The International Journal of Digital Curation
Issue 1, Volume 3 | 2008

Defining File Format Obsolescence: A Risky Journey

David Pearson, Colin Webb,
Web Archiving and Digital Preservation Branch,
National Library of Australia

July 2008

Abstract

File format obsolescence is a major risk factor threatening the ongoing usefulness of digital information collections. While the preservation community has become increasingly interested in tools for assessing a wide range of risks, the National Library of Australia is developing mechanisms specifically focused on the risks of format obsolescence. The paper reports on the AONS II Project, undertaken in conjunction with the Australian Partnership for Sustainable Repositories (APSR). The project aimed to refine and develop a software tool that would automatically find and report indicators of obsolescence risks, to help repository managers decide if preservation action is needed. The paper discusses the current mismatch between this objective and the available sources of information on file formats, and emphasises the need to take account of both local and global factors in assessing risk. The paper calls for the preservation community to engage with the further development of thinking about file format obsolescence.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

We know that in our information-obsessed world, change is everything. And yet some information is required to live beyond the moment; some information is valued beyond tomorrow's headlines, and must be managed to be accessible, usable, and understandable in the long term.

Cycles of change in file formats impinge on even the most casual users of digital data. Technological change and format obsolescence are potentially major problems for every repository manager and data user. This is particularly true given the ever-increasing reliance on digital storage and distribution of information, the plethora of file formats, the dynamic nature of computing environments, and the unremitting but often unpredictable drivers that cause formats to become obsolete. In order to ensure the long-term availability and usefulness of digital materials, repository managers need help in managing format obsolescence risks.

This paper focuses on one set of digital preservation questions that repository managers might ask. For example, how do they know if they risk not being able to provide access to content stored in their repository, which would require them to take action to avoid losing access? The paper discusses an approach to such a question, as well as the development of a tool designed to help repository managers confront this problem.

This paper is quite impartial on the questions of when and where format risk assessment is best undertaken; and of when and where preservation action is best attempted. It is almost certainly true that repositories would reduce their format obsolescence risks if content were to be *normalised* to some kind of durable encodings at creation or at ingest; however, this paper, and the work on which it reports, are based on the realistic supposition that many repositories will continue to deal with file formats affected by technological change.

The authors hope that work on format obsolescence risk assessment, and on the associated AONS II (Automatic Obsolescence Notification System, version 2) software might help to bring digital preservation planning to a more practical level. They invite the engagement of others in critiquing and refining this paper's approach to what this preservation risk really means and how to recognise it.

File Format Obsolescence

Obsolescence

Building a tool to help in recognising format obsolescence has involved a lengthy, at times contentious, and as yet unfinished process of understanding the nature of file format obsolescence and how it might be recognised and measured.

An important starting point is to distinguish between *physical* format (storage media) obsolescence and *file* format obsolescence. While it is easy to focus on the former, as the dated characteristics of old physical format carriers immediately suggest access problems (Figure 1 below), yet file formats supersede one another almost invisibly. Both kinds of format obsolescence are important: if one can not access

either the physical or logical format, then access to the content is lost. This paper assumes access to the physical carrier, and concentrates on the dangers to file formats.

Figure 1. Just a few of the physical format carriers likely to cause access problems (Dinosaurs, media and image courtesy of National Archives of Australia).

A file format is a particular way to encode information for storage and use[1]. While some file formats may be intended to hide information rather than making it retrievable, most file formats are designed with the purpose of allowing the encoded data to be re-presented, given the right means.

For as many years as digital content has been encoded in diverse file formats, access to the content has depended on the availability of means for understanding the encoding. The idea of file format *obsolescence* is related to two often observed phenomena:

- the development of new format encodings that take the place of already existing formats in the marketplace of use; and
- changes in the availability of presentation tools, generally (although not exclusively) in the direction of decreasing availability, for any particular file format.

The basic premise of preservation as it relates to digital information is the maintenance of an ability to provide meaningful access. When file formats can no longer be reliably read, access is effectively lost. This problem is clearly understood by virtually all users of digital information, well beyond the confines of the preservation community. For many data archiving environments, preservation of data in a form that does not compromise the authenticity of those data is also critically important.

