Format Identification, Validation, Characterization and Transformation in DAITSS

Carol C.H. Chou; Florida Center for Library Automation; Gainesville, Florida, USA

Abstract

Rapid technology advancement and obsolescence present a constant challenge for preserving digital objects in a digital repository system. To ensure the long-term preservation of archival content, DAITSS, a digital repository system developed for the Florida Digital Archive, implements a scheme to automatically identify, validate, characterize and transform the format of digital objects in its repository. This scheme purposes to fulfill the unique requirements of the Florida Digital Archive.

Introduction

The Florida Digital Archive (FDA) was established in 2005 to serve as a digital preservation repository for the libraries of the eleven public universities in Florida. Partially funded by U.S. Institute of Museum and Library Services (IMLS), the FDA has developed an in-house digital repository system called Dark Archive in the Sunshine State (DAITSS).

The implementation of DAITSS follows the functional model of the Open Archives Information System (OAIS) [1]. It provides four services in accordance to OAIS: Ingest, Data Management, Archival Storage and Dissemination. The Ingest service processes the submission information package (SIP) sent by the submitting library and populates the preservation database with preservation metadata and provenance information. Then Ingest constructs an Archival Information Package (AIP) and sends the AIP to archival storage. The Dissemination service accepts dissemination requests from the participating libraries and creates a Dissemination Information Package (DIP) consisting of the original submitted content and the last best version of that content. Each information package (SIP, AIP and DIP) includes a METS-format XML document describing the content of the information package [6].

Ingest is the core service in DAITSS, providing the majority of the system's preservation functionality. Ingest identifies and validates each digital object in the SIP, extracts the technical metadata from the digital object and performs the applicable format transformations. The rest of this paper describes the requirements and processing of the format automation scheme adopted in DAITSS in the areas of format identification, validation, characterization and transformation. It also compares the DAITSS scheme with other comparable format software applications including JHOVE, DROID and NLNZ metadata extractor. This paper further describes the format transformations in DAITSS and provides an analysis of the considerations in evaluating format normalization processes, especially for video and audio archival collections.

Requirements

One fundamental building block of a digital preservation repository system is a method to automatically identify, validate, characterize and manage the format of digital objects. DAITSS has several unique requirements for format automation imposed by the FDA preservation policy. One of these policies is to accept any package with a valid SIP descriptor even if individual files within the SIP are faulty. Assuming all the files are virus-free, a SIP is considered valid if the SIP descriptor is valid, contains proper administrative information, and contains a valid checksum for each digital object in the SIP. This is based on the premise that a participating library can fix a faulty descriptor but will not have the resources to fix each invalid digital object. In addition, much of the FDA archival collection are Electronic Thesis and Dissertations (ETDs) where the original content creator (the student) may no longer be reachable by the time the ETD is ingested into the FDA. Rejecting any SIP containing invalid digital objects could cause a large backlog, thus is not favorable. Therefore, DAITSS requires a more tolerant model for format validation where the format validation status of the digital object is not used as the acceptance criteria for the SIP.

Digital objects submitted to the FDA may contain references to external digital objects that may not be available in the future. For example, a XML file may contain references to external schema on the web required for format validation. In addition, the FDA implements a migration-on-request strategy where the requested information package is processed to create the latest representation during the dissemination process. To ensure a descriptor remains usable regardless of the availability of its referencing schema on the web, DAITSS adopts a localization process during ingest where every external object referenced is be recorded and downloaded if applicable. These unique FDA requirements dictate the design and implementation of format identification, validation, characterization and transformation in DAITSS.

Related Software

Several applications have been developed recently to aid in digital preservation. The British National Archive has developed a format identification software, DROID (Digital Record Object Identification), under its PRONOM technical registry umbrella [2][5]. DROID uses a format signature model using both internal signatures and external signatures (file extension). A format can be registered with more than one internal signature and/or multiple external signatures.

JHOVE (JSTOR/Harvard Object Validation Environment) is a Java library for comprehensive format identification, validation and characterization developed by Harvard University [4]. Similar to DAITSS, JHOVE uses the internal signature for format identification. The format validation process in JHOVE is quite stringent where any non-conformance invalidates the digital object. JHOVE extracts an exhaustive list of metadata from the digital objects.

The National Library of New Zealand has developed a metadata tool, the NLNZ metadata extractor, for extracting the technical content of digital objects [7]. The NLNZ metadata extractor uses the file extension to identify the format, which is perhaps its biggest drawback. The tool adopts an attractive pluggable model where each format is self contained as an adaptor. In contrast to JHOVE, the metadata extracted by NLNZ metadata extractor are fairly limited. The following table provides an overall comparison of these three tools and the format automation in DAITSS.

