Pathways to Preservation

Digital Curation Strategies in North Carolina State Government

North Carolina State Archives and State Library of North Carolina

April 2, 2009  DigCCur 2009
Brief History of NC State Gov’t Information Curation

- **Two divisions within the Department of Cultural Resources**
  - **The State Library of North Carolina**
    - Targets a variety of statewide needs (Library Development, LBPH, G&H)
    - The Government and Heritage Library provides direct services to state government employees and researchers
      - State Publications Clearinghouse collects and processes state agency created publications
      - Digital Information Management Program develops recommendations and tests solutions that support the identification, collection, cataloging, storage, and preservation of state government information for permanent public access
  - **The North Carolina State Archives**
    - Provides services to state government and public
      - Designated the official archival agency of the State of North Carolina;
      - Responsible for providing advice as well as records management program
      - Public records include all formats regardless of “physical form or characteristics”
      - Collects variety of formats of archival records
  - Separate mandates, missions, administrative structures
  - Mandates overlap, often work together to solve challenges

- **Information Technology**
  - DCR-IT provides desktop support
  - ITS provides network support
  - No defined support for long-term management of born-digital state records at this time
Why is digital curation our problem?

- **Statutory mandates**
  - Gather, provide access to, and permanently store state agency electronic publications and records in North Carolina
- **Trained professionals (Librarians & Archivists)**
- **Administrative support**

<table>
<thead>
<tr>
<th>Records</th>
<th>Pubs</th>
<th>Websites</th>
<th>Geospatial Data</th>
<th>E-mail</th>
<th>Databases</th>
<th>Other Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archives</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Library</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>

**Challenges:** Statutory mandates, but no enforcement authority and no programmatic IT support
What we’re doing today to meet the challenges of digital curation
Websites: Harvests

- Archive-It (Internet Archive)
  - Pilot partners in 2005
  - Harvests began in fall 2005
  - Currently collected nearly 3TB
  - This spring: combining our harvests with relevant content harvested by Internet Archive between 1995 and 2005

- Challenges
  - Agency web design vs. tool limitations
  - Don’t have 100% buy-in
  - Crawl analysis
Websites: Access

- Access (short/long-term)
  - ARC/WARC file format
  - Annual backup to tape; stored locally
  - Multiple access points
    - NC State Gov. Web Site Archives, Webfeat search, Wayback machine, Archive-it page

- Anecdotal support
  - Governor Hunt speeches
  - Governor Easley executive orders
State Publications

- **Acquisition**
  - Via e-mail, CD, or any other way we can get it

- **Access & Storage:**
  - Hosted access & storage – the OCLC suite of tools
    - Began using CONTENTdm (CDM), Connexion Digital Import (CDI), and the Digital Archive (DA) summer 2007
    - Current storage: ~1.5GB
    - CDI
      - the WorldCat record is automatically modified to include a link to the CDM record in the 856 field
      - a copy of the file is automatically passed to and stored in DA
      - the WorldCat record is automatically crosswalked from to DC and the file and the new DC record are added to CDM
        - DC metadata is manually edited to include additional info. (e.g., creating/submitting Agency)
  - Records findable via local website, search engines, Webfeat, WorldCat

- **Local storage**
  - Local Storage Area Network (SAN)
  - Limited space, funds for growth
Publications: Pre- and Post-Processing

Connexion (MARC) view

CONTENTdm (Dublin Core) view

April 2, 2009

DigCCur 2009
Publications: Answering the Challenges

- Grants participation:
  - Washington State Digital Archives
    - NDIIPP grant through Oct. 2009, extension requested
    - Establish the Washington State Digital Archives (WSDA) model as one that can be replicated in other states and/or as a viable hosting service
    - No North Carolina publications submitted yet
  - iRODS & DCAPE (NHPRC grant through May 2011)
    - Build a distributed production preservation environment that meets the needs of archival repositories for trusted archival preservation services
    - Currently identifying required rule set for distributed data management environment and attempting to load Archive-It and CDM files
State Records

