AONS II: continuing the trend towards preservation software 'Nirvana'

David Pearson
Web Archiving and Digital Preservation Branch, National Library of Australia,
Canberra ACT 2600, Australia.
dapearso@nla.gov.au

Abstract. File format obsolescence is a major risk factor threatening the sustainability of and access to digital information. While the preservation community has become increasingly interested in tools for migration and transformation of file formats, the National Library of Australia is developing mechanisms specifically focused on monitoring and assessing the risks of file format obsolescence. This paper reports on the AONS II project, undertaken by the National Library of Australia (NLA) in conjunction with the Australian Partnership for Sustainable Repositories (APSR). The project aimed to develop a software tool which allows users to automatically monitor the status of file formats in their repositories, make risk assessments based on a core set of obsolescence risk questions, and receive notifications when file format risks change or other related events occur. This paper calls for the preservation community to develop a co-operating file format obsolescence community which includes registries, software tool creators and end users to effectively curate digital content in order to maintain long-term access.

1. Introduction

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 volume of digital materials, the plethora of file formats, the dynamic nature of computing environments, and the unremitting but often unpredictable drivers that cause formats to become obsolete. The high business value of much digital information requires that access be maintained for extended periods of time. In order to ensure the long-term availability and usefulness of digital materials, repository managers need help in managing format obsolescence risks.

More than two decades into a recognisable discipline we call 'digital preservation', we are still far more advanced in creating digital information resources than we are in taking concrete action to preserve them. There are at least two reasons for this.

Firstly, the juggernaut of technological change has been somewhat slower than expected in running down meaningful access - change has happened with all its predicted vigour, but vendors have come to recognise that there is business value in maintaining some level of format compatibility at least over the short-term.

Secondly, much of the thinking and talking about digital preservation has tended to focus on high level issues, avoiding more concrete confrontation with what might be needed to make decisions and to take real action in order to maintain meaningful access.

To help address some of these issues, the National Library of Australia (NLA) [1] and the Australian Partnership for Sustainable Repositories (APSR) [2] have been collaborating on a software development project called AONS II (Automatic Obsolescence Notification System, version 2). The NLA's role in this endeavour has been to produce an open source, platform-independent, configurable and downloadable tool that allows users to automatically monitor the status of file formats in their repositories, make risk assessments based on a core set of obsolescence risk questions, and receive notifications when file format risks change or other related events occur.

The need for a tool like AONS to aid repository managers to monitor file formats is apparent. However, 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 undertaken. It is certainly true that repositories would minimise 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 reality that many repositories will continue to deal with file formats affected by technological change.

This paper outlines the AONS II project, and discusses some of the architectural and design principles and possible future development paths.

2. AONS II - Antecedents

2.1 PANIC

The most important direct antecedent for AONS II is the PANIC (Preservation Webservices Architecture for Newmedia, Interactive Collections and Scientific Data) model proposed and explored by Hunter and Choudhury [3] [4] [5]. This 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. The 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 for the benefit of repository managers in their preservation planning:

- Information registries which store useful information about file formats[1];
- Preservation action tools (such as migration services, emulation services, etc) that may pre-empt, circumvent or remedy the impacts of these changes[2]; and

[1] such as GDFR, PRONOM, LCSDF, Version Tracker

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

- A global information network in which it should be possible to look for relevant indicators of file format obsolescence, and to promptly bring that information to the attention of collection managers so that they might make informed decisions about the need for preservation action. The same network could also allow them to look for and access preservation tools and services to address their needs remotely.

The PANIC model was explored by Dr Hunter and her colleagues, who prototyped an environment in which it would work.

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 recognized 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.

2.2 AONS I

In 2003, the NLA joined with three Australian universities and the Australian Partnership for Advanced Computing to form APSR, a project funded by the Australian Government's Department of Education Science and Training (DEST) under the Systemic Infrastructure Initiative [6]. 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 an "obsolescence identification and notification" element of the PANIC model.

In 2006, the NLA in collaboration with the Australian National University (ANU), the software developer, built the AONS I prototype [7] [8] [9]. 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" [8].

In order to determine this, AONS I used information obtained from the PRONOM [10] and Library of Congress Sustainability of Digital Formats (LCSDF) [11] registries, which it periodically checked against the contents of the repository. When the repository was found to contain objects in danger of becoming obsolete, 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 [12] digital repository environment at the ANU. Similarly, there was a largely successful attempt to make AONS work in a Fez-Fedora repository environment at the University of Queensland [13]. However, the two different repository structures highlighted the need for a repository-agnostic product [8].

At the end of the project the AONS I software could be characterized as follows: it was Java based, had a command-line interface, monolithic architecture, and limited retention of state between invocations (i.e. it had no application memory so it did not build on previous results). It was also non-interactive, offering no repository owner workflows. Similarly, in the prototype that was built, if a format was unidentified the application had no way of dealing with it. Risk identification was based on designated preferred formats in a registry. The prototype also illustrated that a usable Graphical User Interface (GUI) and notification mechanisms other than email could be useful.

