Technical Premises and Digital Preservation
Thinking Ahead, Designing Now

Caroline Arms
Office of Strategic Initiatives
Library of Congress
caar@loc.gov

Planning for the “system”

- What does it have to do?
  - Aggregate, Serve, Preserve
- Guiding principles
  - Consistent with mission and philosophy of NEH and LC
    - Must demonstrate good use of taxpayer $
    - Take care to preserve the asset that NDNP builds
    - Openness – “We the people”
    - Serve scholars and general public
    - Must deliver service to current users of old newspapers
    - Must allow for new services, new users, new expectations
  - Design with a 20-year program and perpetual access in mind
  - Don’t try to do everything at once, but do not close off options

What is certain?

- Technology will change
  - Technology for processing will improve -- and get cheaper
    - OCR
    - Automated segmentation into articles
  - Interface conventions and constraints will change
  - Techniques for indexing and retrieval will improve
- Expectations of users will change
  - Cannot predict what scholars of the future will want, or when
    - Text mining for topics; analysis by time and place
  - New user communities will need new services
    - PDA access; integration with learning management

Starting with a clean slate

- What does “do it right” mean in this context?
  - Content is more important than today’s system
  - Design system to be upgradeable
    - Modular
  - Assume interoperability is a requirement
    - A resource that stands alone but plays well with others
  - No out-of-the-box solution exists
  - Explicit incorporation of testbed phase
    - Opportunity for learning
    - Validation of assumptions
    - Develop best practices (perhaps leading to standards)
    - Build corpus that is of value for technical experimentation

Learn from experience

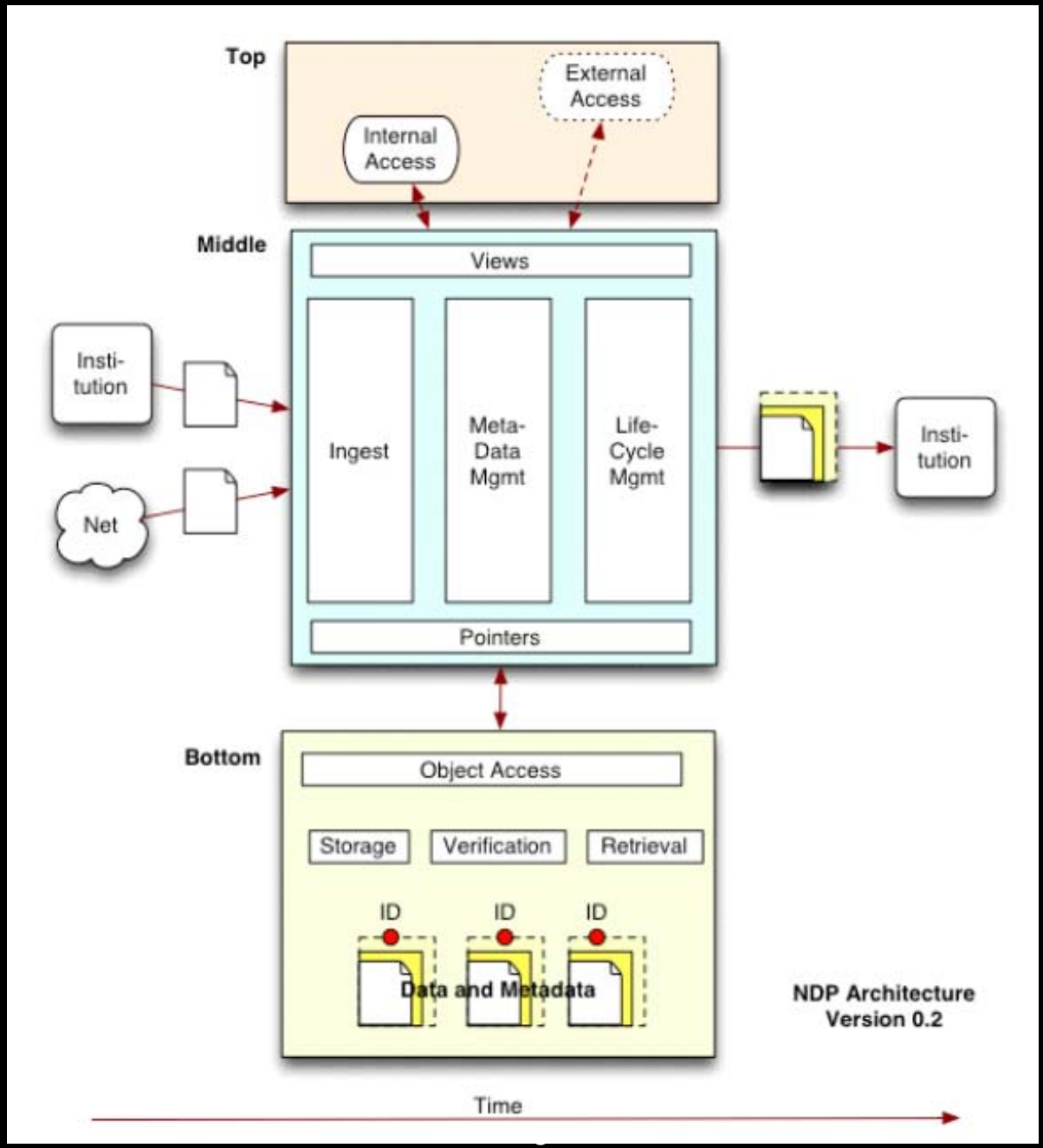
- Think carefully about choice of formats but expect change
- Specifications must be detailed to assure consistency
  - Formats becoming increasingly complex
- Need to validate technical integrity and conformance to specs.
  - Quality control by humans, with automated support
  - Fully automated validation prior to ingest
- Need metadata to understand what rules were in play when digital content was created
- Build on development skills and experience at LC
  - Using METS for compound objects
  - Initial testing of FEDORA promising
  - JPEG 2000 for zooming view
  - What worked well with *Stars and Stripes*
- Expect to learn from awardees

