Search Framework for a Large Digital Records Archive

Electronic Records Archives

DLF SPRING 2007

April 23-25, 2007

Dyung Le & Quyen Nguyen

ERA Systems Engineering
National Archives & Records Administration

Agenda

- ERA Overview
- Access Requirements
- Search Framework
- Search Technologies
- Digital Asset Catalog and Topic Maps

Agenda

➡ ERA Overview

- Access Requirements

- Search Framework

- Search Technologies for ERA
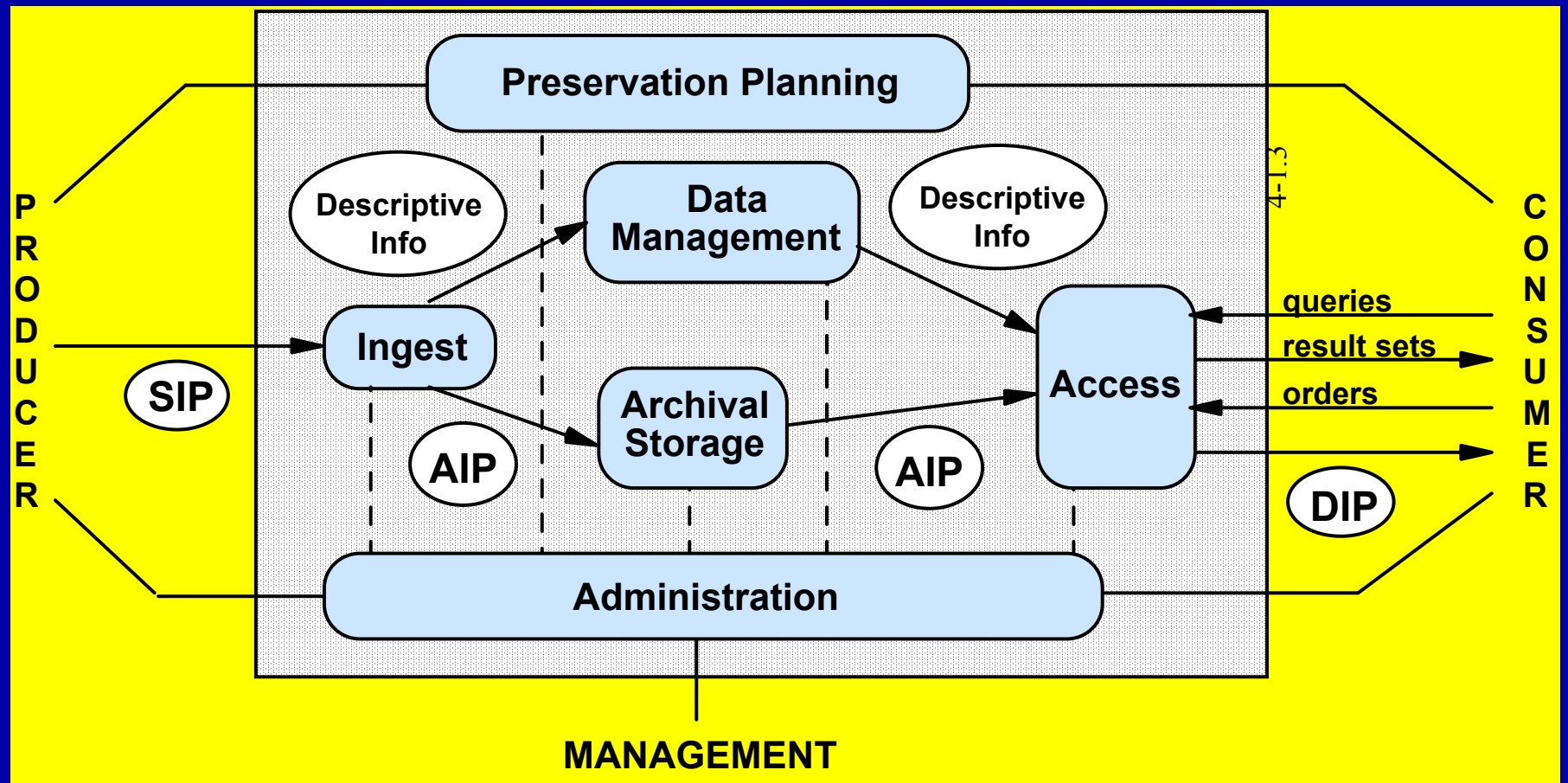
- Digital Asset Catalog and Topic Maps

NARA Mission

- 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*."

ERA Vision

- 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*".

OAIS Functional Model

Agenda

- ERA Overview
- ➡️ Access Requirements
- Search Framework
- Search Technologies for ERA
- Digital Asset Catalog and Topic Maps

ERA System Challenges

- **Scope** — Digital records from the Entire Federal Government

- **Complexity** — Records in different formats, which may be obsolete.

- **Volume** — Enormous amounts of records

Electronic Records Wave

State Department   **(.5 TBytes)**

– 25 million electronic diplomatic messages

Bush Administration **(12 GBytes)**

– 100 million email messages

Clinton Administration   **(6 TBytes)**

– 400 million email messages

9/11 Commission   **(1.2 TBytes)**

– 500 to 600GB emails, plus pdf, GIS, databases, etc.

