TRANSFORMATIONS

A. Introduction

Transformation Assumptions

In addressing the issue of how best to provide electronic records for researchers, NARA will employ several different strategies to meet the needs of its researchers, including providing access to records in their original formats through the use of current viewers or browsers, or by transforming records to current ubiquitous data file formats. Since the transformation option will necessarily be the most expensive, the balance of this white paper focuses on the circumstances around which NARA elects to transform electronic records.

The ERA Requirements Document defines transformation, as follows: "The process or the results of a process, of reformatting or otherwise changing the way an electronic record is digitally encoded in order to reduce or eliminate dependencies on specific hardware or software, while preserving authenticity."

In making transformations, then, NARA makes the following assumptions:

1. NARA will always retain the original bit stream of archival materials and the researcher may always get access to it, even when NARA's primary delivery method for the record is another version.
2. NARA will select which version it wishes to use to make future transformations, which could be either the original bit stream or a subsequent version that NARA had created.
3. A cost consideration in making transformations is to create a new version in a data file format that is prevalent throughout NARA's holdings, is in wide use in the IT market, and is in a sustainable format[1]. This will probably reduce maintenance costs, since fewer data file formats will need to be tracked and maintained over time.
4. NARA might transform all or some records in a series/transfer group or all records in a specific data file format across all series/transfer groups, depending on the needs of the research community and what is most cost effective.
5. Making transformations for preservation or reference purposes might entail different strategies.

[1] See Appendix C.

When To Transform?

Because we expect there to be costs associated with transformations, NARA will probably not perform transformations on a routine basis, unless it is cost beneficial to do so. NARA may elect to transform on ingest, transform later in the lifecycle, transform when a record is requested, and NARA may or may not save every new version created. Engineering/cost/customer experience considerations will determine how we choose. For this reason, then, NARA will need to decide when electronic records need to be transformed. The decision to transform records is driven by two events:

- That the data file format is at risk of obsolescence or
- NARA's research community requires an enhanced level of access, or
- A combination of both.

B. Transforming Records For Reasons of Technical Obsolescence

Obsolescence

NARA understands obsolescence of electronic records in the way that the National Library of Australia defines it. Obsolete records, according to the NLA, are "those [on data file formats] that can no longer be rendered. . . . In preservation terms, a file format only becomes effectively obsolete when access is no longer possible." In other words the question about whether a data file format is obsolete is a yes/no question.

The NLA goes on to note that obsolescence, however, "describes a state of becoming obsolete, rather than a state of already being obsolete." In other words, unlike the concept of obsolete, obsolescence is not a yes/no question, but is better understood as a sliding scale with some records being at low risk of obsolescence, while others being in a middling level of vulnerability, while for others we are nearly unable to effectively render them at all. The NLA concludes: "The purpose of obsolescence risk assessment is to inform decisions about the need to take action."[2] The following figure is meant to display graphically the relationship between obsolescence and obsoleteness, and the strategies that NARA might pursue depending on the level of obsolescence risk associated with the data file format in question.

[2] David Pearson and Colin Webb, "Defining File Format Obsolescence: A Risky Journey," The International Journal of Digital Curation, 2008. p. 93.

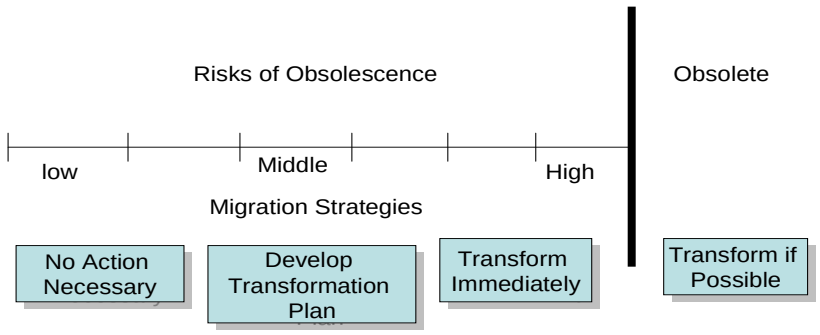


Figure 1. Degrees of Obsolescence

In applying concepts of obsolescence, then NARA adopts the following guiding principle in determining when records should be transformed:

> Are members of the research community who are likely to want to use a record, reasonably able to render and use it to meet their research needs?

NARA believes that it has met this goal if the records are available in a data file format that:

- is ubiquitous in the marketplace,
- is affordable to the research community that wants to use the records,
- effectively conveys the significant properties of the records, and
- is trusted by the research community as authentic (see preservation Principle 1, Authenticity).

This aligns with NARA's Preservation Principle 4: "NARA must make reasonable efforts to provide access to its electronic records holdings via data file formats that are ubiquitous in the marketplace and affordable to its customers."

Conducting Risk Assessments of Electronic Records

Records Preservers are responsible for determining the extent to which the data file formats in NARA's holdings are both ubiquitous in the marketplace and that the tools to effectively render them are affordable to the research community most likely to use the records. Records Preservers monitor the marketplace and the efforts of other like minded institutions to determine the current status of data file formats. As NARA gains more experience in assessing electronic records, it may wish to expand the list of criteria that it wishes to evaluate.