The National Library of Australia (NLA) has been in the business of digital preservation since the early 1990s. Over this time a number of useful paradigms have appeared which have informed the Library's thinking about digital preservation. Two

[1] File Format Definition, Wikipedia http://en.wikipedia.org/wiki/File_format

are seen as being particular relevant to the question of format obsolescence:

- The *performance model* developed by the National Archives of Australia (NAA) is a useful way of thinking about accessing digital content (Heslop, Davis, & Wilson, 2002). This model states that encoded content acted upon by a specific process creates a presentation performance. Different processes (specific combinations of software, hardware and other dependencies) may create essentially similar and acceptable performances. Using this model, meaning can be re-presented regardless of the means of providing access. However in order for this model to work, the required characteristics that must be preserved for each information resource (*essence* in the NAA model, *significant properties* in wider preservation parlance) must be defined, and each performance outcome tested against it. The performance model tells us that it may be possible to break the link between a particular file format and the original means of providing access to it, and still deliver an acceptable presentation of the meaning of the data.
- The *view-path model* developed by the Koninklijke Bibliotheek (KB - National Library of the Netherlands) records the chain of elements (protocols, software, hardware) that allow any specific encoded content to be made understandable to the user (Sierman, 2007). The KB aims to record view-path information and multiple alternative view paths for each file format. Again, the view-path model asserts that useful access to content in a file format is a function of both the file format and the available view-paths.

We have tried to reflect these understandings in a way that might inform thinking about format obsolescence, in the following diagram (Figure 2):

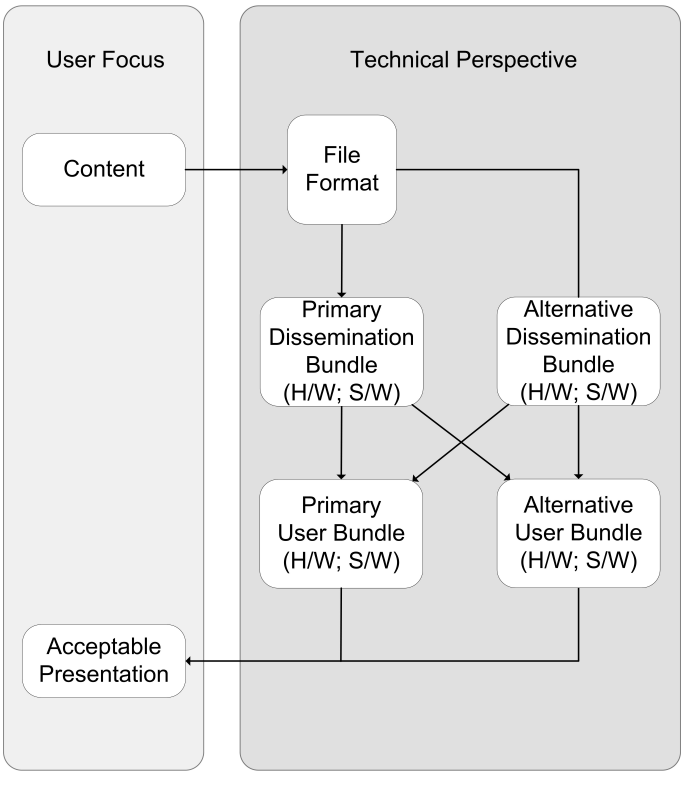

Figure 2. The diagram asserts that the data user, whether human or machine, is primarily concerned with the content (the understandable meaning of the data), and its

presentation. From a technical perspective, however, those responsible for delivering the content have to be concerned with a number of other elements as well. The way the data is encoded in a file format is one obviously critical element. The bundle of hardware and software used to disseminate the encoded data to the user may involve those intended by the creator (the primary dissemination bundle), but they may also involve alternative bundles. In some environments, the user may also have bundles of hardware and software that have to be navigated; again, a number of other bundles may provide access, not just the primary bundle. Managing the risk of file format obsolescence must take account of all of these elements in seeking to ensure digital content remains available for acceptable presentation to users.

Assessing File Format Obsolescence Risk: Some Definitional Issues for our Project

We believe it is important to minimise any vagueness in considering how to assess format obsolescence risks. Vagueness potentially detracts from the objective of making informed, timely decisions about the need for preservation actions.

Obsolescence describes a state of beoming obsolete, rather than a state of already being obsolete. For our purposes, however, we are interested in identifying file formats affected by both - obviously related - conditions:

- those that can no longer be rendered (and are therefore *obsolete*); and
- those that are likely to become unrenderable within a timeframe demanding action (and are therefore *obsolescent*).

For convenience, we refer to both conditions with the label *file format obsolescence*, which we believe is in line with common usage in the digital preservation community, even though it may invite some ambiguity.

