

Risk Management of Digital Information

Case Study for Image File Format

Prepared as a part of the CLIR-funded "Risk Management for Digital Information Project"

Oya Y. Rieger and Anne R. Kenney

1. COLLECTION AND ANALYSIS OF SOURCE AND TARGET FILE FORMAT RELATED INFORMATION

Investigation Test Bed

To assess the risks associated with file format migration for digital image collections, the project team selected one of Cornell University Library's digital image collections as a test bed. The Ezra Cornell Papers consists of correspondence, financial and legal records, court proceedings, and other documents pertaining principally to the Cornell family, the telegraph industry, and the founding of Cornell University. The collection is composed of 30,000 images stored on SCSI disks. They are scanned as 600 dpi, 1-bit TIFF 5.0 ITU Group 4 images. Tagged Image File Format (TIFF) is one of the most popular raster image file format, and is often used as the format of choice for master image files. It is platform-independent, and supports 1-bit to 24-bit imaging using a variety of compression methods.

The Ezra Cornell materials were scanned in-house using a XEROX scanning system. This system organizes and stores the structuring information (e.g., page number, folder number, etc.) in a format called Raster Document Object (RDO), which is Xerox's adoption of the International Office Document Architecture (ODA) and Interchange Format.¹

Goals of the File Format Migration Investigation

The file format migration investigation for image files had the following goals:

- Identify the TIFF file format attributes at risk during migration
- Assess the need to move these TIFF 5.0 image files to the current version (6.0)
- Evaluate the risks involved in converting TIFF 5.0 files to TIFF 6.0 files
- Investigate the status of upcoming revision to TIFF (7.0)
- Assess the risks involved in skipping a generation (TIFF 6.0) and waiting for the release of TIFF 7.0
- Assess risks and data loss associated with converting from RDO format to the open Cornell Digital Library format

¹ ODA, which became an ISO standard in 1988, has been developed to represent and allow the interchange of office documents. It contains facilities that allow both the structure and content of complex multimedia documents to be represented. Although ODA is an open standard, specifications for the RDO architecture are proprietary.

Collection and Analysis of Source and Target File Format Related Information

To identify digital image format attributes at risk, the project staff collected and analyzed information on different versions of TIFF file format. The research process included:

- Conducting a literature search on digital archiving issues pertaining to digital image collections with a specific focus on migration and the effects of file format choice in the migration chain.
- Investigating new digital preservation research and initiatives, such as ISO's Open Archival Information System (OAIS) [ISO, 1998], WGBH's Universal Preservation Format (UPF) [Shepard and MacCarn, 1999], Stephen Robertson's Rosetta Stone model [Robertson,], among others.
- Conducting a literature and projects survey to determine the extent of work performed on developing risk analysis based on image files.
- Reviewing different risk assessment tools developed for various purposes -- focusing on the form and functionality of these tools and how they can be adapted for the purposes of this project.
- Exploring the dependencies that extend beyond basic image file format attributes, such as internal and external relationships between images and their accompanying metadata files (viewing images as "digital objects" and examining their metadata, associated scripts and programs, etc.).
- Identifying the attributes of digital images that are at risk during format migration, including the effects of migration on metadata, and various scripts and programs that support retrieval and management of the collection.
- Investigating the existing and emerging bitmap image file formats with a focus on their longevity and other archival attributes.
- Exploring vulnerabilities associated with file format migration and identifying risks associated with "migrating" or "not migrating" these files with a focus on TIFF files.
- Analyzing the factors involved in decision-making in migration projects, such as reformatting a collection of images from TIFF4 to TIFF5 format.
- Examining and comparing the TIFF file format specifications for Version 4.0, 5.0, and 6.0.
- Exploring the future of TIFF as a file format, with a focus on the characteristics of the upcoming TIFF 7.0.
- Investigating the issues introduced by storing structuring metadata in Xerox Raster Document Object (RDO) format.
- Identifying the risks involved in converting RDO files to Cornell Digital Library (CDL) format (<http://andrew2.andrew.cmu.edu/rfc/rfc1691.html>).

