Difficulties in Electronic Publication Archival Processing for State Governments

Jackson, Larry S.

(Graduate School of Library and Information Science, University of Illinois,
Urbana-Champaign, Illinois 61820, USA)

E-mail: lsjackso@uiuc.edu

Abstract: Government publications in electronic form exist in a gap between differing expectations. Traditionally government publications have been formal, and deliberately retained for many years. However, materials posted on Websites are often ephemeral. Both government Website staffs and users seem unsure what is to become of an electronic publication after a small number of years. Further, there are many open questions concerning how electronic publications can be economically gathered, processed, retained, retrieved, and utilized, when the retrieval and utilization is projected to occur many years in the future.

We have been deeply involved in the preservation of State of Illinois web materials for 46 months, and in similar work with six other US States for 12 months. This paper reviews the several categories of problems we have encountered in attempting to identify, iteratively gather, retain, index, search, and use these electronic publications. Most broadly, these problems include; incomplete conformance with many kinds of standards or conventions involving Websites, reliance on technologies and formats which are likely to become unavailable, issues affecting disk space consumption in the archives, insufficient quantity and quality of metadata for use in search, philosophical differences in how the contents of a Website ought to be accessed, and incomplete support for document access by a harvesting program.

Shortcomings we experienced in harvester-based, and human-selected acquisitions are discussed. However, the State Libraries in all seven States we work with need to use external harvesters rather than working in a more tightly coupled arrangement involving the webmasters of many agencies. Electronic document archival work, conducted in large part from outside the many small government agencies, seems likely to be a continuing concern.

Keywords: Web archives, Government documents, Electronic publications
doi: **Document code**: **CLC number**:

INTRODUCTION

At the request of the Illinois State Library (ISL) in the spring of 2000, we began design work on a very pragmatic and very low cost computerized system which would be capable of harvesting and retaining the vast majority of the contents of state government Websites. The Capturing E-Publications of Public Documents (CEP) National Leadership Grant by the US Institute of Museum and Library Services (IMLS) [1] supported this system [2]. We monthly process 228 Websites for ISL. Using that harvested material, we operate the Illinois Government Information (IGI) search engine for all Illinois State Government Websites [3] and an Open Archives Initiative (OAI) metadata server [4]. Six additional States (Alaska, Arizona, Montana, North Carolina, Utah, and Wisconsin) committed to being partners in this research. Arizona and Alaska quickly acquired their own host computer, and a software installation and data transfer was done to their installations over a year ago. Independent CEP operation by those two States and our operation of web archives for five States contributed greatly in identifying desirable functionality and control features. North Carolina and Utah are in the process of acquiring their CEP host computers. Combined, these CEP systems now archive 2.6 million files. The comparative size of the web archives of the seven States is listed in Table 1.

We have encountered many systemic problems in Website harvesting for archives. This paper describes classes of problems which currently limit the abilities to automatically harvest web materials, to process materials for retention, and to utilize the archive information in the future. Some of these problems can be at least partially addressed. However, the most damaging of these problems cannot be overcome without changes to document source materials, the way in which source materials are prepared, or changes to the document retrieval mechanisms and practices of the original Website.

Table 1. Comparative Sizes of the Web Inventory for the Seven CEP Project States

State	Web Sites	Files	Giga-bytes	Date of Harvest
IL	228	727,271	169.3	07-27
AK	84	184,618	42.5	06-06
AZ	139	261,786	26.1	07-29
MT	83	148,181	35.1	07-08
NC	184	528,787	88.2	07-25
UT	244	321,695	44.3	07-17
WI	81	429,815	58.4	07-06
Total	**1,043**	**2,602,153**	**463.9**	n/a

All Website content might not be desirable to archive. There are unresolved questions concerning the longevity of certain data formats, and the age-old question of substance versus form. Websites could choose to emphasize the delivery of factual information. Or, they could choose as a goal to have users feel good about the associated agency, its products, services, and personnel. Governmental Websites can be of either type. On the web, where the formal and informal intermingle, it is difficult to automatically exclude unwarranted items from a State's historical collection.