However, in suggesting that these states should be monitored together as contiguous risks, it is important to remember that the distinction between them can be critical. There *is* a significant difference between a file format which can no longer be rendered and one which can, even though the latter may be at high risk of soon becoming unrenderable. It is the difference between the dead and the dying (as our editor put it). The importance of the distinction lies in understanding that we are not just shamans communicating with the walking dead, but also doctors looking for early warning signs of illness.

In the absence of another convenient label to cover both the obsolete and the obsolescent, *file format obsolescence* seems reasonable so long as we recognise it as a dynamic continuum which reaches an ultimate point when a file format is no longer readable and therefore becomes functionally obsolete.

Taking this approach, some other definitional points are also worth noting:

- We are not making judgments about which formats should be used for archiving digital objects. While the level of obsolescence risk is very often a factor in such decisions, there are usually many other factors also at work.
- Similarly, we are not making judgments based on how hard a format will be to deal with once preservation action is needed. For example, a proprietary format that is covered by closed specifications may well be

more *risky* because of potential difficulties in taking effective preservation actions. However, these *post-risk* issues are not generally good indicators of when action is going to be needed.

- The concept of *risk* must be clearly defined. Virtually all file formats, from the time they are first released, are *obsolescent*, since they will almost certainly be superseded at some stage. However, it is not particularly helpful to decide a format is at risk of obsolescence because it will at some future time be superseded. We have had to remind ourselves that we are looking for indicators of different levels of risk over particular timeframes. Deciding on an appropriate timeframe in which to access risk is likely to be tied to the level of readiness to take action once it is needed. We eventually decided that in the context of format obsolescence, *risk* is about the impending loss of *the means of providing access*. In preservation terms, a file format only becomes effectively *obsolete* when access is no longer possible. If we are looking for indicators of change that suggest this state is imminent, we will have to look for changes in the availability of the means of providing access, rather than just changes in formats themselves. At the same time, we decided that the release of new format versions and support end-dates for file formats should be taken as potentially important warning signs, even if they are not necessarily sure signs that content in a format is about to become inaccessible.
- We assume that the same format may well have different levels of obsolescence risk in different repositories, depending on the availability of software and hardware dependencies. Format obsolescence assessment must take account of both global and local indicators. A wise repository manager would look for any external indicators that access may become more difficult; but must also consider the local availability and sustainability of the means of providing access. If a repository maintains software that provides non-problematic access to content in a given file format, it might well rate the risk of obsolescence for that format as considerably lower than other repositories which have abandoned the software. A format may be considered *obsolete* in the broader community of users but remain viable in local conditions. The reverse may also be true, where one repository may have an unusually high obsolescence risk for a format because it lacks appropriate processing software that is available elsewhere.
- In line with both the *performance* model and the *view-path* model referred to above, it is perfectly reasonable to take into account not only the primary means of providing access, but also other combinations of access dependencies that might be available. A repository manager left with only one access path to content should be much more nervous than one still able to choose from many viable options.
- The purpose of obsolescence risk assessment is to inform decisions about the need to take action. The action required to maintain access is often not trivial, and may involve the risk of changing the content unacceptably. Repository managers can expect to face difficult decisions about whether to accept less than perfect replacements, or to hope for a reverse of obsolescence – which may happen if enough interested stakeholders can generate a market for new means of keeping an otherwise superseded format accessible.

- We do not mean to imply that format obsolescence is always about to overwhelm us. As Chris Rusbridge reminds us, the juggernaut of technological change is sometimes slower in its negative impacts on access than is popularly believed (Rusbridge, 2006). Nevertheless, it is much better to know what you can about the route it will take, as well as its expected time of arrival in your part of town.

Format Obsolescence Risk Assessment Tools

Digital Preservation and Risk Assessment

In the latter part of the 1990s, thinking about threats to ongoing digital access began to coalesce into a number of approaches to identifying risk levels as a guide to planning and setting priorities. At the turn of the 21st century researchers at Cornell University reported on a risk assessment methodology (Lawrence, Kehoe, Rieger, Walters, & Kenney, 2000); in 2003 the NLA itself undertook what at the time was believed to be a comprehensive risk assessment of its digital collections (Dack, 2004); in 2004, OCLC released its INFORM methodology of building risk statements based on the judgments of a community of experts (Stanescu, 2004, 2005); a year later the British Library publicised details of a risk identification methodology applied to its digital collections (Woodyard, 2005).

The authors of the present paper see their work as a development of this tradition, based on a similar appreciation of the importance of providing tools that will help managers recognise, assess, and make decisions about preservation risk factors.

