

**Capturing E-Publications (CEP) of Public Documents  
Third Report: October 2004 to March 2005**

<b>Organization Name:</b>	<i>Illinois State Library (ISL)</i>
<b>Project Name &amp; Number:</b>	<i>Capturing Electronic Publications (CEP) LG-02030120</i>

<b>Program Influencers</b>
<ul style="list-style-type: none"> <li>• Graduate School of Library and Information Science (GSLIS);</li> <li>• Participating state libraries;</li> <li>• State of Illinois agency personnel including webmasters and public information officers;</li> <li>• Illinois Librarians, including government document librarians;</li> <li>• ISL and University of Illinois at Urbana/Champaign (UIUC); and</li> <li>• State government information seekers</li> </ul>

<b>Organizational Mission</b>
<p>It is the mission of the Illinois State Library to provide state government officials and employees with the information they need to make informed decisions as well as to develop and promote libraries in order to enrich the quality of life for the people of Illinois.</p> <p>Inherent in this mission is the State Library's advocacy of the right of Illinois citizens to read and have full access to information.</p> <p>The mission is accomplished by sharing library resources through the Illinois Library and Information Network.</p>

<b>Program Purpose</b>	
Summary of key proposed services	<ol style="list-style-type: none"> <li>1. Refine existing methodology for permanent retention and public access to state government electronic publications;</li> <li>2. Facilitate the replication of this methodology by other states; and</li> <li>3. Create increased access to Illinois government documents through MARC records.</li> </ol>
Target population	Illinois State Library; other state libraries and state archives; citizens seeking information; and Illinois librarians
Target outcomes	<ol style="list-style-type: none"> <li>1. Other state libraries and archives will implement the system.</li> <li>2. Illinois Document Depository librarians will use the public access of the ISL system.</li> <li>3. Other state libraries and archives will use the system.</li> <li>4. MARC records will be created for Illinois state documents in electronic format.</li> </ol>
<b>Planned Program Activities</b>	<b>Progress in the third quarter of the grant period</b>
<i>Creation of the metadata generator</i>	The metadata generator has been refined. Mechanisms for the depository system to communicate to the generator system have been developed to coordinate the systems through the ingest process.
<i>Hiring and contracting staff</i>	A technician has been hired in ISL Technical Services, who will edit and approve agency metadata and create metadata for the deposit of retrospective electronic documents. New contracts are needed for fiscal year 2006.
<i>Change monitoring and caching e-publications</i>	Change monitoring and caching e-publications is ongoing in Illinois for Illinois, North Carolina and Wisconsin Web publications. Implementation has been completed in Utah, Arizona and Alaska, so change monitoring and caching of

	state Web publications is ongoing in those states.
<i>MARC cataloging</i>	Serious health problems have sidelined our cataloger for this project. Alternatives are being investigated.
<i>Maintaining server list</i>	Maintenance of the list of agency servers is on-going in Illinois, North Carolina, Wisconsin, Alaska, Arizona, Utah and Montana.
<i>State agency training</i>	A training session is scheduled for late April. Illinois Secretary of State employees will attend the morning session, including ISL employees. ISL Technical Services staff will attend the afternoon session.
<i>Recruiting other states</i>	All six participating state libraries have implemented the system in some form. Other states that have shown interest: Arkansas, California, Colorado, Connecticut, Kentucky, Missouri, Pennsylvania and South Dakota
<i>Mentoring other states</i>	Support in terms of encouragement, technical advice, hosting some state systems and sharing via the steering committee continues as needed. Members of the steering committee have been very helpful to each other.
<i>Coordinating steering committee</i>	The steering committee was unable to find a single date and time to meet, so a series of conference calls took place March 8-11 with a sharing of notes with everyone.
<i>Keeping statistics</i>	Mechanisms for computer statistics collection have been developed and installed on Illinois, Wisconsin, North Carolina, and Alaska. Latest Reports: <b>Alaska</b> <a href="http://pep.library.uiuc.edu/~cep/stats/AK/LatestStats.html">http://pep.library.uiuc.edu/~cep/stats/AK/LatestStats.html</a> <b>Arizona</b> Statistics package to be installed in the fourth quarter. <b>Illinois</b> <a href="http://history.lis.uiuc.edu/~cep/stats/IL/LatestStats.html">http://history.lis.uiuc.edu/~cep/stats/IL/LatestStats.html</a> <b>Montana</b> <a href="http://depo.library.uiuc.edu/~cep/stats/MT/LatestStats.html">http://depo.library.uiuc.edu/~cep/stats/MT/LatestStats.html</a> <b>North Carolina</b> <a href="http://depo.library.uiuc.edu/~cep/stats/MT/LatestStats.html">http://depo.library.uiuc.edu/~cep/stats/MT/LatestStats.html</a> <b>Utah</b> <a href="http://pep.library.uiuc.edu/~cep/stats/UT/LatestStats.html">http://pep.library.uiuc.edu/~cep/stats/UT/LatestStats.html</a> <b>Wisconsin</b> <a href="http://pep.library.uiuc.edu/~cep/stats/WI/LatestStats.html">http://pep.library.uiuc.edu/~cep/stats/WI/LatestStats.html</a>
<i>Reporting and evaluation</i>	This document constitutes the third report using the OBE method.