There is a general agreement among the State Libraries involved that not all web materials should be archived. Identification of materials warranting preservation must respect budgetary limits. ISL has chosen a two-pronged approach, wherein Websites are harvested in their entirety using a high level of automation. This step is correspondingly inexpensive. Higher value publications are then human-selected, by either the authoring agency or by ISL staff, and copied to a second facility, the Illinois Electronic Documents Initiative (ILEDI) [5] for permanent public web access. In other work, under a National Digital Information Infrastructure and Preservation Program (NDIIPP) grant from the US Library of Congress, a larger research team is exploring developing software tools to assist State Librarians in the discovering significant materials in Websites, and in discovering heretofore unrecognized state government Websites, as summarized by Cobb, et al [6].

STATE LIBRARIES LIVE AT THE INFORMATION CROSSROADS

The information management problem domain of state libraries is rich in research opportunities. It necessarily draws heavily from Library Science, but needs to adapt these practices for a situation where the boundaries of activities between author, editor, publisher, librarian, archivist, computer technology provider, and end user are very blurred. By comparison to the processing of print publications, processing of materials in electronic form both greatly increases the extent to which computer technologies must be involved. The electronic documents archives problem space leaves state library staffs in the situation of being the final quality control authority for all the activities which replaced the phases of print publishing. In electronic documents archives, corrections within document files, corrections to webpages used for document access, assignment or quality control of metadata, the provisioning of Websites for collection publication, and making provisions for adequate search facilities for collection Websites are now matters for state library staff to address.

It has frequently been necessary to return to discussions of first principles of librarianship and how those ought to be understood to apply in a different publishing medium and under wholly different mechanisms of collection access. For example, how should principles such as subject heading assignment to documents apply in an environment where most resource discovery within the collection is projected to be via web search engine rather than via our electronic approximation of a library catalog? And, design of the web-based access provisions to our catalog also melded catalog usability concerns with webpage accessibility requirements. Metadata scarcity, in turn, motivates exploration of both collection-level metadata measures and automated techniques for metadata extraction or inference. While many related matters are interesting enough as computer science problems, the situated nature of this work within the context of complex state government information management activities adds complexity in attempts to well support human and inter-organizational information flow. And, severe budgetary limitations give rise to considerable scrutiny of every subsystem or function in a design.

State libraries and/or state archives, depending on the State, function somewhat differently than typical libraries. Requirements to preserve the official publications and records of the State are rather unique. Very often, these materials were created as required by some law. Rather than judge the retention worthiness of materials from other agencies, the documents are usually retained.

Agency Websites can include generally informative materials which were not expressly mandated. Additionally, webpages often use a conversational or informal tone. And, webpages may incorporate items which, as a result of their implementation technology, are probably unsuitable for use even a few years in the future. So, what is to be done with the webpages of state government?

SYSTEMS ARCHITECTURE OVERVIEW AT ILLINOIS

Very briefly, the system we developed for Website archives harvests Website content using GNU wget, performs post-acquisition standardization, metadata extraction, and statistics generation, and stores the files under the Concurrent Versions System (CVS). A different CVS "repository" is used for each Website archives. CVS retains all prior versions of a given file, and has proven to be relatively economical of storage space, especially with text files including HTML. Provisions for remote system control and access to archive documents are provided via CEP web scripts. System metadata is stored in XML. Our Website archival software is available for download under open source license [2].

We have also developed software specific to the needs of ISL which is not currently available for download. Our second ISL system, the Illinois Government Information (IGI) search engine [3], provides search and browse facilities for all Websites, or any one Website of Illinois State Government. IGI uses metadata derived from our Website archives to construct surrogate WebPages for indexing, processed by the open source search engine "Simple Web Indexing System for Humans—Enhanced" (SWISH-E), and supports search of over 500,000 documents.

Our third ISL system, the Illinois Electronic Documents Initiative (ILEDI) [5], provides ISL its legally mandated mechanism for permanent public access for those official publications of the Illinois State Government which exist in electronic form.

FUNDAMENTAL PROBLEMS IN ELECTRONIC DOCUMENT ARCHIVES

Managing a decentralized web

State governments operate very many Websites—a mean of 149, across the seven States in this project. By broad consensus, the agencies of state government are not generally closely coupled with one another in the performance of their daily tasks. Inter-organizational relationships are generally perceived as weaker the farther away two agencies are from one another in the government hierarchy.

ISL contacted multiple webmasters before this project concerning existing provisions for Website archives. In general, little provision existed. Some webmasters did produce backup tapes, but these were collected in an ad hoc manner and differed in format, frequency of collection, degree of thoroughness, and in suitability for long-term retention and information recovery.

CEP gathers documents into central archives, and seems to be successful in obtaining the vast majority of web documents. However, there are many problems in harvesting, which can cause the historical record to be incomplete.