PANIC

An important direct antecedent for the work reported in this paper is the PANIC (Preservation Webservices Architecture for Newmedia, Interactive Collections and Scientific Data) model proposed and explored by Hunter and Choudhury (2004, 2005, 2006). The PANIC model recognises that there are many elements in the process of providing meaningful access to digital materials, and that almost all of them are subject to change. This approach grew out of a perception that it can be difficult for collection and repository managers to keep themselves fully informed of changes that might threaten the accessibility of their collections. Development of PANIC was based on the emergence of three potentially powerful components that could be brought together to help repository managers in their preservation planning:

- Information registries which store useful information about file formats[2];
- Development of preservation action tools (such as migration services, emulation services, etc) that may pre-empt, circumvent or remedy the impacts of these changes[3]; and
- A global information network in which it should be possible to look for relevant indicators of format obsolescence and to bring that information to the attention of collection managers as they consider the need for preservation action.

[2] Such as GDFR, PRONOM, LCSDF, Version Tracker

[3] Such as Typed Object Model (TOM), IBM's UVC Emulation Project and National Archives of Australia's XML Electronic Normalising of Archives (Xena)

Many collecting institutions responsible for managing digital data for long-term accessibility, including the NLA, were excited by the potential of the PANIC model for reducing duplication of effort in managing preservation systems. While format obsolescence was recognised as just one of many risks to be negotiated, it did seem to be one that was both particularly critical and particularly amenable to the kind of approach PANIC was exploring.

AONS

In 2003, the National Library joined three Australian universities and the Australian Partnership for Advanced Computing in forming the Australian Partnership for Sustainable Repositories (APSR)[4], a project funded by the Australian Government's Department of Education Science and Training (DEST) under the Systemic Infrastructure Initiative[5]. APSR partners all shared an interest in exploring the viability of the PANIC model, and on the NLA's initiative, agreed to fund further exploratory work focused on the obsolescence identification and notification element of the PANIC model.

In 2006, NLA, in collaboration with the Australian National University (ANU), built the AONS I prototype[6] (Curtis, 2006; Curtis, Koerbin, Raftos, Berriman, & Hunter, 2007). The AONS I software:

> "is a system [designed] to analyse the digital repositories and determine whether any digital objects contained therein may be in danger of becoming obsolescent. It uses preservation information about file formats and the software which supports these formats to determine if the formats used by the digital objects are in danger" (Curtis, 2006).

In order to determine this, the AONS I system used information obtained from the PRONOM[7] and Library of Congress Sustainability of Digital Formats (LCSDF)[8] registries, which it periodically checked against the contents of the repository. When the repository was found to contain objects in danger of becoming obsolescent, a notification report was sent via email to the repository manager. At the conclusion of the AONS I Project, the software code was supported in a DSpace[9] digital repository environment, and less successfully in a Fez-Fedora repository environment[10]. Experience with the two different repository structures highlighted the need for a repository-agnostic product (Curtis, 2006).

In 2007 the NLA and other APSR partners collaborated in the AONS II software development project, to refine and expand the functionality of the prototype AONS I software[11, 12, 13].

---

[4] Australian Partnership for Sustainable Repositories (APSR) http://www.apsr.edu.au/
[5] Department of Education, Science and Training (DEST), Systemic Infrastructure Initiative http://www.dest.gov.au/sectors/higher_education/programmes_funding/programme_categories/research_related_opportunities/systemic_infrastructure_initiative/
[6] APSR AONS (Automatic Obsolescence Notification System) I http://www.apsr.edu.au/aons/
[7] PRONOM http://www.nationalarchives.gov.uk/pronom/
[8] Library of Congress Sustainability of Digital Formats (LCSDF) http://www.digitalpreservation.gov/formats/
[9] DSpace http://www.dspace.org/
[10] Fedora http://www.fedora.info/
[11] APSR AONS II, Home Page http://www.apsr.edu.au/aons2/
[12] APSR Wiki – AONS II http://pilot.apsr.edu.au/wiki/index.php/AONS
[13] AONS II Development Blog http://aons2dev.blogspot.com/

The AONS II Project has produced an open source, platform-independent, configurable, downloadable tool[14] in prototype form that is capable of providing information from authoritative international registries.

How AONS II Works.

AONS II can be deployed as a part of a workflow or as a stand alone application to:

- AONS II can be deployed as a part of a workflow or as a stand-alone application to:
- Check files some time after they have been ingested, either on a one-off basis or on a regular monitoring schedule.