<b>Planned Program Services</b>	<b>Progress in the third quarter of the grant period</b>
<i>Metadata generator</i>	The issues of communication between the generator and the depository appear to be resolved. Final testing is expected to occur before the end of April.
<i>Website</i>	<a href="http://www.cyberdriveillinois.com/departments/library/who_we_are/cep.html">http://www.cyberdriveillinois.com/departments/library/who_we_are/cep.html</a>
<i>Agency training sessions and materials</i>	Web-based training tools for Illinois have been modified in response to feedback from the pilot agency session and amendment of the generator. Agency training has been delayed because technical development required more time than estimated.

	The CEP project manager has requested and received an amendment to extend the grant period for another year for the sole purpose of agency training.
<i>Mentoring</i>	ISL Illinois Documents Depository Program ( <a href="http://www.cyberdriveillinois.com/departments/library/what_we_do/ildocdep.html">http://www.cyberdriveillinois.com/departments/library/what_we_do/ildocdep.html</a> ) staff members are eager to use the database of agency contacts for publications in both tangible and digital formats created for the grant. They expect it to be helpful in locating paper publications as well as digital versions.
<i>Steering committee meetings</i>	Conference calls for representatives from every participating state have been held to share experiences and discuss issues or questions. Details follow.
<i>Refinements to the software</i>	The statistical package has been developed and is operational in Illinois, Wisconsin, North Carolina, Alaska, and Utah. Arizona is scheduled for software update following their next tape backup of data.
<i>Implementation in other states</i>	Illinois is currently hosting North Carolina, Utah and Wisconsin's data and doing monthly up dates. Implementation has been completed in Arizona, Alaska and Montana.
<i>Reports and presentations</i>	A presentation to the Annual Illinois State Library Government Documents Conference is scheduled for May 6. Project staff have accepted an invitation to present to GODORT at the annual ALA meeting in June.

<b>Target Population</b>
<ul style="list-style-type: none"> <li>• State libraries and state archives charged with preserving state government documents in electronic formats;</li> <li>• Citizens seeking state government information;</li> <li>• Illinois librarians using state government information to meet patron needs.</li> </ul>

**Outcome #1** Other state libraries or archives will implement the system.

Indicator	Data Source	Applied to Whom	Data Intervals	Target
Number of state libraries and or archives that complete two Web harvests of their state's documents and cache the resulting files.	Project staff will contact participating institutions by email or telephone.	All institutions that return signed participants' agreements.	Three months after system implementation	At least three of the participating institutions

Progress:

- Utah has operated their PEP system since December 2004. Utah requested the hardware requirements for the PEP system, and intends to sustain the program after the grant period.
- Montana began monthly operations in February 2005. Montana committed to evaluating the safety net system and making decisions about hosting their own system at the end of the grant period.