Ubiquity of a Data File Format

Records Preservers examine such issues as the market share of software products and the extent to which they were being supported by the vendor that produced them or other third parties. For example, Word 2007 is currently ubiquitous in the marketplace, is supported by Microsoft, and can also be viewed by third party software, such as Open Office. In addition, Microsoft produces and supports the freely available Word Viewer software.

Affordability of Tools

Records Preservers also are responsible for understanding the tools that a researcher might use to render records in NARA's holdings. Tool sets might include, for example, the original software that produced the record, viewers, or emulators. To continue with the example of a data file in Word 2007, the original software is moderately priced (around $150), but Open Office is freely available software and provides most of the functionality of Word 2007.

Analyzing Risk Issues in Electronic Records

For all of the data file formats that are in NARA's holdings, then Records Preservers analyze the technical characteristics of data file formats and develop alternative transformation strategies. Initially they focus their attention on those data file formats that are obsolete and then those that have the highest risk of obsolescence. They record and share the results of their research with other NARA Records Preservers, as well as Records Preservers at like minded institutions, in a knowledge management system. The Florida Center for Library Automation provides an excellent example of the format and content these research papers might take.[3]

Determining Levels of Risk

Based on their research, Records Preservers record risk of obsolescence scores in a Risk of Obsolescence Tool (ROT) for each data files format and version in NARA's holdings. The ROT tools exist among the Preservation Strategies Components (PSC) tools set (See Appendix A: NARA Electronic Records Preservation Model).

Figure 2 below provides a graphical example of what the ROT might look like. In this example we are using a scale of 1 to 10, where 1 for ubiquity means that the data file format is widely ubiquitous in the marketplace. A score of 1 on the affordability scale means that the tools the research might need to render the data file are free or very inexpensive. A score of 10 means that the data file format occupies a very small market share and a score of 10 for affordability means the tools are is expensive or perhaps not

[3] See http://fclaweb.fcla.edu/content/fda-format-information-and-action-plans

even available any more for the research public to obtain. The tool would multiply the two scores together and the resulting product would give Records Preservers a Risk Score and an indication of the actions needed, based on the score.

Data File Formats	Ubiquity (1-10)	Affordability to the Research Community (1-10)	Risk Score	Actions Needed
WordPerfect 4.2	8	8	64	Transform Immediately
PDF 1.1	5	5	25	Develop Transformation Plan
TIFF6.0	2	3	6	No Action Necessary

Figure 2. Sample Risk of Obsolescence Tool Report.

The Risk Scores could be adjusted by Records Preservers, so that when periodic reports are run (say once a month), then attention can be focused on data file formats that are most at risk.

Risk Score	Actions Needed
50-100	Transform Immediately
25-49	Develop Transformation Plan
0-24	No Action Necessary

Figure 3. Risk Score Table

Developing Preservation Driven Transformation Plans

NARA's Records Preservers develop Preservation Driven Transformation Plans for those data file formats that have a risk score in the 25-100 range. The Preservation Driven Transformation Plan would include a technical analysis of the data file format (see the Florida model) and several recommended transformation strategies along with the costs associated with each. In addition, each option would indicate the kinds of behaviors, appearance, and structural characteristics that proposed target formats will support (see white paper on Significant Properties). A sample outline of a Preservation Driven Transformation Plan is provided in Appendix B.

Usefulness to the Research Community

At this point Reference Archivists evaluate how the options presented in the Preservation Driven Transformation Plan will impact the researchers who are most likely to want to use the information in the records. Reference Archivists monitor the relevant research communities for trends in research to inform their determinations, based on the current and projected use of records. A Data File Format Specific Report is run for the data file format that is deemed at risk from a technical perspective that identifies the specific record groups and series or transfer groups where data files exist in that data file format among NARA's holdings. A sample of the report is presented in Figure 4 below.

DATA FILE FORMAT REPORT

Data File Format: **WordPerfect 4.2**

Risk Score: **64** (Ubiquity = 8; Affordability = 8)

For this data file format, data files exist in the following Records Groups and series/transfer groups:

Record Group	Series/ Transfer Group Title	Number of Data Files	Number of Data Files with Compound Records	Record Type	Significant Properties
RG 1	Reports	10,000	5	Textual	Default for Textual records
RG 2	Press Releases	20,000	50	Textual	Default for Textual records
RG 3	Studies	30,000	70	Textual	Default for Textual records
RG 4	Subject Files	20,000	1	Textual	Default for Textual records
RG 5	Reading Files	20,000	0	Textual	Default for Textual records
Total Num. of Data Files		**100,000**	**126**	Textual	

Figure 4: Sample Report from PSC Querying Tools: Data File Format Report

Making the Decision to Transform

Records Preservers and Reference Archivists collaborate to decide the best preservation strategy for each of the series/transfer groups that are impacted by the high risk data file

format. Considerations of cost and the extent to which the target data file format will meet the needs of the user communities are paramount in making this decision. Note that a different strategy might be made for different series/transfer groups, depending on the properties that archivists determine are significant to making the records usable to the research community. Once a decision is made it is recorded on the plan and the transformation is carried out. (See a separate white paper on the Preservation and Access Components).

