

Preserving State Government Digital Information California Partners Meeting



Minnesota Historical Society

Tuesday, March 25, 2008
Bion M. Gregory Conference Room, 925 L Street
Sacramento, California

The Minnesota Historical Society's Nancy Hoffman, Bob Horton, and Jennifer Jones met with California partners for the NDIIPP-sponsored project, Preserving State Government Digital Information.

Participants included Stephen Abrams and Tricia Cruse (California Digital Library); Bill Behnk, Diane Boyer-Vine, Abby Cole, Linda Heatherly, Mendora Servin, and Daniel Zavoiu (California Legislative Counsel); Chris Garmire and Nancy Lenoil (California State Archives); and Kris Ogilvie (California State Library).

DZ: Question on cost/benefit analysis of e-legislature.

BH: We looked at four options – SDSCC off-site broker; in MN, OET 2 options, 1: disaster recovery option and 2: standard storage within SAN; also, MHS SAN, and what is going on at Revisor's Office, but limited there.

Discussion about SDSCC and the problems that we ran up against with e-legislature; SDSCC is abandoning the Storage Resource Broker (SRB) – it was expensive, above and beyond everything else you are doing and SDSCC abandoning it; next IRODS, let's you select certain routines but someone still has to write the routines.

TC: Mentions that SRB comes out of Research Side.

DBV: Does Washington State have something? Involved with Microsoft.

BH: Using Biztalk, have a wide-ranging digital preservation program; they got a huge amount of money up front from their legislature, & Microsoft threw in lots of free help; they hope to roll out something that is portable or distributable; it does work.

MS: Asks about progress on the Bush Proposal.

BH: Mapping has taken longer than expected; in implementation phase; should begin working with local partners and ND and SD partners later this spring.

DBV: In MN worked with bills because available, what about other material?

BH: The bill drafting system can accommodate the other material; it can be captured in the system; House and Senate have different approaches; House research closer to Revisor's office; Senate less so; some material simply going to be outside the system – all the web stuff; audio tapes/video now in digital form and being web cast; very complicated to include in this system, material essentially 'ungoverned'; legislators don't want to preserve audio/video of floor sessions for purposes of legislative intent.

DBV: Where is California going in terms of this kind of thing?

BB: Don't talk about archiving/preservation, talk about disaster recovery to sell it to the legislators.

BH: Ties to funding sources; can't sell preservation, but can sell use and disaster recovery.

DZ: Microformats answer to making data available in a wider way.

BH: We can map out to other kinds of legislative content, bill, the act, the newspaper article, etc.

DZ: Dynamic collaboration map (?), tools and strategies, rules for engagement.

BH: Basecamp is the tool; action items develop state by state.

General Conversation – whatever people want to talk about, how we might work together.

DBV: Looking for \$14M to upgrade legacy data but instead got a \$8.8M cut; starting to get codes into XML dealing with legislative history system.

MS: Two projects: LIS project, developing some internal research capabilities; the other is a standard web site portal, put all public information out on web sites; taking 5 web sites and in the process of redoing them. Trying to standardize all web sites and content; add search capability; includes redoing legislative web sites that have bills and other legislative documents – a multi-year initiative; architecture is pretty much done, MS in charge of it; bill drafting was done.

BB: Vignette is the product, comes with IDOL search engine.

DBV: Getting old opinions going back to 1928 (onion skin!) online and searchable; a lot since 1982, but hard to scan onion skin; looking at some outsourcing.

AC: Interested in Omeka tool for selection and annotation [<http://omeka.org/about/>].

MS: What worked and what didn't?

BH: MN150 example; nobody comes to share memories.

BB: MS will build a new legislative site by end of the year; top 500 web site; started in 1993, still going strong with 1993 software; people use it; content is very accessible.

KO: State Library, staff changes; a few projects: California State Publications, with UC-Davis, making accessible governor's executive orders; fish & game; constitution documents; sent to OCLC to scan; hoping to get second-year grant to use ContentDM to figure out how to make accessible to the public and where it might go, where will it go after the grant? Probably come

to the state library; will be using ContentDM to provide access to items in the OCLC Digital Archive; using OCLC Digital Archive for preservation, that is splitting into two separate entities; Digital Archive only for dark archive preservation; if you want access, have to use ContentDM; have about 5,000 objects, only one cataloguer to do everything, digital and paper; government documents folded into a new group; unclear where digital items will end up; for now, KO is handling it, and it is evolving. Observer on eLeg, not an active member. Entering a pilot project with UC-Fresno using LOCKSS protocol; focus on items normally catalogued anyway. Couple of other things: accessibility issues, state-wide group focusing on this, headed by undersecretary of what was Health and Human Services and the state CIO, will be creating standards and guidelines to ensure that web sites will be acceptable; looking at PDF in particular, because of a lawsuit brought against Target Corp, based on ADA because of inaccessibility of their website. Also an eServices web group; own the content on the State of California website; also State Web Accessibility Team (SWAT); because accessibility standards are being looked at and changed; lots of collaboration; one collaboration is a best practices web site with State and Consumer Services Agency to launch on web site on April 14 (few other partners); best practices for IT, consumer services, and Green California – link to a Federal wiki and allow state employees go in and add info and best practices for state employees only, state library part is already going; State Archives had a marketing plan before they even started, announcement on Govs blog, letter from gov, not on paychecks; spreading the word in various presentations and meeting. Also, user website evaluation; involved in web 2.0 stuff, how they can use it in their libraries, etc. All of this done with collaboration to get everything done.

