



Portico Format Monitoring and Migration Policy

1. Policy Statement

1.1. Portico's primary preservation methodology is migration, which involves transitioning content from one file format to another as technology changes and as file formats become obsolete.

- In all cases where a migration is performed, the original asset will be preserved along with the transformed asset.
 - All assets produced by the migration will be validated and assigned a format preservation level.
 - Any source asset which cannot be transformed into a valid instance of the target format will continue to be preserved at bit-level.
 - All metadata associated with the original and the transformed asset, the event metadata associated with the migration, and the association between the original and the transformed asset will be also be preserved.
 - Portico can imagine scenarios where, in the course of a sequence of migrations, intermediate migration artifacts might not be retained. For example, a later migration might take the original artifact, which has already undergone migration one or more times, and directly migrate that original to a new format. In such a case, Portico might choose not to retain the products of the earlier migrations.
- In general, a format typically means a file format as commonly understood. However, should we find that the same file format necessitates different preservation approaches depending on either the context (provider or consumer community) or the content-type (e.g. journal article, digitized book) of the asset, we will develop a separate format preservation strategy for different context and content types in the same file format.
- Although Portico currently does not intend to pursue emulation as a preservation strategy, we will continue to monitor the efforts of other digital repositories that are developing emulation technology as part of their preservation strategies.

1.2. Portico will decide when and how to migrate files based upon a variety of criteria including, but not limited to:

- Preservation commitments made to the content provider.
- Whether the format is in broad use in the general public.
- How other preservation services are managing files of this format.
- The preservation needs of the content-type.
- The functional unit of the file.
- The percentage of the archive impacted.
- Portico's need to manage the content in the archive.



As of July 2009, the only two migrations routinely performed on content are:

1. To migrate supplied mark-up files (e.g., XML) in the e-journal content-type to a Portico version of the NLM Journal Archiving and Interchange Tag Suite. This migration allows Portico to manage the corpus of e-journal content uniformly and to quickly trigger this content and make it available for use. For the first several e-book publishers, we are migrating the supplied mark-up files as well – however, when e-book content is supplied in one of the standard e-book formats (e.g., EPUB), Portico will not migrate those files.
2. To migrate the content provider's packaging to the Portico content model.

Both of these migrations are documented in the Turn Over Document (a type of Format Action Plan) associated with each stream of content from a specific content provider.

- 1.3. Portico will monitor technical developments in the digital preservation community in file format characterization, automated risk assessment, and emulation as a mode of format preservation. Portico will monitor the status of formats for which instances exist in the archive, and will periodically:
 - o assess the current risk status of the file format, and the impact of that risk on the archive as a whole
 - o review migration options available for the format, and assess the risks and benefits of each option
 - o recommend that format action be taken (Note that research into the file format and alternative preservation choices might lead us to determine that no action can or should be taken.)
 - o assign responsibility for the format action to be taken
 - o review the results of the format action taken
- 1.4. The following criteria will form the basis for selection of a target format for migration of an at-risk format:
 - o Portico will prefer a target format that most completely captures the content, structure, behavior, and other features of the source file format.
 - o Portico will prefer file formats with complete specifications and with tools for validating technical conformance of a file to the format. These specifications and tools should be freely available, with a preference given for open standards with a reference implementation.
 - o Portico will prefer file formats well-established in the communities of those who create and/or use the assets of the archive.
 - o Portico will prefer stable file formats, not subject to frequent version changes, with backward-compatible new versions.
 - o Portico will prefer file formats for which there is readily available expertise.
 - o Portico will prefer file formats in which the digital representation is open to direct human inspection or to evaluation with basic tools.



- Portico will prefer file formats with self-documentation or inherent metadata.
- Portico will prefer file formats with the fewest external dependencies (e.g. operating system, rendering tools) and with the greatest ease of interchange amongst different systems.
- Portico will prefer file formats unencumbered by patent, license, or other legal restrictions.
- Portico will prefer file formats unencumbered by encryption or other technical protection mechanisms.
- Portico will prefer file formats for which a cost-effective migration engine exists or can be developed. Portico will prefer that such engines be open and freely available.

