Portico Technical Metadata Tool 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.

- Such mitigation does, upon occasion, include a feedback loop to content providers who are willing and able to provide corrections for technically defective files.
- Technical metadata also inform the setting of preservation levels for digital assets and the determination of appropriate migration paths for assets in formats, or with features, deemed at risk for obsolescence.
- Technical metadata comprise a key component of the *representation information* crucial to the management and preservation of digital assets in an OAIS-compliant archive.
- 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 technical metadata tools

- The significant format-centric managerial operations are identification, validation, characterization, and assessment. These are the functions provided by a combination of JHOVE, ExifTool, and Siegfried.
- 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 Andrew W. Mellon Foundation. It is currently maintained by the Open Preservation Foundation and is available under the GNU LGPL license. It is widely deployed by preservation institutions and programs.[c]
- The JHOVE tool is packaged with modules for the following formats: AIFF, WAVE, GIF, JPEG, JPEG2000, TIFF, PDF, ASCII, UTF8, HTML, XML, EPUB, PNG, and WARC. 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 a format other than those for which there exists a JHOVE module.
- The standard representation information reported by JHOVE for files in any format includes: file path or URI, last modification date, byte size, format, format version, MIME type, format profile(s), CRC32, MD5, and SHA-1 checksums (optional), plus additional format-specific representation information, where available.

- JHOVE also reports the degree of conformance of an asset to its format specification ("validation"), returning one of three values: "Well-formed and valid," "Well-formed, but not valid," "Not well-formed."

- ExifTool is a free and open source tool (GPLv1+ License) for reading, writing, and editing metadata in a wide variety of file formats. It was originally released in 2003, developed by Phil Harvey. It continues to be maintained and has a large community of users.[d]

- For all files, ExifTool can extract standard representation information that includes: file path or URI, byte size, and last modified date. For supported formats, it can also extract format-specific representation information, including: format name, file extension, mime type, and wide variety of format-specific metadata that is embedded in the file, with support for embedded metadata formats including EXIF, GPS, XMP, JFIF, and more.

- Siegfried is a free and open source tool (Apache License 2.0) for signature-based file format identification. It can use the National Archive UK's PRONOM file format signatures for identifying file types. It was developed by Richard Lehane, who continues to maintain it.[e]

2. Implementation Examples

2.1. As part of the preparation of the Submission Information Package (SIP) by the Portico ConPrep system, either JHOVE (if there is a module for the file format) or ExifTool is invoked on all digital assets supplied by the content provider and created by Portico during the ingest process.

- 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.

- 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 Siegfried file format identification tool. This format identification information is stored in the technical metadata. Based on the format identification, if there is no matching JHOVE module for that format, the ExifTool is invoked on the file to extract any available metadata. If no metadata is available, the basic representation information (checksum, file name, file uri, last modified) is stored and the validation status is marked as "Not Determined."

- Portico maintains a format registry that enables cross-walks between format names in the Portico namespace, format names in JHOVE, and the PRONOM identifier provided by Siegfried. The Portico format registry also links Portico format names to mime type names. It is updated periodically to reflect the latest PRONOM data.

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). A status of “Not Determined” is recorded if no JHOVE module is available to validate the format.

2.3. A subset of the technical metadata returned by JHOVE or ExifTool 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. Representation Information: 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. Technical Metadata: 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. References

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) https://public.ccsds.org/publications/archive/650x0b1.pdf accessed 2009.06.03

b. Abrams, Stephen, DCC Digital Curation Manual Installment on "File Formats" October 2007 https://era.ed.ac.uk/handle/1842/3351 accessed 2024.02.12

c. Open Preservation Foundation, JHOVE – JSTOR/Harvard Object Validation Environment https://jhove.openpreservation.org/

d. ExifTool, Phil Harvey. https://exiftool.org/

e. Seigfried, Richard Lehane. https://www.itforarchivists.com/siegfried

5. Document History

5.1. Approved by: Kate Wittenberg

5.2. Last Review Date: 2/12/2024

5.3. Reviewed by: Amy Kirchhoff, John Meyer, Sivaram Challa, Karen Hanson, 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
1.3*	8/7/2023	Added new modules that were enabled	Karen Hanson
1.4*	2/12/2024	Policy Renamed to reflect retirement of BSDTool and introduction of Siegfried and ExifTool to the workflow. Details for these tools added. Links updated.	Karen Hanson

* An approved version of this document.