Metadata Implementation Perspectives for the ERA System

DLF Fall Forum 2005

November 8, 2005

Quyen Nguyen

ERA PMO Systems Engineering Division
National Archives and Records Administration

Agenda

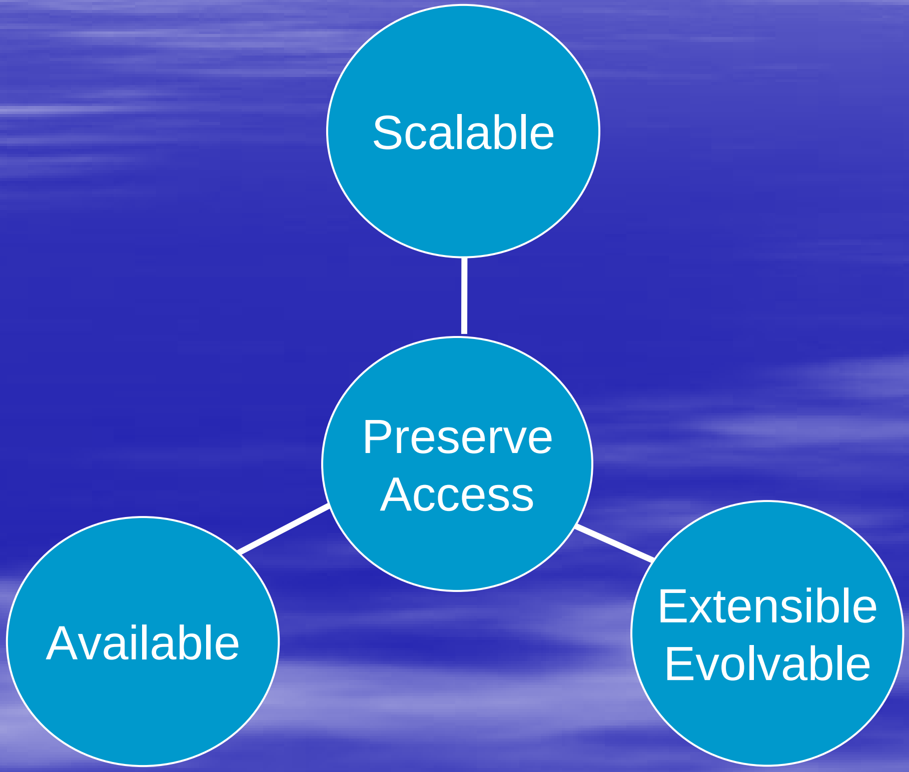
- What is ERA?
- ERA and Metadata
- System design issues and perspectives:
    - XML schema
    - Storage
    - Data management
    - Search and Access
    - Security
    - Performance.

NARA Mission and ERA Vision

- NARA mission is to provide “*for the citizen and the public servant, for the President and for the Congress and the Courts, ready access to essential evidence.*”

- In order to fulfill NARA mission in the 21st century and future, NARA created the ERA program. The product will be a system that “***will authentically preserve and provide access to any kind of electronic record, free from dependence on any specific hardware or software***”.