Table 1: Overall Comparison

	DAITSS	DROID	JHOVE	NLNZ
identification	Internal signature	both	Internal signature	External sig.
validation	tolerant	no	stringent	no
characterization	detailed	no	detailed	limited
external links	recorded	no	no	no
pluggable	semi	yes	yes	yes
output	xml, database	xml	xml, text	xml

With over fifty supported formats and the backing from the National Archive in UK, DROID appears to be a good candidate for automating format identification in DAITSS. However, the absence of MIME type reporting in DROID proscribes DAITSS from incorporating it. The use of external signatures for format identification and the limited characterization output in NLNZ metadata extractor are perceived to be inadequate for DAITSS. JHOVES appears to be the most suitable one for DAITSS, but its stringent validation model and the lack of recording external references prohibit DAITSS from adopting JHOVE for format identification, validation and characterization. The proposed JHOVE 2 project allows a more flexible format validation model which could facilitate DAITSS in adopting JHOVE 2.

Identification

One of the issues challenging the preservation community today is the lack of standardized format identification scheme and a globally adopted format identifier. PRONOM's persistent unique identifier (PUID) is a step toward this direction but it is not widely adopted yet. The ongoing Global Digital Format Registry (GDFR) project is working on providing a standardized format identifier that can be universally adopted [3]. However, before the GDFR is in place, MIME type appears to be the most widely accepted identifier and the most feasible one for the FDA. Recognizing that MIME types alone may not be suited for the purpose of digital preservation due to their non-uniqueness and coarse granularity, DAITSS supplements its use of MIME type with format version and variation codes.

In DAITSS, each format is associated with a unique format identifier consisting of a MIME type, version and/or format variation code. The MIME type for each format is carefully decoded from the list registered in the Internet Assigned Number Authority (IANA) MIME type registry and from the format specifications. For example, DAITSS uses APP_PDF_1_5 as a format identifier for PDF 1.5 documents. Variant files that are not fully conformant to the file format specification are considered new formats and thus associated with different format identifiers supplemented with a format variant code. For example, DAITSS uses IMG_JPEG_ADOBE as the format identifier for Adobe-JPEG images because Adobe-JPEG is not compatible with JFIF specification. Compatible format variations are recorded as profiles under the same format identifier. For instance, GeoTIFF and TIFF/EP are compatible with TIFF 6.0 specifications, thus are recorded with the same format code as TIFF 6.0 but with TIFF_GEO and TIFF_EP profiles. The use of MIME types, versions, format variant codes and profiles allows the FDA to uniquely identify a format and apply different preservation treatments accordingly.

A format identification scheme is typically performed via the external signature (file extension), the internal signature (magic numbers), or both. The internal signature is a list of byte codes characterizing the format structure according to the format specification. Using file extensions has long been perceived as an unreliable scheme for format identification. Therefore, DAITSS uses the file extension only for possible expedition of the identification process; that is, the format for the specified file extension is evaluated first. The actual outcome of format identification is based on a complete match with the internal signature. For formats with no required internal signature such as XML or ASCII Text, DAITSS parses the entire digital object to determine its format.

The process of format identification in DAITSS is implemented by evaluating a digital object from the most specialized to the most general formats. DAITSS currently supports fifteen formats. In order to record a MIME type for current unsupported formats, DAITSS complements its format identification process by using an expandable, light weight JAVA library, *ffident*, which identifies formats in a manner equivalent to the UNIX *file* command.

Validation

Format validation is the process of assessing whether a digital object conforms to the corresponding format specification. Format validation involves a complete parsing of the digital object to check for its conformance with the format specification. Because DAITSS cannot use validation status as a criterion for accepting a digital object, it adopts a tolerant validation model. Rather than invalidating the digital object for any non-conformance, it records instances of non-conformance uncovered during the validation as anomalies and stores them in the preservation database.

Although the validation status of a digital object is not used as an acceptance criterion for the SIP, DAITSS categorizes all observed anomalies into two classes, tolerable or downgrading. If the anomaly is tolerable, indicating it does not affect the preservability of the digital object, the anomaly is noted in the preservation database and the AIP. A PDF document with an invalid page mode is an example of a tolerable anomaly. However, if the anomaly could reduce the preservability of the digital object, such as a PDF document without the required trailer dictionary, the preservation level of the digital object is downgraded. DAITSS preserves the downgraded digital objects "as-is", with no format transformation.

In addition to recording anomalies, DAITSS also identifies and records any attribute in the digital object that could hinder the preservation of the digital object, for example, encryption. These attributes are recorded as inhibitors. All defined inhibitors in DAITSS cause the preservation level of the digital objects to be downgraded.