A decentralized architecture for web archives is possible, where each Website operates archives of its own content. However, coordinating the installation and operation of a systematic and interoperable archival mechanism will have its own difficulties. Many Websites have proven to be relatively technologically simple, displaying documents but not evidencing complex computing. Asking these webmasters to incorporate more complexity may be asking too much. Additionally, state government Websites appear, disappear, and change host machine name much more frequently than one might imagine for a government publishing medium. If comparable volatility applies to archives managed by this same staff, both the current content and the archives for that Website might be at risk. At least in the seven CEP project States, centralized, non-invasive harvesting appears for some time to come to be preferable to attempting coordinated overhaul of numerous Websites.

Fundamental limitations of the harvesting paradigm

Although harvesting Websites to central archives provides at least a partial mechanism for the capture of this information content and does not rely on close cooperation with the Webmaster, probing the Website from afar via harvester program is not a perfect mechanism. Some materials cannot be acquired via harvester.

Over-reliance on web addresses

The problem area with the most widely distributed impacts seems to be an over-reliance on web addresses as document identifiers within archival systems. In compiling web archives, consideration must be given to how these files will be accessed in the future. In CEP, that access is assumed to be primarily via browsing through the archives, in a click sequence highly like that of browsing the original Website. As such, hyperlinks embedded within harvested files generally need to be changed so as to point to the copied files within the archives rather than to the original Website. Hyperlinks pointing to yet another Website are problematic in that this additional Website is not part of the archives, and having a clicked hyperlink return the archives user to the “live Web” would probably be confusing.

The wget harvester performs hyperlink revision, but imperfectly. Many of these imperfections have their origin within document retrieval scripts of the harvested Website. The wget harvester uses the web address of a harvested file to form the path to the storage location for that file within the UNIX file system. This has proven problematic for archive use for multiple reasons.

First, if all the documents of an agency are retrieved by one script, wget will store all the harvested documents within the directory where the script appears to reside (e.g., “cgi-bin”). Storing more than 1000 or so files in one directory will greatly impede attempts to use that directory. This impacts the harvesting process, post-harvest processing, retention processing, backup copy generation, and the serving of files to users via the archives web server.

Second, differences in the set of metacharacters (characters having special meaning to a computer) between the operating systems of different vendors cause problems. All metacharacters within file or directory names must be masked (escaped) to avoid activating their special function on the archives host computer. Problems of this nature can limit the usefulness of common utility programs (e.g., file backup utilities) on the archives host computer.

Third, operating systems have limitations on the length of names of files. Long file names from the original Website can result in illegal file names on the archives host.

Fourth, in document retrieval using scripts, there may be nothing comparable to a file name or file name extension. The script output itself is usually the document of interest, and no named file is involved in the harvesting of that file.

Fifth, Websites often have their contents moved. To avoid inconvenience to users who have previously recorded document web addresses elsewhere, mechanisms exist within the web server software whereby a request for the former name results in the user or harvester being sent instead to the new document location. Depending on how archival storage is managed, the use of a different host name or directory path may cause the harvested documents to be perceived as being a different set than those of earlier harvesting.

Sixth, web addresses may equate to scripts, and scripts may parse the balance of the web address string in unique ways. For example, the Utah Department of

Transportation Website (e.g., http://www.udot.utah.gov/index.php/m=c/tid=1) uses the slash character ("/") as a separator between name-value pairs in script parameters instead of the more common use of the ampersand ("&"). Unfortunately, the slash character is the UNIX metacharacter indicating a subdirectory.

Seventh, script invocation with parameters may or may not produce results dependent on the order of parameters. Where parameters are passed to a database, something like an implied Boolean AND or other symmetric function is often used between all the parameters in retrieval, making the order of parameters immaterial. However, parameter order might matter, if the processing being done is not a symmetric function, or if the order of the parameters being supplied is somehow interpreted by the script to convey a meaning.

In practice we have found very many examples of the use of the implied AND. Unfortunately, the web addresses the harvester processes are those it has found stored inside the previously harvested files of the Website. If the authors working on that Website do not observe a canonical ordering in the way they form the sequence of name-value pairs in a web addresses, multiple web addresses will be encountered. This will cause the harvester, under conditions where parameter order is unimportant, to download multiple copies of the same script result (document). This duplicated document would then be stored under each of the file names derived from all the permuted orders of the parameter list. As, for a parameter list of n parameters, there are n! (i.e., n factorial, i.e., n*(n-1)*(n-2)*...(1)) possible orderings, scripts with even a few parameters can give rise to extreme amounts of redundant storage. Detecting redundant processing during the harvesting phase, particularly when the harvested files contain some nominal difference such as an embedded statement of the time they were produced, is computationally expensive.