ERA System Requirements



Scalable

Preserve Access

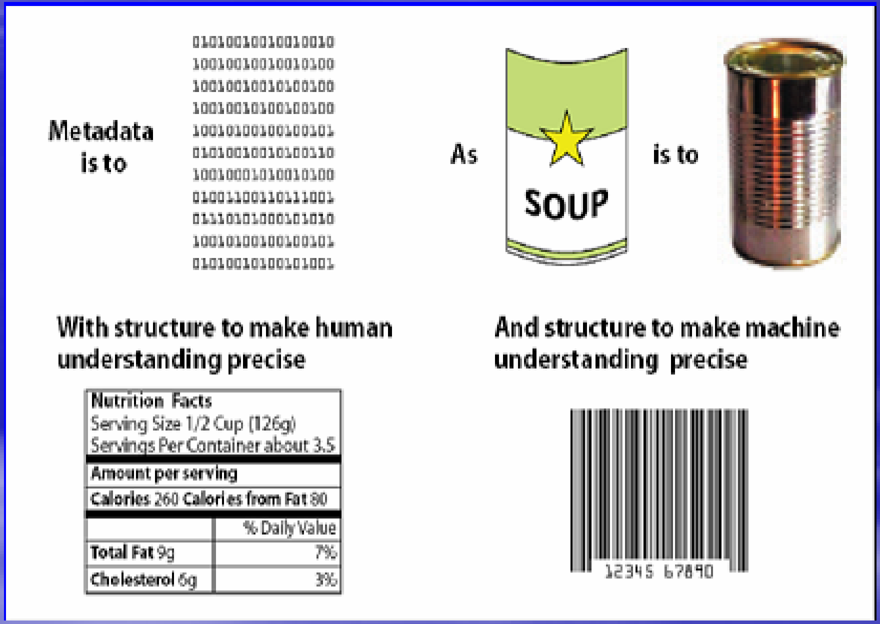
Available

Extensible Evolvable

Metadata Role

- Main functionalities of ERA:
  - Preserve electronic records over time.
  - Provide access to electronic records over time.
- System characteristics to support these two main goals:
  - High availability: key functions available more than 99%, no single point of failure.
  - Scalability: adapt to record volume and user community growth.
  - Extensibility: record types, data types, and services could be added without extensive redesign.
  - Evolvability: new technologies could be inserted using standards APIs and interfaces.
- Metadata is an integral element in order to satisfy these main ERA functionalities.
  - Different preservation strategies.
  - Persistent Object Format. Examples: use of XML, XSLT for text documents; use of GML for GIS records (research by SDSC).
- Metadata implementation will have to adhere to these system characteristics.