- Alaska’s first month of operation occurred September 2004 and every subsequent month. Hardware was purchased expressly for this purpose. Alaska plans to continue operating the PEP system after the grant period.
- Arizona’s system has experienced successful harvests, change monitoring and caching every month beginning July 2004. Arizona plans to maintain the PEP system after the grant period with support from the grant steering group.
- North Carolina, whose system is hosted at UIUC, has successfully operated on a monthly basis beginning August 2004. Sustaining the program in North Carolina after the grant period is in doubt. One agency asked about preserving the state Web. When told about this project, the agency representative thought it was great. While the Library continues to work toward a plan to sustain the program, the IT department has not seriously considered it, citing “security issues.” Having NC IT talk to Larry didn’t seem to help. Daniel Cornwell, who personally runs the Alaska system, offered to talk to them.
- Wisconsin, whose system is also hosted at UIUC, has successfully operated on a monthly basis beginning November 2003. Wisconsin is unable to devote resources to a PEP system after the grant period. However, they are looking at uses for the data gathered during the grant.

**Outcome #2** Other state libraries and archives will use the system

Indicator	Data Source	Applied to Whom	Data Intervals	Target
Number of state libraries and archives that place access to the system on their website	Project team verifies with each state by looking on the web	All institutions that return signed participants’ agreements.	At the end of the grant period	At least two of participating institutions

Progress: Access tools remain to be developed during the remainder of the grant period. However, some states are already finding the system useful.

**Alaska** checks for monthly differences in PEP harvests to know when to contact agencies about new publications. They have found that close to half of the Web publications are still printed, and agencies will send paper copies when reminded.

**Arizona** is feeling the need for the access tool scheduled for development in the remainder of the grant period.

**Illinois** uses the PEP harvest to populate its state government search engine, <http://findit.lis.uiuc.edu/cgi-bin/search.cgi>.

**Montana** is evaluating the system reports for possible uses.

**North Carolina** plans to analyze their data to determine usefulness.

**Utah** has been pleased with the system reports. The Library is getting a new search engine for the “live” Web that they hope will index the safety net files as well.

**Wisconsin** discovered Web sites previously unknown to the library because of the PEP system. They hope to get previously Web-published documents from the data gathered by PEP software to deposit or use the PEP software as a harvester.

**Outcome #3** Illinois Document Depository librarians will use the public access of the ISL system for patron information needs.

Indicator	Data Source	Applied to Whom	Data Intervals	Target
Number of Illinois Document Depository	Email survey	Illinois Document Depository Libraries staff	At the end of the grant period	55% of the Illinois Document Depository

Libraries which have staff who use the public access interface				Libraries
--	--	--	--	-----------

Progress: The Illinois Document Depository Libraries will meet at ISL in early May. They will be polled as to use of the ISL search engine based on the PEP system harvest and for their feedback on the new electronic depository. The depository interface will be demonstrated.

**Outcome #4** MARC records will be created for Illinois state documents in electronic format.

Indicator	Data Source	Applied to Whom	Data Intervals	Target
Number of repository URLs in 856 fields of ISL MARC records	Bibliographic database report	ISL catalogers	For each grant report after the first	50

Progress: Final testing is scheduled for mid-April. Training is scheduled for late April. Despite the illness of the assigned cataloger, ISL still is hopeful that this goal may be met by the end of the grant period.

**Alaska** has found the PEP harvests useful. In analyzing Alaska's PEP Spider results for 82 spiders in December alone, sixteen previously unknown state publications from seven agencies were discovered. In addition, URLs were identified for five publications in the print collection that did not have a note of Internet availability in the catalog.

Examination of the January harvest uncovered sixteen previously uncataloged documents and URLs for two existing paper items. The library sent requests for paper copies of all sixteen titles. The first response indicated that one title, *Alaska Employer's Handbook*, was available only via the Internet for budget reasons. If other documents are not received, they will be printed by the library and cataloged. Extensive use of ColdFusion is found in Alaska State Web Sites. These tend to be very large files, which drives up the statistics

The Alaska State Library experienced an unfortunate flood due to burst pipes. During clean up, Daniel Cornwell realized that he was keeping one back-up tape in the room with the computer and the other in his office in the room right above the computer room. Only luck prevented all copies being ruined. He advises that all back-ups should not be stored in the same building! Alaska's compressed repositories currently take up 35GB.

