Successful Scalability Techniques for Illinois Web Archive Search

Larry S. Jackson & Huamin Yuan
UIUC GSLIS Tech Report **UIUCLIS--2007/1+EARCH**
April 27, 2007

Abstract

The Capturing Electronic Publications (CEP) web archive assembled since 2002 by the Electronic Archive Project group of Graduate School of Library and Information Science (GSLIS) at the University of Illinois, Urbana-Champaign (UIUC), for the Illinois State Library (ISL) currently contains over 37 million files and is increasing by over 900,000 files per month. In order for ISL to utilize this collection effectively in identifying, selecting, and migrating specific documents to permanent storage, some form of search mechanism had to be provided. However, the file inventory far exceeded the capacity of open source or freeware search tools. Detecting those files which had not changed between harvests allows the suppression of search surrogate generation for those files. With that substantial reduction in search surrogate count accomplished, existing provisions of the SWISH-E open-source search engine to use multiple search databases sequentially did not impose noticeable delays on search engine users. Combined, these approaches enable SWISH-E search across the entire collection, despite an assumed initial design limit of one million files.

Problem Description

The Capturing Electronic Publications (CEP) web archive assembled since 2002 by the Electronic Archive Project group [Jackson 2002] of Graduate School of Library and Information Science (GSLIS) at the University of Illinois, Urbana-Champaign (UIUC), for the Illinois State Library (ISL) currently contains over 37 million files and is increasing by over 900,000 files per month [Jackson 2007]. Files from over 450 websites are included in the collection. This web archiving system, and the files thereby obtained was implemented for Illinois and six other states under two National Leadership Grants from the Institute of Museum and Library Services [ISL 2001] [ISL 2003]. While harvester-based file acquisition is inexpensive per file, it does not produce levels of quality control typical of human-generated collections, libraries, or archives. Like is happening in many other states, ISL will then select harvested materials for long-term retention in a more extensively cataloged and indexed facility. For ISL, that facility is the Illinois Electronic Documents Initiative website [Jackson 2006].

In order for ISL to utilize the CEP collection effectively in identifying, selecting, and migrating specific documents to permanent storage, some form of search mechanism had to be provided. Reports were generated on a per-file basis, tracing the dates of existence

and change, but some form of search facility was also needed. The project team had also produced the Illinois Government Information (IGI) web search website [Kwong, Yuan, Jackson], encompassing all the websites of Illinois state government, utilizing the SWISH-E open-source search engine. Rather than index documents directly, and utilizing metadata and linguistics processing inherent in CEP, document surrogates were produced for IGI, monthly, from CEP harvests. However, the file inventory of the complete CEP archive far exceeded the capacity of open source or freeware search tools. Earlier experiments with mySQL database searches met even more severe limitations on the extent of the searchable document set.

A means had to be developed to provide ISL the capability to search the entire CEP archive. However, completely re-indexing the entire CEP archive under SWISH-E monthly needed to be avoided as it would consume too much processor time on computer systems which were already heavily loaded.

Scalability Techniques Employed

The CEP web harvesting and retention system employs the Concurrent Versions System (CVS) [CVS] as its file storage mechanism. CVS is typically used by programmer teams to coordinate simultaneous changes to sets of program code files. As believed, and as confirmed in five years of operation, the vast majority of Illinois state government website files are HTML. The text-oriented CVS software deals with HTML very efficiently, consuming additional space with each subsequent harvest as a function of the degree of change to the websites, and not of the number of harvests conducted. Data storage capacity required by CVS-based retention is thus far less than that required for retention of the downloaded materials themselves, as illustrated in figure 1.

Duplication Detection and Surrogate Generation Preclusion

Alerted by the relatively slow increase in CVS data storage requirement, it was considered likely that individual web files were not undergoing frequent or extensive change. If true, the CEP archive should then contain a great number of duplicates, re-harvesting identical content across a series of months. And, if that is true, there would be no need to provide search engine access to duplicate files. Indeed, having all the duplicate files show up in the search engine results would also be overwhelming and confusing to the ISL user attempting to locate files for long-term retention.