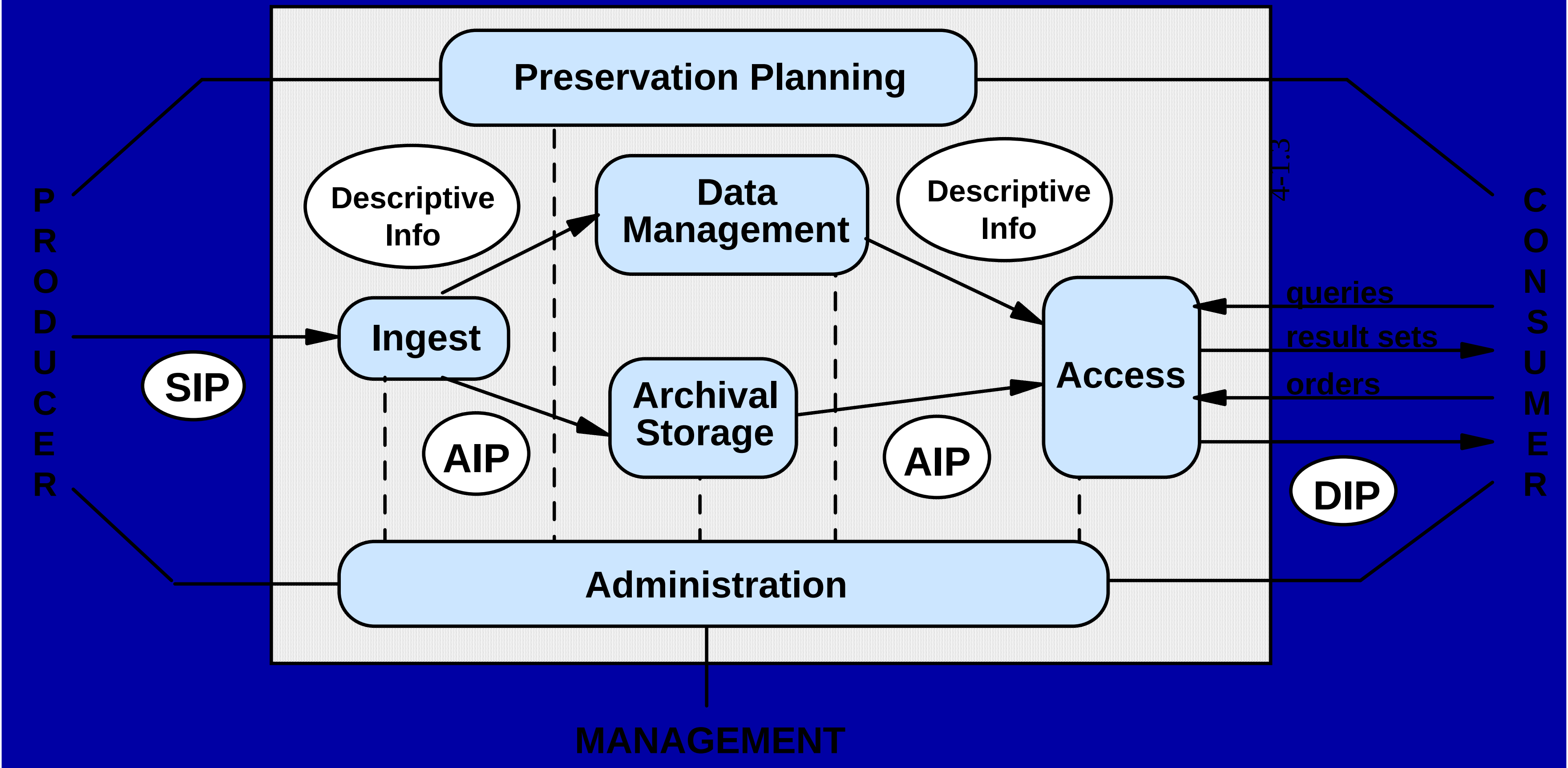
An Analogy



http://www.adobe.com/products/xmp/pdfs/xmp_news-mags_wp.pdf

- "How would one know which cans in the cupboard to use if they did not have labels on them? You would not know the flavour, contents or composition of any of the cans without opening them individually if they were simply bare cans. Metadata works for records as labels work for cans in a cupboard: both tell you what something is, allowing you to make meaningful decisions about it."
  - VERS (Victorian Electronic Records Strategy)

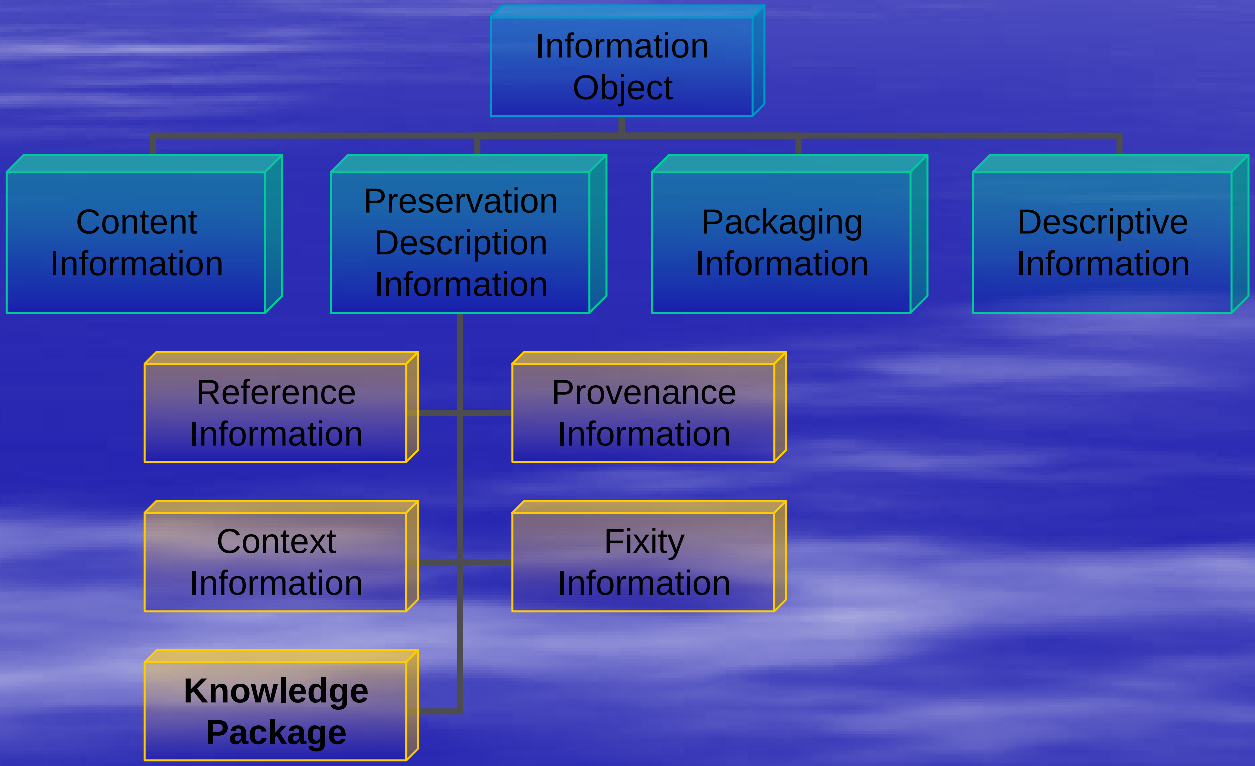
OAIS Functional Model



ERA Framework

- System scope and boundary follow OAIS model.
- System will implement the entities described in the OAIS (Open Archival Information System) model from CCSDS (Consultative Committee for Space Data Systems): Ingest, Storage, Preservation, Access, Data Management, and Administration.
- Metadata flow in the model from archive Producer to archive Consumer.
- Metadata encapsulated in various information packages: SIP (Submission), AIP (Archive), and DIP (Dissemination).
- Service-oriented Architecture paradigm
  - Archival functionalities encapsulated in loosely coupled services.
  - Achieve flexibility.

