ERA Briefing:

Indexing and Search

Dyung Le
Director of System Engineering, ERA
September 8, 2005

Presentation Outlines

- Overview of ERA program
- High level requirements
- COTS product capabilities
- ERA Challenges and Mitigations
- Design Approach
- Areas for Investigation
- Q&A

NARA's Strategic Response

The Electronic Records Archives (ERA) Program is NARA's strategic response to the problem of electronic records.

Its goal is to enable NARA to preserve and provide access to any type of electronic record created anywhere in the Federal Government.

ERA Vision Statement

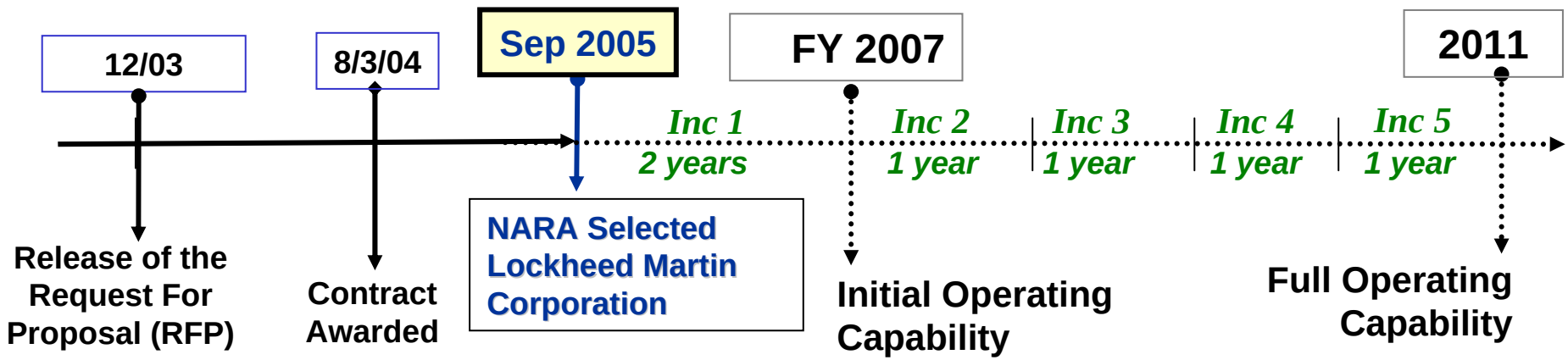
"ERA will ***authentically preserve and provide access to any kind of electronic record***, free from dependency on any specific hardware or software, enabling NARA to carry out its mission into the future."

The ERA Program: Where we are now

- 12/ 03 Release of the RFP
- 8/3/04 Awarded Two Design Contracts
- **9/8/05** **NARA Selects a Single Developer**

2005 - 2011 Five Increments (Inc) w/ Multiple Releases:

- FY07 Initial Operating Capability
- 2011 Full Operating Capability



Examples of how ERA will facilitate the Storing, Preserving and Accessing of Permanent Electronic Records

- Access electronic records that have been migrated to new formats,
- Output copies of electronic records in appropriate formats,
- Output electronic records on a variety of media,
- Search for records using a variety of search paths,
- Request access to or return copies of their own records,
- Receive online transfers of electronic records from NARA,
- View electronic records stored at NARA online,
- Authorize third party access to agency records held in Federal Records Centers, and
- Collaborate with NARA or other equity holders on review and redaction of records with restricted content.

ERA User Search Needs

- Both NARA and public users need search capability
- Public Users will expect:
  - Google-like capability for casual users
  - Researchers also need hierarchical search and browsing capability
- NARA Archivists and Agency Records Officers needs oriented to job function
  - Same capability as public, plus:
  - Searches in metadata, records lifecycle artifacts (records schedules, appraisal reports, transfer agreements, etc.)