Searching Multiple Indexes Sequentially

In separate work, SWISH-E provisions for searching multiple index databases sequentially were investigated, and prototype capabilities incorporated into a test version of IGI. While the implementation was straightforward enough, these databases are searched sequentially. There was a concern that, unless the number of search surrogates constructed for previous months could be substantially reduced in comparison to the IGI

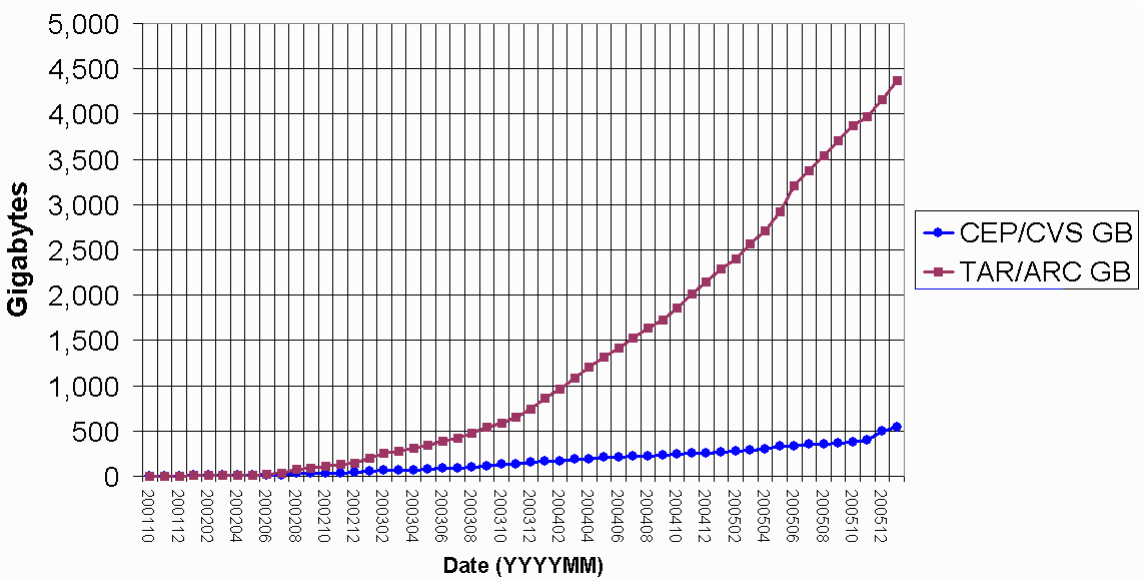

Figure 1. Storage consumption requirements for tar/arc-based design vs. CEP's CVS-based design.

indexing required to support the current month, a sequence of such searches, spanning 60 months or more, would be too slow for user acceptance.

Implementation

Programming was done to examine the extensive set of metadata retained by CEP harvests, and to thereby determine if each individual file had changed from one harvest to the next. Metadata, specifically the file URL and size, were used as the indicator of change, rather than calling the operating system to either re-determine file size or to perform a bit-level comparison (UNIX diff), in the interest of speed of program execution. This simplification does not detect file changes where the contents of a file change, but continue to consume exactly the same number of bytes (e.g., if "April 27, 2007" becomes "April 28, 2007"). However, it was considered that such changes would probably result from script outputs involving page headers or look-and-feel formatting rather than an author's modifications to natural language text. Resultant processing speed was thereby sufficient for overnight generation of the new or revised surrogate sets, using our existing computer hardware.

A search surrogate is only generated, within a directory named for a given month, if that month was the most recent time that exact version of that harvested file was seen to exist. Each new month therefore requires the generation and indexing of surrogates, for all harvested files, as all the current month's CEP harvest are, by definition, seen to exist in the current month. The scope of the current-month processing is comparable to that done for IGI, but with an added pair of metadata fields specifying the earliest and latest CEP-observed dates for the existence of a particular version of a file.

As a file may have existed longer than the current month, it is necessary to check back in time to determine the range of dates for which a specific URL and file size value are found. However, no directory associated with a month earlier than the immediately previous month will have its search surrogates affected. If a file has a search surrogate in a month older than the immediately previous month, then, by definition, that file did not exist in exactly that version in the immediately previous month. Surrogates older than the immediately previous month therefore represent versions which were already known to have changed or disappeared, as of the time of the immediately previous month.

