ERA Challenges

Draft Discussion Document for ACERA: 10/7/30

ACERA asked for information about how NARA defines ERA completion. We have a list of functions that we would like for ERA to perform that we're managing as we prioritize corrective and adaptive maintenance in FY 2012 and beyond. However, since we are closing the books on the original set of requirements (68% complete as of September 2011) we do not currently have a document that defines completion. In fact, many features of ERA were designed for maximum flexibility and adaptability. We are planning for continuous change in the system over time as the nature of electronic records change and as researcher expectations for how those records would be delivered change.

That being said, NARA understands that there are a few significant unsolved problems that stand between us and the ideal state of well-managed, preserved, and accessible permanent electronic records. In order to benefit most from the expertise of ACERA, we are sharing these challenges with you at the earliest stage when we are still defining the problems. This is the point when your expertise can help us the most: we have not yet framed long-term solutions to these problems and your input can help us decide the best path forward as we try to address them.

NARA would welcome discussion and input from ACERA members on any of the topics listed below. ACERA subcommittees might also find topics of interest here to include in their recommendations.

Overall

Need for Scaled Down ERA - Is there still a demand for a scaled down version of ERA for other institutions to use, given that there are now COTs (Rosetta, Tessella) and GOTs (Washington State) products and services that offer digital preservation to smaller institutions? What would other institutions want to receive?

General Scalability of Processes – One of our overall engineering challenges is how to develop new functions fast, do it well enough, do it so that it scales, and do it with limited resources.

How could NARA move toward a different way of doing business to avoid being overwhelmed with the workload and falling further and further behind? How could we develop much faster, probably leveraging off FOSS products, simplify the requirements process, accept a good enough nimble and quick solution. Could we move our architecture to a point where it becomes even lighter, simpler and easier to modify than it is right now (taking advantage of scripting, instead of Java, for example)? Can we convert ERA to open source, so that we can share with the rest of the community, and benefit from their add-on values?

ERA Interoperability – How could or should ERA evolve to better integrate into the larger framework of information and service sharing? What types of APIs/services make sense to expose to the public to be consumed? What interoperability standards (data or services) should NARA be targeting to best support information sharing?

Managing Records in Originating Offices

Records Management Services - Can records management services, now an approved OMG standard, be used in a practical setting to achieve greater rates of appropriate capture and disposition than traditional RM processes with a Records Management Application? What should the relationship between RMS and ERA be?

Getting Records In

Avoiding Physical Transfer **–** Currently, the most practical way to transport large volumes of data between agencies and ERA is by loading servers or storage devices onto a truck. Is there a way to avoid this process altogether? One idea is to encourage a shared Federal cloud storage environment that is appropriate for records still under agency control and also appropriate for NARA's archival storage. If such a cloud environment could be established, and a standard set of records management metadata control could be applied to all records in the environment, NARA could accession and take custody of electronic records by flipping a metadata switch. The records themselves wouldn't have to move anywhere. What steps could NARA take to move toward a model like this? Are there other ways of approaching the problem of transferring large volumes?

Crawling for Records Capture – There are advantages to capturing electronic records by crawling the environment where they are stored rather than relying on the agency to correctly file and transfer everything in that environment. Web sites are an obvious application of this, but there are challenges related to the deep web associated with this approach. What can be harvested this way, and what is missed? How could more be harvested? A related idea is crawling or harvesting content from agency share drives or related storage facilities, and then applying categorization tools to establish series or other bodies of permanent records. How could NARA best use the tools available now to capture records?

Finding people to write data transfer scripts - We anticipate encountering permanently valuable records in a variety of native applications in Federal agencies. These records, which have their own internal storage and metadata models defined by the applications in which they were created, will need to be transferred to ERA as automatically as possible. Accordingly, the export methodology and associated metadata of these native records should be closely aligned with the receiving, or import, methodology and metadata of the ERA Asset Catalog Entries as this record material is ingested to ERA. Examples of such native applications might be MS Outlook or GroupWise e-mail, MS SharePoint, records management applications, or Digital Asset Management systems. In order to transfer records in the most reliable, authentic, and efficient way, scripts or other automated data-exchange mechanisms may have to be written for each application using the existing ERA data model and corresponding application APIs. NARA is not currently staffed to do this kind of coding, and this kind of technical expertise changes over time. Are there ways for NARA to partner with university computer science departments, specialized volunteer communities, or other sources to get help with these kinds of needs?

