“Building for the Future…”

The National Digital Newspaper Program

Deborah Thomas

US Library of Congress

DigCCurr 2007

Chapel Hill, NC – April 19, 2007

What is NDNP?

- Provide access to historic newspapers
  - Select historic American newspapers, published 1836-1922
  - Newspaper Directory, 1690-present, representing 138,000 titles (USNP)
- Leverage USNP investment (data, film, networks)
- Long-term partnership (LC and NEH) to build sustainable digital resource

Guiding Principles

- Aggregate, serve, and preserve
- Consistent with missions and philosophies of NEH and LC
    - Open and perpetual access to the general public and scholarly community
    - Take care to preserve the asset that NDNP builds
- Phased development
    - Build incrementally – don’t close off options
- System will change, content won’t

Experience Informing the Future

How can we plan for the Future?

- Content is more important than today’s “system”
- Design “system” to be expandable
  - Modular and upgradeable
- Assume interoperability is a requirement
  - A resource that stands alone but plays well with others
- Explicit incorporation of development phase
  - Opportunity for learning
  - Validation of assumptions
  - Develop best practices (perhaps leading to standards)
  - Build corpus that is of value for technical experimentation

Design to be Open

- A “system” that is open in many senses
  - freely accessible (a public resource)
  - available to use and re-use
  - deep linking and persistent identification to support citation and scholarship

Incorporates:

- open technical formats
- interoperable through support for standard protocols
- modular architecture
- software based on open source code as possible.

Interoperability

- Standards-based
  - JPEG
  - XML-METS/ MODS/ALTO
  - TIFF
  - JPEG2000
  - PDF
  - JPEG

- Web exposed API’s*
  - Search API- SRW** with extensions
  - Access API
  - Ingest API
  - Relations API
  - (tbd) Export API
  - (tbd) OAI-PMH to enable Metadata Harvesting **
  - (tbd) Client library

•* Abbreviation of application program interface, a set of routines, protocols, and tools for building software applications. A good API makes it easier to develop a program by providing all the building blocks.

•**SRW is a variation of SRU. Messages are conveyed from client to server, not by a URL, but instead using XML over HTTP via the W3C recommendation, SOAP, which specifies how to wrap an XML message within an XML envelope. The SRW specification tries to adhere to the Web Services Interoperability profile.

•***Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

•*Abbreviation of application program interface, a set of routines, protocols, and tools for building software applications. A good API makes it easier to develop a program by providing all the building blocks.

Archival Needs, Data Needs

- Open Archival Information System (OAIS) Model
- Archive interacts with:
  - Producers, Managers, Consumers
- Archive needs to:
  - Ingest, Manage, Distribute
- OAIS vehicles:
  - Submission Information Package (SIP)
  - Archival Information Package (AIP)
  - Distribution Information Package (DIP)

Information Object, Data Object

- Information object: microfilm, original newspaper
- Data Object is the digital surrogate
  - TIFF (6.0, 8bit, 400 dpi, uncompressed, tags)
  - JP2 (Part 1, 8:1, no tile, RDF/Dublin Core metadata in XML box)
  - PDF (hidden txt, 150dpi, XMP/RDF/Dublin Core metadata)
  - Optical Character Recognition (OCR) text (ALTO schema, bounding-box coordinates)
  - Structural Metadata (METS, ALTO)
  - Preservation Metadata (PREMIS, MIX)
  - Need to know the complete data object is valid

Structural Metadata

- XML based - Metadata Encoding and Transmission Standard (METS)
- Title METS Object
  - Bibliographic and holdings data
  - Corrections, additions (essay, geographic coverage)
- Issue METS Object
  - Issue & page data
- Reel METS Object
  - Reel data & Technical target data

Validating the SIP

- Digital Viewer and Validator (DVV)
  - Java library for validating SIPs (self-referring batches of data)
  - Built on JHOVE validation, extends capabilities
  - Command line or GUI viewer interface
  - Digital signatures for validated SIP
  - Adds preservation metadata to METS
  - Valid SIP ingested to repository

Data Workflows

- Acquire –
  - awardees produce and validate (repurpose?),
  - awardees send to LC,
  - LC verifies, transfer, backup
- Ingest – transfer, verify (& index)
- Disseminate – search, browser app (APIs)
- *Manage – create, read, update, delete, navigate, monitor, report*
- Sustain – bit preservation, architecture layers

Infrastructure

- Memorandum of Understanding btwn agencies (long-term commitment)
- NEH/LC Program group
- LC Technical Support
  - Program management team
  - Dedicated acquisition/qr and technical ops
  - Shared expertise with other repository projects
  - Integration with LC digital product operations

Thank You

- **NDNP Public Web**
  http://www.loc.gov/ndnp/

- **NDNP Web Service**
  Chronicling America: Historic American Newspapers
  http://www.loc.gov/chroniclingamerica