Portico JHOVE Usage

1. Policy Statement

1.1. Technical metadata is one element of Portico's arsenal of tools to assess and ultimately to mitigate risks to assets in the archive.

- o Such mitigation does, upon occasion, include a feedback loop to content providers who are willing and able to provide corrections for technically defective files.

- o Technical metadata also inform the setting of preservation levels for digital assets and the determination of appropriate migration paths for assets in formats deemed at risk for obsolescence.

- o Technical metadata comprise a key component of the representation information crucial to the management and preservation of digital assets in an OAIS-compliant archive.

- o As with the other elements of preservation metadata – descriptive metadata, event metadata, administrative metadata, and rights metadata, technical metadata enables Portico to manage the corpus of content preserved within the archive.

1.2. Functions and features of the JHOVE tool

- o The significant format-centric managerial operations are identification, validation, characterization, and assessment. These are the functions provided by the JHOVE tool.

- o JHOVE (the JSTOR/Harvard Object Validation Environment) is an extensible, open-source Java-based framework for format-specific digital object identification, validation, and characterization. It was developed at Harvard University Library in conjunction with the JSTOR Electronic-Archiving Initiative (now known as Portico), with funding provided by the Andrew W. Mellon Foundation. It is available under the GNU LGPL license, and is widely deployed by international repository and preservation institutions and programs.[c]

- o The JHOVE tool has modules for the following formats: IFF, WAVE, GIF, JPEG, JPEG2000, TIFF, PDF, ASCII, UTF8, HTML, and XML. Portico has created additional modules for SGML, TAR, ZIP, and GZIP. JHOVE also provides a default BYTESTREAM module to provide standard representation information for files in format other than those for which there exists a JHOVE module.

- o The standard representation information reported by JHOVE for files in any format includes:

    - ▪ File pathname or URI
    - ▪ Last modification date
    - ▪ Byte size
    - ▪ Format
    - ▪ Format version
    - ▪ MIME type

- Format profile(s)
- CRC32, MD5, and SHA-1 checksums (optional)
  - o Additional format-specific representation information is also reported.
  - o JHOVE also reports the degree of conformance of an asset to its format specification, returning one of three values: "Well formed and valid," "Well formed," "Not Well Formed."

2. Implementation Examples

2.1. As part of the preparation of the Submission Information Package (SIP) by the Portico ConPrep system, the JHOVE tool is invoked on all digital assets supplied by the content provider and created by Portico during the ingest process.

  - o This includes Portico-created assets, such as the output of normalization of XML and SGML to Portico NLM XML, and business artifacts such as contracts with publishers.

  - o Selection of the appropriate JHOVE format module is directed by information in the profile associated with the content stream of which the digital asset is a part. For files for which there is no JHOVE module, or for files which fail validity and well-formedness tests for that module, additional confirmation of the file format identity is provided by the BSD FILE tool. This format identification information is stored in the technical metadata.

  - o Portico maintains a format mapping registry to enable cross-walks between format names in the Portico namespace, and format names in the JHOVE and BSD file namespaces. The Portico format registry also links Portico format names to mime type names. As Portico determines that other public format registry namespaces (for example, the UDFR) are sufficiently well developed and in general usage, Portico will extend this cross-walk capability to these namespaces.

2.2. The validation status returned by JHOVE is used in the determination of the preservation level of each asset (see the Portico Format Monitoring and Migration Policy for further information on preservation levels).

2.3. A subset of the technical metadata returned by JHOVE in XML format is included in the Portico metadata file associated with each archival unit.

2.4. The Portico metadata file itself is validated using the JHOVE tool before ingest into the archive.

3. Definitions

3.1. Content Information: The set of information that is the original target of preservation. It is an Information Object comprised of its Content Data Object and its Representation Information. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures, but excluding the documentation that would explain its history and origin, how it relates to other observations, etc.[a]

3.2. <u>Representation Information:</u> The information that maps a Data Object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e., a Data Object) is mapped into a symbol. [a]

3.3. <u>Technical Metadata</u>: Information about how a digital object was created and stored, including, for example, checksum, file creation type, file size, file format, and any salient format-specific properties

4. <u>References</u>

a. OAIS (2002) CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. January 2002 (ISO 14721:2003) http://public.ccsds.org/publications/archive/650x0b1.pdf accessed 2009.06.03

b. Abrams, Stephen, DCC Digital Curation Manual Installment on "File Formats" October 2007 http://www.dcc.ac.uk/resource/curation-manual/chapters/file-formats/file-formats.pdf accessed 2009.06.03

c. Harvard University Library, JHOVE – JSTOR/Harvard Object Validation Environment http://hul.harvard.edu/jhove/

5. <u>Document History</u>

5.1. Approved by: Amy Kirchhoff

5.2. Last Review Date: 3/27/2016

5.3. Reviewed by: Shelia Morrissey, Stephanie Orphan, Kate Wittenberg

5.4. Change history:

Version	Date	Change	Author
0.1	06/04/2009	Draft created	Sheila Morrissey
0.2	07/07/2009	Edited and formatted to new template	Sheila Morrissey
0.3	07/27/2009	Incorporate Stephanie's edits	Sheila Morrissey
1.0*	7/28/2009	Replaced references to PMETs with metadata file and some other minor edits.	Amy Kirchhoff
1.1*	8/5/2009	Added reviewed by line.	Amy Kirchhoff
1.1.1	4/4/2011	Updated the logo (approved on 3/16/2016)	Amy Kirchhoff
1.2*	3/27/2016	Slight tweaks	Amy Kirchhoff

* An approved version of this document.