Scenarios: Transformation For Preservation

At Ingest Scenario

- A creating entity transfers to NARA an accession of textual electronic records in the data file format, WordPerfect 4.2.
- During the ingest process, the data file format and version, along with a variety of other technical and archival data are captured and recorded in the ACE. (See the ID Records and Formats step in the Accessioning workflow in Appendix D.)
- Records Preservers generate a ROT Report for WordPerfect 4.2 that indicates that data files in WordPerfect 4.2 have a high risk score. This means that Records Preservers have already analyzed the data file format in terms of its ubiquity in the marketplace and affordability of tools available to the likely research community and have determined that the research community will have a high degree of difficulty using the information in the files. (See the Risk of Obsolescence step in the Accessioning workflow in Appendix D.)
- A Preservation Driven Transformation Plan already exists with options on how to transform the records.
- The Records Preservers refer the Preservation Driven Transformation Plan to Reference Archivists who examine it to determine the best option that will most align with the needs of the research community. (See the Access Questions step in the Accessioning workflow in Appendix D.)
- Selecting from among the options presented on the Plan, Reference Archivists and Records Processors determine that transforming the data files to PDF1.6 will preserve the significant properties of the materials. The research community will, for example, benefit from being able to use features such as "find" and the physical format of the data files will be preserved.
- The Reference Archivists and Records Processors notify the Records Preservers that they should transform the data files into PDF 1.6.
- The Records Preservers take steps to transform the files to PDF 1.6. The new version in PDF 1.6 is ingested into ERA with references in its ACE to the original native version that indicate that the PDF files are a preservation version of the original.
- Researchers who want access to these records will be provided the PDF version, but could get a copy in WordPerfect 4.2, if they requested it.

Post Ingest Scenario

- Fifteen years ago, NARA accessioned a series of textual electronic records in the data file format, WordPerfect 4.2.
- Last month Records Preservers completed a technical paper on the properties of WordPerfect 4.2 indicating that it was not longer supported by the original company or any third party. Its risk of obsolescence score was in the high range.
- The Records Preservers develop a Preservation Driven Transformation Plan with several options: transform from WordPerfect 4.2 to PDF 1.6, or transform from WordPerfect 4.2 to Word 2007.
- They run a Data File Format Report for all records in NARA's custody in WordPerfect 4.2 and refer the report and the Preservation Driven Transformation Plan to Reference Archivists who have custody over the records.
- Reference Archivists review the significant properties of each of the series identified in the report against the proposed transformation options.
- Reference Archivists determine that transforming the data files for most of the series to PDF1.6 will preserve the significant properties of the materials. The research community will benefit from being able to use features such as "find" and the physical format will be preserved. The Reference Archivist reject transforming to Word 2007, because the cost of the software to the research community exceeds that of the Adobe Reader and that the functionality that could be gained from Open Office was not needed by the research community. The costs of the two options are about the same.
- The Reference Archivists therefore notify the Records Preservers that they should transform some of the data files into PDF 1.6.
- The Records Preservers take steps to transform the files to PDF 1.6. The new version in PDF 1.6 is ingested into ERA with references in its ACEs to the original native version that indicate that the PDF files are a preservation version of the original.
- Researchers who want access to these records will be provided the PDF version, but could get a copy in WordPerfect 4.2, if they requested it.

C. Transforming Records to Enhance Reference Service

The previous section of this paper described the circumstances under which NARA would transform records to address issues of technical obsolescence. However NARA might also wish to transform electronic records for purposes of providing enhanced reference service to a body of electronic records whose risk of obsolescence might be low.

Currently, for example, NARA has taken its most popular databases and entered the entire schema and supporting code lists from the databases into Access to Archival Databases system (AAD). Researchers can then locate information they want by using a free text search box. Note that this kind of transformation activity is not needed to

address issues of obsolescence, since the data files could be rendered by the research community. Researchers could take a copy of the raw data and search for the information they wanted using their own software. So from a preservation perspective, this example would not trigger a need to develop a Preservation Driven Transformation Plan. Nonetheless, NARA needs to be able to account for the need to transform records to meet the needs of the research community for enhanced levels of access to electronic materials.

Levels of Service

Several years ago, NARA developed a paper on "Levels of Service" that addressed the broader issue of how NARA would address the needs of its research communities.[4] NARA argued that it would offer one of three levels of service to the records it provides to researchers. At the first level NARA would provide just the original records in whatever format NARA received them, while for records at the second level of service, NARA would transform the records into something that would provide a significantly higher level of functionally, such as is done for records in AAD currently. At the third level of service, NARA would provide specialized tools to enable researchers to analyze the content of the records. Broadly speaking NARA still adheres to this concept, but NARA has evolved in its thinking to believe that a better way of conceptualizing the process is not just three levels of service, but as a menu of many options that depend on factors such as the record type, data file format, and significant properties of the electronic materials, as well as the needs of the researcher community. The number of options available to NARA in this iteration of this concept is greater than the three levels of service that we had originally conceived and should provide for NARA and its research clientele a more nuanced approach to meet the needs of researchers.

