "a not-for-profit organization helping the academic community use digital technologies to preserve the scholarly record". [http://www.jstor.org/]

[5] The Committee on Institutional Cooperation (CIC) is "a consortium of the Big Ten universities plus the University of Chicago . . . [serve] the common good by sharing expertise, leveraging campus resources, and collaborating on innovative programs." [http://www.cic.net]

faculty require students to access and reference print materials, even if the library can provide access to the same materials in digital formats. Students increasingly express a preference for digital formats.

From libraries' perspectives, maintaining duplicate formats is often no longer an economic option. Collaborative management of print collections among libraries, for example, across a state or within a multi-campus system, decreases duplication of print materials, frees up space, saves money, and, in some cases, results in an increase in the digital and print content available through inter-library loan services.

SPACE UTILIZATION

*"Space in general is a premium."*

Space is often constrained in libraries and economic utilization of space is a primary consideration. Digital resources are considered economical in terms of physical space and storage costs. Conversely, space not utilized for physical materials can be used for other purposes, such as study areas in academic libraries and DVD collections or coffee shops in public libraries.

## *Web-published Content*

TYPES OF CONTENT

Participants completed the questionnaire in Appendix A at the end of the group discussion. Table 1 summarizes the types of digital content participants select for the collection(s) they manage. All participants collect three types of digital content:

1. web-published reports
2. agency/organizational websites
3. statistical databases

| Q7. Please indicate the types of digital content that you select for the collection(s) you manage: | **# Yes** | **# No** |
|---|---|---|
| web-published reports | 6 | 0 |
| agency/organizational websites | 6 | 0 |
| statistical databases | 6 | 0 |
| video recordings | 3 | 3 |
| audio recordings | 2 | 4 |
| blogs | 1 | 5 |
| wikis | 1 | 5 |

Table 1. Digital Content Types Selected for Collections (*N*=6)

About half select video and audio recordings. Only one participant selects blogs and wikis for their collection(s). Participants added four digital content types to the list of materials they select: maps, web published materials, images, and serials.

### *Issues*

SELECTION

*"One of the challenges we're having is finding new sources of digital content."*

Identifying sources of new digital content is a problem. Subject selectors are often unaware of possible sources of digital content. Some primarily select from known sources, to the exclusion of materials that might be available, for example, from societies, universities, and government agencies. When using Google to identify new materials, materials that have not been indexed by Google will not be discovered. In terms of selecting books for their collection, one library is considering a patron-driven approach that involves establishing buying thresholds for new e-books. A purchase would be made when the number of times a book is selected by patrons exceeds the threshold.

LEVEL OF ACCESS

Access costs for digital materials from some content providers correspond to the feature richness of the access provided. For example, limiting user access to "viewing" an eBook is less costly that allowing users access to the textual and statistical content. However, such a limitation may not meet users' needs.

## Acquisition

### *Models*

Four acquisition models were identified. These are:

1. Access. A library provides discovery services to users, via its card catalog or web pages, and network services for resources that are served by another entity. This model applies both to licensed and freely-available materials.
2. Purchase. A library acquires materials and typically provides the following services: storage, maintenance, discovery, and access. These services may be provided to users who are associated with the library (e.g., students, faculty, and staff), to other libraries via consortia and consortia-like arrangements, or to web users of any ilk.
3. Production. A library digitizes materials and collections or acquires born-digital materials and collections. The same services identified in the purchasing model apply to this model.
4. User-driven. For particular materials or collections, such as eBooks, a library follows an access model. Based on usage, a library purchases specific materials.

ACCESS

The trend seen in the acquisition of serials, that is, away from ownership of the physical materials and towards providing access to digital materials, may predicate how other classes of digital materials are "added to collections" in the future. Two types of arrangements characterize current access models: license agreements (e.g., agreements with vendors and aggregators[6] for serials and eBooks) and institutional collaborations (e.g., inter-institutional collaborations, such as the Center for Research

[6] "A type of vendor that hosts content from multiple publishers, delivers content direct to customers and is paid for this service by customers." [The COUNTER Code of Practice, Journals & Databases, Release 3, August 2008: http://www.projectcounter.org/r3/Release3D9.pdf]

Libraries[7], and collaborations among multiple libraries within a university system in which one institution may "own" or "license" materials and provide access to other libraries). One participant suggested that future collection policies might explicitly identify materials that a library provides access to, in addition to those materials the library holds and owns.

PURCHASE
Some libraries acquire born-digital materials that are: (a) essential to their collection and (b) no longer produced in an analog format. Other libraries purchase collections in order to provide users and researchers the flexibility to re-purpose the materials, for example to manipulate the data or extract the text contained in materials.

PRODUCTION
In addition to accessing and purchasing digital content, some libraries are producing digital collections comprised of materials the institution owns or creates. The following content categories for collections were identified:

- Digitized formats of analog collections
- University publications
- Data sets in support of faculty research

USER-DRIVEN
As content aggregators and library system vendors offer new content discovery and delivery products a new user-driven acquisition model may emerge. The new product features promise: (a) discovery of content across a spectrum of content providers, including the library via its catalog and digital repositories, (b) delivery of content, including licensed journal articles, and (c) standardized usage data across content providers. These products enable a user-driven acquisitions model, in which users are given access to articles in journals, including those a library does not license, for a set number of instances and the "loan" cost is charged to the library. Based on user demand for particular materials or titles, a library can modify its licenses from "loan" to "license" or "purchase" for those titles that reach an economical threshold of demand.

## *Issues: Serial Acquisition*

Content aggregators, or vendors who package content from multiple content providers, are driven by their agreements with those providers. This stands in contrast to their being driven by libraries' requirements, in particular in two areas: serial selection and usage statistics.

JOURNAL AND DATABASE SELECTION

*"It makes us look good for ARL status."*

In practice, selection of an aggregator involves identifying the unique content offered, the number of titles included, and the broadness of the contents' applicability to meet research needs across university departments. In order to acquire specific serial titles, entire packages must be acquired. In one case, moving from print selection of serials to electronic selection generated an estimated 50-75% increase in

[7] "An international consortium of university, college, and independent research libraries" that acquires, preserves, and provides access to "newspapers, journals, documents, archives, and other traditional and digital resources from a global network of sources." [http://www.crl.edu/]

the libraries' number of serial titles. The implication is that "more is not necessarily better". However, libraries have had little choice in the matter.

VENDOR PRICING / CONTRACTS

*"It used to be you paid a premium for the electronic version. Now, you're paying a premium for the print."*

In one case, duplicate electronic titles are retained in order to gain price discounts on print versions of journals from the publisher. A few people reported they receive both microfilm and electronic versions of some journals because it is the more economical choice, despite the fact that they do not want the microfilm. Another person reported that vendors use pricing to influence format choice, for example, by setting very high prices on print journals versus electronic formats of the same journals.

NEW ACQUISITIONS

*"You're really stepping into the unknown, partially, with the actual acquisition."*

To get a good price for new acquisitions, it is helpful to be an early adopter. However, the downside is that usage data is unknown and it takes time for new acquisitions to become known and used.

## *Benefits*

ALTERNATIVE CONTENT PROVIDERS

As libraries produce digital collections and make them publicly available, other libraries can provide access to the materials. This is highly desirable.

DISCOVERY SERVICES

Providing discovery services can be an attractive and economical option compared to purchasing and/or holding materials. Discovery services involve making materials "discoverable" (i.e., visible or findable) within a library catalog or database or web page. Library system vendors and content aggregators include discovery services in their products.

## *Acquisition Budgets*

Acquisitions budgets at universities tend to be political and classist and difficult to reallocate. One person reported that at two very different universities (one state-funded and another with large endowments), budget allocations among departments for acquisitions were a legacy from almost 30 years ago and were quite difficult to reapportion in a manner that better matched current needs. Some collections are individually endowed and acquisitions continue to be made without regard to comparative use of the collection.

## *Acquisition and Web Archives*

AGGREGATOR ACQUISITION MODEL

Perhaps the criteria libraries currently employ in selecting an aggregator's products are applicable to selecting a web archive. These criteria include:

- Broadness of Applicability:
    - Scope or breadth of material coverage to serve the "broadest possible group of users"
    - Promotes buy-in from multiple departments
- Number of "titles"
- Unique Content:

    - Number of unique items in the archive, i.e., materials not available elsewhere
- Duplicate Content:
    - The "titles" (or materials) in the existing collection that are duplicated in the archive

FOCUSED SUBJECT AREA

Criteria for acquisition of materials from a web archive in a particular subject area might include:

- Broadness of coverage
- The degree of "completeness" that is appropriate or needed for this subject at a particular university
    - Assessing completeness in regard to government information is difficult; there are generally not comparisons among content providers that can be made in this regard

ACQUISITION MODEL

*"We could look at the biennial survey from depository libraries . . . there was a question: If there was digital content available would you pull it down and put it on your own server? And about thirty percent of the libraries, out of the twelve hundred that responded, said "yes they would"."*

Thirty-seven percent of respondents to the last biennial survey of depository libraries (416 libraries) indicated they would download content to their own servers. However, there was general consensus in the group that:

- Many libraries would be content to primarily or exclusively access a trusted web archive, such as the EOT Archive at UNT.
- The scope a library would be interested in downloading varied from quite narrow to very broad.

One person suggested their library would only download and host specialized materials that the library or its users were interested in repurposing for local use. Another person thought there would be "mixed bag" of interest among libraries regarding owning (or downloading) or simply accessing materials in web archives.

DISCOVERY TOOLS

*"If we don't acquire, discovery becomes all important."*

Libraries create or augment discovery tools to enable their patrons to find materials in remote archives ad repositories. Typically, the tools are catalog entries in the OPAC or web-based resource pages.

## Metrics

### *Statistics Gathered*

REQUIRED STATISTICAL REPORTING

*"We are driven by our masters. Much of what we collect, we don't see the end to [it]: We don't know why they want it; we don't understand what it's for. It's probably a relic from the past and they just haven't gotten around to changing it. But there it sits and we continue to collect [it]."*

Libraries collect the data that they are required to report: to the government (i.e., IPEDS[8]); to library associations of which they are members (e.g., ARL[9] or ACRL[10]); and to their administrations. These data are not always useful to the library for analyzing collections or for making rational collection management decisions.

USAGE STATISTICS

*"Because it's the most important title in the field doesn't mean that an institution such as mine should purchase it if it's not being used."*

*"Usage is really about retention because you don't know what usage is before you buy it."*

Usage statistics are heavily used in making retention decisions for both print materials (circulation data) and electronic resources (usage data). Reliable data is a must!

DIGITAL MONOGRAPHS

Some statistics are provided for eBooks but for other electronic books, the statistics are not adequate.

DIGITAL REPOSITORIES & ARCHIVES

*"If you want to choose the best resources for your users and some of those resources are free, then usage statistics should be part of your consideration."*

Libraries and other content providers who make their repositories and archives freely accessible do not provide other libraries with usage data, for example data from the providers' server logs. Obviously, the providers themselves are knowledgeable of archive and repository usage and access data from their own server logs.

HYPERLINKS IN LIBRARY CATALOG

One library collects statistics regarding users' link selections within the library's catalog to measure users' differential interest in the content of Federal government agencies; however, the data collected does not generally extend to specific titles. One person added that most libraries do not have the specific titles of publications in their Federal Depository Library (FDL) collection in their library catalogs.

REPORTS FROM GPO

Libraries that do include FDL titles in their catalog, and register their domain(s) with the Government Printing Office (GPO), can get statistics for their domain(s) regarding: (a) the number of times users clicked through the library's catalog to the GPO PURL server and (b) the number of items downloaded from the server to the library's domain. Although libraries do not pay for these materials, gathering the

[8] Integrated Postsecondary Education Data Systems (IPEDS) "is a system of interrelated surveys conducted annually by the U.S. Department of Education's Institute of Education Sciences' National Center for Education Statistics (NCES). IPEDS gathers information from every college, university, and technical and vocational institution that participates in the federal student financial aid programs. [http://nces.ed.gov/ipeds/]

[9] Association of Research Libraries (ARL) [http://www.arl.org/]

[10] Association of College & Research Libraries (ACRL) is the largest division of the American Library Association. [http://www.ala.org/ala/mgrps/divs/acrl/index.cfm]

usage data provides a more complete usage picture of their collection and helps in justifying personnel resources engaged in collection management.