While the search tools are under development for the repositories created by PEP software, Alaska has found an interim solution. Using the Samba software that came with their Redhat Enterprise Linux operating system, they have created a READ-ONLY shared volume consisting of just the files in the compressed CVSRepositories (Controlled Versioning Software) directory. Daniel has also created a new user on the Linux server, called "searcher", so the searching activity can be distinguished from regular management activities. This allows the PEP/CEP server to be accessible from the Network Neighborhood within the Library, but not in the wider Internet.

Library staff can use Window's native search function to search the repository by file name or containing text without having direct access to the PEP/CEP server. The repository can also be browsed.

If staff identifies an HTML or text file that fits their criteria, they can look at it unaided. If they find a binary (Word, PDF, etc) file that looks promising, they need to request a "CVS checkout" of the entire spider's Web site so the specific file(s) they are seeking can be emailed to them. The date of the CVS repository to check is determined by the date and version information within the file.

So far, the experience has been that searching the entire repository for a file name takes five minutes or less. Searching the entire CVS archive for a "containing text" match does not seem to be feasible. Searching a single spider's archive for "containing text" does work and seems to take between three and ten minutes, depending on the size of the archive. The staff users only have READ-ONLY access to the files.

**Arizona** is happy with the system, but reports that it has not been easy to host their own system. While Arizona will be able to manage and maintain their system with little to no outside help after the grant period because they ran their own system from the beginning, the lack of UNIX experience has been challenging. They hired a UNIX consultant, who has proved able and competent. Certain issues involved in the system are unfamiliar to him, however, especially adding a data back-up mechanism.

The tape-drive back-up mechanism has now been installed and has proven effective. The system is now ready to be upgraded and statistical functions added. Richard Pearce-Moses expressed appreciation for the group because of the mutual support.

Arizona has been feeling the need for the access tool scheduled for development in the remainder of the grant period. Richard is one of the prime movers in the NDIIPP ECHO grant project, and has plans to employ the tools under development in that project to manage the PEP system-harvested files from this project.

**Montana** shared some statistics, especially Web site growth, from their online reports with webmasters, who probably don't receive that kind of information from other sources. Jim Kammerer asked a number of state agency webmasters for their comments about the monthly web statistics, specifically what they liked, disliked, and their suggestions.

While he liked the idea of archiving the websites with dependent files for archives sake, one person thought the statistical reports should be "tweaked" a bit to make them more useful. He wished that the reports would tell the number of updated or changed pages. He worried about a substantial load being placed upon the web server by a spider that aggregates content. (However, no such problem has been experienced unless bandwidth was insufficient for public viewing.)

The state has had a change of administration, so Web sites have undergone massive updates. The transition period is not over because the Montana information technology office announced plans to put uniform headers and footers on all of the State Web pages. Those additions will show up as huge statistical change in the harvested files without content change.

Another Montana webmaster remarked that the statistics do give some information that she did not usually get from her web trends reports, like number of pages, and types of pages. She wished the formatting for printing were better. Otherwise, she thought it useful--maybe on a quarter or semi-annual basis-- and that she would continue to be interested in seeing those reports. Based on her interest, Jim posted the URL for the reports on Montana's State Publications Center website.

**North Carolina** appreciates the system as a safety net against losing information, but they are not publicizing it because sustaining the program after the grant period is in doubt. A number of barriers with the implementing the PEP system have materialized. Because North Carolina has a large, decentralized state government and a small IT staff within the Department of Cultural Resources, neither the State Library nor the state itself has the resources or the expertise that Illinois enjoys through Larry Jackson at the University of Illinois. There seems to be a gap between what Jackson can accomplish at the University of Illinois and the ability of state institutions in North Carolina to replicate that process within state government. This may be an issue for some other states as well.