Like its predecessor, AONS II is intended to work by identifying the file formats found in a digital repository, and seeking information on obsolescence risk indicators by reference to external registries of file format information. Where relevant indicators are detected, the tool generates a notification to a designated person. Unlike its predecessor, AONS II recognises the need to refer to internal information as well, and engages the manager more actively in determining the apparent level of risk based on both external and internal indicators.

Once a risk profile has been established for a particular repository format profile, the software can be configured to look regularly for changes in the targeted indicators, generating an automatic notification that either a new risk assessment should be carried out, or that preservation action may be needed.

Recognising File Formats and Building Collection Profiles.

AONS II builds a profile of the formats within a specified set of files (which can range from a whole repository to a single file). The profile is constructed as an XML metadata summary, which can be sourced from any existing compliant metadata summary, or from a repository crawl using purpose-built AONS adaptors designed for a given repository type (DSpace, Fedora, etc). Crawl results may be obtained from existing repository metadata or automated format recognition tools (such as DROID[15], JHOVE[16]) used to identify the file formats.

This approach differs from other format profiling systems which rely on downloading content files in order to identify them and build a format profile, or which use generic harvesting tools (Hitchcock, Hey, Brody, & Carr, 2007).

Format Identifiers.

A comparison tool like AONS II depends on being able to distinguish accurately between different versions of formats, in order to identify relevant risk levels. Format identification is generally an ambiguous exercise. Files may be labeled with misleading extensions; different sources may refer to the same format under different names. So that it can bring together relevant information from disparate sources, AONS II creates an internal format identifier for each apparent format found, and then tries to map it to the likely matching format identifiers used by external registries.

[14] AONS II download from SourceForge http://sourceforge.net/projects/aons/
[15] Digital Record Object Identification (DROID) http://droid.sourceforge.net/wiki/index.php/Introduction
[16] JSTOR/Harvard Object Validation Environment (JHOVE) http://hul.harvard.edu/jhove/

Based on the repository formats found, AONS II may classify formats as identified, and matched with format information held in external registries, or as unidentified. As part of this process, a repository manager could:

- Decide to link an unidentified format to an existing AONS internal format;
- Create a new internal format with links to external format information;
- Create a new internal format with no links (not a particularly desirable option but a valid use case because a format might not yet be recorded in external registries, given the ever-expanding superset of file formats); or
- Simply leave the format as unidentified.

Once the formats have been established in the repository or collection profile, the AONS II software compares the list of formats and versions with information on formats mapped as equivalents derived from external registries. For efficiency, AONS stores format information from the target registries in local databases. Users can also add other useful links and access them through the Graphical User Interface, without using a local cached copy.

A feature of AONS II is its adaptability. Users can configure it to target authoritative sources of format information as they emerge or are found to be useful. Currently the target registries included are PRONOM and LCDSF. As these registries change over time and as new registries are created and become stable, such as Global Digital Format Registry (GDFR)[17], new adapters can be created with minimal effort. This ability to configure the targeting of registries is considered critical; during the development of this tool it became apparent that there was still no single definitive source of information on file formats.

Adapters.

AONS uses repository/registry adapters which are abstracted from the core software for interfacing to different repository and registry types. This keeps the core code isolated from the adapters so that the basic business logic does not need to be modified when creating or modifying adapters. Potentially anyone with a new repository type can write an appropriate adapter and share it with the user community on SourceForge. Currently the repository adapters which have been written include generic file system, RESTful-pull[18], DSpace version 1.4, Fedora version 2.2, and NLA Pandora[19]. Similarly, registry adapters include PRONOM, and LCSDF.

Notification.

The notification part of AONS II is configurable and based on either a change in state within the system, such as the end of a repository crawl, a change in the information about a format in an external registry; or the expiry of a time-sensitive trigger, such as a format risk re-assessment period ending. Notification can occur in a number of forms: via email, RSS feed, and task boxes via the Graphical User Interface.

Checking for Obsolescence Risk Information.

Critically, AONS II software aims to help in assessing levels of obsolescence risk,

[17] GDFR (Global Digital Format Registry) http://hul.harvard.edu/gdfr/
[18] RESTful Definition, Wikipedia http://en.wikipedia.org/wiki/REST
[19] PANDORA, Australia's Web Archive http://pandora.nla.gov.au/

with a view to informing decisions about the need for preservation action.

An initial business driver for the project was a perceived need for a tool which could use standardised metrics that would support machine-formulation of recommendations on risk levels. This approach presupposed access to relevant authoritative and machine-usable information about a wide range of file formats, including information that might offer warnings about format obsolescence risks.