- Acquisition
  - Via external media or hosted service

- Access & Storage:
  - Dark Storage— stored off-line or on external drives
  - Access to catalog via Manuscripts and Archives Reference System
    - Current storage: ~2 TB
    - Access provide upon request. No on-line access at the moment to the record. Access to the metadata

- Local storage
  - Local Storage Area Network (SAN)
  - Limited space
  - Identifying funds for growth and to provide access

- Pursuing Grants to build programs and test tools

- Government Initiatives—Senate Bill 6.12
Records: Answering the Challenges

- Grants participation:
  - GeoMapp
    - NDIIPP grant through December 2009, extension requested
    - To do a demonstration of transfer of archival geospatial data as well as test data from counties.
    - Begin building capability within the State Archives to ingest and provide access to superseded data
    - Moving into the phase of transferring data
  - iRODS & DCAPE (NHPRC grant through May 2011)
  - EMCAP (NHPRC grant through June 30, 2009)
    - Development of email collection and preservation tools
    - Builds on existing protocols and behaviors
    - Keeps originals and transforms them to XML
# Geospatial Data--GeoMAPP

## Table 3: Archival Activity by Data Type and Organization Type

<table>
<thead>
<tr>
<th>Counties (64)</th>
<th>Archiving Raster</th>
<th>Archiving Vector</th>
<th>Archiving Both</th>
<th>Archiving Either</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>63%</td>
<td>73%</td>
<td>47%</td>
<td>89%</td>
</tr>
<tr>
<td>No</td>
<td>38%</td>
<td>27%</td>
<td>53%</td>
<td>11%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Municipalities (40)</th>
<th>Archiving Raster</th>
<th>Archiving Vector</th>
<th>Archiving Both</th>
<th>Archiving Either</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>65%</td>
<td>68%</td>
<td>48%</td>
<td>85%</td>
</tr>
<tr>
<td>No</td>
<td>33%</td>
<td>30%</td>
<td>50%</td>
<td>13%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>All Respondents (104)</th>
<th>Archiving Raster</th>
<th>Archiving Vector</th>
<th>Archiving Both</th>
<th>Archiving Either</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>63%</td>
<td>71%</td>
<td>47%</td>
<td>88%</td>
</tr>
<tr>
<td>No</td>
<td>36%</td>
<td>28%</td>
<td>52%</td>
<td>12%</td>
</tr>
</tbody>
</table>

## Table 5: Archival Rates of ‘Focus’ Data Layers

<table>
<thead>
<tr>
<th>Layer</th>
<th>County</th>
<th>Municipality</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parcel</td>
<td>63%</td>
<td>45%</td>
<td>56%</td>
</tr>
<tr>
<td>Street Centerlines</td>
<td>52%</td>
<td>43%</td>
<td>48%</td>
</tr>
<tr>
<td>Jurisdictional Boundary</td>
<td>44%</td>
<td>60%</td>
<td>50%</td>
</tr>
<tr>
<td>Zoning</td>
<td>38%</td>
<td>53%</td>
<td>43%</td>
</tr>
<tr>
<td>Orthos</td>
<td>66%</td>
<td>60%</td>
<td>63%</td>
</tr>
</tbody>
</table>
# Geospatial Data

## GeoMAPP Infrastructure Inventory: Geospatial

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Storage Type (SAN/ Direct Attached (RAID)/ External drive)</th>
<th>If yes, Compression Technique (zip, rar, mrsid, etc)</th>
<th>Uncompressed (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector (NCOM prodN SQL Server DB)</td>
<td>SAN and RAID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vector (NCOM other prodN SQL Server DBs)</td>
<td>SAN and RAID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vector (NCOM download - shapefile)</td>
<td>HD</td>
<td>ZIP</td>
<td>7.3</td>
</tr>
<tr>
<td>Raster (1993 DOQQ)</td>
<td>HD</td>
<td>JPEG</td>
<td>195</td>
</tr>
<tr>
<td>Raster (1998 DOQQ)</td>
<td>HD</td>
<td>MrSID</td>
<td>586</td>
</tr>
<tr>
<td>Raster (DRG)</td>
<td>HD</td>
<td>MrSID</td>
<td>29.3</td>
</tr>
<tr>
<td>Raster (Landcover)</td>
<td>HD</td>
<td>MrSID</td>
<td>13035</td>
</tr>
<tr>
<td>Raster (Local Imagery) Available/unallocated (free) storage</td>
<td>HD</td>
<td>MrSID</td>
<td></td>
</tr>
</tbody>
</table>