Criteria for Making Reference Driven Transformations

The criteria NARA would employ to determine if a body of electronic records needed to be transformed to meet researcher's needs for enhanced access include:

- Increased reference demand for the materials,
- The needs of the research community that uses the materials,
- Available transformation options, and
- The resources needed to perform the transformation

This aligns with NARA's Preservation Principle 4: "NARA must make reasonable efforts to provide access to its electronic records holdings via data file formats that are ubiquitous in the marketplace and affordable to its customers."

[4] See Appendix E for the full text of the report.

Research Demand

Reference Archivists monitor the needs of the research communities that use the records that the Reference Archivists service. When records are accessioned into ERA, Reference Archivists and Records Processors note the expected research communities, the level of use that might be expected, and any needs to transform records. They also note whether the data file format that the creating entity is sending to NARA needs to be changed to better meet the needs of the research community. If they believe that some other data file format might be beneficial, they need to complete a Reference Driven Transformation Plan. See the section below on developing the Plan.

In some cases, where we know in advance that we always want to perform a transformation under certain conditions, Reference Archivists could prepare a Reference Driven Transformation Plan that could apply to all records of a certain record type and data file format. For example, we may know in advance that for all transfers of photographs, NARA will always want to produce thumbnail images, if the creating entity did not include them in the transfer. In this situation, Reference Archivists could prepare the plan once and Records Preservers could execute on it for any transfer that came in that matched the criteria.

Once records have been ingested into ERA and are available to NARA only or public access, Reference Archivists monitor the extent to which the data file formats are meeting the needs of the research community. They are supported by Preservation Strategies Components (PSC) Reference Reports, which provide Reference Archivists information such as the numbers of times that each series is queried and downloaded by researchers. Figure 5 provides an example of what such a report might look like.

REFERENCE REPORT

Time period: First Quarter, FY 2009

RG	Series/TG	Num of Ref Requests	Num of Hits	Num. Downloads
4	19th century immigration records	200	200,000	20,000
1	WWII draft cards	150	100,000	50,000
2	Reports	6	5,000	5
3	memos	0	2	0

Figure 5. Sample Reference Report

As NARA gains more experience in using the reports, Reference Archivists can set parameters so that the reports can provide information about only those series that have more than a certain number of hits or downloads in a set period of time (like a month). Reference Archivists use information from the Reference Reports and their knowledge of

the relevant research communities to investigate whether NARA should consider enhancing the level of service. So for example, NARA may decide that all series that have more than 10,000 hits or 5,000 downloads in a quarter need to be investigated. Records represented by RGs 1 and 4 from the fictitious report shown in Figure 5, for example, would then trigger an action on the part of Reference Archivists to inquire into what level of service might better support the needs of the research community.

Needs of the Research Community

Regardless of how Reference Archivists are made aware of the possible need for enhanced access to a body of materials, next they need to have a clear idea of the research requirements of the community. This could be accomplished by holding focus groups or conducting surveys with members of the research community to better understand the needs of the community and how NARA might support those needs.

Developing Transformation Plans

If Reference Archivists are persuaded that NARA needs to transform records to meet researchers' needs, then Reference Archivists generate a Reference Driven Transformation Plan. A copy of what such a plan might look like is provided in Appendix F. On the plan, Reference Archivists note the record identification information, reference demand, and the results of any focus groups or surveys they held with the research community. The plan is forwarded onto the Records Preservers to evaluate technical options that might meet the needs of the research community. In addition, each option would indicate the kinds of behaviors, appearance, and structural characteristics that the target format will support (see white paper on Significant Properties). Their recommendations are recorded in the Reference Driven Transformation Plan, along with the resources required to perform each transformation option. One option for example, might be to transform records on the fly when researchers want to download them.

Making the Decision to Transform

Records Preservers and Reference Archivists collaborate to decide the best transformation strategy to employ. Considerations of cost and the extent to which the transformation strategy will meet the needs of the user community are paramount in making this decision. Once the decision is made, it is recorded on the Plan and the transformation is carried out. (See a separate white paper on the Preservation and Access Components).

Scenarios: Transformation For Enhanced Reference Service

At Ingest Scenario

- NARA accessions a database of records of 19th century immigration records in ASCII. Since the risk of obsolescence for this data file format is low, Records Preservers do not develop a Preservation Driven Transformation Plan.
- Reference Archivists however believe that these kinds of records in the past have been of high interest among genealogists looking for individual ancestors. They believe that providing the records in their native format will not be what the research community will want or expect. They develop a Reference Driven Transformation Plan.
- Records Preservers explore the options and recommend that NARA consider preparing these records for an AAD type tool. The Reference Archivists agree.
- The schema of the database is entered into ERA along with the associated code lists.
- The series is available to the research community via an AAD like application.
- The transformed reference version is ingested into ERA and its ACEs indicate that it is a reference version and note the associated preservation version.
- The transformed records are a "reference version" of the series, while the ASCII version is considered both the original bit stream, as well as the "preservation version" of the series.