Behind this was an assumption about the state of development of format registries that might offer warnings about format obsolescence risks. Development of the project has involved close study of the information offered by known target registries, and their likely ability to support automated format risk judgments.

It became apparent that in the short-term – certainly within the funding life of the AONS II Project – the intended international target registries would present some problems in the support they could offer to a tool like AONS. This certainly applied to any expectation of support for a risk metrics approach. However, there were, at the time of the AONS II Project, a number of other constraints in using the targeted format information registries:

- Information is typically limited to common formats, and is unlikely to include reference to obscure formats. This constrains their usefulness as sources of information about less common formats, but also their usefulness in supporting automatic identification tools such as DROID, JHOVE;
- Information is often limited to notional suggestions or recommendations without supporting reasons or evidence;
- There is no commonly agreed vocabulary;
- For some formats, fields are not filled in; the format has been included as a place holder for later entry;
- Occasional collisions in format identifiers occur between registries.

Even when data are available, they are not sufficiently structured to be useful in a system-automated context without considerable human input to make the content understandable.

Given that the target registries were not designed with tools like AONS in mind, it is not surprising that there are some frustrations in seeking to achieve automatically derived risk metrics or even consistent, machine-usable information from them. However, it would be pleasing to see file format registries interested in automated obsolescence notification as a critical use case.

As discussed earlier, we have come to recognise that the information from a format registry can only ever provide partial guidance on obsolescence risks. Repository managers must also take account of the local circumstances upon which access depends: A format vendor's support end-date is only a partial guide to whether a repository manager will have to manage short- or medium-term loss of access, or even whether access has already been lost locally.

Discussion of the Assessment Questions

At the current state of development, the project offers a software tool which is designed to work in multiple deployment modes, is open source, Java-based, reusable, adaptable and extensible. However, the rule set on which we believe risk assessments should be based (a core part of the AONS process) has not been automated. Instead, the rule set has been manifested, at the current stage of the project, as a set of questions (see Appendix A). On a practical level, the project needs community feedback about the usefulness and appropriateness of the questions before hard coding workflows metrics into the software.

Devising this rule set has been a major piece of work, and it has resolved itself into a series of questions we believe provide an effective basis for judging the level of obsolescence risk for a given file format at a particular time.

The risk assessment questions seek answers that will indicate the likely stage of obsolescence for a file format in a specific real world repository. As a consequence of having to cater for potentially thousands of possible file formats, the questions need to be generic and somewhat simplistic. However, the questions still aim to allow a repository owner to build specific risk profiles of an individual file format.

The risk questions are classified into two general groups: *Community environment* (which should be answerable by reference to the digital preservation community) and *My repository environment* (which relate specifically to an individual environment and depend on the sustained availability of combinations of software and hardware required as view-paths).

At a community level, the questions assume certain generic information might serve as useful indicators:

- The current level of support for rendering the format, indicated by the tools available and the technical support offered by vendors, creators or others in the preservation community.
- How long it has been since the format version was first released. This is based on an assumption that the software industry has cycles of change that affect both the release of new format versions and the development of tools to render formats.
- How many versions have been released since that time. This is based on an assumption that the greater the gap between the format version held by a repository and currently released versions, the more likely it is that the backwards compatibility of current rendering tools will not help.
- The range of view-paths that could be used for acceptable presentation of content.

It is recognised that these factors may not always serve as good indicators.

At a local repository level, the questions assume that it is possible for a repository manager to determine whether required view-paths for access are locally available and workable.

Other issues where subjective judgments may be needed include:

- Decisions about how much notice is needed in order to take manageable action.
- The degree of rendering difficulty that the repository owner and users are willing to bear.
- The degree of loss that is acceptable.
- What constitutes a "base format" unlikely to require repeated assessment (because it can be expected to be readable in all expected computing environments).
- Whether there may be other sources of information worth checking for indications of a looming accessibility problem.