Independence from web address

Harvesters should not store documents at locations which are functions of the web address harvested. Some form of indirection should be used. This indirection can be via a database of arbitrary complexity, or could be as simple as a pair of lists, giving the ability to translate from the original web address to the storage location within the archives, and the reverse. Storage locations within the archives host computer could then be changed as necessary to distribute the data inventory for expeditious processing. Edited hyperlinks within archive document files would correspondingly need to point to the actual storage location within the archives, or the archives web server would need to employ a redirection mechanism.

File format proliferation problematic

In order for a harvester to be able to process a file and discover embedded hyperlinks, it must parse the file. As companies continually create software tools for document production, proliferation of file formats results. Harvester programmers cannot be expected to keep up.

There is much concern about future unavailability of all the computer programs needed to accessing information in all these formats. There is also somewhat of a feeling of helplessness in that libraries and archives are unlikely to wield enough fiscal power to be able to cause software vendors to respect the need for continued support for very old file formats. Someday the retirement of a file format will be announced, and the libraries and archives of the world dread a rush to examine their holdings and convert from the retired format to the new format, while software tools which can speak both dialects still exist. This will unpredictably introduce a large, mandatory expense.

Another path to avoiding format obsolescence would be the potentially expensive emulation of old software. Intellectual property rights may encumber the old format, making it illegal to develop software capable of processing proprietary formats. Legal or not, the threat of lawsuit may be enough to discourage the attempt.

Archiving snapshots of application and operating system software is also now more difficult, as frequently installation materials, and especially updates, may arrive via Internet, leaving no artifact (e.g., CD-ROM) to keep in secure storage. Fabricating a CD-ROM containing the combined effects of multiple patch updates may violate copyright laws.

Or, perhaps we can start persuading society to not produce its official records and publications using volatile encodings such as those employed in proprietary software.

Invalid HTML syntax

When webpage authors fail to use syntactically correct HTML, or use document production tools which have this failing, an HTML dialect is a file format all its own. Other tools, including harvesters in archiving, may not be able to correctly process the differences.

Checking HTML files produced by Illinois agencies, a small sample was analyzed using the markup validation service Website provided by W3C (http://validator.w3.org/). Using the "validate by URL" option, the syntax checker acquires the copy of the document for analysis directly from the agency Website. A list of 300 document web addresses was drawn from the contents of 487,646 documents searchable via the July 2005 database update of IGI. Of those, 26 links were already broken, 69 produced word processor documents, and 205 (68%) produced HTML. Of the HTML, 10 files caused the syntax checker to fail, so no analytical results were available. Of the 195 HTML files successfully analyzed, we found a mean of 48.0 HTML syntax errors per file.

With so many syntactical errors per file, it is not reasonable to expect a harvester to work perfectly. If the harvester errs, the web archives will be incomplete. However, hyperlinks are embedded within only a small fraction of the HTML tokens (tags), so it may be that syntactical damage involving other parts of a file might not confuse the parsing of hyperlink tags themselves. The degree to which syntactic errors interfere with successful harvesting is therefore not fully known.

Multiple character set options

In addition to syntactical tokens within a file, the character set chosen for use within the file is another source of variation an archival system must support. The syntax checker work above also identifies the character set used within the file. (If the character set is unspecified, HTML analysis proceeds, assuming UTF-8.) In the 205 HTML files analyzed; 23% are unspecified, Windows-1252 accounted for (38%), UTF-8 17%, and ISO 8859-1 22%. The 10 files which caused the HTML analysis to fail did not affect character set determination.

Inaccessible designs

Another topic related to the construction of HTML webpages is the degree to which the webpages are accessible to people with differing abilities (e.g., in vision and motor control). The US Rehabilitation Act, as amended [7] requires certain steps be taken in providing information to the public so as to minimize the exclusion of a portion of the population. Although a federal law, States and projects receiving federal funds must comply.

