Digital Repositories

Babak Hamidzadeh
Office of Strategic Initiatives
(October 2005)
Outline

• Repository: definition, scope, services
• Development environment
• Case: National Digital Newspaper Program (NDNP)
Repository

- Software, hardware & processes that enable deposit, retrieval, & *preservation* of digital *objects*. 
In the grand scheme of things!
Bit preservation characteristics

- Files & directories
- Storage management
- Data replication & backup
- Checksuming
- Storage media refreshing
Repository characteristics

- Objects & relationships
- Content management
- Meta-data
- Context
- Migration (format)
What is this photograph showing?
Who took it?
Who owns it?
When was it taken?
Preservation

- Identity
- Integrity
  - Fixity
  - Completeness
- Understandability
  - Readability
  - Intelligibility

Authenticity
How is this document related to the previous photograph?
Content Model

• Objects
  – Identity
  – Intellectual content
  – Description (attributes)
  – Behavior

• Relationships
  – Identity
  – Definition
Services and functionalities

• Unique, persistent, global identification
• Object inventory & registration
• Object representation
• Continuous, bulk ingest (validation, tagging, registration)
Services and functionalities

- Automated migration (w/ validation)
- Version management
- Rights management
- Meta-data management (updates, content association, preservation)
- Search and retrieval (content & metadata)
Repository Development Center (RDC)

- Architecture
  - Commodity HW
  - Inexpensive
  - Scalable
  - Configurable
- Open Source
  - Content format
  - System software
  - Application software
  - Interoperable
RDC Architecture

- Internet
- Firewall / VPN
- Application Net Ethernet Switch
- Storage Servers (Linux/x86_64)
- Storage (SATA, 50TB+)
- Fiber Channel Switches
- Storage Network (Fiber)
- GigE Storage Network (Copper)
- Management Network (Copper)
- Ethernet Switch
- 802.1x/WPA-secured
- Laptops
- Printer(s)
- Workstations
- Linux/x86 + x86_64 Development Servers
- Drive Tower(s)
- IEEE1394b
- GigE Application Network (Copper)
- GigE Storage Network (Copper)
- Management Network (Copper)
Case: NDNP

• NEH/LC collaborative program
  – NEH: Funds the program (“We the People” initiative)
  – Awardees: Select and convert
  – LC: Aggregates, preserves and serves

• Content:
  – Granularity: Newspaper page
  – Model: Page, Section, Issue, Title
  – Materialization:
    • OCR’d text
    • Page image
Case: NDNP

• Volume:
  – 147,000 Newspaper titles
  – 63 MB/Page
  – 750,000 Pages
  – Total: ~50 TB’s

• Delivery to LC
  – 500 Gb hard drives (approx. 8000 pages)
Current generation

• Formats
  – Page image: TIFF 6.0; JPEG 2000; PDF
  – OCR: XML, ALTO Schema

• Pre-Ingest:
  – Submission package tool
  – Validation tool
    • Visual
    • Automated
    • Authentication
    • Adding preservation metadata
Current generation

- Search and access
  - Full-text
  - Hit highlighting
- Repository prototype:
  - Ingest
    - Unique ID assignment
    - Automated file placement
  - Search
  - METS/MODS objects
  - Metadata
Next generation

- Delivery through Internet2
- Article-level granularity
- Object representation
- Preservation
  - Plans
  - Functions
Next generation

• Metadata
  – Support different standards
  – Preservation
• Scale
• Automated
Impact

• Modes of partnership
  – Funding
  – Production
  – Preservation
• Use & adoption of standards & best practices
• Open source architecture
• Large-scale preservation
More Information

• Website: http://www.loc.gov/ndnp
• Email:
  – babak@loc.gov
  – ndnptech@loc.gov