OAIS Information Model



OAIS Information Model (cont)

- Content: content data object and representation information. Representation information facilitates the structural and semantic interpretation of the stream of bits of records.
- Preservation
  - Provenance: history and chain of custody of the records.
  - Reference: record identifier.
    - Need to be unique.
    - Version independent.
    - Location independent.
  - Context: for the creation of records.
  - Fixity: for authenticity mechanisms of records such as digital signature, checksum, etc.
- Descriptive: for resource discovery and search facilitation of records.

Metadata Categories

- Metadata can be classified into the following functional categories:
  - Descriptive metadata for resource discovery and record search/retrieval.
  - Structural metadata for relationship between components.
  - Administrative metadata for management of objects.
- Preservation metadata use elements from all categories, but mainly from administrative and structural metadata.
- How will ERA facilitate the creation and maintenance of these types of metadata?

Metadata Creation Service

- extractMetadata(): is automatic extraction of metadata possible?
  - For ERA system, the entry point of records is via a SIP. Is it possible to extract metadata from SIP? What kind of metadata in SIP?
    - Would a Transfer Schema that Federal agencies should follow help? A Transfer Schema would specify the archival metadata elements that come with the package of electronic records transferred to NARA.
    - Would Templates help? Template is a concept which contains “specifications about a type of electronic document, record, donated historic material or an aggregate of such electronic documentary materials”.
    - How to extract metadata embedded in records? Example: Word document metadata, Adobe XMP. Is there any standard?
- generateMetadata(): is system generation of metadata possible?
  - Administrative data: lifecycle data, transformation, audit trail, etc.
  - For descriptive data: some could be generated out of lifecycle data.
- validateMetadata(): Validation service for system-generated metadata against NARA standards.

Metadata Creation Service (cont)

- Manual creation should be supported.
  - How would ERA support archive Description of intellectual nature that are dear to archivists?
  - Manual creation has to follow NARA business process (draft, validation, review, approval, etc.), and should involve team work.
  - Provide aid for completeness, consistency, and compliance with LCDRG (Lifecycle Data Requirements Guide).
    - LCDRG: set of specifications of metadata elements used by NARA staff to provide descriptive metadata for records. Description could be performed at various levels: collections, series, file units, and items.
- Other management functions: update(), delete(), import(), export(), version().
- Supported by the service that manages authority sources.
- Consider to implement metadata creation, both automatic and manual as web services.
- BPEL (Business Process Execution Language) with its service orchestration is a candidate to support business workflow.
- Provide flexible environment to manage business rules.

Metadata Expression

- Potential use of XML for encoding metadata due to the benefits:
  - XML is human and machine readable.
  - XML can sustain software and hardware evolution.
- But, the questions are:
  - How to harmonize multiple existing standards?
    - Don’t want to start from scratch.
    - Must be able to take advantage of vast amount of work that has been done by experts and institutions.
    - Handle overlapped elements, and synonyms.
  - How to make room for emerging standards, thereby adding more elements and attributes?

Generic XML Standards

- XML schema and instance standards from W3C allow different schemas to be used in one instance, thanks to XML namespace.
- An example is the METS schema, which allows multiple standards to be used in one XML document, via the element <mets:mdWrap>. At the same time, any new standard could be incorporated.
    - <xsd:element name="mdWrap" minOccurs="0">
    - <xsd:complexType>
    - <xsd:choice>
        - <xsd:element name="binData" type="xsd:base64Binary" minOccurs="0“ />
          <xsd:element name="xmlData" minOccurs="0">
            - <xsd:complexType>
            - <xsd:sequence>
            - <xsd:any namespace="##any" maxOccurs="unbounded"/>
            - </xsd:sequence>
            - </xsd:complexType>
        - </xsd:element>
    - </xsd:choice>
    - </xsd:complexType>
    - </xsd:element>