NL: Difficult context of three elections, State Archives, beginning to see state agencies identify electronic records on their retention schedules; starting to have talks with these agencies; no consistency among state agencies; legislation that requires General Services and Sec. of State to develop standards, but there has not so far been real pressure to do that; concern that with inexpensive storage that agencies will keep records rather than transfer them PLUS SA does not infrastructure to receive the records; unlikely that in this budget climate that they will get some kind of digital content management; website “Learn California.org” different topics relating to CA history plus exhibits; nearly ½ million hits in January; averaging 360,000 hits a month; work on web page to make it ADA compliant; launch of online catalog Minerva; still have a number of collection on OAC, but processed records; but Minerva handles things through the entire process, before it arrives, through processing to access by researcher; anyone can see anything whether processed or not processed.

CG: Geospatial, involved with SDSCC with State Library on NHPRC grant to experiment on managing, distributing geospatial records; different agencies can contribute layers for existing maps; interesting archival opportunity, multiple record creators in dynamic GIS system, challenges in appraising and accessioning; possibility of using web 2.0 to have users tag items that might be able to give some feedback about the process; RSS feeds, to notify if different agencies create new layers; way of creating these appraisal teams around the 2.0 technology; use social networking tools as part of the appraisal process; will be using the Storage Resource Broker and some iRODS [SRB's new data grid software system], the latter will facilitate automatic ingest; in the experimental phase, also dealing with the three elections.

BB: Bob said SDSCC expensive; B. has always thought they were freebies; but not anymore.

BH: Their pricing has become more expensive and comprehensive; every service is priced out.

TC: Need to deal with production side of the house first.

DZ: Have we looked at any other online service?

BH: No; MN bandwidth primitive; makes more sense to transfer using portable hard drives; but serious limitations of network prevent production phase.

NL: Keep alluding to 3 elections; the IT folks put a freeze on all but emergency needs 60 and 30 days out from election; very tiny windows to get things done. As far as budget goes, SA OK. Fully staffed, constitutional office, permanent staff of 28, not put in the position of absorbing big budget cuts. Only four divisions in Sec. of State's office that are general fund; none have staff of more than 40. They will be looking to other places to cut. SA will maintain what they are doing without having to reduce activities or services.

TC: CDL has five separate programs: digital special collections, mass digitization group, bibliographic services team, scholarly publishing group, and TC's digital preservation focus. Just finishing 3-year project (NDIIPP) to create a web archiving service, just moving to production in June, calling it a soft release with partners; working with 40 curators at 15 institutions around the country but mostly CA; using a Heritrix crawler and putting a front-end based on how a librarian or curator wants to collect information; curator can sit down and say I want to collect this agency's web site every day for six months; will provide tools to provide info on what changes; putting some preservation services around this web content. More money from LC for next phase: what access is going to look like for that content; right now access is for curators and the people who are going to use this material. Tools to look at the data in the aggregate, trends, etc. Part of work, how can we use more partners for using web archiving service, useful not just to UC system but also the State of California. Working with archivists in the state of California. Not just librarians but also archivists. Lots of learning tools, YouTube videos, etc. Preservation program 5 years old. How much is it going to cost, how will we sustain it? A question for everyone. Need to systematically understand costs, taking it in, data consultancy, where we might leverage partners to make storage cheaper. What are the preservation services? Not just storage, and what is it going to cost? Moving big chunks of content around. Working with LC and SDSCC. Working on how to break down barriers to move big chunks, what kind of wrapper to put around it so it can make sense to the receiver "bag it and tag it"; haven't had a lot of success working with content providers; never got to the single person to authorize taking it; decided to grab publicly accessible data on the web.

DBV: Versioning is a huge issue.

TC: Careful to put dates up, but tell people that they are not at the agency's web site; put responsibility on the researcher; hopefully agencies will be responsive and label content, but it seems they don't necessarily per DBV's experience; on web archiving, can't solve every problem first or we'll never get off the ground, so issues need to be solved as they go.

BH: Interesting problem, hope we can work with them. Example, GIS community, evaluation of users to figure out lowest common denominator of metadata.

TC: If we set standard for metadata too high, we are going to lose content.

SA: Potential for automatic extraction of metadata from the content; working on a project to extract technical metadata; this is an active research area in computer science.