North Carolina continues trying to work out ways to have the PEP system up and running in that state after September 2005, when UIUC will no longer be hosting their system. Kristin Martin is gathering some more information about the size of North Carolina websites for the IT department. The State Library is currently undergoing a building renovation, which is planned to include a refrigerated server room. IT is not eager to set up a new service until after this change. The renovation most likely will not be completed before the end of September, after which the University of Illinois no longer will support crawling NC websites and hosting their archive, so the library will need a plan to transfer the existing cached files to North Carolina.

One agency asked about preserving the state Web. When told about this project, the agency representative thought it was great. Another agency webmaster who learned about the project said, "It sounds like you've got a great project going utilizing an effective tool that saves time, money and resources. I hope this project works out--for everyone's sake!" While the Library continues to work toward a plan to sustain the program, the Information Technology Department has continued to cite "security issues."

Larry Jackson reported that he was contacted and asked the sizes of the NC agencies' Web presence. The caller was working with the NC mainframe people to set up the CEP/PEP hosting. However, this conversation didn't seem to help. Daniel Cornwell, who runs the Alaska system, offered to talk to them

**Utah** has been pleased with the reports from their PEP system. They have requested specifications for purchasing the necessary hardware to continue maintaining the program after the grant period. While Utah has a new Information Technology Department, there have been no indications that the PEP system would not be supported.

**Wisconsin** has never viewed hosting a PEP system as a practical possibility. They simply don't have the personnel or other resources necessary. However, Sally Drew reported that working with GSLIS has been great. Not only did Larry and his graduate assistants operate the system for Wisconsin, but they also helped identify Web sites previously unknown to the State Library. Sally hopes to get past Web-published

documents from the data gathered by PEP software to deposit or use the PEP software as a harvester. She plans to arrange digital long-term storage with the OCLC Digital Archives.

## **Statistics**

The CEP Web site, which is located at [http://www.cyberdriveillinois.com/departments/library/who\\_we\\_are/cep.html](http://www.cyberdriveillinois.com/departments/library/who_we_are/cep.html), lists participating states. All, except Arizona, have links to the corresponding statistical tables. These tables list each state Web site harvested by agency, Web site size, file types found, and metadata type and quantity used. Arizona statistics will be added when the statistics package has been installed on their software suite. To understand the advantage of the PEP software, please find a chart at the end of this report section illustrating the growth of Illinois Web sites and the rate of growth in CVS storage of the files harvested by PEP spiders.

Each month the system sends out “spiders,” which are browsers that copy and return all the files from each targeted Web site. This is called a harvest. The first month a website is processed, the CVS copy is a slightly formatted version of the full inventory downloaded by that spider. In the subsequent months, the spider again harvests a complete copy of the website, which is then put into CVS.

However, CVS compares the new version against what it already holds—what we refer to as “change-monitoring.” Then, for text files, instead of copying in the entire new version, CVS either (1.) notes that a file did not change, thus hardly increasing disk consumption at all, or (2) records a series of editor commands sufficient to change its old version to match the text of the current version. For changed binary files, the entire new version is stored along with previous versions.

Therefore, the size of the CVS cache of a website starts out basically the same as the first or baseline size of the first harvest. In subsequent months, the CVS size only slowly increases with each harvest of the website because the increase is basically due to the amount of changed material within the website, not the total size of the website. So, if a website does not change, there is no increase in the cache size. The second chart emphasizes the space thus saved by storing only the changes to Web sites, not all of the files each month.

## **Challenges**

Server issues have hampered progress by absorbing the time and energy of our principle investigator, Larry Jackson.

In mid September 2004, a hard disk within the RAID array of the UIUC-hosted PEP system failed. Because this RAID was redundant, operations continued as before, and no one noticed. A new machine, an Apple X-Serve computer, acquired in July, was just completing local RAID testing.

In late September 2004, a second disk within the RAID array of the PEP system failed. This exceeded the redundant processing capability of the RAID, so the system halted. Repairs were performed using identical spare disks that were purchased at the time of acquisition of the RAID array. Western Digital replaced the failed disks under warranty, at no charge.