Domain XML Standards

- Dublin Core provides basic descriptive metadata elements (15):
  - Title, author, creator, subject, type, format, source, language, etc.
- METS can provide description for information packages to be transmitted between OAIS entities.
  - A METS document would include header, description, and file structure.
- PREMIS (Preservation Metadata) can support long-term preservation of digital materials: software, hardware, etc.
- NISO MIX (NISO Metadata for Images in XML):
  - Technical data for digital still images such as colorimetry, calibration, image quality attributes, etc.
- EAD (Encoded Archival Description)
- MARC (MAchine-Readable Cataloging).
- …

NARA Guide

- Lifecycle Data Requirements Guide (LCDRG) for Descriptive metadata:
  - Framework for all descriptions of permanent archival materials at various levels: record group, series, file unit, item.
  - Help create complete and consistent descriptions.
  - Define elements to be used for description.
  - Define when authority sources should be used.

XML Schema Design

- The ERA system needs an XML schema for metadata.
  - Will it be based mainly on NARA LCDRG?
  - Which elements from existing standards could be reused?
  - Which are the new elements that are specific to ERA?
- Which is the approach to mix existing and new elements?
  1. Will we start from the METS model so that new elements from the ERA schema could be embedded in a METS document?
  2. Conversely, will the new schema import existing metadata standards, such as DC, PREMIS, MIX, etc?
  3. Could we use RDF (Resource Description Framework)?
     - simple model and syntax based on triple (subject, predicate, object).
     - describe relationships using graph-like model. Useful for hierarchy?
     - flexible to insert new nodes, and schemas.
     - based on URI.

Example 1

- Example taken from a record found in NARA AAD system (Access to Archival Databases): http://www.archives.gov/aad/

```
<!-- Declaration of namespaces to be used -->
<mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:era="http://www.archives.gov/era"
      xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd
      http://purl.org/dc/elements/1.1/ http://dublincore.org/schemas/xmls/qdc/2003/04/02/dc.xsd
      http://www.archives.gov/era era.xsd" OBJID="123456" TYPE="" LABEL=“Record 1">

<!-- Header -->
      <mets:metsHdr CREATEDATE="2005-08-15T15:00:00“ RECORDSTATUS="Complete">
        <mets:agent ROLE=“Archivist“ TYPE="Individual">
          <mets:name>John B</mets:name>
        </mets:agent>
      </mets:metsHdr>

<!-- Descriptive Metadata using Dublin Core -->
<mets:dmdSec ID="dmd001">
      <mets:mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Basic Metadata">
        <mets:xmlData>
          <dc:title>Records About Community Action Program grants and Grantees , 7/1/1964 -
      9/30/1981</dc:title>
          <dc:creator>Community Services Administration</dc:creator>
          <dc:rights>Restricted - Partly</dc:rights>
          <dc:type>text</dc:type>
        </mets:xmlData>
      </mets:mdWrap>
</mets:dmdSec>
```

Example 1 (cont)

```
<!-- Descriptive Metadata using new elements -->
<mets:dmdSec ID="dmd002">
    <mets:mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL=“ERA Archival Metadata">
    <mets:xmlData>
      <era:era>
          <era:description-level>Series from Record Group 381: Records of the Community
        Services Administration</era:description-level>
          <era:location>NWME Electronic and Special Media Records Services Division National
        Archives at College Park, 8601 Adelphi Road, College Park, MD 20740-6001 (phone)
        301-837-0470 (fax) 301-837-3681 (e-mail) cer@nara.gov</era:location>
          <era:part-of>Record Group 381: Records of the Community Services
        Administration</era:part-of>
           <era:function-use>The agency created these files to track grants made to local
        community action programs and the organizations receiving grants</era:function-use>
          <era:scope-content>This series contains two types of files. The Grantee Organization
        Master Files provide the names, addresses, and target areas of organizations that
        received grants-in-aid through the Community Action Program (CAP). </era:scope-
        content>
      </era:era>
    </mets:xmlData>
    </mets:mdWrap>
    </mets:dmdSec> …

</mets:mets>
```

Example 2

```xml
<era:era>
    <era:basicDescription>
        <dc:title>Records About Community Action Program grants and Grantees , 7/1/1964 -
        9/30/1981</dc:title>
        <dc:creator>Community Services Administration</dc:creator>
        <dc:rights>Restricted - Partly</dc:rights>
        <dc:type>text</dc:type>
    </era:basicDescription>
    <era:description-level>Series from Record Group 381: Records of the Community Services
    Administration</era:description-level>
    <era:location>NWME Electronic and Special Media Records Services Division National
    Archives at College Park, 8601 Adelphi Road, College Park, MD 20740-6001 (phone) 301-
    837-0470 (fax) 301-837-3681 (e-mail) cer@nara.gov
    </era:location>
    <era:part-of>Record Group 381: Records of the Community Services Administration
    </era:part-of>
    <era:function-use>The agency created these files to track grants made to local community
    action programs and the organizations receiving grants</era:function-use>
    <era:scope-content>This series contains two types of files. The Grantee Organization Master
    Files provide the names, addresses, and target areas of organizations that received grants-in-
    aid through the Community Action Program (CAP).
    </era:scope-content>
</era:era>
```

Example 3

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/“xmlns:era="http://www.archives.gov/era/">

<rdf:Description rdf:about=“urn:gov.archives.era.123456">
    <dc:title>Records About Community Action Program grants and Grantees , 7/1/1964 -
    9/30/1981</dc:title>
    <dc:creator>Community Services Administration</dc:creator>
    <dc:rights>Restricted - Partly</dc:rights>
    <dc:type>text</dc:type>
</rfd:Description>

<rdf:Description rdf:about=“urn:gov.archives.era.123456 ">
    <era:description-level>Series from Record Group 381: Records of the Community Services
    Administration</era:description-level>
    <era:location>NWME Electronic and Special Media Records Services Division National
    Archives at College Park, 8601 Adelphi Road, College Park, MD 20740-6001 (phone) 301-
    837-0470 (fax) 301-837-3681 (e-mail) cer@nara.gov</era:location>
    <era:part-of>Record Group 381: Records of the Community Services
    Administration</era:part-of>
    <era:function-use>The agency created these files to track grants made to local community
    action programs and the organizations receiving grants</era:function-use>
    <era:scope-content>This series contains two types of files. The Grantee Organization Master
    Files provide the names, addresses, and target areas of organizations that received grants-in-
    aid through the Community Action Program (CAP). </era:scope-content>
</rfd:Description>

</rdf:RDF>
```

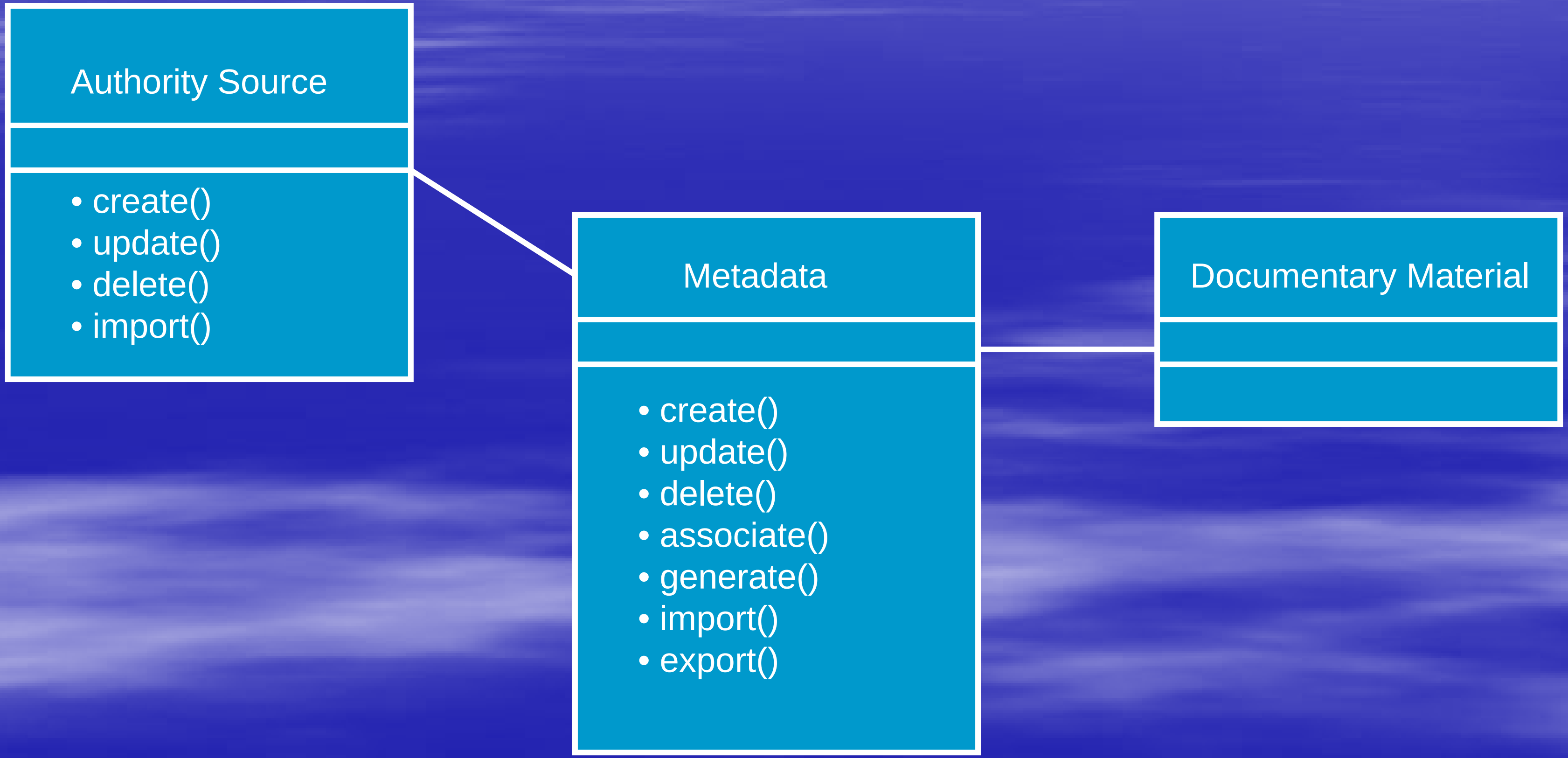
Storage Service

- How will ERA store metadata?
  - a) Metadata separate from records. Metadata could be stored in an RDBMS with pointers to actual records. Thus facilitates index mechanism for fast search and retrieval.
  - b) Metadata embedded in records. Make records self-describing.
  - c) Both. This option enjoys benefits of the above two. While helping the availability, and data migration, the redundancy requires more storage space.
- Given that metadata implemented in XML, should we store all metadata types in one document, or separate documents?

Data Service

- Data Modeling
  - How to model metadata versus record depends on decision of how to store metadata vis-à-vis records:
    - a) Metadata as a separate class.
    - b) Metadata as property of DocumentMaterial class.
    - c) Metadata class as associated in DocumentMaterial class.
- Data Replication
  - metadataReplicate(): How would metadata replication be performed, as it is necessary for service availability and load distribution?
    - Replication can be done in "one shot" with b), but in "two steps" with a) or c).
  - metadataSync(): Synchronization issue if metadata stored simultaneously within records, and outside of records.
- Metadata Hierarchy
  - Need to support at all levels of assets, according to current NARA guidance: record groups, series, file units, and items.
  - Maintain hierarchy, so that metadata at aggregate level can be shared among items at lower level.
    - metadataGet() and metadataSet() at each level.

Conceptual Object Model



Authority Source

- create()
- update()
- delete()
- import()

Metadata

- create()
- update()
- delete()
- associate()
- generate()
- import()
- export()

Documentary Material

Search and Access Services

- Metadata is the driver for Search and Access processing capability.
- Metadata Indexing service creates finding aids for Search.
  - Which database model to use for storing XML encoded metadata, relational or XML document model? Potential impact on performance.
  - From the metadata, could we build topic maps to facilitate search by domain of knowledge?
- Access service uses metadata which contains information for preservation, and access methods, and control.
  - For example, document viewer(s) could be included.
    - checkAccessRights()
    - listDocumentMaterialViewers()
    - selectDocumentMaterialViewer()
    - presentDocumentMaterial()
  - What is the framework to exploit this information, and accommodate different preservation planning and access methods?

Security Considerations

- Need to secure records. System must have mechanism to maintain security-purpose metadata from the point records enter ERA, and got disseminated to consumers.
  - Access privileges. This access right metadata will be processed for every access to a record.
    - Processing should navigate metadata hierarchy for access determination. For example, a record in a series might have different access rights than other records.
  - Audit trail, including transformations during the lifecycle of the records (Provenance).
  - Checksum and digital signature (Fixity).
- Need to secure metadata:
  - Who is authorized to modify description? Set up authorization to manage metadata based on user roles.
  - Any modification to metadata must be loggable for audit trail.
  - Checksum and signature may be used for integrity of metadata, especially manually created description.

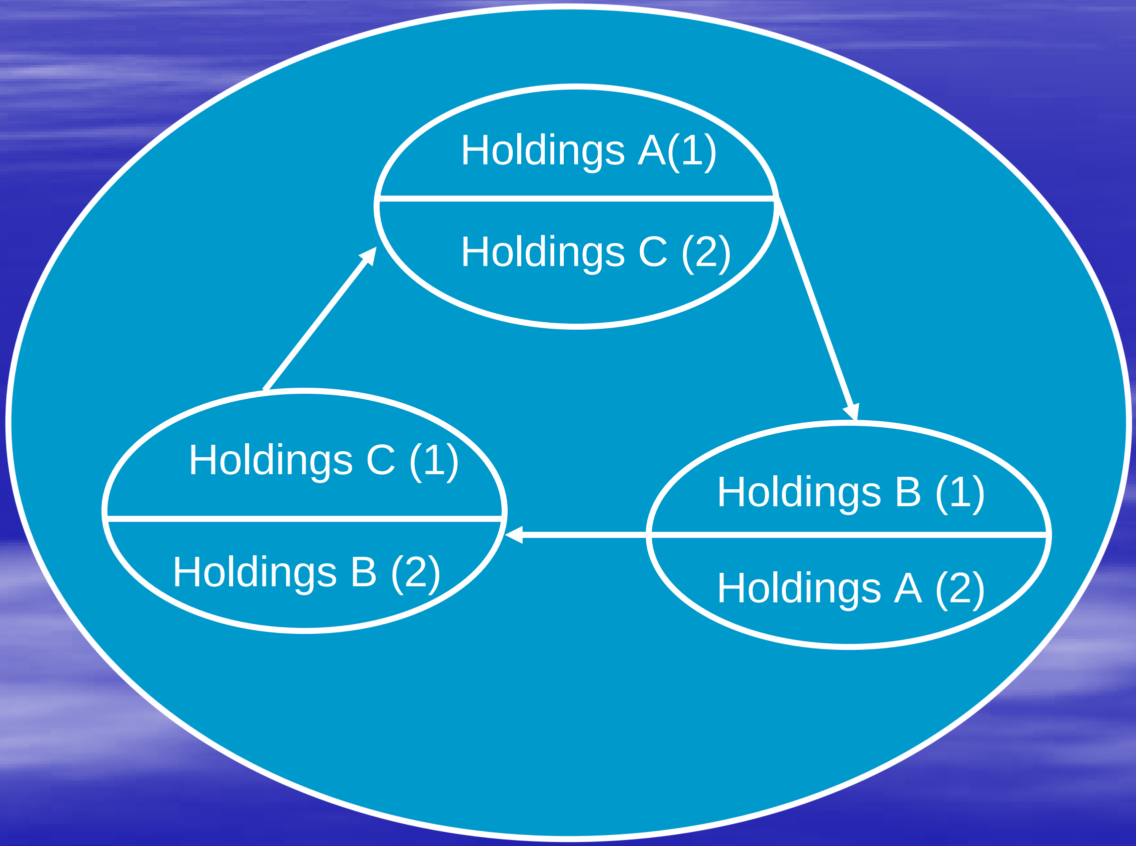
Security Considerations (cont)

- For ERA, Data Security requirements:
  - Records will have different classification levels.
  - Different levels must be completely air-gapped.
  - For the majority of highly classified records, although their access must be tightly controlled, the knowledge of their existence is not.
- Therefore, how would system architecture accommodate various scenarios?
  a) Metadata available at high classification level only.
  b) Different versions and copies of metadata at different levels. Some versions may be redacted.
  c) Same version, and different copies of metadata at all levels.

System Performance

- Service Load Balancing
  - In order to satisfy performance, load balancing is considered, especially for search.
  - This highly depends on metadata distribution.
- Service Availability
  - Search functionality must be 99.93%, which depends on availability of descriptive metadata.
  - Active Safe store applied not only to records but also to metadata.

Active Safe-store



- Active Safe-store applied not only to records but also to metadata.

Conclusion

- We have discussed a various technical issues of implementing metadata in ERA.
- Although there were issues, solutions do exist, that satisfy the core functionalities as well as the characteristics of the system.
- Design decisions will have to be made in the near future.
- Continue to monitor evolution of metadata standards.

Thank You

ERA

Electronic Records Archives

http://www.archives.gov/era

email: quyen.nguyen@nara.gov