TC: SA's project is a big project to deal with something like this. With mass digitization, loaded with government documents that are going to be publicly available.

After Lunch Conversation (*go through this and ID the deliverables/initiatives*)

BH: Picking up on the conversation, some common themes that emerge, big picture items; then try to come up with some action items: collaboration, access/use, accessibility, tools to enhance value such as social tagging, search tools (Idol), audiences, authentication, sustainability—particularly in terms of costs and benefits to justify funding.

DZ: Microformats and metadata for search; should talk about deliverables for the project and how the partners contribute to those.

NL: Cost benefits; not just the costs but you have the capacity to do it as well.

BH: Part of the sustainability – ongoing costs and initial investments.

NL: Need capacity to do it and capacity to sustain it.

BB: Are you finding people are less inclined to buy in if the costs are too high and not sustainable?

BH: We phrase it as a transition from project to program, which is complex.

BB: Private sector efforts (Google and Microsoft), and how those efforts tie in, whether they can be leveraged, Thomson-West is another one.

TC: Cost of storage, but we need more and it stays really expensive; thinking about private sector and its capacity in this area; also SRB, thinking about what its functions are and ultimately, what is preservation.

BH: LC is really interested in access, so one deliverable is something that is a tangible online product that provides some enhanced access to the kind of content we can pull from legislative sources; MN will use our version of IDOL along with some social tagging directed at a legislative audience; CA seems to be far ahead of us with accessibility (ADA) issues.

BH: What are CA's priorities?

DVB: Would like to see storage capacity pursued a little more; sharing info with National Conference of Commissioners on Uniform State Laws (NCCUSL) and American Association of Law Libraries (AALL).

BH: What are the functions of preservation vs. just storage and elaborate on costs, partners, and models.

BB: Some of this will be in our existing eLeg project.

BH: That had a tighter focus; in NDIIPP we pull it out a bit more; what we and NC are doing with content is complementary to what somebody like LOCKSS is doing; we create content; AZ works on a place to store it.

DZ: Difference between preservation and storage.

BH: DZ is introducing archiving, records management, and appraisal issues; preservation in libraries is not focusing on these issues (SA, TC); according to SA, much more about content; they don't evaluate.

DBV: Can't divorce it when you are dealing with legislation; issue of authentication comes into play.

BH: There are cost, investment implications; there are unanswered questions; but the perfect is the enemy of the good; we cannot define preservation as an overly expensive or complex benchmark, we won't be able to do anything.

MS: We are content creators; we could add information that allows info to be saved, searched, etc. Could be done at the outset if we knew what to do upfront (sentiment echoed by DBV).

SA: Concept gaining currency, significant properties; any piece of content can have unlimited properties; but, from curatorial perspective, what are the critical properties that I need to capture to preserve it? Take an expansive definition that is customized to specific communities so that they can be defined, preserved.

TC: We try to work with people so that things are "born archival."

BH: What can we do at the beginning to make things easier down the road of preservation? We can define these as schemas, etc., that feed into microformats, etc.

DZ: We need to define preservation for the whole group: "Preservation is long-term, error-free access as long as it's required."

BH: We need to define it for a critical audience: legislators and legislative researchers; project focus is on born-digital materials.

BH: (summarizing) Storage, preservation defined for particular audience including authentication, recommendations we can make at point of creation.

DZ: Storage of materials with legal implications, and without legal implications.

BH: Technological consideration and legal consideration; we have to understand implications of the things we are doing (e.g., the implications of authentication). We should figure out what the most useful definition of authentication is; sounds like one of the things we want to look at quickly is authentication and how it plays into preservation; we will post that white paper/research on our website. We should start looking at IDOL/access with Mendora.

BH: We would like to work closely with CDL on web harvesting initiatives. MN can be beta users?

BB: National Conference of State Legislatures (NCSL) is working on database of statutes for all 50 states; put aside. StateNet is third-party vendor (they do bill tracking).

BH: Underscores need to continue working with NCSL. Also, they are the best way to do outreach and promotion.

BH: (summarizing) Preservation/storage; authentication; continue to look at use and various tools – connect IT staff with Mendora, particularly with IDOL, web harvesting with CDL, continue to work with NCSL with outreach and education (4 tasks).

Storage/preservation: Daniel, Diane, and Trisha

Audiences/Search/Accessibility: Mendora, Kris

Web Harvesting: Trisha, Stephen

NCSL/Outreach/Education: All

BH: This is a year-long plan to help us frame a work plan with deliverables; we will also share reports as we visit other states to begin fitting it all together; we will ask people to fit into the framework we are developing; there will be modifications, but we will keep you all informed.

JJ: Basecamp overview, please use it.

BH: Project proposal, a light reworking of the original proposal. No time for partner input in the revision, so it is a work in progress. LC said fine. It will be fleshed out as we define objectives and zero in on results. An all-partner meeting is probably prohibitively expensive. We will probably do a lot of communication through Basecamp and website, and we will come back here to California at some point later on.