Upon completion of repairs to the hardware, the process of restoring the PEP data for Illinois, Utah, and Wisconsin from backups began. During this process, two backup files on another machine, one of the offsite backup computers backing up Safety Net archives of all States, could not be read, apparently due to disk faults (e.g., bad sectors), necessitating use of another backup copy from a third machine. Backup copies were shuffled to other locations and re-generated, before any other failures could occur.

As all the data had to be restored; CEP promised to include still further Web site archives for other states; and one of the computer vendors, Apple, promised to make 2.9TB of additional disk available very soon, Jackson decided to relocate the Illinois Safety Net processing to the Apple computer. All of the Illinois data was ported. The PEP software required a few changes due to the difference between Apple's version of BSD UNIX and RedHat Linux. In some cases, the changes have caused the Illinois software to diverge from the multi-State deployment of the RedHat version.

Apple's use of case-insensitive file systems, which is thoroughly untypical for a UNIX operating system, was not recognized until too late, rendering the IL Safety Net archive case-insensitive from this point on. Upon reflection, this is probably not a problem, though the existing means of case substitution based on Apache's ModRewrite module will probably have to be greatly amplified upon if these archives are to be publicly browseable.

In October 2004, a GSLIS student noted that running ever-increasing processing jobs on the Apple server triggered multiple reboots. While growing, these processes were still very small, in comparison to project processing on other machines. As CVS work builds on former jobs, failures can give rise to very great amounts of reprocessing. Backup files are necessary to support the reprocessing in case of failures like these.

When GSLIS could not locate the cause of the problem, contact was made with Apple. The default limit of processes on the X-Serve is that of a desktop Macintosh -- extremely low for a rack-mounted data server with multi-terabyte disk space. That limit was raised, but, upon further testing with slowly increasing load, random reboots continued. Apple advised where to look in log files for hints at a cause, but those log files contained no trace messages at all.

Because of these difficult conditions, the primary investigator, Larry Jackson, took over Illinois processing from the student, and individual websites were processed using command line activation of programs. Further, as random reboots were frequent, archive copies of data had to be made prior to attempting each job. If the job succeeded, the just-produced copy could be discarded. If the job failed, the just-produced copy was used to re-create the website's files, and the job could be attempted again. This highly manual processing, conducted roughly 16 hours a day, was just able to keep up with the retention of IL spider-harvested data. Assembling a complete system capable of producing statistical reports and DVD backups was becoming impossible.

Jackson placed a rush order for a computer with an operating system known to be able to handle roughly 1000 PEP processes simultaneously. Illinois spider jobs are run on the original machine, and the data then moved to the Apple machine for CVS retention. Further software installation work on the Apple computer was cancelled.

In early January 2005, a disk in the RAID array of another computer failed, in part due to cabling and connection problems of the enclosure holding that disk within the larger cabinet. This machine is the PEP system host for Montana and North Carolina as well as the host for Illinois' Electronic Depository Initiative, in which files of state electronic documents are assigned persistent URLs for cataloging records. This computer is located in a machine room at the UIUC Main Library, and is operated by the Library Systems Office.

RAID redundancy was not able to prevent loss of data, so the entire CEP processing of North Carolina, which had been completed only hours before, was lost and had to be regenerated. One very large website did not restore correctly from the \*.tar.gz copy, so the uncompressed DVD copy had to be used, with the loss of the data for changes that occurred after the DVD was burned. Hardware repairs were performed using identical disks, as before. Data was shuffled, and backups re-done. Montana processing, only started in November, was not yet on an automated backup schedule, and so lost one agency's data for its December spider completely. This disk was out of warranty.

Work began on installing yet another computer, a new RedHat Linux computer with an Apple 2.9TB RAID attached, in a machine room already strapped for electrical power and air conditioning capacity at GSLIS.