High Level Formal ERA Requirements (paraphrased)

- 19.1: User selection of search characteristics (criteria)
  - Search by subject, time, agency/function, asset type, record type, etc.
  - Search using keywords, exact phrase, Boolean, proximity, wildcards, etc
  - Search using Question and Answer format
  - Search limited to descriptions, Records Lifecycle data, Authority Sources, records within a transfer

High Level Formal ERA Requirements (paraphrased)

- 19.2: Search by Content
- 19.3: Hierarchical searching
- 19.4: NARA created default searches
- 19.5: Selectable search complexity level
- 19.6: Control of search run times
- 19.7: Search progress indicators

High Level Formal ERA Requirements (paraphrased)

- 19.8: Display of search results
  - Special treatment for withheld restricted data
  - Identification of record versions
  - User selectable results size and level of detail
- 19.9: User capability to refine a search
  - Search within search results set
  - “More like this” search
- 19.10: Ability to select an asset from the search results set

High Level Formal ERA Requirements (paraphrased)

- 19.11: User ability to save search criteria
  - Ability to select and run a saved search
- 19.12: User ability to save search results sets
- 19.13: Management of Mediated Searches
  - User can request a mediated search and have a dialog with assigned searchers
  - System management and prioritization of mediated searches

High Level Formal ERA Requirements (paraphrased)

- 20.1: System must be able to to present all electronic record types
- 20.2: User ability to request copies of assets
- 20.3: System capability to output assets to media, and to print printable assets
- 20.4: Access to electronic records independent of hardware they were created on
- 20.5: Access to electronic records independent of software they were created on

High Level Formal ERA Requirements (paraphrased)

- 20.6: Capability to access an entire electronic record
- 20.7: Capability to access a set of electronic records
- 20.8: Capability to access a portion of an electronic record
- 20.9: Capability to access all components of an electronic record

High Level Formal ERA Requirements (paraphrased)

- 20.10: Output of assets in user selectable formats and media (from those available)
  - Ability to provide certified copies of electronic records
- 20.11: System will maintain authenticity of electronic records during access
  - Includes content, context, behavior, structure, presentation
  - Present or output all record components individually

Security Issues

- Access to records limited by classification, privacy, and other restrictions
  - Public limited to unclassified, non-restricted, and non-sensitive data
  - NARA users have more access, but limited by specifically granted access rights
  - Security features in ERA will manage access rights and control access to data
  - *Unclassified descriptions of classified data* and redacted versions of records will be available to the public

Enterprise Search Product Capabilities (Indexing)

Keyword or Semantic Indexing:

- Full text keyword indexing based on document content
- Works against unstructured as well as structured (e.g. databases) content
- Expanded indexing based on terms or phrases and semantic relationships (intelligent analysis of concepts and relationships)
- Some capability exists to index content of other non-textual data, including images, audio/video, geographic
- Multiple language indexing, with or without translation to English

Enterprise Search Product Capabilities (Indexing)

Categorization and Classification:

- Based on a defined taxonomy (hierarchically structured terms)
- Documents are categorized based on concepts found in the content and tagged to identify categories to which they belong
- Capability to manually force categorization of a document or to force promotion (relevance) of a document within a category

Enterprise Search Product Capabilities (Indexing)

Taxonomy Support

- Ability to create taxonomies from scratch based on analysis of themes and concepts found in the enterprise's documents
- Commercially available pre-packaged taxonomies, including governmental
- Ability to import or convert external taxonomies
- Tools to maintain taxonomies (add, modify, delete, move categories, collaborative workflow)

Enterprise Search Product Capabilities (Indexing)

Metadata Extraction

- Associates words or numbers (entities) in a document with named tags
- Out-of-the-box tags for organizations, people, phone numbers, email addresses, date/time, currency, products, etc.
- Libraries of known entities are available
- User-created tags
  - Ability to train the system by example to identify items via pattern matching
- Automatic extraction of author, date, subject

Enterprise Search Product Capabilities (Search)

Keyword or Concept Search:

- Content based searching against keyword index, based on user entered word string.
- Query terms expanded by synonyms, spelling variations, stemming, etc. to obtain more “hits”
- User input may include Boolean expressions, proximity operators, wildcards
- Use of semantic analysis of meaning or concepts to better determine relevance
- Multiple language search