It may be the case, however, that a file, existing in the immediately previous month, continues to exist unchanged in the current month. Such files should therefore have their surrogates in the directory named for the current month. Those surrogates will have the same contents as previously, except that the metadata field specifying the last observed date will now reflect the current month.

Accordingly, regeneration of the most recent previous month's surrogates, figured individually for each harvester, is required each month, as well as the generation of surrogates for the current month. And, the SWISH-E search databases for those affected months must then be regenerated. However, no older surrogates, nor databases, will be affected. This ensures the monthly calculation load to support SWISH-E re-indexing for the all-CEP search engine is limited to slightly more than that of monthly IGI indexing.

Algorithm Used

1. For all harvests of a single website, load the metadata associated with all harvested files having file types of interest (i.e., those associated with "documents", in the view of the archiving agency, e.g., Adobe PDF, HTML, Microsoft Word).

2. For all harvested files of interest in the most recent harvest, determine how long that file existed, under the same exact file name and size, by examining the harvest metadata in reverse chronological order. Using the metadata, write a surrogate record in a directory structure named for the year and month of the most recent harvest. Mark all prior occurrences of this file in the metadata of earlier harvests so that additional surrogates will not be generated (below).

3. Repeat step 2 for the second-most-recent harvest of the website being processed. (This is the only earlier harvest which might have its surrogates affected by the harvest of the current month. All surrogates written using earlier dates than this will reflect terminations of file availability predating the second-most-recent harvest.)

4. Repeat steps 1 through 3 for all archived websites which are to be searchable.

Results Achieved & Future Work

Detecting those files which had not changed between CEP harvests allows the suppression of search surrogate generation for those files duplicated by sequential

harvests. Accordingly, the number of search surrogates required, as a function of project month, is illustrated in figure 2. With that substantial reduction in search surrogate count accomplished, use of the existing provisions of the SWISH-E open-source search engine to use multiple search databases sequentially does not generally impose noticeable delays on search engine users. However, when searches return a very large number of matching documents (e.g., searches for words like "Illinois", or "information", which return 2,322,448 and 1,231,531 matches, respectively), those results are noticeably slow in arriving. For searches returning fewer than 500,000 matches, the response time is typically less than two seconds.