The application of standards and conventions concerning accessibility across Website designs is difficult to consistently and uniformly judge. Accessibility standards are a mixture of the syntactic and the semantic, and, while syntactic features are relatively easily analyzed with a software tool, the semantic points are subjective. If inaccessible materials are retained in archives, it is likely they will continue to exhibit at least nearly the same degree of inaccessible in the future. While some advances in assistive devices will occur, that wish does not relieve the electronic document archivist of the need to try and obtain the most accessible copy possible at the time of collection.

To obtain an estimate of the degree of compliance with accessibility requirements, we used a simple metric, which, while not completely reliable or indicative, is easy and repeatable, and which discloses some informative, if personally disappointing, results. In many state government Websites, icons are prominently displayed concerning Bobby compliance or Section 508 compliance. While the absence of an icon does not mean a Website is designed without accessibility in mind, frequent practice in prominent State of Illinois Websites is to proudly display such an icon where justified. Expert reviewers may differ on the degree to which a given Website actually meets the requirements, but the presence of the icon seems to at least indicate an awareness of the issue and the requirements, and some attempt to begin to meet them. So, we have used the presence of either one of these icons, or the inclusion of links at or near their homepage (including within “about” pages, Website maps, and other provided introductory information) concerning the availability of some alternative form of rendering of the web materials which may be helpful to users with disabilities (e.g., “printer-friendly pages”, or “text-only versions”), as indicative of a Website deliberately addressing accessibility issues to some degree.

In a survey of all 220 then-known State of Illinois Websites in early July, 2005, 90 (41%) made at least some mention of, or used an icon of the Bobby software for accessibility guidance, or similar icons concerning Section 508 compliance. 14 (6%) of the others had more accessible alternate renderings of Website content in evidence. The remaining 116 Websites (53%) had none of the preceding items in evidence. As Illinois agencies are required to make provisions for access by all, it seems probable most agencies would wish to make their support known to the segments of the population which most benefit from these features. So, the 53% with no such indication readily in evidence is cause for some concern.

Adjunct programs problematic

Additionally, several software adjuncts to web browsers employ programs of their own (e.g., JavaScript, Java, Macromedia, and Active-X). These programs need not communicate using HTTP or encode information as HTML, but may. Harvester programmers cannot be expected to devise a means to recognize what another running program is doing, and to identify those moments during program execution when a file is being loaded. In general, running programs are impervious to analysis by harvesters, although some harvesters such as Heritrix take the partially successful step of searching some downloaded program code files for any character string appearing to be another web address.

Web servers can be operated unreliably, inconsistently

Harvesting systems could be operated more efficiently if they could accurately know exactly which files of a Website changed since the preceding harvest. By only retrieving the changed material, accurate archives could be kept, without having to redundantly process the unchanged material. However, in practice we have found sufficient unreliability in the reporting of the date of most recent file change by web servers so that we do not use that method. We have also found unreliability in the reporting of file size, file type (MIME type), and the character set used to encode the file, so reliance on file metadata reported by web servers does entail some risk.

Web servers are capable of reporting the date a file was last changed, without necessarily sending the file itself. Avoiding sending the file, by detecting that is has not changed, would save bandwidth and processing time for web server and harvester alike. However, if that information is unreliable, no advantage can be gained.

In cases where a web server is not serving files from a static collection, but instead is pulling copies from a document database, a temporary scratch file may be used. The scratch file has a modification date of the immediate moment, resulting in the probably erroneous assumption that the file changed.

CEP always re-harvests the entire Website. With very rare exception, this has not proven to be a resource consumption problem for either the Websites or our harvesting facility. CEP currently processes 1,043 Websites, and in configuring all those harvesters, only one Website has been found to have genuine bandwidth limitation concerns. That Website is using an Internet connection with bandwidth more typical of a home than a business. For a handful of other Websites, we have reduced the frequency with which our harvester requests files, but monthly harvesting continues. Local CEP post-harvesting processing draws its own conclusions concerning whether or not a given file has changed.

What is meant by saying a document has changed? In traditional library contexts, a document may be discovered to have been reissued. If a book is reprinted, even with some minor editorial changes such as the correction of typographical errors, the new edition is not generally purchased by libraries. However, within computer systems, probably the least expensive way to compare two files is to see if they differ at a verbatim level (e.g., UNIX “diff”). Given a pair of files from a document, differing only in the correction of a typographical error, a diff comparison would say the files differ. It would require a much more sophisticated computer program to detect that the differences are “minor” —however minor is to be defined. Thus, computerized archival systems will tend to consider a document to have undergone more revisions than would a manual archival system. Ignoring file storage costs, as they continue to decrease rather rapidly, the prominent negative effect of storing too many versions would be the inconvenience to the user who finally wishes to search this particular document. Search results would presumably match more versions of the document. But, that may not be a significant disadvantage if, say, the oldest matching file is always presented to the user preferentially.