Enterprise Search Product Capabilities (Search)

Hierarchical Search:

- Based on a taxonomy
- User traverses the category tree to refine search scope. Data corresponding to current branch is presented
- Ability to do keyword search limited to scope of current category
- Ability to traverse a second taxonomy with results limited to intersection of both taxonomies
- Hierarchical-like ability to drill down into structured data (databases)

Enterprise Search Product Capabilities (Search)

Dynamic Classification:

- Textual query string returns a list of categories which contain documents that are relevant to the search string
  - Complex logic to determine which categories with “hits” are worth displaying
  - Categories may be from a pre-defined taxonomy or completely dynamically created based on content of relevant records
- User can traverse the categories in same manner as for Hierarchical Search

Enterprise Search Product Capabilities (Search)

Profiling:

- User browsing or consumption patterns are monitored and analyzed
- Similar data given more relevance in queries
- New information meeting the pattern automatically routed to the user:
  - Email notification or placement in user folders
  - Recommendations for related information
- Identification of highly used data at enterprise level:
  - "what's hot" or performance tuning
  - Identification of communities of interest
  - Adaptive ranking for higher relevance

Enterprise Search Product Capabilities (Search)

Other Search Capabilities:

- Federated or distributed searching
  - Ability to integrate results from external internet search engines
- Management of stored queries at the user level
- Searches within user-owned or shared folders of saved documents
- Ability to dynamically jump between keyword and category searching

Enterprise Search Product Capabilities (Search)

Presentation:

- Viewers that can present documents in many native formats (200+) directly in the user's web browser (via HTML or XML)
  - Highlighted search terms
  - Dynamically generated links to similar documents as a document is viewed
- Automatically generated content summarization at document or query results levels

(Some of the) Technology Trends

- More relevant search results:
  - Increasingly human-like automated capabilities for development of vocabularies for and indexing or classification of collections of data.
  - Better understanding of information content, meaning, and relationships when performing searches.
  - More sophisticated searching techniques

(Some of the) Technology Trends

- Text mining: Software that analyzes content and relationships, and presents textual and graphical reports showing relationships between and insights about the information in a body of data.
- Profiling: More emphasis on subscriptions for information rather than direct searches. Google Sidebar is a good example.

(Some of the) Technology Trends

- Metadata: Longer term, can expect ever increasing amounts of standards based, subject specific metadata associated with records (web sites, databases, RMAs)

- Standards: Other than metadata, for which there is much activity, there is little standards development related to Indexing and Search.

ERA Challenges and Mitigations

Index Size

- ERA cannot afford storage or server costs for full text indexing and subsequent searching of all content
  - Approximately 30 petabytes of index for 100 petabytes of archived data, not counting backup of the index
- Some potential mitigations:
  - Only use hierarchical searching for records using combination of NARA generated and commercial taxonomies
  - Limit indexes to words within a defined vocabulary
  - Automatically generate limited metadata from records and constrain searches to that metadata
  - Full indexing of small numbers of high interest collections

ERA Challenges and Mitigations

Search Performance

- Potentially huge number of public users, with expectations of instantaneous service, but limited number of servers dedicated to indexing and/or search
- Single hierarchical storage system is bottleneck when retrieving records, especially if not cached
- Some potential mitigations:
  - Provide copies of high interest collections to 3rd party providers, and redirect queries for that data to those providers
  - Priority for access
  - Caching of high interest records, and/or extract high interest subset of info from larger records and promote its relevance

Record Data Types

- Huge number of electronic data type, from text, email, to pictures, video, GIS, etc…
- Some potential mitigations:
  - Framework design approach which will contain the services necessary to carry Searches and a selection mechanism to choose which are appropriate to fulfill a user's request.
  - Which strategies the design incorporates depends upon which ones are most necessary and which ones have the best balance between cost and performance, and quality of results returned.