1.5. The following criteria will form the basis for selection of tools for migration of an at-risk format:

- Portico will prefer tools which accurately preserve the content, structure, appearance, context, behavior, and ability to render the source file.
- Portico will consider the configuration capabilities of any tool.
- Portico will consider the performance of the tool, both in terms of the size of the resulting file and the time to perform the migration.
- Portico will consider the batch and automation capabilities of any tool.
- Portico will consider the logging features of the tool.
- Portico will consider the transparency of any error messages produced by the tool, and the ability to map those messages to the features of the source and target formats.
- Portico will prefer tools that are transparent (particularly open-source tools), economical, and that have a robust development community.

1.6. Portico will preserve the preservation action event metadata and any metadata necessary to associate the source and target file. The preservation action metadata will detail the configuration, settings, tool version, and environment in which the tool is employed in the course of a migration.

1.7. Portico will retain any relevant format specification documentation used in assessing format risk and target format choice, so long as that retention does not conflict with any legal restrictions on that documentation.

1.8. Portico will retain any relevant migration tool documentation, so long as that retention does not conflict with any legal restrictions on that documentation.

2. Implementation Examples

2.1. Format migration at ingest

- Assets submitted in certain file formats (for example, journal article metadata files in proprietary publisher XML or SGML formats) are migrated during ingestion workflows



before an archive preservation level is set for the asset. The original file is preserved in the archive, along with the migrated file. There is a specification of the migration to be performed for each file format to be migrated in the profile associated with the asset submitted for ingest into the archive. All technical metadata about the original and transformed asset, along with any event metadata associated with the migration, are maintained in the metadata associated with the asset, along with the association between the original and transformed file. All original assets, along with any assets produced as the result of ingest migration, are assigned an archive preservation level upon ingest to the archive.

- In such transformations, any text generated by Portico is clearly marked as such in the transformed asset. Please see the Portico NLM DTD documentation for a complete discussion of Portico's policy and procedures for generated text in normalized XML.
- Portico maintains an archive copy of any publisher DTD or schema of which the publisher-provided XML or SGML object is an instance. The publisher-provided asset is validated against that DTD or schema, and the results of that validation are stored in the technical metadata associated with the asset.

2.2. Technology Watch

- Portico has designated a research developer to perform daily monitoring of technical developments in the digital preservation community in file format characterization, automated risk assessment, and emulation as a mode of format preservation. This developer performs daily monitoring of roughly 100 relevant RSS feeds, as well as of approximately 50 relevant listservs.

3. **Definitions**

3.1. Preservation Levels. Upon ingest into the archive, all file assets are assigned one of three preservation levels at which the archived asset will be maintained:

- fully supported (Portico will maintain the intellectual content of this file for the long term)
- reasonable-effort (Portico will attempt to maintain the intellectual content of this file for the long term, but we cannot guarantee success)
- byte-preserve (Portico will maintain the file for the long term. We will not migrate or normalize the file, but Portico cannot guarantee the usability of the file now or in the future)

The preservation level is determined by format validity and by the preservation policy specified for that format for each provider. A defective file cannot be fully supported; at best, Portico can only promise reasonable effort.

The preservation level is declared and maintained in the metadata associated with the asset.



4. **Document History**

4.1. Approved by: Amy Kirchhoff

4.2. Last Review Date: 8/12/2009

4.3. Reviewed by: Stephanie Orphan, Sheila Morrissey

4.4. Change history:

Version	Date	Change	Author
0.1	06/04/2009	Drafted document	Sheila Morrissey
0.2	07/07/2009	Edited and applied new template	Sheila Morrissey
0.3	07/27/2009	Incorporate Stephanie's edits	Sheila Morrissey
1.0*	7/28/2009	Slight modifications to use the preferred term "migration" over normalization and added in the section on why we might choose to migrate.	Amy Kirchhoff
1.1*	7/29/2009	Synced with E-Journal content type action plan.	Amy Kirchhoff
1.2*	8/5/2009	Added Reviewed by line and slight copy editing.	Amy Kirchhoff
1.3*	8/12/2009	Made EGF's changes.	Amy Kirchhoff

* An approved version of this document.