Comments on the Digital Preservation Repository

Survey of Digital Preservation Systems
Karim B. Boughida and Sally Hubbard
(The Getty Research Institute)

Daniel Davis

DLF Spring Forum 2007: Pasadena, California
April 24th, 2007

Emergence of Digital Preservation Archives

- Most are small institutional or individual implementations
- Emergence of Large Gov't TDR/DPS
  - National Archives and Records Administration
    - Electronics Records Archive
    - Centralized / COTS Core / FOSS Components
  - Government Printing Office
    - Future Digital System
    - Centralized / COTS Core / FOSS Components
  - National Oceanic and Atmospheric Administration
    - Comprehensive Large Array Stewardship Systems
    - Replaces an existing custom production TDR
  - National Institutes of Health
    - National Health Records System
    - Decentralized / To Be Determined / Problematic Business Model
- Corporate DPS are showing up
  - Long tail media value and regulatory compliance

Hard Decisions

- There are always finite resources
  - You cannot keep everything
  - You cannot preserve everything to the same level of service
  - Good, Better, Best approach?
- To what degree do you use commercial products (COTS) and where?
  - COTS at the core may create vendor lock-in (watch indexes)
  - Reduce up-front risk and enable faster development
  - May hide key metadata needed for preservation
  - Are often not well designed to act as components in a service oriented architecture
  - Often lack key preservation features and must be adapted
- To what degree should you use (Free) Open-Source Software (FOSS)?
  - It is not really free but it is cheaper
  - Can enable control of your core mission
  - Most are far from mature
  - Lots of proof-of-concept, few products with sustainable business models

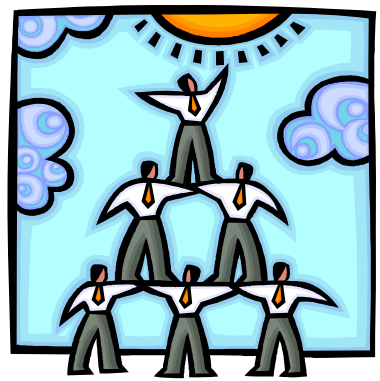
Preservation

- Geographical dispersion is essential (and a way to find the copies)
- COTS products are now supporting integrity and authenticity checks
- COTS products have been supporting media migration for some time
- Format migration is now being introduced
  - COTS and FOSS for format identification, validation, conversion
  - No completed fine-grained format registry for on-line use
- Other than print-equivalent formats and static-web sites, preserving essential characteristics is a work-in-progress
  - e.g. storing algorithms with scientific data sets
- A single, uniform item-level registry/resolver system is unlikely soon
- Innovative authoring and access systems usually ignore preservation
- Links between information must also be preserved
- A comprehensive ready-to-use system (overarching architecture) has not yet emerged but XML and service orientation is the best bet
- A homogeneous (mono-culture) implementation is guaranteed to fail

Items to consider

- If you look hard enough a DPS is needed as a component of any large enterprise system but rarely the whole purpose.
- Institutional questions (NARA-RLG TDR Checklist)
  - What is the sustaining business model?
  - What is the mission, policies, and requirements of the system?
  - What is your tolerance for risk?
- All things will change over a long enough period.
  - Architect for continuous change and non-uniformity
  - The system will never be finished
  - But the information must endure
- All things fail over a long enough period
- There a great economies of scale in infrastructure but little in creativity
- For the foreseeable future it will be a patchwork due to the human nature of funding, competition, cooperation and collaboration

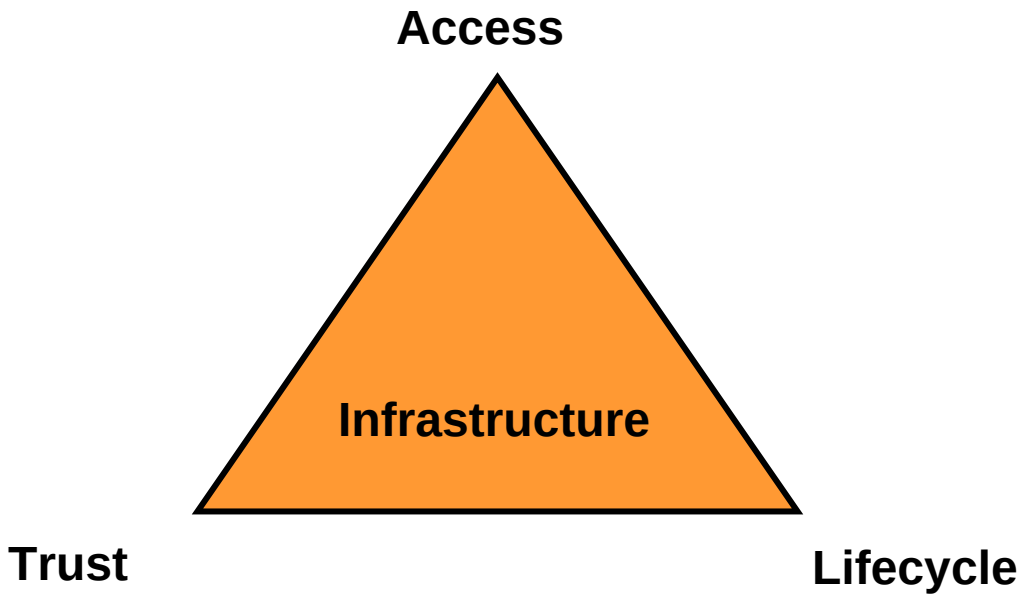
Goals for enabling users



1. Creation and publication of new forms of "***information units***"
2. **Services** to better enable ***business processes***
3. ***Knowledge*** environments that captures semantic and factual relationships among information units
4. Promote information ***re-use*** and ***contextualization***
5. Facilitate ***collaborative activity*** and capture information that is created as a byproduct of it
6. Preserve the information for future use

Sandy Payette - http://www.vala.org/vala2006/prog2006.htm

Requirements Tradeoffs



Overlap between Web and the Trusted Repository Paradigms

- Both the Web and Trusted Repositories are content-driven systems with overlapping needs
  - Content Creation and Capture, Collaboration
  - Content Storage differs on time scale optimizations
    - Content managers – Creation and Collaboration
    - Trusted (preservation) repositories (archives) – Long term storage, integrity, and preservation
  - Both require information lifecycle management capabilities
  - Support for other services and applications
- Both need a well-defined trust and security model

The Repository Dilemma

- Must implement a trusted core because:
  - Trust Model
    - All repositories have a significant trust requirement
    - Low fault tolerance for repository content custodianship
    - Specifiable (policy-driven) fault tolerance for mediation capabilities
    - Create (Ingest), Read, Update, and Delete must be transactional
  - Architectural Fit
    - Fit as a component in a SOA
    - Clustering, High-Availability, Transactions, Messaging
    - Federation
- *But must also support the Web paradigm*

Key Trends

- **While Web 2.0 is a major implementation trend among emerging applications**
- **We can combine the Web, Web 2.0, WOA and SOA into an integrated system that leverages the best qualities of each while providing trusted persistence**

  - Move volatility into data and technology stability into code and content.
  - Empower stakeholders with the (controlled) freedom to responsibly use, study, copy and change the system.
  - Embrace consumers as an integral part of the application and content development process.
  - Embrace Web mashups as important model to create composite enterprise applications and opportunistic user applications.
  - Use SOA to create composite services where high reliability and security is needed.
  - Add a preservation architecture as an overlay to Web/Internet architectures.
  - Add preservation capabilities as services.