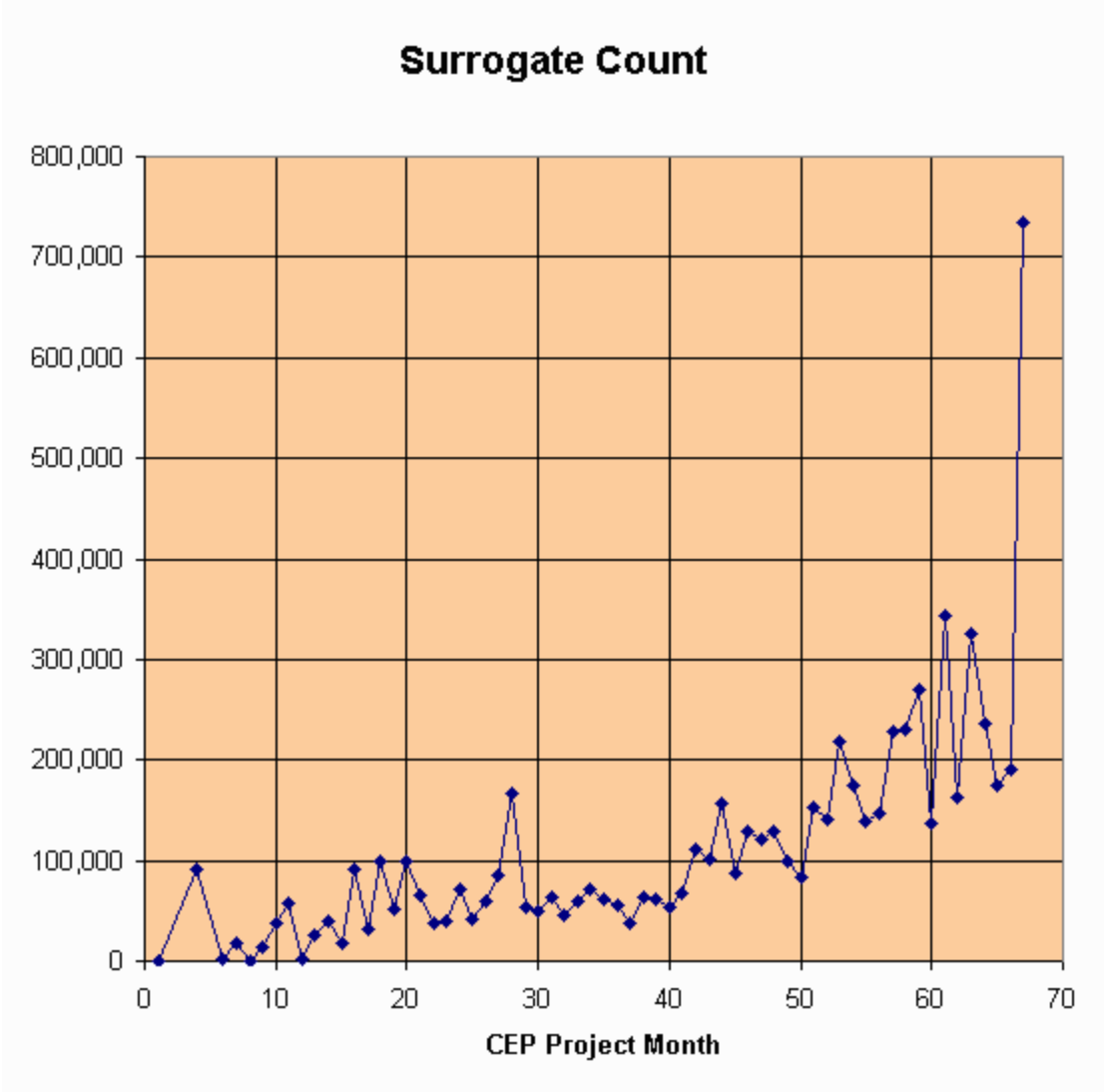

Figure 2. Surrogate count in each CEP project month.
April, 2007 CEP harvest data, totaling 6,865,927 surrogate files.

Further, as the older months have far fewer surrogates, they can be combined into one database. Reducing the number of databases reduces the sequential delay time as successive databases are opened and searched. We currently employ fifteen separate

databases, encompassing the surrogates for 67 months of harvested files. These databases consume a total of nine gigabytes of disk. Combined, these approaches enable SWISH-E to search across the entire CEP collection, despite its assumed initial design limit of one million files. Figure 2, representing the re-indexing using the result of the April, 2007 CEP harvest, totaled 6,865,927 surrogate files. Comparing the search time required here, as opposed to the IGI system which has four databases, 1.2 gigabytes in total size, searching for "Illinois" returns 292,439 matches, in 1.268 seconds. As a simple benchmark, searching all of CEP for "Illinois", employing approximately four times number of databases and seven times the total database size, increased the search time by only about 60%.

Subsequent to implementation, it turned out the number of surrogates associated with the immediately previous month was far reduced compared to those of the current month, that is, most harvested files change infrequently. So, the monthly processing load of the all-CEP search capability is only a few times larger than that of the IGI (current web) search. This makes the all-CEP search capability achievable using the installed base of computer facilities.

ISL has begun evaluating this expanded search capability in locating document files for long-term retention. Unlike earlier reports which only tracked individual files, and were thus comparable to one of the search surrogates here produced, a search engine supports keyword search across file name changes. Websites might employ a file name like "AnnualReport.pdf", with its contents changing annually, but, to keep more than one version of the annual report online simultaneously, it is likely they will employ names containing the date (e.g., "AnnualReport2007.pdf"). However, across the five years of the CEP project, webmasters have come and gone, and new staff may not continue a predecessor's naming conventions or file production techniques (e.g., "AnnualReport2007.pdf" may have been preceded by "06ReportToLegislature.doc". With a search engine available, template text or other conventions may be searched for, facilitating detection of a set of files which form a serial publication.