Automatically Identifying the Boundaries of a Record – The PREMIS metadata model allows us to track relationships among files that comprise an Intellectual Entity (essentially a record), but how can agencies use ERA to transfer records to us in a way that explains these relationships? How can we use ERA transfer processes to populate this PREMIS metadata in an automated way? We cannot think of a way to handle this except developing templates for specific recurring sets of internally consistent, well-defined records. The practical problem is that many of the transfers we get from agencies are not like that; we are increasingly getting mixed collections of office automation records from high-level agency leaders that are miscellaneous in format, complexity, and organization.

Managing Compound Records – A subset of the problem above. In the short term, does it make more sense to focus on responsibly and reliably capturing and preserving aggregates of records (e.g. PST, WARC) for which we know the structure, and worry about disaggregating and defining record boundaries later? What are the implications of this decision for ingest processing, preservation, and access?

Balance between transfer standards and capture of all permanently valuable records – NARA believes that there is tension between our goal of capturing all permanently valuable electronic records and transfer standards that limit and control how records must come to us. It is to our advantage for managing and preserving records to have strict standards for transfer formats, transfer metadata, and transfer structure templates. These standards would allow automatic processing of internally consistent sets of records that met those criteria. However, many permanently valuable records in agencies would not meet such criteria, are too heterogeneous to develop structural templates for, etc. How can NARA best maximize the number of permanently valuable records we capture and can preserve, while also maximizing the amount of automation we can apply to management of the records? Can we apply the 80/20 rule and accept that many records will not be able to use the same streamlined processing paths developed for the most consistent and structured ones? What are the implications of such a double path?

Balance Between Data Wrangling and Archival Purity – Related to the questions above, we have heard colleagues working with other digital preservation repositories acknowledge that the most costly, time-consuming, and human-intensive part of their process is the step between identification of records appropriate for ingest and actual ingest. During this step data specialists (sometimes referred to as “data wranglers”) do whatever is required to get the original data into a form that the repository can accept, process, and make available. Most of these institutions are probably digital libraries, not archives, which may be an important distinction for this problem. However, archival theory generally prohibits alteration of records received from the creators; the archival responsibility is to preserve what the creator created. Archivists could still modify records to increase usability, but what they created in so doing would be merely modified use copies, not the authentic records. NARA has not resolved this tension: the more we modify the records received to standardize them, improve their metadata, etc., the easier and more automated their ongoing management and access could be. How are other archival institutions resolving this tension? Does it matter what researchers expect and prefer given a choice between authenticity and usability? Does the sheer volume of records we will be receiving argue for less human wrangling and more emphasis on providing better access tools to sort through whatever we accept, however difficult it is to use?

Balance between Technical and Legal Risk – In choosing a transfer and preservation format for e-mail, for example, there are technical and archival reasons why the product of an entire e-mail account might be transferred (perhaps as a PST file) but there are legal liabilities associated with accepting information that was not scheduled as permanent and which should have been destroyed. Calendars, deleted files, etc. could all exist in the PST and be FOIA-able or discoverable in a legal action, for instance. How should an archives balance its conservative instinct to keep an exact copy of what was transferred as a safeguard with the risk of possessing or even accessioning information it shouldn't have?

Storing and Preserving Records

Format and software obsolescence - NARA's current understanding of the mitigation strategy for the risk of format obsolescence is tracking the formats we have, monitoring technology changes for impending obsolescence, and developing migration or other preservation strategies for formats that are becoming obsolete. This process could become increasingly difficult to manage as the number of formats and the number of compound records that include multiple formats increase over time. Are there other practical approaches for an archives like ours to take? Or is this a problem that is likely to get easier over time rather than harder, as current and backward compatibility emerge as major themes in technological progression?