Another source of file modification which is generally agreed to not equate to significant change in the content of a document is the continuing modification of features such as a statement of the date and time the file was downloaded, or the incorporation of a “banner ad” image which differs for each successive retrieval. These differences will also cause a verbatim file comparison to report a change.

Perhaps mechanisms outside of the Website computer operating system itself could be used to provide more accurate information. If a document is managed by a document management system, that system could have tracked the various versions and official releases of a document, and could provide dates or identifiers based on the author's decision to declare a fundamentally new version, rather than assuming a change in an operating system file date indicates change in intellectual content.

Revision information could be provided to the archives, either as metadata embedded within the file itself, or extrinsically. An OAI server could be used to provide metadata to harvester systems, and that metadata could contain an authoritative date of last substantive modification. Proposals have also been made for expanding the role of the web server “robots.txt” file to additionally convey revision information. Qualitatively, the important feature in these options would be to accurately inform the harvester which files have changed at an intellectual level, obviating transfer of the files. If the information is kept accurate, externally supplied revision information could reduce harvester resource consumption. However, if Websites are currently operated with many kinds of errors, asking these same webmasters to set up and operate yet another information transfer mechanism may not quickly result in a solution.

Identifying State Websites

Recognizing that a Website is a function of the State can be difficult. The Website host name may fail to indicate an affiliation with the State, or may suggest a generality of scope exceeding the State (e.g., many State Websites now use “.com”, “.org”, or “.net” suffixes). Use of alternative suffixes can obscure the relationship of the Website to a sponsoring governmental unit. Many Website host names involve acronyms, which may not be widely recognized. Obscuring Website ownership or identity also complicates the process of constructing Internet content filters.

ISL previously operated a web search engine, so CEP was initially configured to support those 120 Websites. Subsequently, a mix of library-style research, plus Internet search engines were used to rapidly discover 80 additional Websites.

Absent or poor quality metadata

Very early in this work, a brief survey was done concerning the degree to which embedded metadata was being incorporated into Illinois State Government Websites [7]. Subsequently, CEP monthly statistical reports have included similar information [9]. Both report only a very small percentage of the Illinois State Government web documents have been labeled with metadata by their authors.

Paucity of metadata is a concern for the operators of electronic document archives in that it is desirable to use information retrieval mechanisms from library science in addition to simple matching of keywords. If the authors have not provided metadata, some other provisions must be made. Perhaps state library staff could construct the metadata like they have long generated bibliographic descriptive information for the print publications of the State. That approach is thought to be expensive, per document. While that expense might readily be justified in the case of the most valuable or prominent publications, it will be much harder to justify for the very much large numbers of webpages which already exist (Table 1).

The IGI search engine for all Illinois State Government Websites is based on the SWISH-E search engine, and as such can be operated as "metadata aware", able to benefit from that metadata which is available. We have supplemented the rare author-generated metadata through the inclusion of some plain-text and the extraction of noun phrases. Also, if corporate authors do not identify themselves in metadata, knowledge of the ownership of the Website is used instead to infer authorship.

We are pursuing sources of subject classification metadata other than author-generated. In part this is due to having added a subject heading browse capability to IGI, as well as a capability to search within the documents classified under a specific subject heading. In the absence of subject classification assignments to most documents, those features go underutilized. We intend to try out our experimental sources of subject classification metadata via deployments in IGI.

We are pursuing the use of collection level metadata in default values for subject classifications used across all documents of a given Website. We are also pursuing classifier programs for the assignment of subject classifications. A classifier program could assign different subject classifications to every different document. We have had promising initial results with an Expectation Maximization algorithm, trained to, and therefore operating only within the top two levels of the State GILS subject classification tree.

CONTRASTING APPROACHES TO ELECTRONIC DOCUMENT ARCHIVES

In devising a plan for archives of the important electronic publications of a State, a number of implementation options are found, and a number of local social factors must be taken in to account. In our work with CEP and ECHODep, two different approaches to collection building are being used by those States processing their own web archives. Illinois and Alaska are using a document-oriented approach, while Arizona is pursuing how an archivist would address the acquisition of a large body of hierarchically organized material, as described by Pearce-Moses and Kaczmarek [10]. Segmentation of the storage hierarchy in many cases indicates relatedness of the documents stored therein.