Look forward

- JPEG 2000 as delivery image format
  - Track adoption, tool development, and new features
  - Tap external expertise to develop profile
  - Consider for master format down the road
- Emerging best practices for preserving digital content
  - OAIS reference model
    - Ingest, manage, disseminate
    - Categories of metadata to support preservation
  - PREMIS (PREservation Metadata: Implementation Strategies)
    - http://www.oclc.org/research/projects/pmwg/
    - Core elements -- specification out soon
  - NISO technical metadata for images
    - http://www.niso.org/committees/committee_au.html
    - Revised draft expected soon
  - NDIIPP architecture principles

NDIIPP Preservation Architecture

- National Digital Information Infrastructure and Preservation Program
  - http://www.digitalpreservation.gov/
- Framework to guide development of national preservation network
- Design principles:
  - Support institutional relationships
  - Separate preservation and access
    - *storage and object management independent of search and display*
    - *task support for administrators separate from end user access*
  - Construct modularly
  - Assemble over time, not all at once
  - Upgrade parts without disruption of the whole
  - Use broadly adoptable standards and protocols


Top
External Access
Internal Access
Middle
Views
Insti-tution
Ingest
Meta-Data Mgmt
Life-Cycle Mgmt
Insti-tution
Net
Pointers
Bottom
Object Access
Storage
Verification
Retrieval
ID
ID
ID
Data and Metadata
NDP Architecture Version 0.2
Time

Open in many senses

- We envisage 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
  - corpus for scholarly analysis, encouraging creative use
  - open technical formats
  - interoperable through support for standard protocols
    - e.g., OAI-PMH, SRW/Z39.50
  - transparent modular architecture, extensible by design
  - software based on open source code to degree possible.