Post Ingest Scenario

- Ten years ago NARA accessioned a series of photographs in the JPEG data file format. Records Preservers are still satisfied that this format has a low risk of obsolescence. At the time of ingest, NARA's Reference Archivists had recommended that researchers would be well served by maintaining these records in their original format.
- In the intervening years, however, research demand for the photographs has increased dramatically and has stayed high for several quarters. Reference Archivists decide that they should investigate why interested has spiked and if the research community is satisfied with the current level of service. In talking with the community, they learned that the community would like to have thumb nails available to them on the search results page. This would speed the search process for them. Reference Archivist completes a Reference Drive Transformation Plan noting these facts.
- Records Preservers complete the optional portions of the Plan, noting the costs to develop thumbnails and displaying them to researchers on the search results page.
- Reference Archivists are satisfied with the solution and ask the Records Preservers to develop the thumb nails.
- The transformed reference version is ingested into ERA and its ACEs indicate that it is a reference version and note the associated preservation version.
- The transformed records are a "reference version" of the series, while the JPEG version is considered both the original bit stream, as well as the "preservation version" of the series.

One-Offs Scenario

- A researcher wants NARA to make a copy of some electronic records in a data file format different than the versions NARA has made publicly available.
- Reference Archivists complete a Reference Driven Transformation Plan, noting that this is a transformation based on the needs of a single user. Reference Archivists investigate the extent to which the transformation may have any lasting benefit to NARA or its researchers. The record their findings in the plan.
- Records Preservers need not develop options, but just note the cost to perform the transformation.
- NARA needs to consider the extent to which this is to be provided for a fee.
- If the copy is to be retained by NARA, then the transformed reference version is ingested into ERA and its ACEs indicate that it is a reference version and note the associated preservation version.
- The transformed records are a "reference version" of the series, while the original version is considered both the original bit stream, as well as the "preservation version" of the series.

D. Successful Transformations

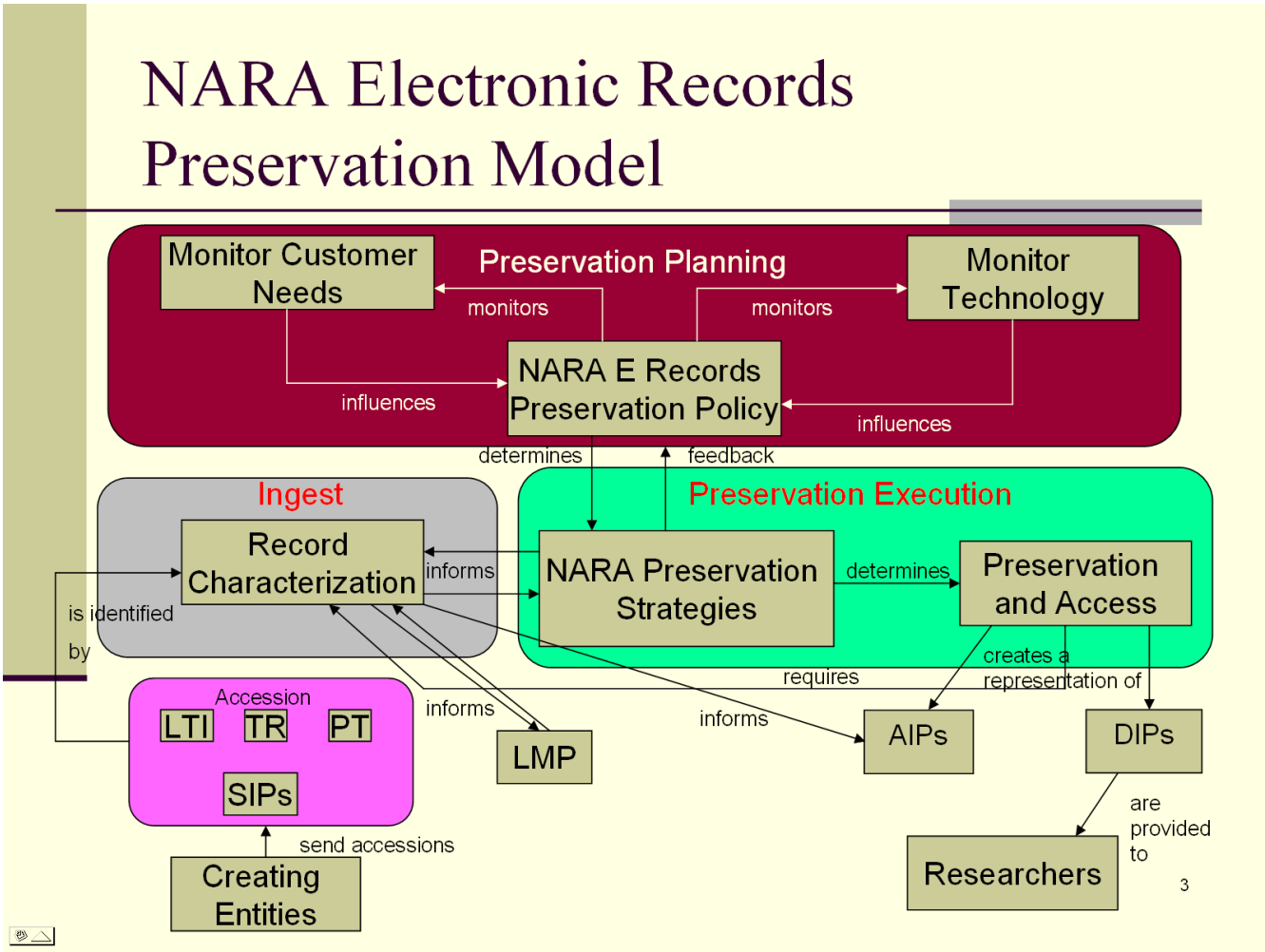
At the point of transformation, the test that any given transformation strategy was successful becomes the extent to which it produced an authentic version of the original. Authentic versions must:

- Be free of corruption,
- Effectively reproduce the significant properties of all records that were transformed,
- Retains all of the attributes and relationships detailed in the AIP of each record, and
- Be able to document the processes and decisions by which the transformation took place to satisfy the needs of official bodies that may need NARA to authenticate the records

E. Follow Up

1) Where will the information referenced in this paper be recorded? How does the LMP, for example, fit into this model? What information is stored in the ACE?
2) Are any of the artifacts mentioned in the paper supposed to be an ERA business object? Examples: The Risk of Obsolescence Tool, Preservation Driven Transformation Plan, Reference Driven Transformation Plan, Reports,
3) What portions of the paper need to be deployed in the prototype before others?
4) Complete a white paper on Preservation and Access Components.

Appendix A. NARA Electronic Records Preservation Model



Appendix B. Sample Preservation Driven Transformation Plan

Transformation Plan - WordPerfect 4.2

Number of data files in this data file format: 100,000

Technical Characteristics of WordPerfect 4.2 – XXX

Market Share of WordPerfect 4.2 - XXX

Tools that users can use to render WordPerfect 4.2

- Tool A – pros and cons
- Tool B – pros and cons
- Tool C – pros and cons

Transformation Strategies:

- Option A: Transform to Word 2007
  - Appearance – pros and cons
  - Behavior – pros and cons
  - Structure – pros and cons
  - Costs to Transform
- Option B: Transform to PDF 1.7
  - Appearance – pros and cons
  - Behavior – pros and cons
  - Structure – pros and cons
  - Costs to Transform
- Option C: Transform to TIFF 6.0
  - Appearance – pros and cons
  - Behavior – pros and cons
  - Structure – pros and cons
  - Costs to Transform

Transformation Strategies for Files Where Compound Documents Exist

- Data Files with WordPerfect 4.2 and Lotus Notes
  - Option A: XXX
    - Appearance – pros and cons
    - Behavior – pros and cons
    - Structure – pros and cons
    - Costs to Transform

Option B: YYY
- Appearance – pros and cons
- Behavior – pros and cons
- Structure – pros and cons
- Costs to Transform

Option C: ZZZ
- Appearance – pros and cons
- Behavior – pros and cons
- Structure – pros and cons
- Costs to Transform

Transformation Decision:

APPENDIX C. Sustainable Formats

Frequently Asked Questions (FAQs) About Selecting Sustainable Formats for Electronic Records[5]

This FAQ is provided to Federal agencies to assist them in meeting their records management responsibilities under 44 U.S.C. Ch. 31. Agencies can use this information when selecting and implementing formats for long-term electronic records, and when business needs necessitate formats not specifically addressed by existing NARA guidance. Visit the Toolkit for Managing Electronic Records to access existing guidance under the topic of "Specific Records Technologies" (e.g., digital photographic records).

1. Why should agencies use sustainable formats for electronic records?

Agencies create and maintain increasing volumes of Federal records in electronic formats. Typically, agencies select electronic formats for Federal records based on business needs and current technical requirements. Once selected, those formats must be sustainable, that is accessible both throughout their lifecycle and as technology evolves. Formats that are not sustainable may cause Federal records to become obsolete and inaccessible before they are eligible for deletion as authorized in the approved records schedule.

2. What is a sustainable format?

The term "sustainable format" means the ability to access an electronic record throughout its lifecycle, regardless of the technology used when it was originally created. A sustainable format is one that increases the likelihood of a record being accessible in the future.

3. What are characteristics of a sustainable format?

When the records need to be maintained over the long term (sustainability), agencies should consider each of the following characteristics of formats:

a. Published Documentation and Open Disclosure

Publicly and openly documented formats adhere to specifications that are published and accessible. Publicly accessible specifications allow developers to create a wide variety of applications to read, process, and validate files. Openly documented specifications assist developers in creating tools to access the information in obsolete formats, and/or assist in migrating files to future formats.

[5] See NARA, "FAQ About Selecting Sustainable Formats for Electronic Records." http://www.archives.gov//records-mgmt/initiatives/sustainable-faq.html

Tagged Image File Format (TIFF) is an example of a format based on a publicly available, authoritative specification for scanned images.

b. Widespread Adoption and Use

Formats adopted for widespread use have a higher probability of being sustainable over time. When a format has been widely adopted by users, multiple software tools are created to open, read, and access the records and the market supports ongoing sustainability of the file format. This extends the time that the information can be maintained in the format using readily-available tools. The adoption of a file format by information creators, disseminators, and users is an indicator of sustainability.

