Chapter 4

Support for Digital Formats

Long-term renderability cannot be ensured without detailed knowledge about and documentation of digital file formats. In this respect, digital formats are at the heart of digital preservation activities.

Significant Properties

The term *significant properties* is used to refer to the properties of digital objects that must be preserved over time through preservation treatments such as migrations or emulations in order to ensure the continued usability and meaning of the objects. (*Significant characteristics, essential characteristics,* and *essence* are less commonly used synonyms). The definition and determination of these properties constitute a critical and mostly unsolved issue in the field of digital preservation.

Significant properties are usually categorized as pertaining to content, context, appearance, structure, and behavior. If, for example, the digital object in question were a chapter of a book in PDF format, the content might be the text and pictures, the context would be the bibliographic description of the book and chapter, the appearance would be the layout of the pages, the structure would include any metadata relating the chapter to the book as a whole, and the behaviors could include internal and external hyperlinks. For this particular PDF, it might be decided that the content, context, and structure must be preserved, but that the appearance and behaviors could be sacrificed in the course of preservation treatment.

Significant properties may adhere to formats, genres, or individual objects, and in some cases may be in the eye of the beholder—the actionable links that you consider expendable may be critical to me. Currently, most research into significant properties is focused on formats. The InSPECT project of the U.K. Arts and Humanities Data Service is investigating the significant properties of raster images, structured text, digital audio, and e-mail messages, and new awards were recently granted to study e-learning objects, software, vector images, and moving images.[1]

Readings

- Andrew Wilson, "Significant Properties Report," Oct. 2007, www.significantproperties.org.uk/documents/wp22_significant_properties.pdf. A cogent review of work to date undertaken for the InSPECT project.
- Margaret Hedstrom and Christopher Lee, "Significant Properties of Digital Objects: Definitions, Applications, Implications," in *Proceedings of the DLM-Forum* 2002, http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/dlm-proceed2002.pdf. Describes preliminary research taking a rather broad view of significant properties, although follow-up appears to be unavailable.

Representation Information and Registries

A good understanding of digital formats is essential for the execution of nearly all preservation strategies. Unfortunately, the concept of format is anything but straightforward. Informally we tend to think of formats as generic file types such as PDF or QuickTime, denoted by MIME type or file extension. These distinctions are not particularly useful for preservation purposes, which

require more specific and granular information about version, compression, profile, and bitstream encoding. Is it PDF 1.5 or 1.6? Is the codec used in this QuickTime file Apple Video, Cinepak or MPEG-1? Detailed representation information (as defined by OAIS), preferably linked to the official format specification, is critical for carrying out preservation strategies such as migration and emulation.

A related type of information is sometimes called "environment" information, and concerns the software capable of creating or rendering a format and the hardware capable of supporting that software. For example, a PDF 1.4 file can be created by Adobe Acrobat 5.0 and rendered by Acrobat Reader 5.0 (as well as dozens of other applications). Acrobat 5.0 can run under Windows 98, Windows 2000, Windows Me, and Windows 95. Windows 98 in turn requires a 486 DX2, 66 MHz or higher processor and at least 16 MB of RAM.

By this point it should be obvious that it takes a huge amount of research and expertise to determine both representation and environment information, and that everyone would benefit from having this information accessible from centrally maintained registries. Currently there are at least three registries in some stage of production or development, all with overlapping scopes and different but overlapping content:

- The PRONOM registry developed and maintained by The National Archives (U.K.) has elements of representation and environment information, and is expanding to include preservation planning information.
- The Format Descriptions database maintained by the Library of Congress has detailed representation information as well as an analysis of sustainability factors for many formats.
- The Representation Information Registry Repository under development by the Digital Curation Centre aims to have extensive representation information closely tied to the OAIS model.

PRONOM Registry
www.nationalarchives.gov.uk/pronom

Format Descriptions Database
www.digitalpreservation.gov/formats/fdd/descriptions.shtml

Representation Information Registry Repository
http://registry.dcc.ac.uk/omar

Global Digital Format Registry
http://hul.harvard.edu/gdfr/

The Global Digital Format Registry (GDFR) initiative aims to unify these and other registries by defining a common data model and a common network protocol for distributed registries to communicate with each other and synchronize their format representation information. Registries such as PRONOM would then become nodes in the global network. The GDFR was initiated by Harvard University and received a major grant supporting development work from the Andrew W. Mellon Foundation in 2005; the technical work will be done by OCLC.[2]

Another use for representation and environment information is to help preservation managers assess the risk of format obsolescence for files in their repositories. This is another time- and knowledge-intensive task that would benefit from centralized support. One approach is the Automatic Obsolescence Notification System (AONS) being developed by the National Library of Australia and its partners.[3] AONS is designed to extract data from existing registries and the future GDFR and provide the data as decision-support information to repository managers. A beta pilot service was launched in 2007. Other studies concerned with methodologies for assessing the risk of format obsolescence are noted below.