An outcome of this research process was the development of the following chart to categorize the risks associated with file format-based migration:

Risks Associated With File Format-Based Migration for Image Collections

RISK CATEGORY	EXAMPLES
content fixity (bit configuration, including bitstream, form, and structure)	bits/bitstreams are corrupted due to software bugs or mishandling of storage media, mechanical failure of devices (heads crash, tape jams), etc.
	file format is accompanied with new compression that alters the bit configuration
	file header information did not migrate, or partially or incorrectly migrated
	image quality is affected (resolution, dynamic range, color spaces, etc.) due to the alterations to the bit configuration -- new format affects quality
	changes to byte order due to the new file format specifications
security	effects of format migration on watermark, digital stamp, or other cryptographic techniques for "fixity"
context and integrity relationship/interaction with other related files (or other elements of the wider digital environment) including hardware/software dependencies	due to different hardware and software dependencies, reading/processing the new file format requires a new configuration
	linkages to other files (metadata files, scripts, derivatives -- such as marked-up or text versions, on-the-fly conversion programs) are altered during migration
	new file format reduces the file size (due to file format organization or new compression) causing denser storage - potential directory structuring problems if one tries to consolidate files to use extra storage space
	media becomes more dense, affecting labels and file structuring (this might be also due to file organization protocols of the new storage medium/OS (refresh & migrate strategy))
references locate images definitively and reliably over time among other digital objects	file extension change due to file format upgrade and its effect on URLs
	migration activity is not well documented -- causing provenance information to be incomplete/inaccurate (potential problem for future migration activities)
cost	long-term costs associated with migration are unpredictable as each migration cycle may potentially involve different procedures depending on the nature of migration (routine migration vs. paradigm shift)
	value of the collection may not be sufficiently determined, making prioritization impossible (how do you evaluate the parts of a collection to make individual migrate or not-migrate decisions -- e.g., going through the MOA1 journals and deciding which ones to migrate -- labor intensive selection process)
	unscalable unless there is a standard architecture (centralized storage, metadata standards, file format/compression standards, etc.) that encompasses the image collections so that the same migration strategy can be easily implemented for other similar collections

<i>staffing</i>	staff turnover and continuity of migration decisions -- long term planning (especially if insufficient preservation metadata are captured & migration path is not well documented)
	full-time, permanent job responsibility or ad hoc, temporary assignments for rescue operations
	insufficient technical expertise
	unpredictability of migration cycles makes staffing requirements (skills, time involved, \$ spent, etc.) planning challenging
<i>functionality</i>	derivative creation, such as printing, may be affected due to the new features introduced by the new file format
	if the master copy is also used for access -- changes may cause decreased/increased functionality requiring interface modifications (e.g., static vs. multi-resolution image, or Web not supporting the new format)
	loss of unique features that are not supported in other files formats (e.g., losing the progressive display functionality when GIF files are migrated to another format)
	artifactual value (original use context) may be lost due to changes introduced during migration -- "preserving the experience"
<i>legal</i>	strict copyright regulations may limit the use of new derivatives that can be created from the new format (e.g., the institution is only allowed to provide images at a certain resolution not to compete with the original)

Conclusions of the Source and Target File Format Analysis

The project staff was able to gather a substantial amount of information about the different versions of TIFF as most of the specifications are publicly available on the Adobe FTP site. TIFF was developed by Aldus and Microsoft, and the specification was owned by Aldus, which in turn merged with Adobe Systems, Incorporated. Consequently, Adobe Systems now holds the copyright for the TIFF specifications. TIFF is a highly flexible and platform-independent file format. It is supported by numerous image processing applications. A great strength of the TIFF file format is its file header option, which enables recording within the file itself of a wide variety of metadata (descriptive, administrative, and structural). The set of fields (or "tags") in TIFF is quite extensive, making it the format of choice for most archival reformatting. But a very large number of TIFF fields are not defined by the standard. This offers the advantage of being open and useable, alongside the dangers that different institutions will define these fields in different ways, leading to problems of compatibility. Another flexibility of TIFF that causes confusion is related to byte order. For example the TIFF format permits both MSB ("Motorola") and LSB ("Intel") byte order data to be stored, with a header item indicating which order is used.