Hyper-text Markup Language (HTML) is an example of a format that has been widely adopted for Internet use.

c. Self-describing Formats

Self-describing formats contain metadata needed to interpret the content, context, and/or structure of the record. Metadata embedded within the format minimizes reliance on external documentation and the risk of disassociation of metadata from the file over time. While self-describing formats provide the capability for including metadata (e.g. in the file header or through tags within the file structure), they may not necessarily mandate it in the format specification. If present, the metadata should be easily accessed. This ensures that descriptive information about the record is sustainable.

Extensible Markup Language (XML) is an example of a self-documenting format because it describes its structure and field names.

When agencies use formats that exhibit these characteristics, they increase the likelihood that the information will be accessible over the long term.

4. How do agencies enhance the sustainability of formats?

When creating electronic records or converting source data, agencies can enhance sustainability by maintaining the original quality of source data. The following methods are typically applied through software settings and vary depending on the format being used.

a. Technical Protection Mechanisms

Long-term records should be unrestricted and/or unencrypted so that user IDs and/or passwords are not needed to maintain the file. User IDs and passwords can be lost over time. For more information, see NARA's Bulletin 2007-02, guidance concerning the use of Enterprise Rights Management (ERM) and other encryption-related software on Federal records.

b. Maintain Integrity of Source Data

When using compression to reduce file size, agencies should use lossless compression to maintain the integrity of source data. Lossless compression produces smaller file sizes without removing any information. Maintaining the original quality of source data can facilitate future migration and conversion. Minimizing subsequent modification of the records after production is also recommended to maintain integrity. See Frequently Asked Questions (FAQ) about Digital Audio and Video Records for a discussion of lossless compression.

5. Will selecting appropriate formats guarantee sustainability?

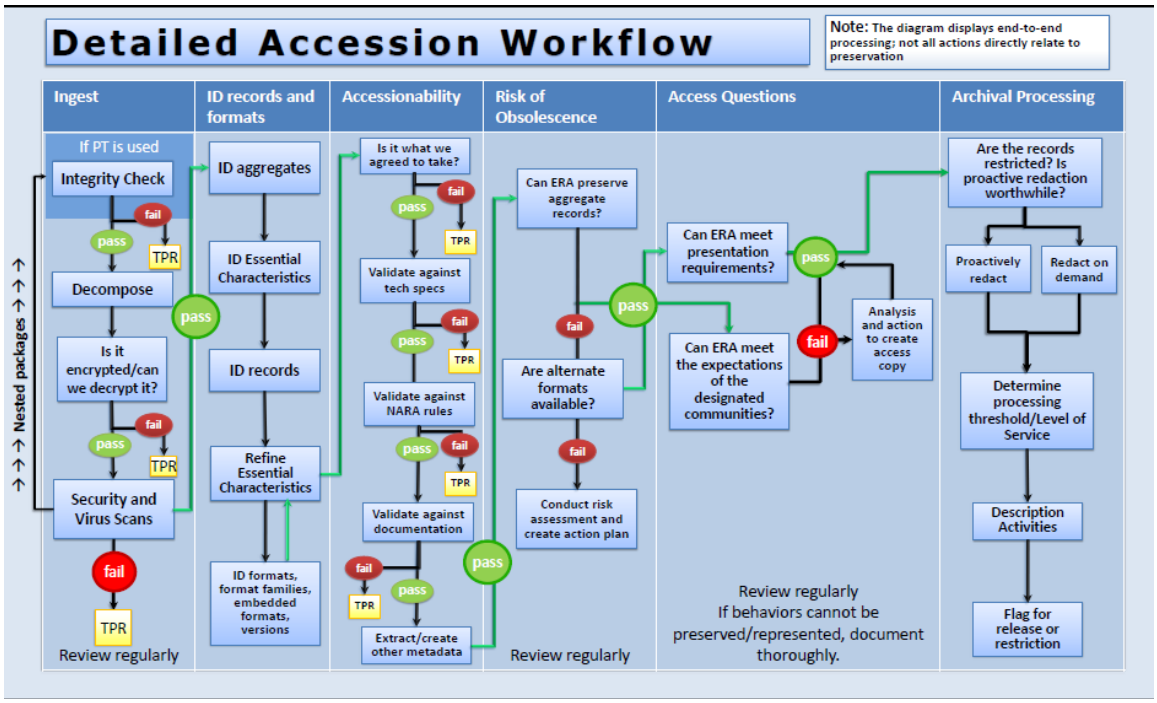
No. In addition, agencies need to follow record policies and procedures governing creation and management of electronic records and follow approved records schedules.

Section 7 of ISO 15489, "Information and Documentation -- Records management -- Part 1: General;" discusses other factors that should be considered.

6. Who can I contact for further assistance?

- In the Washington, DC, area, the NARA RM Life Cycle Management Division can provide assistance. See List of NARA Contacts for Your Agency at http://www.archives.gov/records-mgmt/appraisal/
- Outside the immediate Washington DC area, the NARA Records Management Staff in NARA's regional offices, can provide assistance. A complete list of NARA's regional facilities can be found at http://www.archives.gov/locations/index.html.