Readings

- "Risk Management of Digital Information: A File Format Investigation," Council on Library and Information Resources, 2000, www.clir.org/PUBS/reports/pub93/pub93.pdf.
- Andreas Stenescu, "Assessing the Durability of Formats in a Digital Preservation Environment: The INFORM Methodology," *D-Lib Magazine* 10, no. 11, Nov. 2004, www.dlib.org/dlib/november04/stanescu/11stanescu.html.
- Cornell University Library, Virtual Remote Control Web site, http://prism.library.cornell.edu/VRC.

Tools

Several Java-based open-source tools have been released in recent years to aid in format identification, validation, and characterization.

DROID

DROID (Digital Record Object Identification) is a software tool developed by The National Archives (U.K.) to identify file formats based on their binary signatures. It can be run on single files or batches of files, and has both a graphical interface and a command line interface. It uses signature files (information about the internal and external characteristics of a format) stored in the PRONOM registry and returns PRONOM format identifiers. DROID is available as an open-source application under a BSD license.

DROID Web site
http://droid.sourceforge.net/wiki/index.php/Introduction

Reading

- Adrian White, "Automatic Format Identification Using PRONOM and DROID," 2006, http://droid.sourceforge.net/wiki/images/b/b4/Technical_Paper_1_-_Automatic_Format_Identification_v2.pdf.

JHOVE

JHOVE (JSTOR/Harvard Object Validation Environment) is a software tool that identifies, validates, and characterizes digital files. Like DROID, it has both graphical and command line interfaces. Validation checks that the file conforms to the appropriate file format specifications. Characterization returns technical metadata in an XML format including file name, modification date, byte size, format, format version, MIME type, format profiles, and checksums, as well as more detailed technical metadata for image and audio formats. The original release of JHOVE handles PDF and several open image, audio, and text-based formats. An enhanced and re-architected version called JHOVE2 is under development by Harvard University, Portico, and Stanford University. JHOVE2 will use DROID for format identification and will be designed to be more easily incorporated into other applications. JHOVE is available as an open-source application under the GNU Lesser General Public License.

JHOVE Web Site
http://hul.harvard.edu/jhove

Reading

- Martin Donnelly, "JSTOR/Harvard Object Validation Environment (JHOVE)," Digital Curation Centre Case Studies and Interviews, March 2006, www.dcc.ac.uk/resource/case-studies/jhove/case_study_jhove.pdf.

Metadata Extraction Tool

The Metadata Extraction Tool was developed by the National Library of New Zealand. Like JHOVE, it identifies file formats and extracts technical metadata relevant to preservation, which it outputs in an XML format. Major differences from JHOVE are that this tool does not also perform validation, and this tool includes routines for a number of proprietary Microsoft Office formats that JHOVE does not handle. Version 3 was released as an open-source application under the Apache Public License (version 2).

Metadata Extraction Tool Web Site
http://meta-extractor.sourceforge.net

Xena

Xena (XML Electronic Normalising for Archives) was developed by the National Archives of Australia. Xena will identify the format of a source file and create a normalized version in a more open format. For example, audio files in AIFF, MP3, or WAVE will be normalized to FLAC (Free Lossless Audio Codec); Microsoft Word files are converted to Open Office format; while GIF images are converted to PNG. Xena can be invoked manually or via an API (Application Program Interface). Xena version 4.0 was released in October 2007. It is available under a GNU General Public License version 2 and requires Sun's Java Runtime Environment and OpenOffice.org 2.x in order to run.

Xena Web Site
http://xena.sourceforge.net

Notes

1. InSPECT Web site, www.significantproperties.org.uk (accessed Nov. 17, 2007).
2. Global Digital Format Registry Web site, http://hul.harvard.edu/gdfr (accessed Nov. 17, 2007).
3. Australian Partnership for Sustainable Repositories, AONS Automated Obsolescence Notification System II, www.apsr.edu.au/aons2 (accessed Nov. 17, 2007).