Tracking the TIFF 7.0 development turned out to be a challenging task. The attempts of the project team in contacting TIFF 7.0 developers, Adobe, and even TIFF listserv subscribers were fruitless. The TIFF 7.0 development group seems to be determined not to release any information regarding the development work. Therefore, the project team was not able to do any comparison between TIFF 7.0 and the earlier versions. After conducting an extensive evaluation and comparison of TIFF 5.0 and TIFF 6.0 specifications, several tests were run to compare the quality and utility of a subset of TIFF 5.0 images converted to TIFF 6.0. This exploration revealed no major differences between the different versions of TIFF. The project team concluded that there were no risks involved at this point in leaving the testbed images in TIFF 5.0 format. The team will continue to monitor the development of TIFF 7.0. After reaching this conclusion, the team shifted its focus of the risk assessment study for image files to examining storing structural metadata in the proprietary Xerox RDO format.

Raster Document Object Files

An RDO file contains information about the structure of an image document, as well as a file location pointer for each page image in that document. Each page in the document is represented by a single TIFF file. The TIFF files contain the digital data from the scanned page, along with a TIFF header that describes the characteristics of the image file. Because the XDOD system is proprietary, the structure of image documents can be displayed only through the use of the appropriate Xerox software.

2. SELECTION AND EVALUATION OF CONVERSION SOFTWARE

Evaluation of the TIFF conversion software was not necessary as a decision was made to maintain the files in TIFF 5.0 format. There are several conversion programs in the market for converting TIFF files to various TIFF versions and other file formats (e.g., TIFF to GIFF, TIFF to PNG, etc.). TIFF 5.0 to TIFF 6.0 conversion could be interpreted as an update than a migration process.

In 1994, Cornell undertook a project to convert the proprietary RDO files to an open Cornell Digital Library (CDL) format. The specifications for CDL, which were released in August 1994 through a Request for Comments (#1691), defines an architecture for the storage and retrieval of CUL's image collection. Similar to RDO files, the CDL document structure provides direct access to the components of image collections (pages, sections, chapters, etc.).

While the project team's main interest was exploring the export of files created on XDOD 3.0, the immediate concern was with the older RDOs, especially in light of the Y2K compliance issues (the XDODs may no longer work after December '99 unless an expensive upgrade is implemented).

The XDOD RDO to CDL format conversion involved two steps. Cornell runs a XEROX-supplied conversion tool (XDOD export tool) that converts the RDO files into a series of ASCII metadata files. This tool is old and can only run in Windows 3.1, and its dissemination is authorized "only pursuant to a valid written license from Xerox." Then, through a locally developed PERL script, the ASCII metadata files are converted to Cornell Digital Library (CDL) format (Appendix A). These CDL-formatted structural metadata files are used for navigating through a document (<http://moa.cit.cornell.edu/MOA/EZRA.html>). The ASCII RDO-to-CDL program was written by the CUL Information Technology staff.

RDO-to-CDL conversion can not be achieved through a single software tool as Xerox has not released any RDO specifications.

3. DEVELOPMENT OF TOOLS FOR ASSESSING THE SOURCE-TO-TARGET FORMAT TRANSFER

No specific software tool was developed to analyze the effects of migration from RDO to CDL format, as all files created using the XDOD scanning system possess identical information fields.

4. COMPARISON AND ANALYSIS AFTER CONVERSION TO SOURCE FILE FORMAT

The comparison was done manually by comparing the structural metadata elements that were captured in RDO files to the CDL structure. Basically, the team compared the list of structural metadata elements captured during scanning to the CDL structuring requirements. All the structural elements mapped to the

CDL structure, and there was no loss. Even if there had been a loss, the project team decided that it was much riskier (actually detrimental) to leave the structuring information in an unsupported proprietary format.

5. Releasing the Export Tool to Other Institutions

As part of this project, Cornell investigated the possibility of further developing and making available the export tool to other institutions that have legacy collections in the proprietary Xerox RDO format. This investigation was spurred by two additional concerns. First, several institutions had requested access to the tool over the past three to four years, but only Yale University had secured permission from Xerox to use it. Second, in early summer 1999, Xerox informed Cornell that the XDOD 2.x scanning workstations would not be Y2K compliant without an expensive upgrade. Because Cornell had begun to phase out use of the XDOD systems, and had converted all RDO files to the CDL format, our concerns over the millennium focused on our sister institutions' collections.

We initially considered further development of the Export Tool into more generic software for external use, but quickly concluded that this would be both expensive and time-consuming. Cornell did not receive any specifications from Xerox for the proprietary tool, and the software developer at Xerox indicated that he doubted that the company still had the tools and specifications to make the system work. We concluded then to focus attention on securing permission to release the current version of the Export Tool. A two-year effort to obtain a blanket permission from Xerox to make the tool broadly accessible had stalled, so we turned to documenting the extent of the problem, concluding that Xerox might be more amenable to a very limited release.

In late April 1999, Cornell posted the following announcement on eleven listservs.

Export Tool to Convert Xerox RDO Files to Open Digital Library Format

Has your institution created digital image files using the proprietary Xerox Documents on Demand software that generates Raster Document Objects (RDOs) to store structural metadata? Cornell University is seeking feedback from these institutions to determine what demand there would be for freeware to convert those RDOs for use in other metadata applications. Cornell has used the RDO2CDL export tool to migrate RDOs to ASCII metadata files that recreate the logical and physical structure format of the RDO (called CDL). If your institution is interested in utilizing such an Export Tool, please send contact information and a brief description of your needs to: Anne R. Kenney (ark3@cornell.edu).

The responses received by early June were surprisingly few in number.

Universities with files created on XDOD 2.5 or older included Harvard, Penn State, the University of Tennessee--Knoxville, and Yale. Those responding with files created using XDOD 3.01 or DigiPath included the Hein Publishing Co., Illinois State Library, the National Document Center (Athens), Indiana University, the University of Toronto, and the NOAA Miami Regional Library (which was considering using the technology).

Inquiries to Xerox about releasing the tool to this group resulted in further clarification that the RDO Export software would only work as configured on XDOD Version 2.x systems. The format of the RDO changed slightly from version 2 to version 3, and the Export Tool would not convert the structural data on version 3 or higher systems. The Hein Company had used the tool with version 3 files through a collaborative project with Cornell, but only page labels were exported, not structuring information. William Anderson, the Xerox software engineer who created the tool, did suggest that it would be possible to get the structure information out of the version 3 RDO files, but it would take a programmer with knowledge of the Office Document Architecture (of which RDO is a variant), fair knowledge of Unix tools, and a copy of the RDO Version 3 specification, which Xerox seemed unwilling or unable to make available publicly. Anderson suggested that "If customers are looking to buy DigiPath today, and they need that facility, they should ask for it." Xerox determined that they would only be able to grant access to this software to XDOD 2.x customers who were not migrating to DigiPath.

From June to early September, efforts continued to reach legal agreement with Xerox over the release of the Export Tool software to XDOD 2.x users. Cornell received a copy of the proposed "Software License Agreement" on August 26, 1999, which granted the institution a non-exclusive, perpetual, royalty-free license to use the software and the right to provide a sublicense only to those institutions who had reported using the XD Version 2.x systems, collectively referred to as "Authorized Educational Institutions" (AEI). Lee Cartmill, the chief financial officer at Cornell University Library, expressed concern about the indemnity clause in the agreement, which required Cornell to "defend, indemnify and hold Xerox harmless from and against any and all third party claims that arise from or relate to the Software and their respective use of the Software." Cornell attempted to have this clause modified, but when Xerox remained adamant, Lee Cartmill drafted a "Software Sublicense Agreement," which would require the AEIs to extend the indemnity and limitation of liability to Cornell University. As of this writing, the four institutions have been notified of these stipulations, and copies of the agreements are being reviewed by their legal advisors. It remains to be seen whether any or all of these institutions will agree to these license stipulations, but Cornell will not sign the agreement with Xerox unless they do so.

Bibliography

ISO Reference Model for an Open Archival Information System (OAIS). 1998.

<http://ssdoo.gsfc.nasa.gov/nost/isoas/us/overview.html>

Robertson, Steven. 1998. The Rosetta Stone Model. Proceeding of the Sixth DELOS Workshop, Preservation of Digital Information, June 1998. Available at:

<http://crack.inesc.pt/events/ercim/delos6/papers/rosetta.doc>

Shepard, Thom, and MacCarn, Dave. 1999. UPF: Universal Preservation Format.

Available at: <http://info.wgbh.org/upf/>