Appendix D. Ingest Workflow



Appendix E. Levels of Service

LEVEL ONE REFERENCE SERVICE

At this level of service, NARA will return records to the user in either the original transferred format or another contemporary format. The user of the records will be expected to have software with which to use or view the records. This level of service will be provided for records that have minimal current and expected research interest.

1 At this level of reference service, users will be able to search any metadata captured by the lifecycle systems, e.g., descriptive catalog and template data.

2 Descriptive information will be captured at the series level and include an archival description, technical characteristics of the records, size of the files in bytes, list of the available files

3 Users may order a copy of the records for a fee via an automated ordering system and have the records delivered electronically or on the physical media of their choice.

4 Users will need to have the appropriate software with which to use the records.

5 Users may order technical documentation. If the documentation is in paper form, it will be scanned and sent to the user for a fee. If the documentation is in electronic form, it will be delivered electronically or on the physical media of their choice.

LEVEL TWO REFERENCE SERVICE

At this level of service, NARA will return records in a usable contemporary format. Users will be able to utilize ERA for searching, inspecting, printing, and downloading records. The user of the records will be expected to have software with which to use records outside of the ERA environment. This level of service will be provided for records that have some current and expected research interest. It will be most appropriate for records that identify specific persons, geographic areas, organizations, or dates.

1 At this level of reference service, users will be able to search any metadata captured by the lifecycle systems, e.g., descriptive catalog and template data.

2 Descriptive information will be captured at the series level and include an archival description, technical characteristics of the records, size of the files in bytes, list of the available files.

3 Users will be able to search record-level information in a web environment and have ERA return individual records and a subset of records to a user, similar to the current capabilities of AAD.

4 Users may print or download records. The user of the records will be expected to have software with which to use records outside of the ERA environment.

5 Users may order a copy of the records for a fee via an automated ordering system and have the records delivered electronically or on the physical media of their choice.

6 Users will have online access to technical documentation, if necessary for understanding the records. Otherwise, users may order technical documentation for a fee. If the documentation is in paper form, it will be scanned and sent to the user. If the documentation is in electronic form, it will be delivered electronically or on the physical media of their choice.

LEVEL THREE REFERENCE SERVICE

At this level of service, NARA will return and visualize records in a variety of ways: for example, render geographic data as maps, search transcriptions of sound recordings and return selected sound bites, perform analyses using social science data, and other similar scenarios. Users will work within the ERA environment to locate, return, and analyze records, but would be expected to have software with which to use records outside of the ERA environment. This level of service will be provided for records that have a high level of public interest and lend themselves technically to a variety of visualization scenarios.

1. At this level of reference service, users will be able to search any metadata captured by the lifecycle systems, e.g., descriptive catalog and template data.

2. Users will be able to search record content information using a variety behavioral search tools, such as AAD, and in domains devoted to certain types of data, such as social science data, GIS data, photographic collections, sound recordings, and motion picture files.

3. Descriptive information will be captured at the series, file, and/or record levels and include an archival description, technical characteristics of the records, size of the files in bytes, list of the available files.

4. Users will be able to search record-level information in a web environment and have ERA return individual records, subsets of records, and have them displayed in a variety of user-defined ways.

5. Users may analyze records using on-line software supplied by NARA.

6. Users may order a copy of the records for a fee via an automated ordering system and have the records delivered electronically or on the physical media of their choice.

7. Users will have online access to technical documentation and other finding aids.

Appendix F. Sample Reference Driven Transformation Plan

Records identification

RG:

Series/Transfer Group:

Reference Demand:

Time Period	RG	Series/TG	Num of Hits	Num. Downloads
1Q, 09	4	19th century immigration records	300,000	20,000
2Q, 09	4	19th century immigration records	200,000	50,000
3Q, 09	4	19th century immigration records	400,000	10,000
4Q, 09	4	19th century immigration records	500,000	20,000

Needs of the Research Community:

Data File Format - WordPerfect 4.2

Number of data files in this data file format: 100,000

Technical Characteristics of WordPerfect 4.2 – XXX

Market Share of WordPerfect 4.2 - XXX

Tools that users can use to render WordPerfect 4.2

Tool A – pros and cons
Tool B – pros and cons
Tool C – pros and cons

Transformation Strategies:

Option A: Transform to Word 2007
Appearance – pros and cons

Behavior – pros and cons
Structure – pros and cons
Costs to Transform

Option B: Transform to PDF 1.7
Appearance – pros and cons
Behavior – pros and cons
Structure – pros and cons
Costs to Transform

Option C: Transform to TIFF 6.0
Appearance – pros and cons
Behavior – pros and cons
Structure – pros and cons
Costs to Transform

Transformation Strategies for Files Where Compound Documents Exist

Data Files with WordPerfect 4.2 and Lotus Notes
Option A: XXX
Appearance – pros and cons
Behavior – pros and cons
Structure – pros and cons
Costs to Transform

Option B: YYY
Appearance – pros and cons
Behavior – pros and cons
Structure – pros and cons
Costs to Transform

Option C: ZZZ
Appearance – pros and cons
Behavior – pros and cons
Structure – pros and cons
Costs to Transform

Transformation Decision: