July 2010 Partners Meeting Breakout Session

Case Studies from the Preserving State Government Information Initiative
Breakout Session #11
NDIIPP Partners Meeting
Wednesday July 21, 2010
4:15 p.m. – 5:30p.m

Presenter: Kelly Eubank, North Carolina State Archives
Robert Horton, Minnesota Historical Society
Matt Peters, Utah Automated Geographic Reference Center
Dan Waterbly, Washington State Archives
Pete Watters, Arizona State Library, Archives and Public Record

Attendees: 18

Preservation and Access--Can One Live without the other?
Presenters
Kelly Eubank, North Carolina State Archives
Matt Peters, Utah Automated Geographic Reference Center

Presentation

- GeoMAPP and preserving "at risk" geospatial content
- Building relationship between state GIS and archives staff
- Interstate partnerships
- Have ArcSDE Enterprise database, centralized
- Output a shapefile. Stagings are with ISO catedgories: geosciences, biota, etc.
- Bagged with Bagit spec
- Dataset evaluation, metadata content check. Kentucky does similar thing, but file-based geodatabase output. File structure of open, closed, conditional.
- Raster and vector folder. Snapshots of data in ISO categories by year and quarter. Utah is similar but puts out several formats, file-base geodatabase, shapefile, geospatial pdf.
- Learned about data transfer: metadat is complex, varieties, Orthoimagery poses storage challenge, and requires detailed planning.
- Vector data structure of file tranferred to archive, compare KY and NC
- Metadata is imported, transferred into access
- KY used DSpace for records. Utah has their own system with finding aid, container listing, file fromat, file size and type, etc.
- UT ftp site, into categories broken into county boundaries, online access to this material
- NC got space in ContentDM. No online access. MARS is their thing. Looking into map tool, Mappify, and other tools. Geocommons. ESTI geoportal extension.

Discussion

- Do you have any information about the kinds of users that are using this data? The user base is mostly students, GIS folks. If they use the raw files they'll have to use GIS. They're training staff to handle GIS inquiries.
- NCSU purchased 55 TB of storage and bought space in ContentDM. They are partnering with state libraries. It took the initial frame money to build the infrastructure.
- How did you get that fee through your legislators? The demo sold the project to legislators.
- Have any journalists used this mapping data? NC? Utah? NC journalists are going through contraction, no inquiries. In Utah some journalists will reference maps.
- How much content has Utah collected? It's still just in TB, not PB.
- NC used e911 funds for it. How do we get orthophotos so that it makes sense to us? Student wrote a report so the tiles will make sense when it gets to them (us).
- are both compressed (1.7 TB) and uncompressed (20 TB) for this one collection.

Digital Archives: We have apps for that
Presenter: Dan Waterbly, Washington State Archives

- Data challenge, files are metadata in different formats. Metadata is often incomplete, link between files and is often broken.
- It's tough to upload big datasets, content must be virus scanned and there is no guarantee that files are valid
- Admin challenges. Users do not alway do what we want them to do. You need custom SQL scripts to monitor progress. Many systems are involved in creating config problems.
- Wrote three apps:
    - Archive This. Transmits and validates metdata and files from variety of sources. Fingerprints files for authenticity. Login, select a format of the data to be sent, record template to fill in, drop down menus to standardize data terms.
    - AutoTodd automates ingestion, monitors for newly transferred data from ArchiveThis. Virus scans new data with up to seven virus scanners, verifies the integrity of the transferred data using the digital fingerprint file. Process, monitors, backup to tape, fully automated. Intranet admin site: monitors.
    - Ingestion Coordinator can manage users and access levels, manage collections, control ingestion, move data to backup tape. Generates reports on what was ingested. Start to finish, Ingestion Coordinator adds agency and collections, create a user account, user transforms their data using ArchiveThis and uploads it to the Digital Archives. AutoTodd is alerted when transfer is complete.

Discussion

- Is the code available to general public? Not yet. Check the web site for information though.

Preserving Digital Legislative Content
Presenter: Robert Horton

- Started with assumptions: collaboration, standards, national cyberinfrasturre, appraisal and ROI, education, sustainability. (2007)
- Project partners: MN is lead parter (ROS, LRL, MHS), CA and KS, CDL, NCSL

- National conference of state legislators
- Thomson Reuters, private sector
- Process: lots of meetings, documentation: base camp, research, re-grants, implementation, evaluation
- Lessons we're learning: 2010, change. Perpetual beta, budget, personnel, all changed. "Constant partial attention." User expectations: preservation = access over time; successs = content + funtionality. Access: open content, loosely coupled with specialized needs. Focused on access: open content but not defining how anyone else is going to use it, so users can meet their own specialized needs.
- Collaboration and integration, lower costs, lower barriers, catalysts, business case, mandate, charisma (personality), local knowledge, each environment is different in each state, no single model, no one-size solution. Common problems but not the same solutions.
- Practical outcome: incremental improvement. Move from storage to preservation, but still have to address polices, etc.
- MHS: next steps. Integraion of non-XML content. Import and export, us and CDL, CA and KS. Automating the processes. Education. Will do gap analysis for partners and toolkit. Evaluation of MN approach.
- It's an ongoing process. Standards migration, collaboration.

Discussion

- How much do you think your approach will appeal to other states? Other states' situations are awful, limits their ability to do anything. Focus on one area, take from another. Funding is a problem. MN budget cut 15%, stopped doing some things to focus on technology. The payoff is down the line.

Preserving Email: The PeDALS Approach

Presenter: Pete Watters

- Policy framework, slideshare.net/pweet/ for this presentiaon. PeDALS strives for OAIS compliance. Archivist focus on process, not records. Business rules to create and generate normalized metadata, transform SIPs into standardized AIPs, create DIPS for each record.
- Why email? Suited to Pedals methodology, born digital, potential for historical value, message transmission formation proves a rich source of metadata; all partners had Outlook PST files.
- Project goals: atomize individual messages, store as individual AIPs, disseminate as browser -friendly DIPs
- Create a DB of rich metadata, room process: to support administration from email headers, metadata to put into admin catalog for finding aids, supports discovery. Bagit, New Zealand metadata extractor.
- PeDALS for permanent records, it's not a records management system.
- Why curate PST? When negating with office, archivists encourage weeds PSTs; archivists work with rules rather than records, no time to weed junk. PSTs plucked from hard drives can work, but more likely to generate errors. If you don't curate PST files, will still take it.

- Negotiated with director of a couple of boards, took emails and put them into folders.
- Metadata taken from headers is messy. One way to deal with it is learn to cope with a lack of authority control or possibly correct by data wrangling.
- Senders and recipients can be an email address or display name from one or more contact lists. Variations on a name. Subject line is not reliable source for titles or abstracts. Often blank, repetitive or a remnant from an unrelated message.
- Privacy storm clouds: legislation for privacy.
- First attempts, PST file structure was proprietary, considered third-party. Outlook plug-ins. Adopted open-source PST export utility.
- From within Outlook, could generate human-readable XML of email messages
- Looking at PST, more than just email. Has tasks, calendar, items and contacts? What about viruses, corrupt attachments? We need to give the archivists the ability to decide what to keep.
- Privacy. Email may be open to the public by statute, but some content may be sensitive.
- Personally identifying information, private information, intimate, of no public interest
- Repositories must develop procedures and policies for aggregates that may have some records with sensitive information.
- PST, not archiving a record like databases. XML is the best way to represent.
- Export utilities drawbacks: slow, etc.
- In February, MS release the PST specifiation. 203 page of tech speak with some errors and inaccuracies. Based on the spec, we've been developing a file base tool that doesn't require Outlook. Now have a file-based processor. Good for PeDALS. Will try to make it available as open source in the next few weeks.
- Other lessons learned, test on as many PST samples as possible. PSTs are not an automatic occurrence in Outlook 2010. But they can be generated manually and remain part of a scheduled retention.

Discussion

- Do archivists walk through the PST?? Washington inspires them. They do great stuff.