Census Bureau   **(44 TBytes)**

– 600 to 800 million image files (2000 census)

Load Projection/Accumulated holdings: 2010 (10.6 PB) 2018 (131 PB))

System Requirements

- Scalability
- Extensibility
- Evolvability
- Ingest Preserve **ACCESS**
- Availability
- Security
- **Search**
- Presentation

• Highly Available Search capability to satisfy the need of researchers and archivists community.

• Highly Scalable Search to adapt to huge digital record volume and user community growth.

• Extensible Search Framework to handle various types of digital records.

• Evolvable Search Framework to allow new search technologies be inserted.

• Secure Search to protect assets via role-based access.

Access Use Cases

- A public researcher:
  - Wants to search for a record by using a Google-like keyword search.
  - Browses collections of records.
- An Agency Record Officer or NARA Records Processor searches for an existing Business Object e.g. Disposition Agreement, Transfer Request, Legal Transfer Instrument, etc. from which to create a new one or to update.
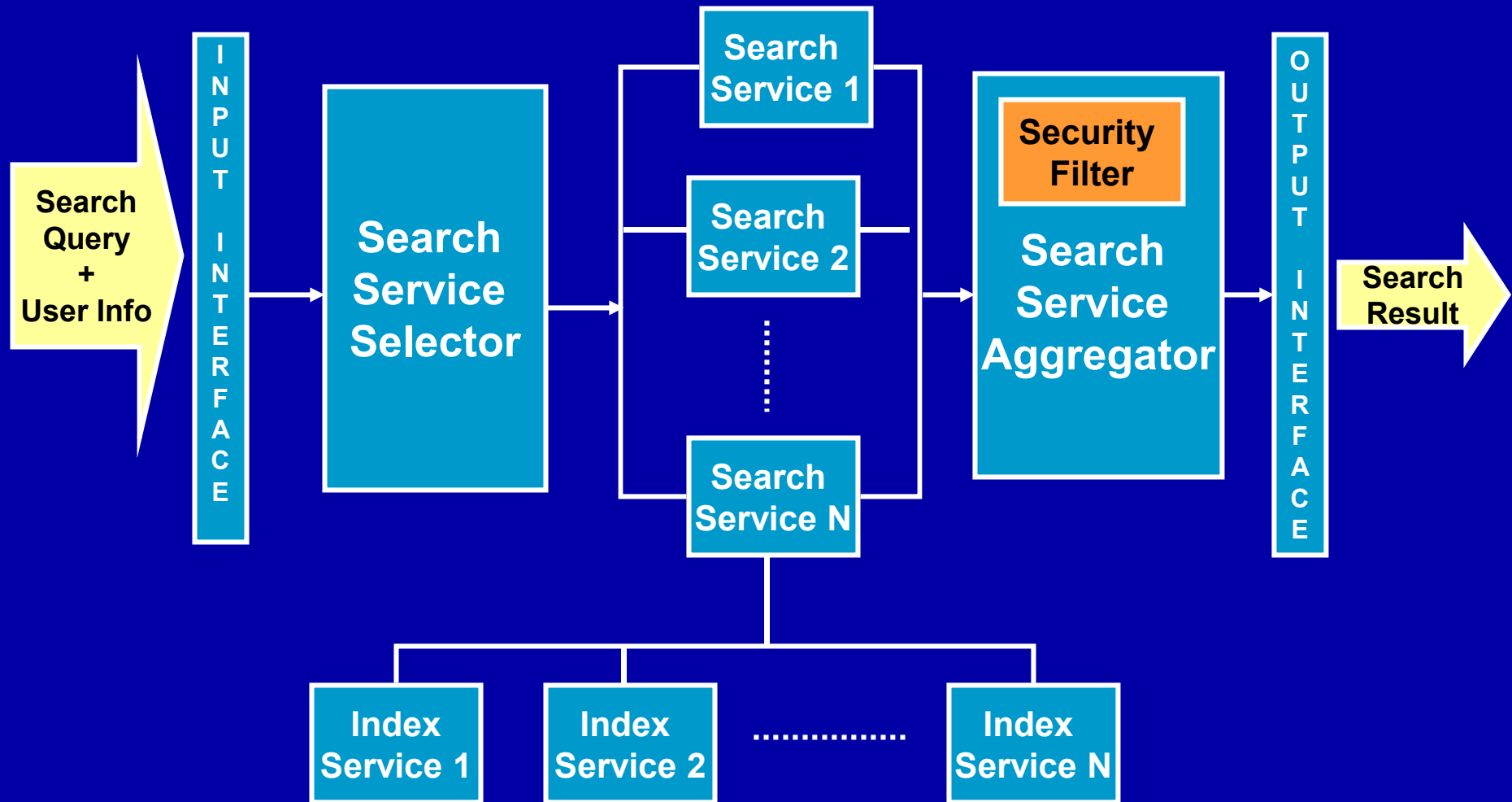- A NARA Access Reviewer searches for a record under FOIA (Freedom of Information Act) request.