Discussion

Some interesting points have already arisen in trying to apply the questions:

1. The basic risk point the approach tries to identify is the need for a decision to take preservation action, in order to regain or maintain access. Once access is lost, it usually remains lost until some action is taken. The action could be as trivial as finding a piece of downloadable software via a Google search, or as significant as designing transformation software or developing an emulator.
2. It seems reasonable to assume a file format is heading for obsolescence when a large part of the community of users cannot access it, or have decided to move content away from it.
3. As Paul Wheatley[20] of the British Library has suggested, it may be useful to think of "a scale of obsolescence that begins with inconvenience to users and ends in the digital black hole of loss".
4. It may be necessary to consider granular differences that may not be evident from a generic perspective. For example, TIFF, which may be identified as the file format of a range of subtly different file types created using different software products.
5. Another aspect of the above point is the issue of compendium (or container) files that may include a number of different file formats within them. Besides the obvious issue of format identification, risk assessment identification and action could be made more problematic. In the case of compressed files, the method used to uncover this information should not vary from the method of dealing with *normal* files.
6. Open source renderers are a good thing, but they may not obviate the need to take preservation action. Instead, their main benefit may be in making action easier to take. The consequences of obsolescence may be less severe because the format is in an open and documented form – so long as the open specification is maintained and accessible.
7. *My repository environment* may be quite idiosyncratic. For example, a repository might have data stored on physical carriers such as 5½" Floppy Disks, formatted in the Burroughs B20 Format, created circa 1982. Presumably this would be a problem format for most of the digital preservation community. However, for an individual repository manager who has the view path required to recover the data from 5½" Floppy disks (if the disks are readable), as well as the means of accurately reading that

[20] Paul Wheatley, personal communications (2007)

data (or an alternative such as InterMedia media data conversion software), this might be less of a problem. Be this as it may, in the above example it would probably be prudent for this manager to recover the data from the physical carrier and maybe consider taking some further action regarding the format before the means to access it becomes lost.

Possible Futures

Over time, we hope that feedback and use will help to refine the questions we have proposed as ways of finding useful obsolescence indicators. While our own institution is committed to trialling and developing this approach to preservation planning for its digital collections, we are sure that the approach and its questions would benefit from the insights and experiences of others.

The search for obsolescence risk indicators has stimulated thinking about how such an approach can be made to work efficiently. At present, some of the questions will be very hard to answer, but over time (if indeed they are the right questions), they can help in focusing discussion on the true nature of the risks. They can also provide a spur to the building or adapting of better tools to find the information needed to answer them.

Ultimately, it will be as a preservation community that such an approach will be developed to its optimum utility. We will all need tools to make assessments, but we will also need each other.

We believe it would be a welcome development to be able to share the results of local risk assessments from tools like AONS through something like a central web service. Such a service could generate federated risk metrics based on members of an active community sharing their individual risk findings. This would allow relevant global registries to draw on the experiences and expertise of the contributing preservation community and add considerably to their usefulness.

Because file format obsolescence affects virtually all long-term digital repositories to some degree, there appears to be great potential for creating a community which includes repositories, registries, software tool developers, format developers, and other end-users and stakeholders. A similar community model and approach would also benefit other digital management needs such as file format recognition.

There are, of course, many precedents for sharing information that may alert others to danger, creating virtual communities through information that addresses common needs over distance and time. Seeking a course through the murkiness of file format obsolescence has reminded us of ancient (and not so ancient) mariners who reportedly annotated the uncharted parts of their maps with the words: “Here be dragons.” [21]

[21] Here be Dragons, Wikipedia definition http://en.wikipedia.org/wiki/Here_be_dragons

Acknowledgements

The authors wish to thank APSR and DEST for supporting this project. At the NLA the authors would like to recognise the assistance given by Gerard Clifton, Douglas Elford and Nicholas del Pozo on the definition of file format obsolescence, and David Levy and Matthew Walker for their work on AONS II. We are also indebted to Dr Jane Hunter for the PANIC model and for providing project assurance. We would also like to thank our colleagues at the National Archives of Australia's Digital Preservation Section, ANU Division of Information's Digital Resources Service, and The Library Technology Service at the University of Queensland Library for providing a sounding board on our Risky Journey. A special thanks to Paul Wheatley from the British Library for his thoughts on the topic.

References

Curtis, J. (2006). *AONS system documentation revision 169*. 2006-09-29. Canberra, Australia: Australian National University.

Curtis, J., Koerbin, P., Raftos, P., Berriman, D., & Hunter, J. (2007). AONS – An obsolescence detection and notification service for web archives and digital repositories. Special issue on Web Archiving, *New Review on Hypermedia and Multimedia, (JNRHM)*. Retrieved July 3, 2008, from http://www.informaworld.com/smpp/content~content=a780448483~db=all~order=page

Dack, D. (2004). *An assessment of the risks to the National Library of Australia's digital collections*. Unpublished. Canberra, Australia: National Library of Australia.