Characterization

One essential process in digital preservation is to perform format characterization to extract technical metadata associated with each digital object in the preservation archival collection. The technical metadata are important attributes for understanding and managing the digital archival collections, especially for format monitoring and researching format transformation procedures.

Format characterization routines in DAITSS traverse through the structure of a digital object to extract a detailed list of its technical metadata. An analysis called a "background paper" for each format provides a summary of the technical metadata extracted by DAITSS [8]. The extracted metadata are then compared with any metadata submitted by the depositor. Discrepancies are noted and reported to the submitters. The extracted technical metadata are stored in the preservation database and also in the AIP descriptor.

To support the localization process required by the FDA preservation policy, DAITSS identifies and records any external URI reference contained in the digital object during the format characterization process. All identified external references, including ones that cannot be downloaded (broken links), are described in the report sent back to the submitters. Although DAITSS does not retrieve and download every external reference due to potential copyright infringement, recording of external references provides valuable information for describing the complexity of digital objects. By reporting the external references to the submitters, the submitters are aware of unpreserved external references and hence may avoid possible future disputes about preserved materials.

Format Transformation

To ensure the preserved digital objects remain usable, a common practice in the preservation repository is to transform formats from obsolete, non-standardized, or harder-to-preserve formats to more current, standardized, stable and preservable formats. The Action Plans on the FDA web site describe the current applicable format transformations for the supported formats in the FDA [8]. If a digital object is transformed from one format to the other, DAITSS records and maintains the relationships between the original and the transformed objects in its preservation database.

DAITSS implements the format transformation using a plugin model which carries out the transformation via local or third party (preferably free and open-source) software. Currently there are three format transformations in DAITSS: forward migration, normalization and localization.

Forward Migration

Forward migration is a transformation to convert one obsolete file format to a successor format. During the initial ingest and subsequent reingest of a SIP, DAITSS applies applicable forward migration on the digital objects in the SIP.

Perhaps the most critical challenges to preservation planning are monitoring and determining the formats obsolescence. Both PRONOM and the proposed GFDR plan to include format monitoring information. Format monitoring in the FDA is provided by periodical review of the format action plans. As none of the supported formats in the FDA has yet been replaced by a successor format, DAITSS does not yet implement any format migration on any current supported format.

Localization

Localization is a transformation of a file referencing other external files into a file with references to the downloaded local files. There are two primary purposes for localization. One is to guarantee that the information package descriptors remain usable. The other is to ensure that the materials submitted to the FDA will be preserved and disseminated as a complete entity in the repository. Originally, the FDA planned and implemented localization for all supported file formats that could contain external links, including XML, PDF and Quicktime. However, issues with potential copyright infringement when downloading and preserving external objects emerged during implementation, and the localization in the FDA was subsequently limited to XML schema.

Normalization

Format normalization in DAITSS is defined as the transformation of one format to another format which is perceived to be more stable and easier to be preserved. The FDA currently implements normalization for PDF, WAVE, AVI and Quicktime.

Any PDF file preserved in the FDA is normalized into a set of uncompressed TIFF images where each TIFF represents a page in the PDF. The relationships among the page-image TIFFs are maintained by creating an XML file describing those relationships. The PDF-to-TIFFs normalization is performed via the free Ghostscript software. Now that PDF/A has become an ISO standard, the FDA hopes to implement PDF-to-PDF/A normalization when a Linux based PDF-to-PDF/A converter becomes available.

In addition to file-based normalization as for PDF, the FDA also has bitstream-based normalization in place. Both AVI and Quicktime can contain multiple audio and/or video streams that are encoded in a variety of formats. With the rapid obsolescence and adoption of audio/video codecs and with limited resources for supporting each individual video/audio encoding format in the FDA, the FDA decided to perform bitstream normalization for compressed audio and video. By maintaining software for normalizing the encoded audio/video stream, the FDA demonstrates its decoding ability on the audio/video stream in the preserved multimedia wrapper formats. The normalized video/audio stream format may become needed if it ever becomes infeasible to transcode the original bistream format to a successor one in the future.

