ARTICLE EXCERPTED FROM:

ISQ INFORMATION STANDARDS QUARTERLY

SPRING 2010 | VOL 22 | ISSUE 2 | ISSN 1041-0031

SPECIAL ISSUE: DIGITAL PRESERVATION

DIGITAL PRESERVATION METADATA STANDARDS

TRUSTWORTHY DIGITAL REPOSITORIES

UNIFIED DIGITAL FORMATS REGISTRY

AUDIO-VISUAL DIGITIZATION GUIDELINES

DIGITAL PRESERVATION EDUCATION

NISO
How the information world CONNECTS

Digital Preservation in Capable Hands:

Taking Control of Risk Assessment at the National Library of New Zealand

KEVIN DE VORSEY AND PETER MCKINNEY

New Zealand's digital documentary heritage is encoded according to a diverse array of file formats. Identification and characterization of the formats is a constant challenge. This challenge makes it difficult to establish an accurate risk view of the content to mitigate format obsoleteness.

The National Digital Heritage Archive (NDHA) of the National Library of New Zealand Te Puna Mātauranga o Aotearoa has concluded that the measurement of conformance of files to a format standard for such risk analysis is at best insufficient and at worst harmful. For the digital documentary heritage of New Zealand, the ideal is the measurement of individual file profiles against application specifications. This gives a meaningful and actionable risk view of our content.

With no limitations or control over the format of the content that is collected and preserved, the Library has issues to resolve before the long-term preservation of digital collections can be assured. There are many significant obstacles that make the term "permanent access" an almost meaningless catchphrase when applied to such a collection of digital content made up of disparate file formats. Solving these and other problems is the responsibility of the National Digital Heritage Archive (NDHA) and a significant step has been taken through the development of the Rosetta preservation repository system in conjunction with Ex Libris Group.

While the life-span of content stored on physical materials such as paper, glass, wood, and stone can be accurately predicted based on hundreds of years of experience, backed by scientific research into material composition and the effects of environmental conditions like temperature and humidity, the best that the preservation community can do with digital material is to make educated guesses based on a few decades of mostly anecdotal experience. The concept of information encoded according to a file format has only been in existence since about the 1950s and therefore the field of digital preservation must be considered as being still in its infancy. Happily, significant advances have occurred in the area of data storage and management that permit cultural heritage institutions to manage enormous digital collections of permanently valuable material in online (or nearly online) repositories of spinning disks and/or robotic tape libraries. Through the use of checksums to detect format rot or corruption, virus scanning to protect against malicious code, robust network and physical security, and comprehensive disaster planning, it is not too far-fetched to believe that it is now possible to guarantee bitstream preservation—which is to say, preserving deposited files perfectly in their original form. We view this as "passive preservation" that is

CONTINUED »

Through the use of checksums to detect format rot or corruption, virus scanning to protect against malicious code, robust network and physical security, and comprehensive disaster planning, it is not too far-fetched to believe that it is now possible to guarantee bitstream preservation—which is to say, preserving deposited files perfectly in their original form.

The range of formats the National Library accepts is very wide:

- wave, broadcast wave
- Wordstar
- TIFF, JPEG, GIF
- ARC format
- Sibelius music composition
- MacWrite
- text, mp3, flac
- unidentified

foundational to digital preservation. Unfortunately, while the perfect preservation of a human-readable format such as a paper manuscript is usually synonymous with access to its content, bit-preservation of electronic formats is not. The inevitable obsoleteness of the hardware and software components necessary to interpret and render files in a usable form makes it necessary to complement perfect but passive preservation with some form of active, managed preservation. (We are painfully aware that we do not discuss in more detail our use of the word "render." It is a loaded term with many levels of interpretation. We are currently defining this internally as it is critical to our risk analysis. Space deters us from exploring it further in this paper.) This demands an accurate risk view of the repository. This risk view is the mechanism that offers enough warning to the NDHA in order that action can be taken to allow continued access to the material.

The problem with format specification adherence as an indicator of risk

From our reading, the primary methods currently being suggested for predicting this type of risk involve the comparison of files to a format specification, which in turn is graded against agreed-upon sustainability criteria. There exist many misunderstandings around format sustainability that have contributed to the idea that there are "archival" or "preservation" formats. The NDHA is uncomfortable with the concept of inherently preservation-worthy formats. It is our belief that while sustainability factors may prove useful for a forensic understanding of formats in the future and in interpreting files that are discovered after a period of benign neglect, there are other more practicable methods for identifying risk that are better suited for supporting active preservation in a repository.

Along with bit preservation, accurate identification of a file's encoding (its format) is foundational to preservation. In response, the preservation community has developed utilities that identify files by format and "validate" them; that is, measure their compliance with the format specification with which they have been associated. It is certainly useful to possess the information that a string of bits is an audio file that is encoded according to the Broadcast Wave EBU Specification and that they failed JHOVE's measures of well-formedness or validity. But, it is more important to know that the bits were written according to a profile that has been associated with a particular legacy application, and that this application is known to encode in a non-standard way due to an aspect of the specification that was originally open to interpretation. The NDHA has encountered this phenomenon with a number of formats including Rich Text Format, Tagged Image File Format, and Broadcast Wave.

The content the NDHA preserves

The National Library can, and does accept all formats. It collects content, not "perfect" formats. All materials selected as Library collection items are ingested into the preservation repository essentially as is. The current policy of the NDHA is not to transform content into preferred formats on ingest, but this may be considered after additional research is conducted. In order to actively preserve this content, we must first understand exactly what form it is in. Every file must therefore be identified by its format and where possible, a picture created of its characteristics. Once this is done, different management views can be taken. The range of formats is very wide. We have Sibelius music composition files, web harvests in ARC format, Wordstar, MacWrite, TIFF, JPEG, GIF, text, mp3, flac, wave, broadcast wave, and a whole host of format unknowns.

Our experience with New Zealand's documentary heritage is that files contain multifarious properties. These are based on the world of possibilities that the format standard describes, but can also include non-standard properties. The range of possibilities and relationships between them is such that it is quite meaningless to purely measure a file's adherence to the format standard.

We can take PDF files as an instructive example. Adobe has clear standards for versions of PDF. However, the wide range of applications that can create PDFs do not always stick to this standard. Indeed, non-adhering PDFs can be made by Adobe's own suite of applications. What does it mean if a PDF is invalid because it does not have tags for the images as the standard requires? Should we base risk on this non-compliance?

In addition, consider this: it is not a bold statement to suggest that the majority of the world-wide Web is written in non-conforming code. Would this content not be at risk if it was written in perfect code? Is the conformance of the code really the biggest risk facing this material?

We can take PDF files as an instructive example. Adobe has clear standards for versions of PDF. However, the wide range of applications that can create PDFs do not always stick to this standard. Indeed, non-adhering PDFs can be made by Adobe's own suite of applications.

NDHA Risk Analysis

Within the NDHA, we base risk on practical capabilities; risk analysis through tracking format standards is too abstract for us. There has been a degree of literature about risk analysis of collections. Our understanding of the body of work is that many have coalesced around the utilization of what are described as "sustainability factors." Depending on the source, these number from seven to fourteen. The original study by Arms and Fleischhauer identifies seven factors. These were put in place to assess the sustainability of formats for preserving content. Further work at the Dutch National Library moved this work into the area of risk assessment for their own very specific circumstances (regular access is not offered by the institution to the materials this was applied to). In essence across all the literature on this, the criteria remain essentially the same but are given different situational groupings and nomenclature. They include factors such as the level of documentation for a format, a format's backwards compatibility, and its complexity.

We do not believe these factors belong in the area of risk assessment of collections. Within our repository, we do note sustainability factors against formats and applications. But these are not used to determine any view of risk to content due to format obsoleteness. (For an excellent discussion on the terms obsolete, obsolescence, and obsolescent, see Pearson & Webb 2008.) We will use those factors to offer decision-making information when selecting formats to migrate content into (i.e., dealing with the risk). However, it is unlikely that they will be key elements of decision-making as the most important input will be the new format's ability to render and our level of comfort with that rendering.

The risk analysis method the National Library employs measures each ingested file against a format and application relationship. Simply, our view of risk is that if the National Library cannot render it, it is at risk. We use a format, application, and risk library within Rosetta to identify risks. As the first stage of risk analysis, the format and application libraries use two levels of relationship—an association and an activation. A format can be associated with an application. For example, we can link PSD files with Adobe Photoshop CS3. However, even with this association, any PSD files we receive will still be classed as at risk. The second level of linkage is an activation of the association. An activation is the institution (in this case, NLNZ) declaring that they have the application within their own environment. ("Environment" here means "within the institution." The activated application could be embedded within the preservation system, deployed as part of the institution's IT infrastructure, or even held on a stand-alone PC within a relevant business unit. Critically though, the application is under the control of the institution.) Once this activation is made, then PSD files in the repository are no longer viewed as being at risk: the National Library can render the content.

What these relationships between the format and application libraries offer is a description of the universe of rendering possibilities and the ability to define a world view upon which an institution's risk is determined. This is the most basic level of risk analysis employed in Rosetta.

A more detailed interpretation is the next layer of risk analysis. This layer looks at the characteristics of the files themselves. What we do is understand the capabilities of the applications we have and determine if there are certain

CONTINUED »

What does it mean if we have 10,000 images in a format that a risk matrix based on sustainability factors tells us are at risk because documentation on the format is incomplete? If we can happily render these files, this analysis is unhelpful.

properties that will cause them to "reject" a file that is otherwise in the correct format. These properties are then noted in the risk library. It is important to understand that we do not maintain a registry of all characteristics that are possible within a format; we only note exceptions that break the format/application relationship.

For example, the Library receives a number of MP3 files each week. We know currently that we have the rendering capability for MP3s encoded with LAME and Fraunhofer methods. However, we cannot reliably render MP3s that are created using the Xing encoding method. If a Xing MP3 is deposited, the relationship at the format and application level determines that the file is not at risk, because at this gross level, we are happy with the file. However, the extracted metadata (from the NLNZ metadata extraction utility) contains the troublesome encoding. This does not stop the file from being passed to the permanent repository, but the next day, when the risk report is re-run it identifies this particular file as matching the risk profile and reports on it as being at risk.

Conclusion

Where does this leave us? If we rely on identification and characterization tools that measure a file's risk through adherence to a standard, we are basing our risk on what to us, is relatively meaningless information. What does it mean that our TIFF is invalid? What does it mean if we have 10,000 images in a format that a risk matrix based on sustainability factors tells us are at risk because documentation on the format is incomplete? If we can happily render these files, this analysis is unhelpful. The most worrying end of this particular road is that it could very well guide us to a course of action when none is actually required.

The risk we have been discussing is the risk of what the community terms "digital obsolescence." It is our view that risk is situational; it is not a statement of fact. At risk is not an inherent state of files and formats, it is an institution's view of its content determined by the policies, guidelines, and drivers it has at any one point in time. We have included no discussion on our "control" of the applications used to render files. Tracking the contract dates and review dates for all applications is our method of analyzing "obsolescence" (the march to being obsolete). This will give is adequate time to plan for action that is truly required.

The National Library of New Zealand, by basing its risk routines on institutional rendering capability, creates a view of its repository that gives accurate and meaningful information on what can and cannot be rendered. To us, this is the essence of obsoleteness. | IP | doi: 10.3789/isqv22n2.2010.06

KEVIN DE VORSEY <Kevin.DeVorsey@natlib.govt.nz> is an Electronic Formats Specialist at the National Records and Archives Administration, but at time of writing was the Digital Preservation Analyst in the NDHA. Peter McKinney (Peter.McKinney@natlib.govt.nz) is NDHA Policy Analyst at the National Library of New Zealand Te Puna Mātauranga o Aotearoa.

RELEVANT LINKS

Arms, Caroline and Carl Fleischhauer. Digital Formats: Factors for Sustainability, Functionality, and Quality. IS&T Archiving 2005 Conference, Washington, D.C.
memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf

European Broadcast Union Specification of the Broadcast Wave Format
tech.ebu.ch/docs/tech/tech3285.pdf

Lawrence, Gregory W., et al. Risk Management of Digital Information: A File Format Investigation. Council on Library and Information Resource, June 2000.
www.clir.org/pubs/reports/pub93/contents.html

National Digital Heritage Archive (NDHA)
www.natlib.govt.nz/about-us/current-initiatives/ndha

NLNZ Metadata Extraction Tool
meta-extractor.sourceforge.net/

Rog, Judith and Caroline van Wijk. Evaluating File Formats for Long-term Preservation. National Library of the Netherlands, 2008.
www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf

Rosetta system, Ex Libris
www.exlibrisgroup.com/category/RosettaOverview

Strodl, Stephan, et al. How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure. In: International Conference on Digital Libraries, Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries, Vancouver, BC, Canada: 2007.
www.ifs.tuwien.ac.at/~strodl/paper/FP060-strodl.pdf

Pearson, David and Colin Webb. Defining File Format Obsolescence: A Risky Journey. The International Journal of Digital Curation, 3(1), 2008.
www.ifs.tuwien.ac.at/~strodl/paper/FP060-strodl.pdf