In February, operations began on the new machine, including moving the data via a private gigabit network. However, the new machine began to lock up under what was identified as times of heavy input/output load. The extremely high bandwidth of the fiber channel I/O card of the Apple X-RAID is found to have a design deficiency in that only a tiny fraction of the heat sink is actually making contact with the chip it is intended to protect. Inspection of several such cards at GSLIS produces identical results. GSLIS rigged an improved heat sink mounting, and ordered a superior heat sink with integral fan.

Apple, under some pressure from the Dean of GSLIS concerning the continued inability of the Apple machine to support processing loads typical of even a simple RedHat workstation, found that a sub-function of the screen saver initiates system reboots if less processor time is available to it. The screensaver concludes that the system is stalled. So, when the multi-terabyte server is computationally busier than typical for a Macintosh desktop, the screen saver reboots it. This function is shipped with the default as "on." Further the screensaver software makes no log file entry saying that the reboot was deliberate, or which program initiated it. All of this was particularly ironic and frustrating because no screen was being used.

Apple explained how to disable this function, and the GSLIS crew subsequently successfully operated the Apple computer to workload levels typical of PEP processing. Roughly four months of very extensive manual processing was the cost of this strange, and unknown, sub-function of a screen saver.

The Apple X-RAID continued to experience heat-triggered lockups, and GSLIS continued to devise methods to provide more and cooler air to the card. Progress was slow, but steady. Ultimately, GSLIS re-mounted the computer in a cabinet twice as large, and then re-mounted the fiber channel card perpendicular to the motherboard, allowing it to receive far more cooling. Processing of the largest IL websites ran

successfully. CVS processing throughput appears several times faster than on other machines. Jackson transitioned Illinois processing back to a student in early March.

The computer, which proved inadequate for processing needs, will no longer be used for CEP processing. Instead, plans have been made to transition the offsite backup functions to it. The computer with the cooling problems was moved to a new machine room at GSLIS with 65,000 additional BTUs of cooling capacity in March.

### **Lessons learned**

Although RAID arrays can be configured to recover from loss of one disk drive without operator action, they cannot recover from loss of two drives, even with operator help. Accordingly, the operator must detect the first failure, and accomplish those repairs, before the second failure occurs. To detect this failure, the operator needs to monitor log files at least daily, or, preferably, to institute an e-mail notification system.

At the I/O bandwidth the system can generate, that is, after completion of the harvesting phase, in post-spider processing or in system backups that move all the data as fast as possible, normally serviceable I/O cards, disks, and disk RAID arrays may well overheat and malfunction. Our final configuration has replaced the original rack-mounted cabinet with a double-sized unit, and turned the overheated I/O cards perpendicular to the airflow for more direct contact between the air stream and the card. Further, the manufacturer-supplied on-card heat sink was first re-mounted using traditional methods, rather than the peel-and-stick double-sided tape of the original, then replaced by a unit containing an integral fan. Hardware may not have been designed for the level of loads needed for this project.

Multiple backups are mandatory. A backup copy should not be assumed to be 100% perfect, but that it will have some percentage of flaws. Further, it is best if the backups are not all generated with the same technology (e.g., gzip), in case that technology should prove incapable of handling some forms of data (e.g., gzip has occasionally been observed to fail at file sizes of over 10GB, and to frequently fail at 20GB). Further, the backups should span more than one generation of the data. For example, if all the backups are copies of the current data, and the current data itself somehow was corrupted before the backups were made, there is no recourse. To this end, we reinstated the making of DVD backups, every other month, for every state, and storing these in another location. Further, periodic attempts to recover data from the various backups is prudent, to verify that the backups are truly serviceable.