Agenda

- ERA Overview
- Access Requirements
- ⇒ Search Framework
- Search Technologies for ERA
- Digital Asset Catalog and Topic Maps

Search Framework Components

- Search Portal:
  - Allows user to issue single search request
- Query Processor:
  - selector: acts as a mediator between the user and the search engine instances on each server.
  - convert request from common format to query input specific to each of the search engines.
- Result Set Processor:
  - aggregator: remove duplicate results
  - perform security filtering based on user roles and security metadata of result sets.
  - convert the results back into a common format
- Administration & Monitoring:
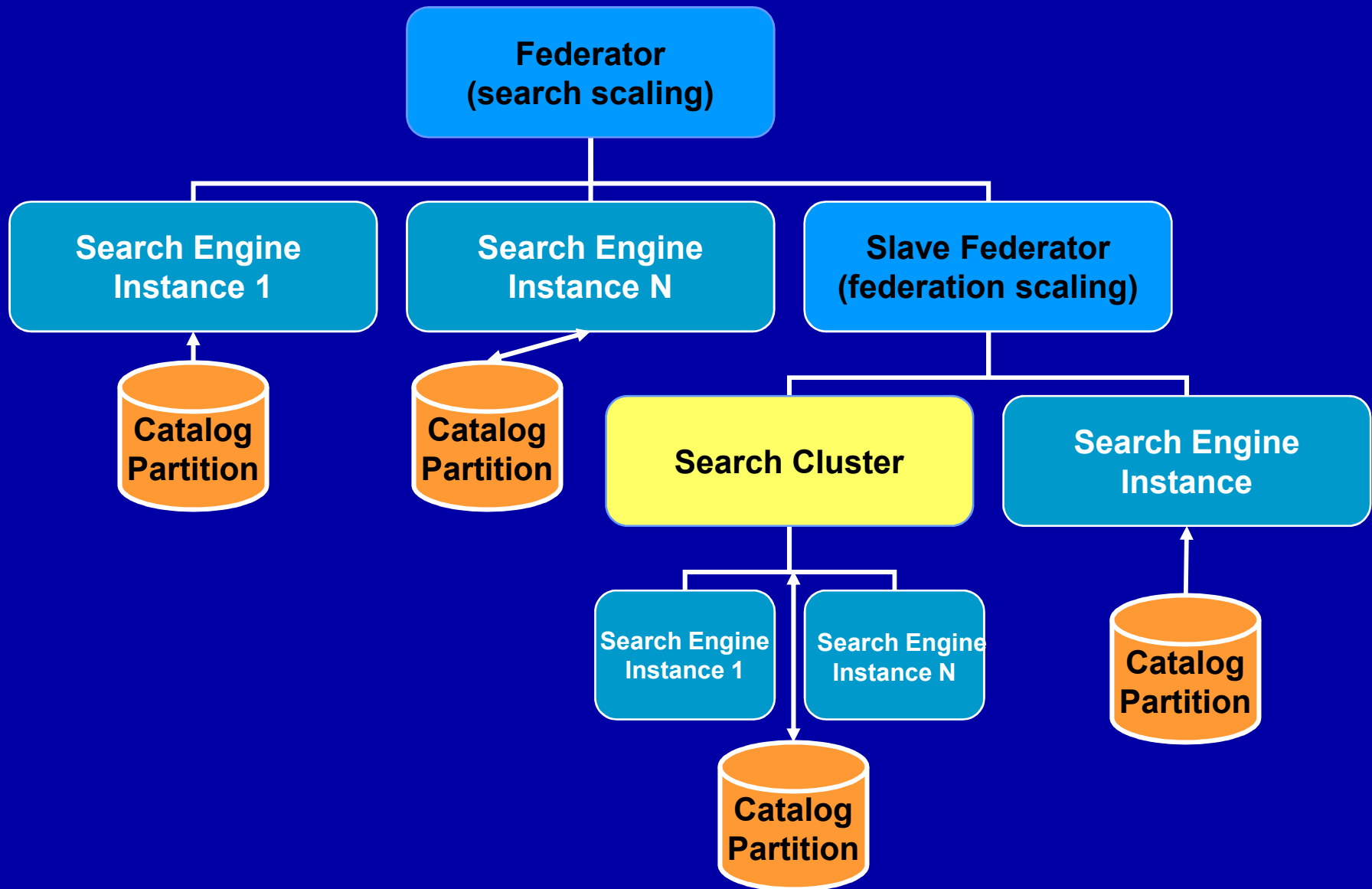 - Control resources used by a search execution based on user profiles and service level agreements.