3. AONS II

The NLA wished to see further development of the AONS tool to test and, if necessary, refine its underlying assumptions so that the methodology could reach its maximum potential as a preservation enabler. In 2007 the NLA and other APSR partners collaborated in the AONS II software development project. This project refined and expanded the functionality of the prototype AONS I software [14] [15] [16].

A number of fundamental principles evolved from the development of AONS I. The AONS II software product was required to:

- Support three different business environments: a national federated infrastructure, enterprise business models, and individual standalone repository sites;
- Be open source using Java code;
- Be modular and have a reusable/adaptable design;
- Be platform independent using a decoupled approach;
- Be interoperable, using common interfaces, protocols and standards;
- Provide service interfaces in a Service-Oriented Architecture based on a RESTful approach (a lightweight methodology for Web Services [17]);
- Provide a core set of functionality, which abstracts repositories and registries functionality away from the core, and would allow new repository and registry adapters must be able to be added without effecting the core; and
- Be demonstrable.

These principles have provided a yardstick and reality check for all development work. In line with the above scope and design principles, AONS II is a workable product available for download from SourceForge [18].

How AONS II works

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

- Check files as they are ingested; or
- Check files some time after they have been ingested, either on a one-off basis or on a regular monitoring schedule.

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

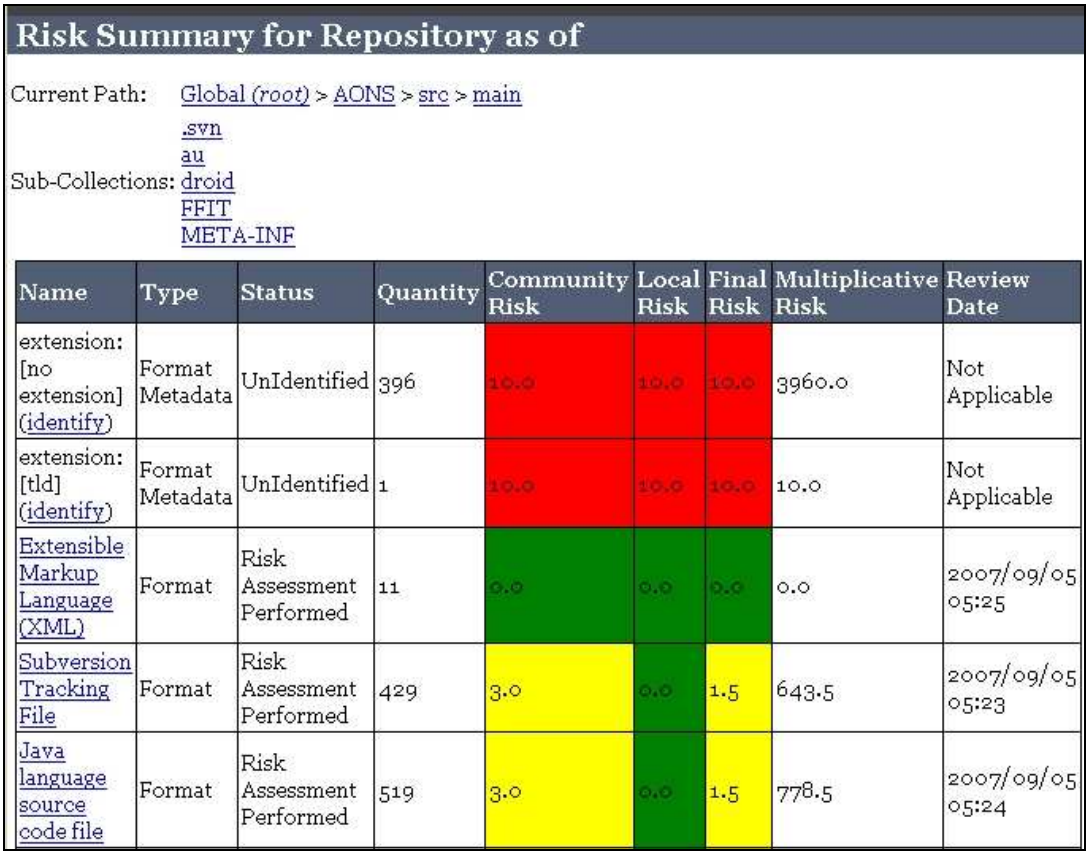
Recognising File Formats and Building Collection Profiles

AONS II builds a profile of the formats in a repository or a subset such as a collection or even a single file. The profile is constructed from 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 are processed using automated format recognition tools (such as DROID [19], JHOVE [20]) to attempt to determine 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 [21].

Format Identifiers

It will be apparent that a comparison tool like AONS II depends on being able to distinguish accurately between different formats, and between different versions of formats, in order to identify relevant risk levels. Format identification is not necessarily an unambiguous 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 a internal format identifier for each apparent format found, and then tries to map it to the likely matching format identifiers used by external registries (Fig. 1).