ERA Challenges and Mitigations

Presentation - Viewers:

- Available viewers handle limited subset of expected data formats (hundreds vs. thousands) and can themselves become obsolete
- Some potential mitigations:
  - Immediately preserve the few hundred data types for which appropriate transformations are readily available
  - Specifically identify all other expected formats and seek out unique transformation software or services
  - Ultimately, the only solution is persistent preservation for all data types (a daunting challenge)

Presentation – Record Summarization:

- Lack of descriptions at the record level limits ability to present meaningful summary information about them in query results sets.
- Some potential mitigations:
  - Use search engine to automatically extract selected metadata for each record (or component document) as it is stored and present that metadata in query results set
  - Use search engine to intelligently summarize document content on the fly (not yet!)

Presentation – ERA Specific Requirements

- Capability to browse record collections based on the collection's original order
- Some potential mitigations:
  - Establish metadata fields for Collection ID and Collection Order, and associate it with each record in the collection during Ingest. Search for items with that Collection ID, and sort in Collection Order
  - Custom code for perusing records in a collection

ERA Challenges and Mitigations

Presentation – ERA Specific Requirements

- How to present complex records comprised of multiple components with different data types
- Some potential mitigations:
  - Use a metadata field to identify compound records by type, such as emails with attachments, tag the records in Ingest, and use custom logic by type, linked to from the search engine presentation module to display them
  - Use custom logic in Ingest to generate HTML or XML to represent the document structure and content, including links to the component documents. Store and index this representation as a document that the search engine can find and present directly.

COTS Design Lock In:

- Real risk of being locked into an enterprise search COTS vendor
- No standards for indexes, taxonomies, or other structures used by search engines. No standards for search engine interfaces. Each product's architecture is unique.

COTS Lock In: Some potential mitigations:

- Use COTS product that explicitly offers service oriented architecture compliant APIs. Isolate the application interfaces with an additional custom layer which always presents the same interfaces to the application regardless of the COTS product
- Limit utilization of COTS-specific capabilities that are entirely unique to that product, and are likely to stay unique
- Do not allow the ERA design to explicitly reflect the architectural peculiarities or limitations of the selected COTS product

ERA Search Design Concept

A Framework approach is envisioned for the ERA Search design. This allows a best-fit search engine to be used depending upon the nature of user's query. For example:

- Search by content vs. search by descriptions
- Text search vs. image search
- Federated search across multiple sites

The ERA System needs to use a combination of techniques in order to strike the right balance of breadth, depth, and response time for each search.

Basic Search Services

- Search by Hierarchy: for experts who understand the structure of NARA's collection and know where to go based on their knowledge and experience
- Search by Description: Often associated with sets of assets instead of with an individual asset
- Search Records Life Cycle Data: targeted toward the context of the records, the circumstances under which it was crated, used and transferred.

Search by Content

- Content Searches Using Performance Buffer
  - At least 50 percent of searches and accesses within the archives are concentrated on the 10 percent of assets that are the most popular.
- Non-cached Asset Indexing
  - Index the assets without adding them to the cache
- Targeted Entity Extraction
  - Index only selected terms instead of all content
- Full Content Search on All Assets
  - Full content searches on the entire archives may be necessary to fulfill court orders for exhaustive searches

Search Performance Efficiency

- Saving Search Indices
  - Saving search indices to deep storage can reduce the required time and processing resources for all agent-based searches.
- Prioritization
  - Based on user role, identity, class, fee for service
- Active Tape Access Management
  - Intelligent management of search requests which require access to the tape repository

Access Framework

- The access design accounts for the wide variety of assets, which need to be available to users by implementing access methods within a framework.
- Depending on the user's option selected, the system may need to transform assets to make an accessible copy of the asset.
- For output processes, the selected access option may include working with physical media.
- The access service also records the statistical information necessary to determine the popularity of assets and how they are accessed.

(Some of the) Areas for Investigation

1. Leverage existing Search infrastructure
   - Reuse of commercial “free” services
   - Reuse existing government Search capabilities
   - Become components of government wide MetaSearch
2. Metadata definition and design
3. Enhanced navigation with technology such as Topic Maps
4. Deep databases

Questions & Answers

dyung.le@nara.gov