NARA Electronic Records Archives
Lessons Learned and Future Direction

a presentation at NIST’s workshop on roadmap development for Digital Preservation Interoperability Framework

by Dyung Le
Director, ERA System Engineering

03/29/2010
Clinton Administration
– 40 million email messages

State Department
– 25 million electronic diplomatic messages

Department of Defense
– 54 million images from electronic official military personnel files annually

Census Bureau
– 600 to 800 million image files (2000 census)
The Challenges of ERA

• How do you build a system when the objects it is meant to process are evolving faster than the system can?
• How do you do that within the constrain of a budget cycle that is relatively rigid?
• While still meeting the business priority of the day?
What does that mean to the design of an interoperable digital preservation framework?

In the following we took the list of ERA design challenges and mark up in green the relevant ones.
ERA Program Business Challenges: External

• Coordination among 300+ user agencies!
  – Data transfer mechanism and packaging tools
  – Data type and Metadata policies
  – Incoming record volume projection

• Common agreement on security models and processes

• Managing public user expectations in the age of Google and YouTube!

• Developing partnerships with value-added private enterprises
ERA Program Technical Challenges: System Architecture

Importance of end-to-end system architecture integrity and vision

- Design for Evolvability and Scalability
- Design for Policy neutrality
- Design for incremental deployment and geographically distributed instances
- “Self-describing” representation format
- Scalable Object Identifier. Global namespace
- Scalable (billions objects) metadata repository
- WOA style vs. WS-* style for Web Service definition
ERA Program Technical Challenges: Data Model and Metadata

- **Record Object and Metadata**
  - Definition and nature of Archival Information Package (AIP or Record Object)
  - Distributed vs. Centralized Asset Catalogs
  - Computer assisted or automated metadata generation for content description

- **Data model for “virtual collection”**

- **Metadata standardization to leverage outside community research**

- **Mechanism for “Rip and Update” Metadata repository on demand**

- **Flexible security model for record object components**
ERA Program Technical Challenges: Ingest

- Ingest transmission performance and bandwidth
- Schema driven Framework for Import and export of records and metadata assets
- Dividing up functionalities to be taken place at Ingest vs. those at Preservation or Access planning
- Architecture for resource specific Ingest services in a logically and physically distributed deployment
ERA Program Technical Challenges: Preservation

- Framework for accommodating future technical approaches
- How to evolve while minimizing rework?
- Model for evaluating the cost associated with various preservation techniques
- Approach for “processing on demand”
- Tracking the simultaneous handling of multiple level of services to the assets
ERA Program Technical Challenges: Search and Access

• **Search framework**
  – Metadata distribution and Search index distribution. Where do the index go? How do they get handled when federated?
  – Integration of handling of data type across multiple specialized search engines. Blended context search
  – Exposing archival information to support external “partner’s” search

• **Leveraging external technology advances**
  – E-Discovery? Semantic web
  – What to keep inside, what to leave out, how to merge the two sources

• **Performance, Performance, Performance!**
ERA Program Technical Challenges: Storage

• Storage scalability is not disk! Or is it?
• Sharing Hierarchical Storage Management (HSM) in a logically and physically distributed architecture
• On-demand Storage provisioning
• Storage software abstraction layer for ease of technology updates

• Scalability, scalability, scalability!
• Evolvability & Extensibility
• Scalability & Performance
• Configurability
• Ease of Use
• Maintainability, Operability & Ease of Deployment
• Unifying themes
  – Common and Virtualized Record Object model
  – Scalable external Object ID and API
  – Standardized internal protocol
  – Only one way to represent an object or class of objects
  – Only one way to manipulate an object or collection of objects
  – Built-in extensibility through framework
  – Clean separation of architecture and design from implementation
Design Approach

• Top Down:
  – Start with OAIS reference model
  – Examine current Base, EOP systems and the current ERA RD
  – Review business requirements from offices and IPT(s)
• Assume SOA paradigm
  – Modular
  – Distributable
  – Swappable & shareable
• Standard Interface is key
  – Use open and scalable standards
  – Standard internal and external interfaces - systems and users
• Bottom Up: review architectural decomposition based on deployment options
• Continuous evolution: Build system platform, add new services and applications, enhance existing ones. A work-in-progress.
• Identify subsystems according to OAIS
  – Loose coupling
  – Autonomous
  – Interactions
  – Interface Objects
• Identify services in each subsystem
• Identify Enterprise Service Bus (ESB) Patterns
• Identify Sub-system software Layer Pattern
• Validate the design using Use cases scenarios
The OAIS Reference model is a framework for digital preservation. It includes the following components:

- **Producer**
- **Consumer**
- **IP**
- **AIP**

The model outlines the processes involved in managing and preserving digital objects, including ingest, management, and preservation planning. The diagram visually represents the flow of information and interactions between these components.
Content Server

- Content Server comprises Storage Object Management, ERA storage and some Access services
- Serves up a set of content (usually) sharing common access or data type characteristics
- Expose a unified and simple interface
- Unified view of object and metadata via AIP
  - Facilitates maintenance and synchronization between metadata and content
- Content server is more logical than physical
  - May have multiple Content Servers mapped to a database and/or storage subsystem
  - Similar to HCAP with respect to handling both storage and discovery

```
<table>
<thead>
<tr>
<th>Put AIP</th>
<th>Get AIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Update AIP</td>
<td>Search/Browse</td>
</tr>
<tr>
<td>Delete AIP</td>
<td></td>
</tr>
</tbody>
</table>
```
ACE and AIP

ACE ID: ERA N-Part ID

AIP Container: AIP-N-part Identifier
AIP Components: PDI Metadata File: Filename: N-Part Identifier(for ACE).PDI.xml
Content Object(s): ERA-supplied Filename(s)

Filename: N-Part Identifier(for ACE).PDI.xml

AIP ID: ERA N-Part ID

Filename: ERA-supplied Filename
Dpofofoul.pckfduN.fubebub

Dpofofoul.pckfduE.;FSB.tvgqmesGjfohn.;
• How are Content Servers aligned?
  - By record group collections (Federal, Legislative, Presidential)
  - By Data Type. Could help in specific storage and service offerings
• Document search vs. database search
  - By LoS. Could help in specific service offerings.
  - Thumbnails in search result.
  - Full content search vs. metadata search
• May need multi-dimensional alignment.
• Specific CS instances can be created as demand arises.
• A “logical Content server” can map into multiple “physical Content servers”.
Search and Access Flow: End-to-End
Business Object Management

Portal

Business Objects Management:
- Manage authority list
- Record Schedule
- LTI
- TR

Review Business Object
Approve BusinessObject

BPM

System Orchestrations (on ESB)

Form Service

Business Object Operations
- GET, PUT, UPDATE, DELETE

Versioning
Metadata Management
Taxonomy Builder
Indexing
Search
Browse

Common Services (Logging, Data Access, Authorization, etc.)

AssignAssetID

Physical Data Layer

Business Object Repository (BOR)
Benefits of the new Reference Architecture

• Infrastructures:
  – ESB: orchestrated framework for configurability and re-usability
  – BPM: shorter feedback loop between business users and system developers
  – XFORM: 47K LOC code saving. Natural code list integration

• Metadata and files:
  – More scalable and flexible ACE structure
  – Unified mechanism for Object access and processing

• Access:
  – Content Server concept fuses metadata and files at a high level of abstraction, allowing for ease of federation

• Preservation:
  – Preservation framework naturally leverages ESB
Benefits of the new Reference Architecture

• Deployment:
  – Ease of deployment into Access only or Ingest only instance, etc.
  – Lend itself to easy platform virtualization
• Clean decomposition
  – Lent naturally to an Open Source approach in order to take advantage of community expertise
  – Allow for ease of absorption and adaptation by other system integrators
• Shareable component/services in a logically and physically distributed architecture
• Back up and restore infrastructure: why back up a read-only asset?
• Legacy systems migration and legacy systems integration
• Updating system software for multiple cooperating instances without breaking!
In Conclusion

• The long term requirements for an Electronic Archives are leading us to an *evolvable framework* that need to support technology and needs that we don’t even know about yet!

• It is best that the Electronic Records Archive be built in such a way so as to fit in a *technology ecosystem that can evolved naturally*, and can be driven by the end users in ways that naturally ride the technology waves.

• The challenge is to co-exist and to leverage what’s going on outside the Archival space.
Don’t these points turn out to apply to an interoperability digital preservation framework as well?
Dyung.Le@nara.gov

System Engineering
ERA Program Management Office

301-837-0740

The ERA Web site:
http://www.archives.gov/era