What benchmarks and metrics exist for success in preserving electronic records? What is the cheapest and most reliable current approach or system for the preservation of digital information? Is any rate of data loss inevitable or acceptable? Must every bit be preserved? Where are the highest costs, and can those costs be driven down? (Is ingest processing, including preparation of the data for ingest, always the highest cost lifecycle stage?)

Managing Metadata about Records

Format Identification - NARA needs to automatically identify the formats of records transferred, record the format in the metadata if known, and allow NARA to provide information from other sources about the format if it cannot be automatically identified. Current tools for format identification lag behind the formats in our actual collection so we are always managing some records of unknown format or records where we have recorded format information as provided by the record creator. As additional format identification tools are added to DROID, for example, we would re-run the programs and get different results than before, in some cases better results, but support for some formats might be dropped. Our current metadata structure and ingest processing tools cannot manage format identifications from multiple sources or provide an indication of whether a format ID should be overwritten by a newly updated tool. How have other archives managed this problem?

Efficient and Robust Metadata Extraction – Record format is one type of technical metadata, but we have a need to expand our ability to quickly and accurately extract technical and contextual/descriptive metadata from the incoming records. Sometimes the line between technical and contextual metadata can blur. Example: extracting out GPS coordinates and timestamps from images, while technical in nature, can provide good contextual information about the image and be very useful for researchers for later discovery. What tools and standards should NARA consider for extracting and recording technical and contextual metadata?

Automating the Generation of Descriptive Metadata - Descriptive metadata, at the item or Intellectual Entity (record) level, allows for very powerful discovery by the researchers. Using tools that parse text to create abstracts or summaries in an automated manner is an example of descriptive metadata generation that may be useful in allowing researchers to find relevant records. An example of how this might be useful is in the case of a user's search that retrieves a very long list of responsive items with numeric file names. A researcher might have to open or view every file to see what it was. However, if automatically generated names or short descriptions could be provided along with the file names, the researcher could select records of interest more easily. This approach raises issues about the quality of such summaries, the need or practicality of human review, the actual tools used for the summarization, etc. Are there examples of other institutions successfully incorporating tools of this type into their processing streams?

Normalization of Vocabulary and Metadata – If NARA is successful in automating metadata extraction or descriptive metadata generation, we will run into the issue of normalizing metadata terms from divergent record sets and sources. What technologies should NARA explore to help with thesauri, synonyms, controlled vocabulary and search terms? Are there examples of large scale applications for these technologies we could study?

Getting Records Out

Human Review and Redaction Bottleneck - Most series of electronic records we receive have the potential to contain information that should not be released to the public because of the legal requirement to protect personal privacy, confidential business information, national security information, and other specifically protected categories of information. In order to release any information from such series, NARA currently has to do human page-by page review to ensure that protected information is not released. The review and redaction process is currently a significant bottleneck between bringing records in and making them available.   We can't take all records uncritically. We can't **not** take them. We want to explore using the auto-discovery and auto-classification tools that are available today to process massive volumes and speed the human part of review. Such tools are less than perfect. What are the implications of using imperfect tools? How are other institutions balancing the risk of disclosure with the need to review huge volumes of records?

Researcher Use of E-Records Collections - How do researchers expect to interact with large bodies of miscellaneous electronic records (not just data sets)? What new kinds of research could be done with such collections of Federal records (textual analysis, analysis of patterns of communication, topic clustering to find dominant issues in a Federal agency during a time period)? What new tools for analyzing such collections might be useful to researchers? How could the archives share information with researchers about ways to interact with the collections?

How can E-Records Increase Citizen Engagement with Archives? - How could electronic records available online help interest citizens in the history and government of the United States and in the value of archives? How can we use the easier accessibility of electronic records online to increase citizen engagement with records and the importance of records in democracy? NARA is already exploring ways to do this, including making sure our records are on popular social media site where more people will find them, but there may be additional ways to do this that we haven't considered.