Heslop, H., Davis, S., & Wilson, A. (2002). *An approach to the preservation of digital records*. Green paper. Canberra, Australia: National Archives of Australia. Retrieved July 2, 2008, from http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm2-888.pdf

Hitchcock, S., Hey, J., Brody, T. and Carr, L. (2007). *Laying the foundations for repository preservation services – Final report from the PRESERV project*. University of Southampton, UK.

Hunter, J., & Choudhury, S. (2004). A semi-automated digital preservation system based on semantic web services. In *Joint Conference on Digital Libraries, JCDL, 2004*, pp. 269-278. Tucson, AZ.

Hunter, J., & Choudhury, S. (2005). Semi-automated preservation and archival of scientific data using semantic grid services. In *Proceedings of the Fifth IEEE International Symposium on Cluster Computing and the Grid (CCGrid'05). vol. 1,* pp. 160--167. Cardiff, UK.

Hunter, J., & Choudhury, S. (2006). PANIC: An integrated approach to the preservation of composite digital objects using semantic web services. *International Journal on Digital Libraries, Special Issue on Complex Digital Objects. vol. 6, no. 2*, pp. 174-183.

Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., Walters, W. H., & Kenney, A. R. (2000). *Risk management of digital information: A file format investigation*. Washington, D.C.: Council on Library and Information Resources.

Rusbridge, C. (2006, February). Excuse me… Some digital preservation fallacies? *Ariadne, 46*. Retrieved July 2, 2008, from http://www.ariadne.ac.uk/issue46/rusbridge/

Sierman, B. (2007). *Enhancing our data model with PREMIS*. The Hague, Netherlands: Koninklijke Bibliotheek.

Stanescu, A. (2004, November). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *D-Lib Magazine, vol. 10*, *(11*). Retrieved July 2, 2008, from http://www.dlib.org/dlib/november04/stanescu/11stanescu.html

Stanescu, A. (2005). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *OCLC Systems & Services: International Digital Library Perspectives, vol. 21 no. 1*, pp. 61 - 81.

Woodyard, D. (2005). Risk analysis of digital library material. In *Society for Imaging Science and Technology, Archiving Conference*, pp. 215-221. Washington D.C.: IS&T.

Appendix A

File Format Obsolescence Risk Decision Support System - Version 1.1 (released November 2007)

Note: The questions below have been formatted for use in the AONS II software. Within the software, each of the questions has a more detailed help text. However, due to space limitations, this help text cannot be included in this paper. Please contact the authors for more details[22].

Step. 1: **Community Environment**	**Step 2:** **My Repository Environment**
Q1. Is this a base format? (a ubiquitous format which is likely to be rendered by most applications; e.g. EBCDIC, ASCII, Unicode) If yes, consider the format low risk and go to the end of Step 2. If unknown, state "Unknown".	Q1. The original primary rendering software has been identified as... (see Step 1 - Q6.). Is this primary rendering software available to you?
Q2. Is this file format and version referenced in any searched information resources?	Q2. The following hardware and software dependencies have been identified for effective rendering of this format using the original software… (see Step 1 - Q7). Are these critical dependencies available to you?
Q3. Is there a known support end date for this format version? If yes, how many years to that support end date?	Q3. The following alternative software options have been identified for safe and effective rendering… (see Step 1 - Q8). How many of these alternative rendering options are available to you?
Q4. How many years since this version was released?	Q4. For the alternative rendering options, the following critical dependencies have been identified… (see Step 1 - Q9). Are these critical dependencies available to you?

[22] See Reading Tools column, right, on IJDC platform: "For this peer-reviewed article: See the authors' bio"

Q5. How many new versions have been released since then?	Q5. Do you have any other means of providing safe and effective access? (e.g. custom designed applications, scripts, emulators). What are they?
Q6. Is the primary rendering software for this format version identified? What is it? If unknown, state "Unknown".	Q6. Overall, how many access options are effectively available to you (i.e. how many can you make work), including the original rendering software? If none – consider access lost. If one, consider high risk.
Q7. Are there critical hardware and software dependencies for effective use of the original rendering software? What are they? If unknown, state "Unknown".	Q7. Do you have any other information that would exacerbate or mitigate the level of technical obsolescence risk? (i.e. information which might indicate a change in access to this format).
Q8. How many alternative software options for safe and effective rendering can be identified? What are they? If unknown, state "Unknown".	
Q9. For each alternative, are there critical hardware and software dependencies for effective use of the alternative rendering software? What are they? If unknown, state "Unknown".	