To ensure that the FDA maintains the capability to decode the audio streams without degrading the quality of the audio, every compressed audio stream in WAVE, AVI and Quicktime is normalized into a linear pulse-code modulation (LPCM) audio stream, the uncompressed audio format. Using an uncompressed video like RGB24 for video normalization is not appealing due to its space requirement [Table 2]. The FDA has identified several

selection criteria for video normalization including no inter-frame compression, standardized non-proprietary format, and software availability. Considering the potential quality degradation of normalized video, the FDA has decided to only normalize to video streams with no inter-frame compression. JPEG2000-based video, a.k.a. Motion JPEG 2000, appears to be the most suitable video normalization format due to its support for lossless intra-frame compression. In addition, part 3 of the ISO/IEC 15444 specification [11] standardized the Motion JPEG 2000 video stream for wrapping inside MJ2 files (.mj2). Governed by the rule for only normalizing into those file formats supported by the FDA, plus the lack of software support for direct transcoding of video encodings in AVI or Quicktime into MJ2 format on the Linux platform, every video stream in AVI and Quicktime are instead normalized into Motion JPEG. The normalized audio/video streams are wrapped back to their original wrapper formats (AVI or QuickTime) as they are the current supported multimedia wrappers in the FDA.

Table 2: Space requirement for one-minute video

	RGB24	4:2:2 (YUY2)	Motion JPEG	MPEG II [1]
720*480, ~30 fps	1800 MB	1200 MB	66 MB	6.2 MB
800*600, ~30 fps	2500 MB	1700 MB	100 MB	6.4 MB

Video and audio in AVI files are transcoded via the *mencoder* software [9] loaded with required video and audio codecs for the FDA. Likewise, the transcoding of video and audio in Quicktime files is supported through the *libquicktime* software [10]. Codec support in the multimedia wrapper formats is based on the needs of the FDA, the codec availability in the transcoding software, and a visual/acoustic evaluation of the transcoded video/audio. Files requiring unsupported codec are recorded with a limitation indicating the need for future codec additions in DAITSS.

Conclusion

Format automation including identification, validation, characterization and transformation are critical components for the implementation of any digital repository system. Given the current available technologies and resources, DAITSS has implemented a format automation scheme to satisfy its requirements.

Nevertheless, the question remains how we will ascertain the accuracy of the format automation schemes adopted in DAITSS. DAITSS is designed to allow progressive refinement of the processes of file identification, validation and characterization. Factoring the possibility of previously mis-identified objects, DAITSS re-identifies the re-ingested digital objects using the latest identification scheme upon dissemination. When a standardized file identification scheme becomes available, DAITSS plans to use the standardized scheme instead or reverify the identification result in DAITSS with a comparable scheme. We also hope to re-validate the result of the format validation and characterization with similar software like JHOVE 2 in the next major release of DAITSS. The semi-pluggable format model in DAITSS will also be redesigned to make it easier for others outside of the FDA to add format support.

Format automation for digital repository systems is still in its infancy. As technology on format automation becomes more standardized and mature, we hope to enhance the format automation scheme in DAITSS accordingly.

1 Compression ratio on MPEG II video is varied depending on the motion differentiation among video frames.

Acknowledgments

Priscilla Caplan, Assistant Director for Digital Library Service at the Florida Center for Library Automation, has led the research and development for the FDA and DAITSS. This paper would not be possible without the kind mentoring and strong leadership from her. The author also thanks Chris Vicary and Andrea Goethals for designing and implementing the format automation scheme in DAITSS during their employment with the FCLA.

References

[1] Priscilla Caplan, Building a Dark Archive in the Sunshine State: A Case Study, Proc. IS&T, pg. 9-13. (2005).

[2] Adrian Brown, Automatic Format Identification Using PRONOM and DROID, March 2006, http://droid.sourceforge.net/wiki/images/b/b4/Technical_Paper_1_-_Automatic_Format_Identification_v2.pdf

[3] Stephen L. Abrams, Establishing a Global Digital Format Registry, Library Trends, Vol. 54, No. 1, Summer 2005

[4] JHOVE, JSTOR/Harvard Object Validation Environment, http://hul.harvard.edu/jhove/

[5] PRONOM, UK National Archive, http://www.nationalarchives.gov.uk/pronom/

[6] Metadata Encoding & Transmission Standard, http://www.loc.gov/standards/mets/

[7] National Library of New Zealand, Metadata Extraction Tool Version 1.0, http://www.natlib.govt.nz/en/whatsnew/4initiatives.html

[8] Florida Digital Archive, Digital Archive Information, http://www.fcla.edu/digitalArchive/daInfo.htm

[9] Encoding with MEncoder, http://www.mplayerhq.hu/DOCS/HTML/en/encoding-guide.html

[10] Libquicktime, http://libquicktime.sourceforge.net/

[11] ISO/IEC 15444-4, JPEG 2000 Image Coding System, Part 3: Motion JPEG 2000.

Author Biography

Carol Chou received a MS degree in Computer Science from Virginia Tech. She joined the Florida Center for Library Automation (FCLA) in 2005 where she has worked as the Format Specialist for the Florida Digital Archive. Since joining of the FCLA, she has extended the file format collection in the FDA to handle audio and video collections including WAVE, AVI and QuickTime.