Perhaps not coincidentally, the professional background of the ISL and Alaska State Library staff involved are as Librarians, while Pearce-Moses in Arizona is an archivist. Both approaches seem systematic and likely to succeed, but both also have their points of faith and risk. These approaches, and others, are not mutually exclusive and may beneficially be pursued in tandem.

The ISL approach has been to first put in place a "safety net" archive copy of their entire web, using the CEP system. Then, ISL makes provisions for the long term retention of human-identified valuable documents in ILEDI. ISL has no intention to designate even a majority of the Illinois State Government web contents as meriting full Library-like acquisitions processing. CEP archives will be available, with at least a keyword form of search, for those hundreds of thousands of web documents which appear to lack the formality and permanency typical of the printed publications of the State.

ISL identification of the valuable documents has capitalized on its knowledge of prior print publications to identify titles of interest, and on existing ties to the authoring/publishing agencies. Further, ISL has ties with State webmasters from having long provided a Website search engine for

Illinois State Government. A manual, traditional, library-style search, by one half-time graduate student for less than one semester, identified the disposition of the documents in question, as well as finding many new electronic publications at those locations. In the future, education and outreach initiatives by ISL to all Illinois agencies is intended to raise agency awareness to the statutory requirements that the "official publications" of the State be deposited with ISL. A definition of "official publications" was not provided in the requirement, and is in general left to the authoring agency to define, with the benefit of the educational work soon to be done by ISL.

The "Arizona Model" [10] begins at an earlier point in the process of seeking out valuable materials, and postulates a systematic review of Websites as hierarchical information structures, by an archivist equipped with computerized tools specific to this situation. Such tools are being developed by OCLC as part of the ECHODepp grant. The planned archivist's view of a Website would involve making most retention decisions at a web server directory level rather than addressing individual documents.

Arizona operates its own CEP system, rather than let Website contents simply disappear in the interim. And, ISL will be trying the ECHODep tools as yet another way to seek important documents. CEP has also been expanded to report heretofore unseen Websites, potentially of interest.

The ISL approach to the identification, acquisition, and archiving of high-value publications risks that agency webmasters and publications staff will not respond to the ISL request (and the legislation behind it). The legislation did not provide additional funds to perform the labor involved. Additionally, there are varying results in attempts to have authors create their own descriptive metadata. If agencies do not fully cooperate, perhaps ISL staff could review Websites, compose metadata, and deposit documents under a process much like the Arizona Model.

The Arizona Model risks that Website reorganization or relocation can invalidate the locations portion of previously identified high-value materials. We have seen Website redesign and host computer name changes are very frequent occurrences. However, if in the redesign process, valuable materials are relocated to another host name or directory name, it may be possible to find the new location, provided one has a complete CEP archives to scan with a computer program.

The Arizona Model will probably experience problems in those Websites where the bulk of the content is served via a document management system. When very many documents retrieved by a single script, on the basis of certain parameter values, it can become difficult to recognize document groupings, so group-level decision-making economies will not result. The document at http://wxyz.gov/cgi-bin/GetDoc.cgi?Number=100 may or may not have any relation to the document at http://wxyz.gov/cgi-bin/GetDoc.cgi?Number=101. Further, we have seen Websites where huge numbers of documents are all retrieved by one script, obscuring relatedness in the bulk retrieved.

Another approach to constructing Website archives would be to rely on whole-web harvesting systems like the Internet Archive or Google. The fine tuning of spider parameters to Website configurations has proven very frequently necessary, and will not be possible on that scale of work. However, a subset of interest to a State might be affordable. Additionally, state agencies responsible for the acquisition, management, and continued availability of important State materials are reluctant to abdicate that responsibility in the hope that someone else will happen to do all that is desired.

CONCLUSIONS

Although living at the crossroads of multiple disciplines exposes the electronic documents staff and researcher to the problems of many communities, there are points of hope in answers made possible by synergistic combination of principles and ideas from all these disciplines. One particularly promising avenue at the moment is a triage-like approach of identifying where the most good could be done, in search and retrieval terms, per effort expended. Stratified approaches to metadata generation or inference promise cost reduction, where systemically a combination of collection-level descriptors and item-level descriptors would both contribute to search.

Identified high-value documents or high-value regions of Websites could be targeted for additional metadata writing, perhaps bringing in cataloging experienced librarians to help. If there truly is too much content, of too low an overall quality to warrant the complete set of “best efforts” for long-term survival, let us at least make an effort to do well with the valuable portion.