Detailed investigations are being made of search failures resulting from attempts to use this search facility to determine causes of failures. File retrieval failures from CEP holdings may have been triggered at many points in the chain, dating back to harvester limitations, and including intermediate storage provisions. Metacharacters (characters having special meaning to a particular operating system, but potentially different meaning to another operating system) are known to be problematic to the open-source tools used as modules within CEP.

If these search facilities can be brought up to a level comparable with other web, website, or OPAC search tools, the CEP archival holdings may be opened up for public access. The public may be very interested to learn of prior statements on a topic by a government agency or officer, as compared with the current statements. Other document history aspects seemingly unique to the web may arouse concern for the veracity of the historical record, such as post-publication editing of annual reports and other publications, ostensibly to correct typographical or minor errors.

Additionally, archives search involves the dates of applicability of a document. Rather than present a rank-ordered list of matches to a query, it may be more useful to incorporate a second dimension in the query results, reflecting the dates a document was in effect or existed. It may do no good to locate a document which did not exist, or apply, in the time period of interest. Or, perhaps documents outside the period of interest may be useful in that they could document agency thinking leading up to that period, or document actions subsequently taken, which may reflect a shift in agency thinking. A display along these lines is illustrated in figure 3. Other display options are also being investigated.

A drawback to the approach used is that it does not readily support searches constrained by date. For example, a document which currently exists and which has existed unmodified for some time will result in a single surrogate, generated in the directory for the current month. This surrogate will therefore affect only the search database for the current month. Searches utilizing the search database for an earlier month will not find this surrogate. Accordingly, and as search time has not been constraining, we search the indexes of all months, and then display the dates of applicability in with the search results list, like the example in figure 4. Presently, users must form their own conclusions