## *Common metrics*

COST PER USE

*"Cost per use is very important."*

Cost per use is a basic metric of great importance to libraries for all types of materials they manage. Historically, this metric has been calculated for print books from circulation data. It is the metric that libraries would like to gather for other material types, including materials from freely available archives and repositories.

COUNTER-COMPLIANT STATISTICS FOR JOURNALS, DATABASES, BOOKS, AND REFERENCE WORKS[11]
Obtaining more accurate, comparable, COUNTER-compliant usage data for journals, databases, and eBooks is one promise proffered by vendors as a feature (or add-on) to their electronic resource management systems. One person noted that these systems are an additional expense but may be worth the investment. Many vendors indicate they are or will be COUNTER-compliant.

## *Issues*

ARL STATISTICS FOR SERIAL COUNTS

*"Electronic [formats are] important for counting titles."*

When a long run of a print serial "dies", ARL guidelines specify that it is no longer be counted as a title. However, when an electronic serial "changes its name", both the old and new titles are counted.

STANDARDIZATION OF SERIAL USAGE STATISTICS
Serial database vendors do not report usage statistics for electronic resources in a comparable manner and are often not compliant with COUNTER statistics. This results in libraries being unable to accurately aggregate or compare usage data from multiple vendors. The National Information Standards Organization's (NISO) SUSHI[12] standard is expected to alleviate this issue and may drive vendors to become COUNTER-compliant.

USAGE-DRIVEN RETENTION OF DATABASES
A library's decision to retain a serial publication is greatly influenced by its usage. As usage decreases, retention may be difficult to justify. This decision-making practice can be disconcerting, particularly given the unreliability of some vendor-measured usage data. Mitigating this practice, one participant indicated that their subject librarians do have the discretion to retain journals and databases with low cost-per-use.

[11] Counting Online Usage of Networked Electronic Resources (COUNTER) "is an international initiative serving librarians, publishers and intermediaries by setting standards that facilitate the recording and reporting of online usage statistics in a consistent, credible and compatible way." [http://www.projectcounter.org/index.html]

[12] Standardized Usage Statistics Harvesting Initiative (SUSHI) is a "protocol standard (ANSI/NISO Z39.93-2007) [that] defines an automated request and response model for the harvesting of electronic resource usage data utilizing a Web services framework. It is intended to replace the time-consuming user-mediated collection of usage data reports. [http://www.niso.org/workrooms/sushi/]

INFLATION OF USAGE STATISTICS BY VENDORS

*"There aren't enough people on our campus to have used [that journal] that many times, so [perhaps] they have a massive security breach, which after two years they still haven't recognized."*

Serial publishers and other content providers are interested in their materials being used. Content aggregators are correspondingly interested in demonstrating to publishers and content providers that the aggregators' products promote discovery and use of materials. This can result in vendors inflating the usage statistics they report.

MEASURING RESOURCE DISCOVERY

*"I mean, if you pull a book off the shelf and look at the table of contents and decide it's not what you're looking for and you put it back, it wouldn't count as a usage; but, there's no way to [make] that distinction [for] online [materials]."*

Statistics for electronic resources measure "discovery". Importantly, this does not indicate whether a resource actually satisfied a user's information need. Likewise, resource discovery can be influenced by page layouts that draw users' attention to certain collections or materials, which will influence usage. So, there are questions and cautions regarding how meaningful vendors' usage statistics truly are.

## *Web Archives*

USAGE DATA

*"We don't clearly understand the differences between what searches are telling us and what retrievals are telling us."*

Libraries would be interested in knowing how many of their users, based on IP addresses, are accessing specific pages and collections in a web archive as well as the specific materials they are retrieving. Usage data at this level of detail could inform:

- A better understanding of the content of interest to users
- Decisions to acquire the collection and make it more discoverable on their own servers.

SELECTION CRITERIA

Libraries would like to know how much of the content in a web archive is unique, that is, content that is not available elsewhere. Specific selection criteria for digital content includes: a supplier's commitment to long term preservation of digital content; contingencies for access to digital content that the library purchased, but accessed remotely, should (a) a supplier cease operations or (b) the library cease to pay for ongoing maintenance. Additionally, the search interface should be easy to use.

UNKNOWN MEASUREMENTS

*The EOT Archive is "a unique site on the transition of the Obama Administration. It may not be used for four years or eight years. Then, all of the sudden, everybody in political science wants into that database and then, everybody who has a political science department wants every one of those files downloaded into their system."*

By its nature an "archive" will have historical content. It may include near-past history, but it will not be active web content. It is expected that the content, with its snapshots in time, will be of interest to researchers, and hence libraries, over time. The type of data libraries will need from the archive provider is not known as yet and one person wondered if it might be largely subjective.

# Closing

Future project work will involve an assessment of the degree to which libraries anticipate owning and providing local access to materials obtained from web archives. This will inform the development of service models, which in turn will inform the statistics needed from web archives and web archive providers.

Future work will also involve an assessment of the current statistics reported by libraries. Those findings will inform recommendations for statistical measures for web archives along two dimensions: scope and usage.

In closing, the project staff expresses thanks to the librarians who contributed their time and expertise as Subject Matter Experts. Their contributions will make the end results of this project more pertinent to libraries and librarians.

# Appendix A Participant Questionnaire

1. What is your gender? _____ Female _____ Male

2. What is your age group? (*check one*)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 21 - 30 | | 41 - 50 | | 51 - 60 | | 71 - 80 | |
| 31 - 40 | | 51 - 60 | | 61 - 70 | | 81 - 90 | |

3. How many years have you been working in the area of government information?

4. What is your current job title?

5. Briefly describe your current job responsibilities:

6. Do you select materials for your library? _____ Yes _____ No

7. Please indicate the types of digital content that you select for the collection(s) you manage:

| | **Yes** | **No** | **Not Applicable** |
|---|---|---|---|
| web-published reports | | | |
| agency/organizational websites | | | |
| statistical databases | | | |
| blogs | | | |
| wikis | | | |
| video recordings | | | |
| audio recordings | | | |
| Please list other material types you select: | | | |

8. Do you report statistics for the materials in the collection(s) you manage? _____ Yes _____ No

9. Circle the number between 1 and 7 that best describes your experience with Web archives?

| Novice | | | | | | Expert |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

10. Your additional comments are welcomed. (*Please use back if more space is needed.*)

# Appendix B. Discussion Guide

Opening: As tools emerge for librarians to discover and select materials from Web archives, libraries will be asked to report measures of the scope and quality of these materials. We are interested in your views regarding the statistics and measurements that might best characterize the materials in Web archives. During the discussion you are encouraged to react to the views of others and engage in dialog with one another.

1. Describe yourself: job responsibilities; experience with collection development
2. In general, how would you characterize the impact of digital content on collection development practices at your library?
    a. What is the range of digital content in your library? In your collection?
    b. How has this changed in the past decade? What future changes do you anticipate?
3. What digital content sources or repositories do you select from for your collection?
    a. How important are web-published materials for the collection(s) you mange?
    b. What classifications for these materials are meaningful to you? (For example: discrete publications, serial reports, organizational websites, agency websites, or brochures)
4. What criteria do you evaluate in making acquisition decisions for various classes of digital content?
    a. How do you identify acquisition costs? How do you budget for them?
    b. Have you encountered any problems or issues? With publishers, archives, administrators, IT, other parties?
    c. What solutions have you implemented to deal with problems and issues?
    d. What additional information or services do you think are needed?
5. What areas in your library do statistics impact: acquisitions, staffing, budgets, performance, accreditation?
6. What statistics are typically reported for the digital materials in your library?
    a. Identify the strength and value of these statistics. Provide examples.
    b. Are any materials are not reported? Provide examples.
    c. What issues do you perceive in regard to these statistics?
7. How is the quality of your library, its staff, services, and collections measured?

a. How are digital materials represented in this measurement?

8. What criteria would you suggest for measuring the scope and quality of different classes of digital materials: web-published reports, agency/organizational websites, statistical databases, blogs, wikis, digital recordings (audio, video), other material types?

    a. What statistical measures do these materials share in common?

    b. How meaningful are these common measurements to report?

Closing: Thank you for taking the time to share your views with us. Please feel free to contact us at any time regarding our project.