The IGI search engine, as an information by-product to CEP web archives construction, raises issues of integration of search and archival efforts. With document whole text, embedded metadata, and collection-level metadata available to both systems, we now seek still more metadata and document summaries. With a mechanism to associate extrinsic metadata with high-value documents, a library staff at a central location could help make the most important documents of the State's web more retrievable, without necessitating full cooperation of the webmaster.

And, there is evidence of synergies in the overlap of project and experiment strengths to cover shortfalls in individual approaches. Until all the necessary steps in electronic document archives construction and operation are thoroughly understood, there will be a need for some give and take between approaches and tools. Happily, the parties involved seem comfortable in working in an environment which is not yet fully codified, and staff flexibility is the norm.

ACKNOWLEDGEMENTS

This work has been funded in part by grants from the US Institute of Museum and Library Services, from the Illinois State Library, and from the US Library of Congress. Opinions expressed are those of the author, and not necessarily those of IMLS, ISL, UIUC, or the Library of Congress.

The extensive participation of multiple staff members from each of the seven State Library research partners in the CEP project, but particularly Richard Pearce-Moses of the Arizona Library, Archives and Public Records, and Daniel Cornwall of the Alaska State Library, is gratefully acknowledged.

The following are gratefully acknowledged for their assistance in the UIUC portion of this research; Sai Deng, Tim Donohue, Sarah Eiben, Sungok Hong, Xiao Hu, Kyung-Ja Hyun, M. Sharif Islam, Annette Jackson, Lanie Klinkner, Wing Yee Kwong, Weiguo Liao, Guixian Lin, Edith List, Terry McLaren, Ozwaldo Meza, Brynnen Owen, Jian Wu, Xuan Xie, Peng Xu, Lin Yang, Huamin Yuan, Yiyi Zeng, Yan Zhang, Jing Zhang, and Jing Zhao.

REFERENCES

1. Illinois State Library, 2003. Capturing E-Publications of Public Documents webpage. http://www.cyberdriveillinois.com/departments/library/who_we_are/cep.html
2. Jackson, L.S., 2000. Preserving Electronic Publications project Website. http://www.isrl.uiuc.edu/pep/
3. Kwong, W.Y., Jackson, L.S., Deng, S., Yuan, H., 2004. Illinois Government Information search engine Website. 9 August 2004. http://findit.lis.uiuc.edu/cgi-bin/search.cgi
4. Zeng, Y., 2005. Open Archives Initiative web server for CEP Illinois Websites. Current edition (14 August 2005 or subsequent). http://history.lis.uiuc.edu/cgi-bin/sn/oai/OAI-XMLFile-2.2/XMLFile/history.lis.uiuc.edu/oai.pl
5. Jackson, L.S., and Deng, S., 2005. Illinois Electronic Documents Initiative Website. 1 August 2005. http://iledi.org/
6. Cobb, J., Pearce-Moses, R., Surface, T., 2005. ECHO DEPository Project. IS&T's 2005 Archiving Conference. Washington, DC; April 26, 2005; p. 175 — 178. http://www.ndiipp.uiuc.edu/
7. U.S. Rehabilitation Act 29 U.S.C. 794d, Sections 504 and 508, as amended, 1998. http://www.section508.gov/index.cfm?FuseAction=Content&ID=14 http://www.section508.gov/index.cfm?FuseAction=Content&ID=12 http://www.section508.gov/index.cfm?FuseAction=Content&ID=15
8. Jackson, L.S., 2003. Preserving State Government Web Publications — First-Year Experiences. In National Science Foundation, Proceedings of the National Conference on Digital Government Research dg.o2003. Digital Government Research Center, Marina del Rey, CA. 2003. pp. 109 — 114. http://www.isrl.uiuc.edu/pep/papers/DGO2003_PaperOnPep.pdf

9. Jackson, L.S., 2005. Capturing Electronic Publications. IL Website Statistics, latest edition (27 June 2005 or subsequent). http://history.lis.uiuc.edu/~cep/stats/IL/LatestStats.html

10. Pearce-Moses, R., and Kaczmarek, J., 2005. An Arizona Model for Preservation and Access of Web Documents. DttP: Documents to the People 33:1 (Spring 2005), p. 17 — 24. http://www.ndiipp.uiuc.edu/pdfs/azmodel.pdf