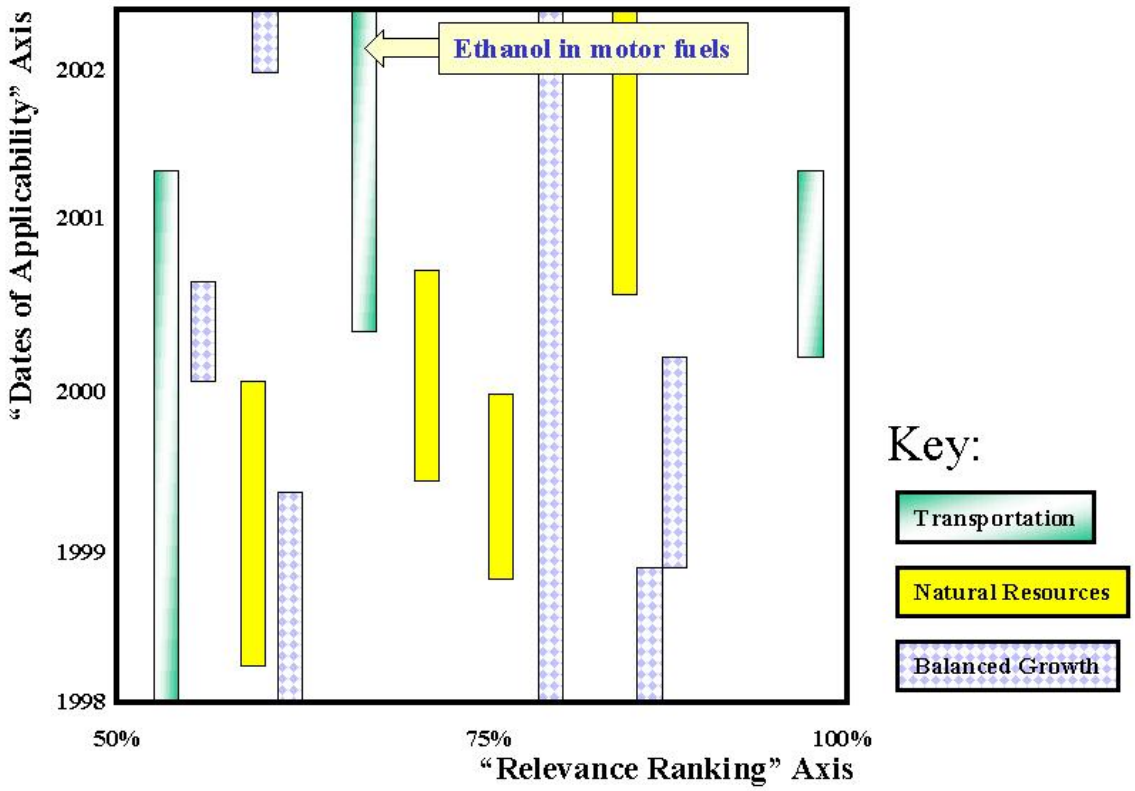

Figure 3. Sample relevance-applicability two-dimensional user interface for search results.

concerning document applicability as a function of these dates, however we will investigate incorporating additional date-based filtering and display adjuncts into the SWISH-E supplied output.

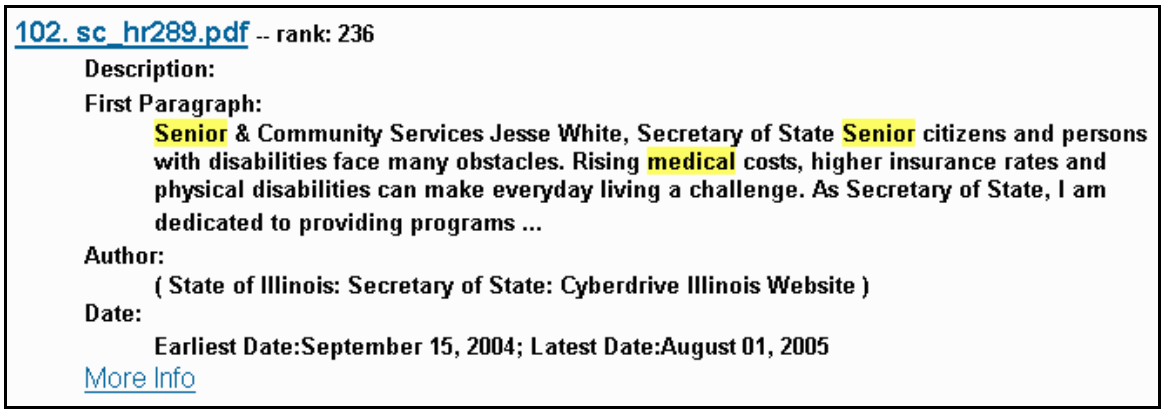

Figure 4. Sample search output showing range of dates the file existed without modification.

Acknowledgements and Disclaimers

This work was supported by a grant from the Illinois State Library, an office of the Secretary of State, Jesse White. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Office of the Secretary of State.

In addition to the authors of cited references, Ms. Sungok Hong created the mySQL-based search prototype mentioned.

References

Concurrent Versions System website, accessed April 27, 2007. *cvs - Concurrent Versions System.* http://www.nongnu.org/cvs/

Illinois State Library, Library Automation and Technology division (2001). *Preserving Electronic Publications (PEP).* October 1, 2001. http://www.cyberdriveillinois.com//departments/library/who_we_are/pep.html

Illinois State Library, Library Automation and Technology division (2003). *Capturing E-Publications (CEP) of Public Documents.* October 1, 2003. http://www.cyberdriveillinois.com/departments/library/who_we_are/cep.html

Jackson, Larry S. (2002). *GSLIS Electronic Archives Project homepage.* http://www.isrl.uiuc.edu/pep/

Jackson, Larry S. (2006). *Illinois Electronic Documents Initiative website.* June 1, 2006. http://iledi.org/

Jackson, Larry S. (2007). *Capturing Electronic Publications -- IL Website Statistics as of 200704270200* (and subsequent). April 27, 2007. http://history.lis.uiuc.edu/~cep/stats/IL/LatestStats.html

Kwong, Wing Yee; Yuan, Huamin; Jackson, Larry S. (2004). *Illinois Government Information web search engine.* August 9, 2004. http://findit.lis.uiuc.edu/cgi-bin/search.cgi

SWISH-E website, accessed April 27, 2007. *Swish-e -- Simple Web Indexing System for Humans - Enhanced.* http://www.swish-e.org/