Portico Content Type Action Plan: D-Collections

1. Content Type Introduction

1.1. D-collection content is initially comprised of digitized historical collections. This content can include full or header-only information in XML, SGML, or other format; page images in TIF or some other format; figure, formula, table, or other graphic images in various still image formats; and media and other supplementary information files in various digital formats.

1.2. D-collection content may sometimes contain content that is not traditional digitized historical collections.

2. Policy Statement

2.1. The fundamental archival unit for d-collection content varies by collection. It is typically a single document, a newspaper issue, an image, or even a dictionary entry.

2.2. D-collection content as received from the publisher is batched and processed into Submission Information Packages (SIP) by the Portico ConPrep System.

- As there is no standard package format for d-collection content, each content provider's content stream package structure is analyzed by Portico staff to determine the directory structure and file naming conventions used by the publisher to associate the components which comprise a single unit.

- The results of this analysis are described in the Turnover Document associated with the content stream. The results are also expressed more formally in the XML profile which informs the ConPrep system as it processes publisher content and assembles the components of each archival unit.

- Analysis sometimes reveals extraneous files that are included in the publisher package, but which do not comprise components of the articles (e.g., thumbs.db files). These files are not ingested into the Portico archive.

2.3. With the exception of these "excluded files," all assets provided by the publisher are validated and characterized with the JHOVE tool, associated with the appropriate unit, and are associated with a Portico metadata file containing technical and descriptive metadata, in the archival unit.

2.4. Should Portico make the determination that any formats of the assets comprising the d-collection content are at risk, these components will be migrated in conformance with the Format Action Plan for that format.

2.5. All d-collection content will be associated with a business artifact comprising the contract authorizing Portico to store and manage those assets.

2.6. Portico will maintain descriptive and technical metadata associated with the technical artifact in the Archival Information Package (AIP) containing the technical artifact.

3. Implementation Examples

3.1. D-collection content is treated differently than e-journal or e-book content. Due to how large the collections are and business reasons, Portico treats the XML formats provided as de-facto standards. During initial analysis, Portico writes a tool to extract all the descriptive metadata

for the archival units from the provided XML files in order to place it into Dublin Core in the Portico preservation metadata file for the item. As with e-journal and e-book content, the publisher's packaging is transformed into Portico's archival unit packaging. The Turn over Document associated with each stream of content from a specific content provider is considered the Format Action Plan for this content. By extracting the descriptive metadata and transforming the packaging into an archival unit, Portico can manage the content in the archive and quickly deliver the content to end users when needed.

4. Document History

4.1. Approved by: Amy Kirchhoff

4.2. Last Review Date: 3/27/2016

4.3. Reviewed by: Sheila Morrissey, Stephanie Orphan, Kate Wittenberg

4.4. Change history:

Version	Date	Change	Author
0.1	3/16/2016	Drafted document	Amy Kirchhoff
1.0*	3/27/2016	Changes as suggested by SO and SM. Approved	Amy Kirchhoff

* An approved version of this document.