Search Framework Data Flow

Search Query + User Info → INPUT INTERFACE → Search Service Selector → Search Service 1, Search Service 2, ... Search Service N → Search Service Aggregator (Security Filter) → OUTPUT INTERFACE → Search Result

Index Service 1, Index Service 2, ..., Index Service N

Framework Benefits

- Address extensibility requirement:
  - Ability to plug-in new search services to support content search of non-textual digital objects: databases, GIS data, images, audio, and rich media.
- Address evolvability requirement:
  - replace or remove old search services.
  - adaptable to technology changes
  - selection of search engine products based on cost/performance
- Address scalability:
  - increase incrementally processor, memory, and storage resources by adding nodes.

Search Framework Deployment

Federator
(search scaling)

Search Engine
Instance 1

Search Engine
Instance N

Slave Federator
(federation scaling)

Catalog
Partition

Catalog
Partition

Search Cluster

Search Engine
Instance

Search Engine
Instance 1

Search Engine
Instance N

Catalog
Partition

Catalog
Partition

Agenda

- ERA Overview
- Access Requirements
- Search Framework
- → Search Technologies for ERA
- Digital Asset Catalog and Topic Maps

Search Characteristics

- Performance Metrics
  - Precision: card (relevant results) / card (resultSet)
  - Recall: card (relevant results) / card (relevant Collection)
  - Example:
    - 100 relevant objects.
    - A query result yields 80 objects.
    - 60 out of 80 in result set are relevant.
    - Then: precision = 60/80 = 75%; recall = 60/100 = 60%.
  - Relevance: ultimate subjective judgment.
- Improvement strategies:
  - Data Query/Retrieval time.
  - Data Preparation
    - Ahead of time, e.g. at time of record ingestion.
    - Ad-hoc, i.e. whenever a user issued a search request.

Search Technologies

strategy \ format	Text	DB	Image	Audio	Video
Metadata	√		√	√	√
Content	√	√	√	√	√
Context	√				

Metadata Search

- Data Preparation
  - Record lifecycle data aggregation during ingest phase.
  - Manual metadata creation usually applied at collection level.
  - Metadata extraction
    - Issues and challenges for each format or media type
- Data Query
  - Keyword: Google type search. Ex: "Thomas Jefferson".
  - Boolean. Ex: "Thomas Jefferson AND Declaration of Independence".
  - Fielded.
  - Parametric. Ex: "Immigration records from 1910 to 1930".
  - These types of retrieval also apply for text content search, after documents have been fully indexed.

Fielded Search

Content Search

- Data Preparation
  - Indexing challenges:
    - Large volume of digital archives
    - Obsolete data type format of digital assets transferred to the system
  - Text Analytics
    - Special modules can be plugged in using IBM's UIMA (Unstructured Information Management Architecture).
    - Use of Natural Language Processing.
    - Thesaurus.
  - Database records:
    - Search engine makes use of connectors to connect to database, and transfer queries.
    - Convert data into XML database, and use XPath, and Xquery.
- Data Retrieval
  - Facets: mutually exclusive containers that contain hierarchies of properties.
    - Ex: UC Berkeley Flamenco; Endeca with "Dimensions" for guided navigation.
  - Dynamic grouping of result sets into pre-configured categories. Ex: FAST.
  - On-the-fly concept abstraction from result sets and dynamic classification into those concept buckets. Ex: AlltheWeb.

Multimedia Content Search

- Data Preparation
    - Prevalent strategy is to translate into a text problem.
    - Audio: Speech-to-text conversion.
    - Image analysis: pattern recognition techniques.
    - Video:
        - combination of various techniques.
        - generate metadata from content analysis.
        - speech-to-text
        - transcript
        - speaker recognition
        - segmentation of a video stream, such as scene selections on a movie DVD.
        - On-screen text recognition.
        - Ex: Blinkx, Virage.
- Data Query
    - Input other than text: tone whispering, hand drawing, etc.
- Use Case: Video clips of Rose Garden ceremonies.

Context Search

- Data Preparation
  - Collect user profile, including interests and roles at registration time
  - Use AI learning agent principles to learn user search patterns
  - Data mining of Query log. Ex: obtain most frequently asked subjects for better guess of intent
  - Save search results for registered users.
- Data Retrieval
  - Query Augmentation:
    - Add context terms.
      - Ex: For Record Officer, ("EPA") → ("EPA", "Disposition Agreement").
    - Add ranking bias.
      - Ex: ("EPA") → ("EPA", ("Disposition Agreement", 7), ("Alaska", 3))
  - Refinement:
    - Search within a result set
   - Search within a category, facet, or concept.

Agenda

- ERA Overview

- Access Requirements

- Search Framework

- Search Technologies for ERA

➡ Digital Asset Catalog and Topic Maps

Digital Asset Catalog

- Each Digital Asset has an Asset Metadata.

- Digital Asset Catalog contains all the Asset Metadata, hence information of all assets in the system.

- Metadata search based on Catalog.

- Physically, a catalog entry is stored in an XML document.

- Conform to XML schema.

Digital Asset Catalog Schema
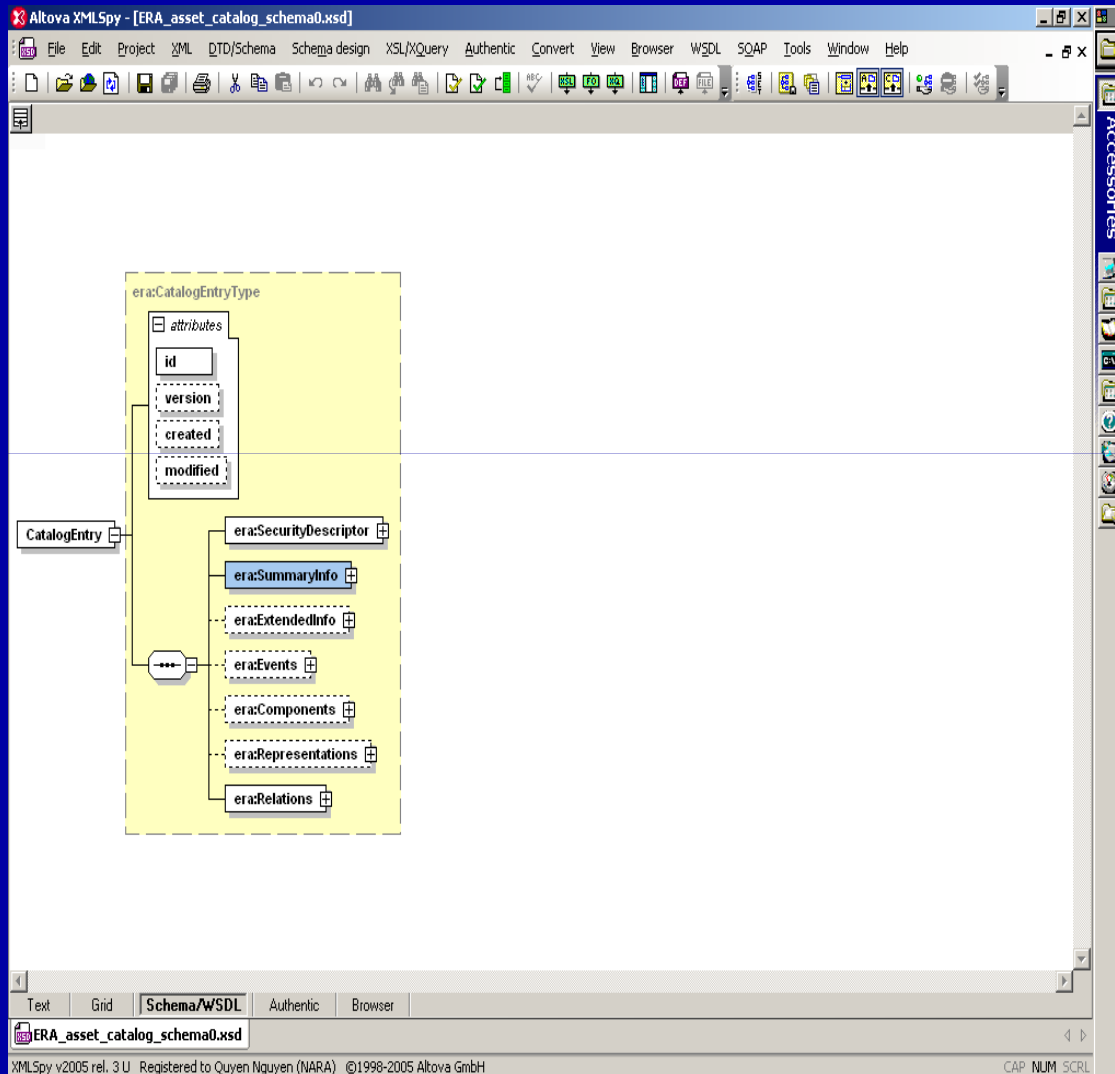
❑ Id: Unique & Persistent Identifier

❑ Summary Info: Information Collected throughout record lifecycle.

❑ Security Info: Access Restriction information, integrity seal.

❑ Extended Information:
  o Bibliography
  o Context
  o Content
  o Archival description

❑ Components: File attributes.

❑ Representation: e.g. transformations, versions.

❑ Relations: associations to other assets

Topic Maps

- The TAO of Topic Maps: A Topic Map is a representation of information based on Topics, Associations, and Occurrences.
- Topic:
  - Subject
  - Abstract concept
- Association
  - relationship between a set of topics.
  - roles involving topics in the relationship.
- Occurrence
  - Instance of a topic.
  - Information resource of a topic.
- Use Case:
  - User can browse topics.
  - Two alternative paths:
    - User can browse topics, then perform keyword search within a topic.
   - User can perform keyword search, then browse topics that group result sets.

AC → TM

- Overlay of multiple TMs on AC.
- Precondition: Process Lifecycle Data and archival description services to populate catalog entries.
- Generate TM from catalog entries.
    - Elements in a catalog entry contain information to define topics, associations, and occurrences.
    - Content in era:SummaryInfo and era:ExtendedInfo could be extracted for forming different topics.
    - era:Representations → occurrences.
    - era:Relations → associations.
    - CatalogEntry.ID, which is a globally unique identifier, will serve to chase the relationships between asset catalog entries.
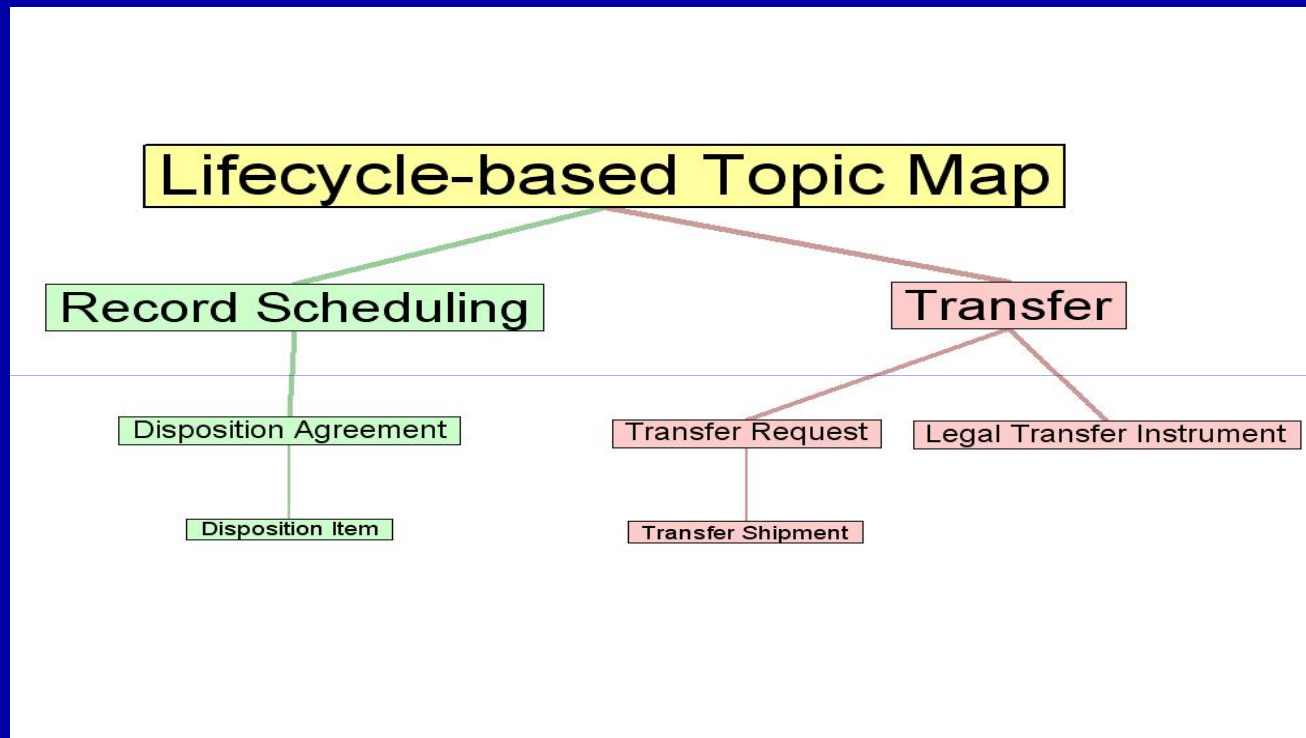
Generating Topic Maps
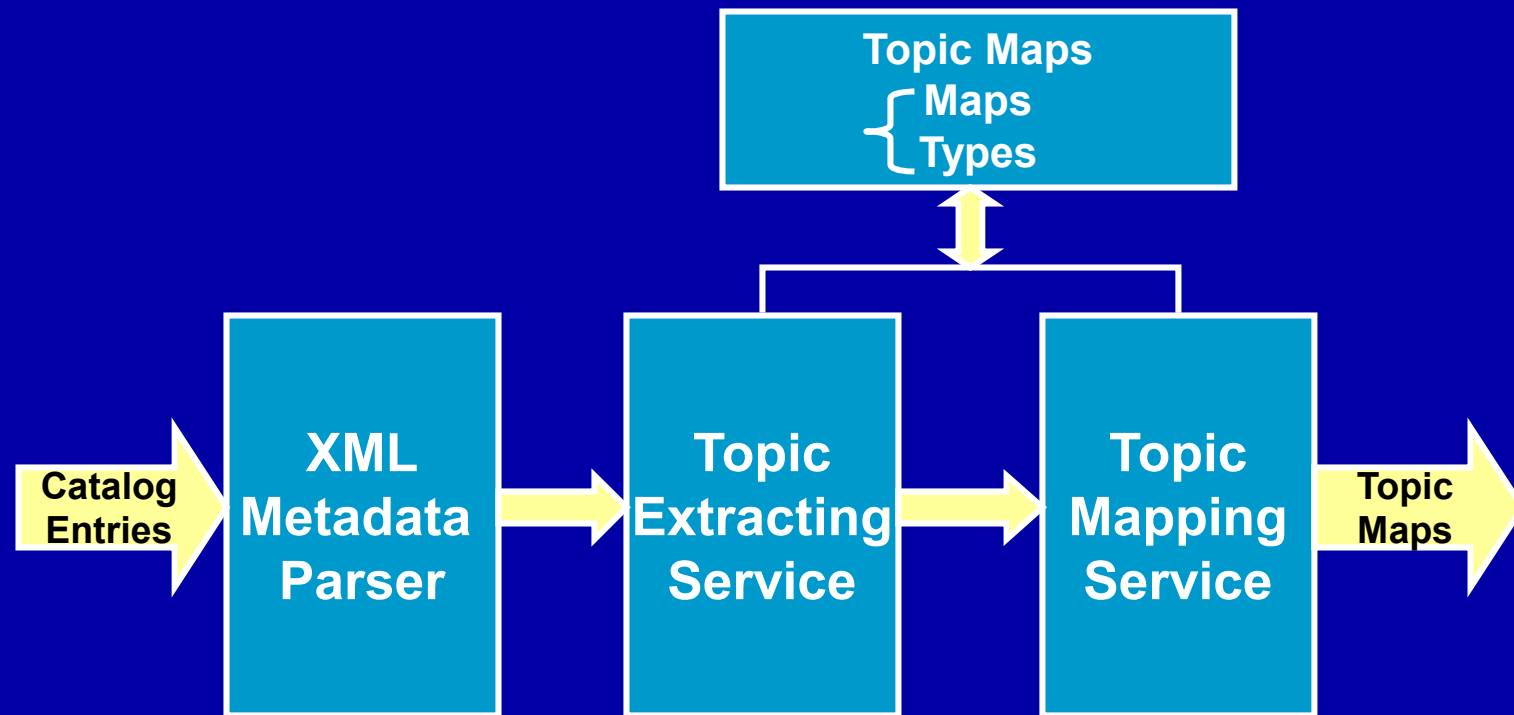
Archival Hierarchy-based TM
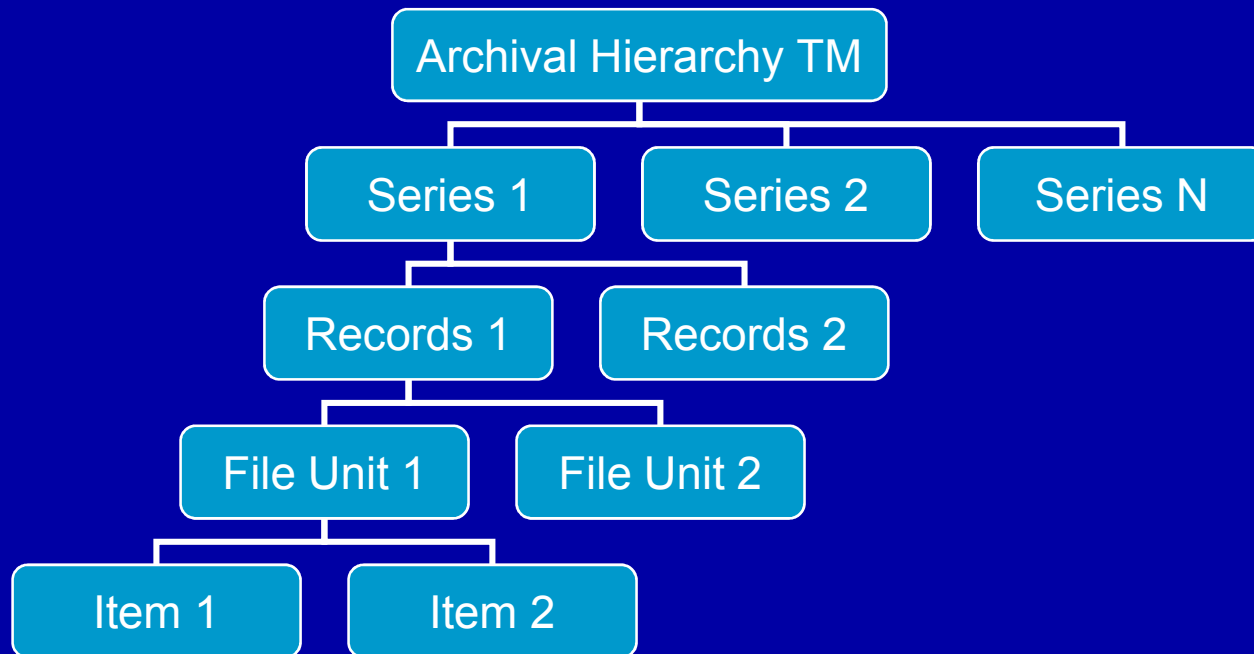
- Archival Hierarchy information could be derived from the data element era:Relations and era:Components populated during ingest phase.

Lifecycle-based TM

Lifecycle-based Topic Map

Record Scheduling

Transfer

Disposition Agreement

Transfer Request

Legal Transfer Instrument

Disposition Item

Transfer Shipment

▪ During the record scheduling workflow process, lifecycle data from the business objects were captured in era:SummaryInfo, which could be the basis of generating the Lifecycle-based topic map.

Subject-based TM
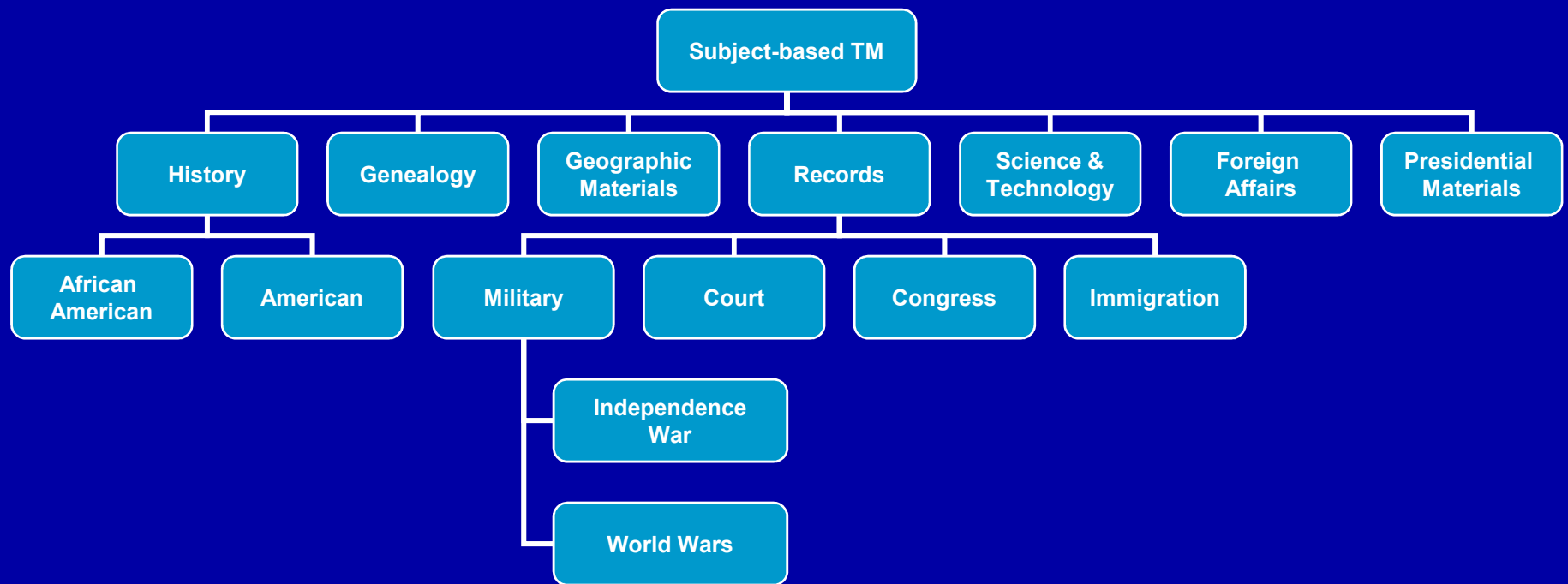
- The archival description contained in the data element era:ExtendedInfo could be used to generate this subject-based topic map.

Summary

- We have shown the requirements and challenges with respect to the access functionality of large digital archive system.
- An extensible and evolvable search framework was designed to handle the above challenges.
- Various search technologies have been investigated with the goal of providing good discovery service to end-users.
- Expect to be able to present presentation issues in the future; those issues are tightly related to long-term digital preservation.

Thank You
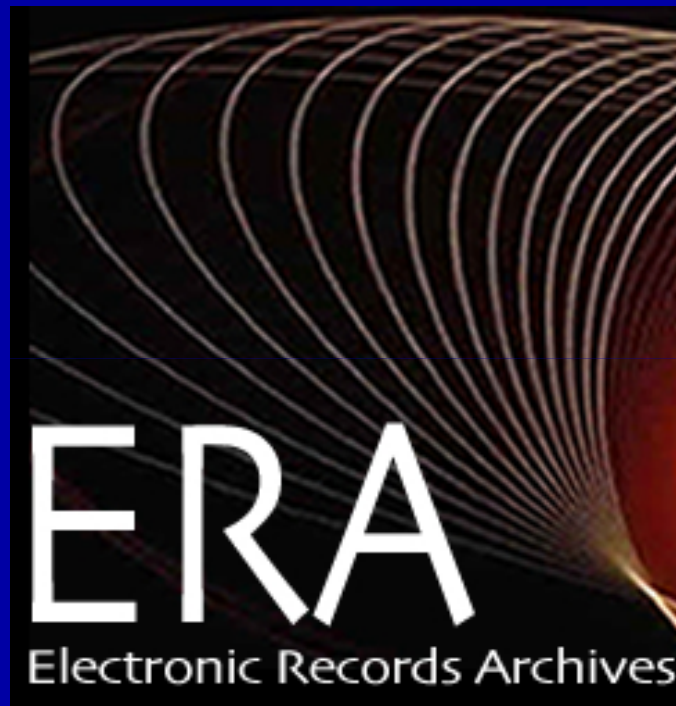
http://www.archives.gov/era
mailto: dyung.le@nara.gov, quyen.nguyen@nara.gov