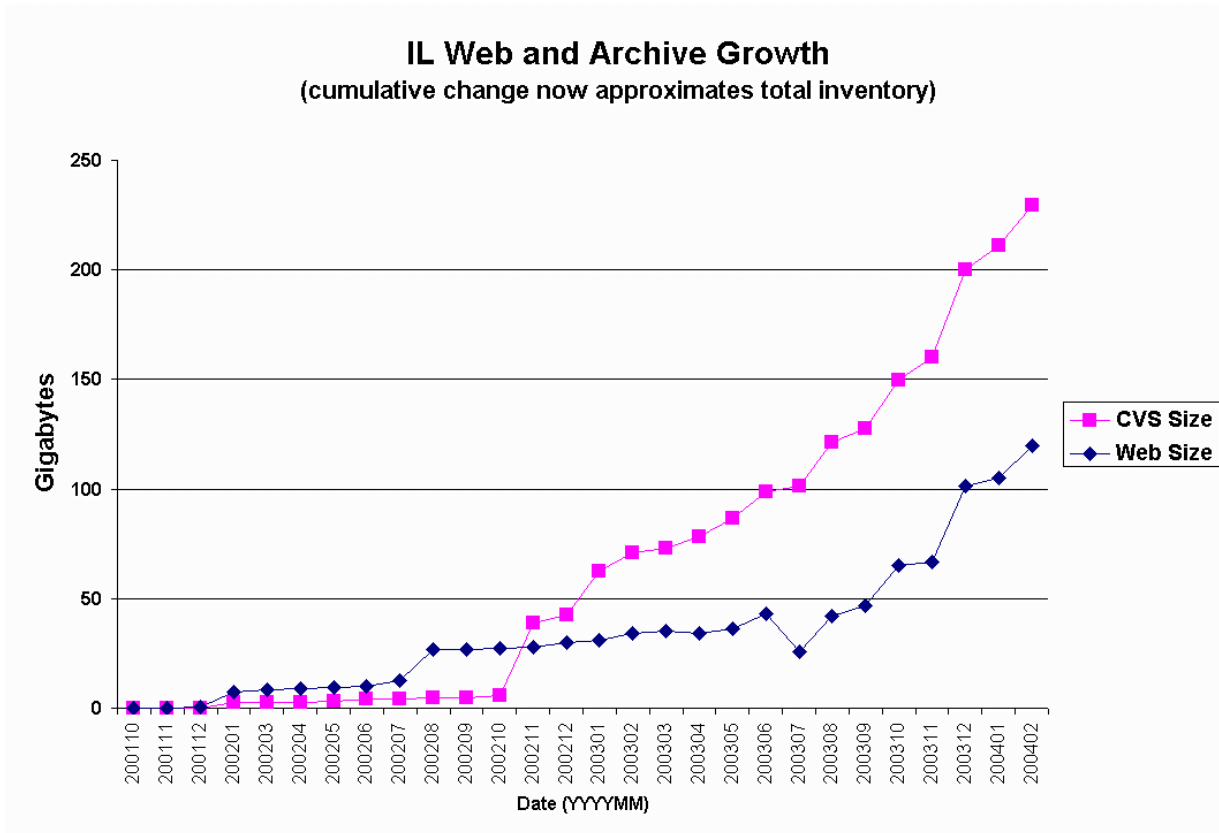
All data, whether originals, backups, or programs, stored within a single computer can be simultaneously lost upon certain types of problems with that one computer. All data stored in one physical location (typically "room", but, as Alaska's State Library has shown, this can also mean "building") can be simultaneously lost upon a physical problem that affects the whole area e.g., fire, flooding, burst pipes, tsunamis, hurricanes, tornados. Backups should be stored at multiple physical locations. GSLIS uses 4 machines in 3 rooms in 2 buildings for backup functions, plus DVD backups stored in a third building. A replacement computer can be purchased, but the data can only be restored from backups that survive.

Time between gathering the original data and making the backup is time during which the data cache is vulnerable to the loss of the only copy of the data. We modified the

project software to immediately produce the first off-site backup copy immediately upon completion of the CVS deposition phase for a given website.

When a failure occurs, one of copies of the data no longer exists. If, initially, only two copies exist, that means there would now only one copy of the data, and if that experiences a fault, the data is irretrievable. So, best practice would be to keep more than two copies at all times and/or generate another backup copy as fast as possible when one backup is lost.

From previous RAID problems, we know that the software supports distribution of data over multiple disks. In most cases, a RAID card is not required, though it is usually easier to administer one large data area than multiple smaller ones.



### IL Web Without CVS (capacity required for a browseable archive)