**Totals:** 13,852.60
Hi, beautiful!

I was lonely without you at my slide during the presentation this morning. Although the video and slides were good, I do not recall I do an adequate job with such a scattered audience of slide. I have been busy and precisethe equipment immediately was boarded with statements that [REDACTED] was concerned about creating such a complex. According to [REDACTED] has been concerned about the cost of an (I am) and the next for a Challenger tanker. We also told the group that we would never get that kind of money from the state.

What really is happening is that (2) they do not give me enough time to tell them the "details" regarding what I (you) get from the needs, plans for tomorrow, etc., and 2) I'm going to keep up with Discovery Place.

Either way, I am not the least bit concerned. They will find out in good time what we can do.

By the way, have [REDACTED] frequently that I have you? Have I told you lately that I need you? Have I told you lately that I miss you? Have I told you lately that I am here for you?

If you still love me, don't you miss me and me?

Please come back to me! I need you!

Love,

E-mail server

Governor Hunt E-mail server

Governor Easley

<table>
<thead>
<tr>
<th>Notes</th>
<th>Folders</th>
<th>Total</th>
<th>Email processed</th>
<th>Attached emails</th>
<th>Emails affected</th>
<th>Duplicate emails</th>
<th>Unique attachments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-Jun 2006</td>
<td>located in governor0607 and governor1206</td>
<td>2</td>
<td>11693</td>
<td>10659</td>
<td>1340</td>
<td>1034</td>
<td>190</td>
</tr>
<tr>
<td>Jul-Dec 2006</td>
<td>located in governor0607 and governor1206</td>
<td>78</td>
<td>13834</td>
<td>13833</td>
<td>2280</td>
<td>1785</td>
<td>167</td>
</tr>
<tr>
<td>Archive 07</td>
<td>located in governor0607, includes entire 12 months, last 6 months of sent mail</td>
<td>96</td>
<td>18979</td>
<td>18973</td>
<td>2145</td>
<td>1624</td>
<td>256</td>
</tr>
<tr>
<td>Jan-Jun 2007</td>
<td>located in governor office email from Jan 07 to June 07; has an additional folder, Agency Responses</td>
<td>4</td>
<td>6219</td>
<td>6219</td>
<td>931</td>
<td>761</td>
<td>84</td>
</tr>
<tr>
<td>Jan-Jul 2007</td>
<td>located in governor0607</td>
<td>3</td>
<td>4795</td>
<td>4794</td>
<td>527</td>
<td>442</td>
<td>11</td>
</tr>
</tbody>
</table>

The main error is the length of the complete pathname - will not convert.
EMCAP—client and administrative console
Records

- IRMC-Information Resource Management Commission

- Governor Easley Files
  - Paper and electronic
  - 90.6 Gigabytes transferred
  - Multiple file formats
The Future

- Continue to monitor and test potential tools that would be a better fit or fill an outstanding need (like preserving databases)
- Continue drafting guidelines and “best practices” documents to provide guidance to state agency staff
  - Seek administrative approval
- Expand current education programs for state agency staff
  - their role in the process
  - importance of following guidelines to ensure that Library and Archives receive “preservation” ready digital info
- Work with IT at all levels to
  - Identify expectations & needs
  - Identify roles & responsibilities
  - Maximize staff & resources
  - Identify efficiencies
Contact Information

Kelly Eubank
North Carolina State Archives
kelly.eubank@ncdcr.gov

Jennifer Ricker
State Library of North Carolina
jennifer.ricker@ncdcr.gov

Amy Rudersdorf
State Library of North Carolina
amy.rudersdorf@ncdcr.gov