Risk Summary for Repository as of

Current Path: Global *(root)* > AONS > src > main

Sub-Collections: .svn, au, droid, FFIT, META-INF

Name	Type	Status	Quantity	Community Risk	Local Risk	Final Risk	Multiplicative Risk	Review Date
extension: [no extension] (identify)	Format Metadata	UnIdentified	396	10.0	10.0	10.0	3960.0	Not Applicable
extension: [tld] (identify)	Format Metadata	UnIdentified	1	10.0	10.0	10.0	10.0	Not Applicable
Extensible Markup Language (XML)	Format	Risk Assessment Performed	11	0.0	0.0	0.0	0.0	2007/09/05 05:25
Subversion Tracking File	Format	Risk Assessment Performed	429	3.0	0.0	1.5	643.5	2007/09/05 05:23
Java language source code file	Format	Risk Assessment Performed	519	3.0	0.0	1.5	778.5	2007/09/05 05:24

Fig. 1 Screen shot of the GUI format summary screen. This screen summarises the status of all file formats found after a repository crawl. It shows whether a format has been identified or not, its details, quantity, risk level and review date (*AONS II Beta 2007-09-05*).

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 classification 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 core software compares the list of formats and versions with information derived from external registries on formats mapped as equivalents. For efficiency purposes, AONS II stores format information from the target registries in local databases. Unlike the AONS I tool, the current software keeps the locally stored registry information from each target registry separate, so that it can be updated, synchronised, replaced or complemented by information from new sources without disrupting the entire database. Users can also add other useful links and access them through the GUI, 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 external target registries include LCSDF and PRONOM. As these registries change over time and as new registries are created and become stable, such as Global Digital Format Registry (GDFR) [22], 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 II 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 (Fig. 2). Having a decoupled approach which uses a new adapter for any new implementation has proven to be very successful in the open community. 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, DSpace version 1.4, Fedora version 2.2, and NLA Pandora. Similarly, registry adapters include LCSDF and PRONOM.

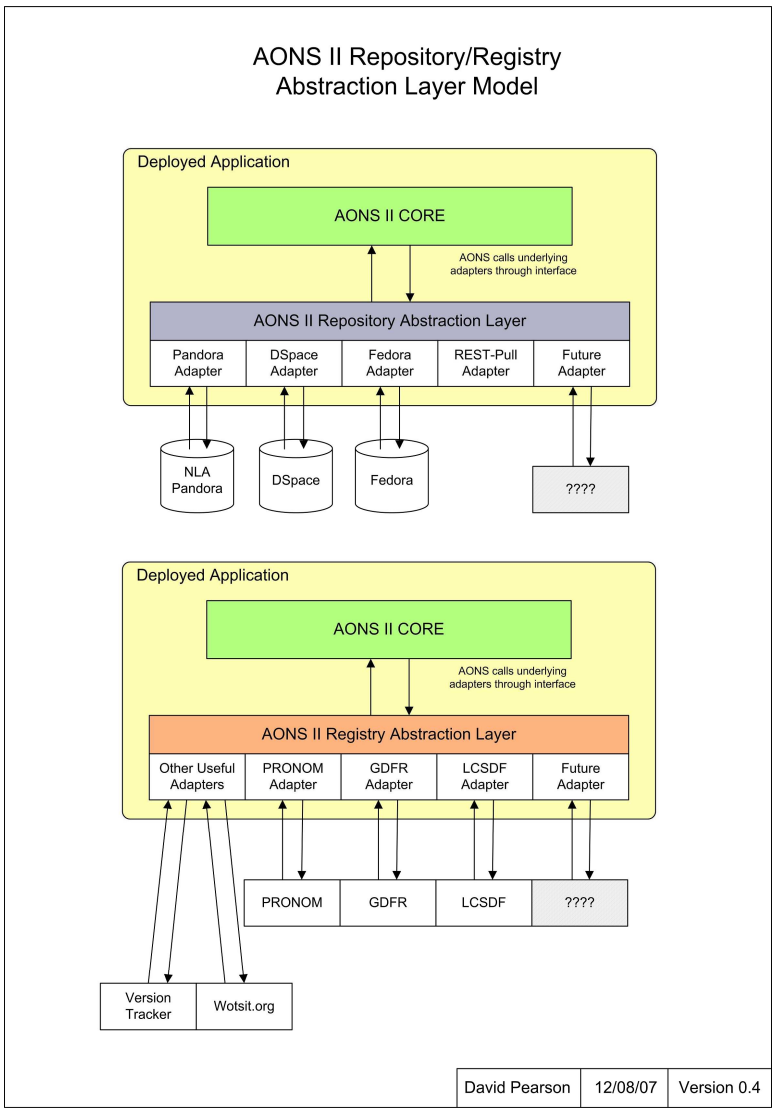


Fig. 2 Diagram showing the Repository/Registry Abstraction Layer Model. This diagram illustrates that the AONS II core code is separated from the various adapters so that the basic business logic does not need to be modified when creating or modifying adapters (*the author*).

Notification

The notification part of AONS II is configurable and based on change in state within the system. Examples of these changes in state are: end of a repository crawl; 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 GUI (Fig. 3).

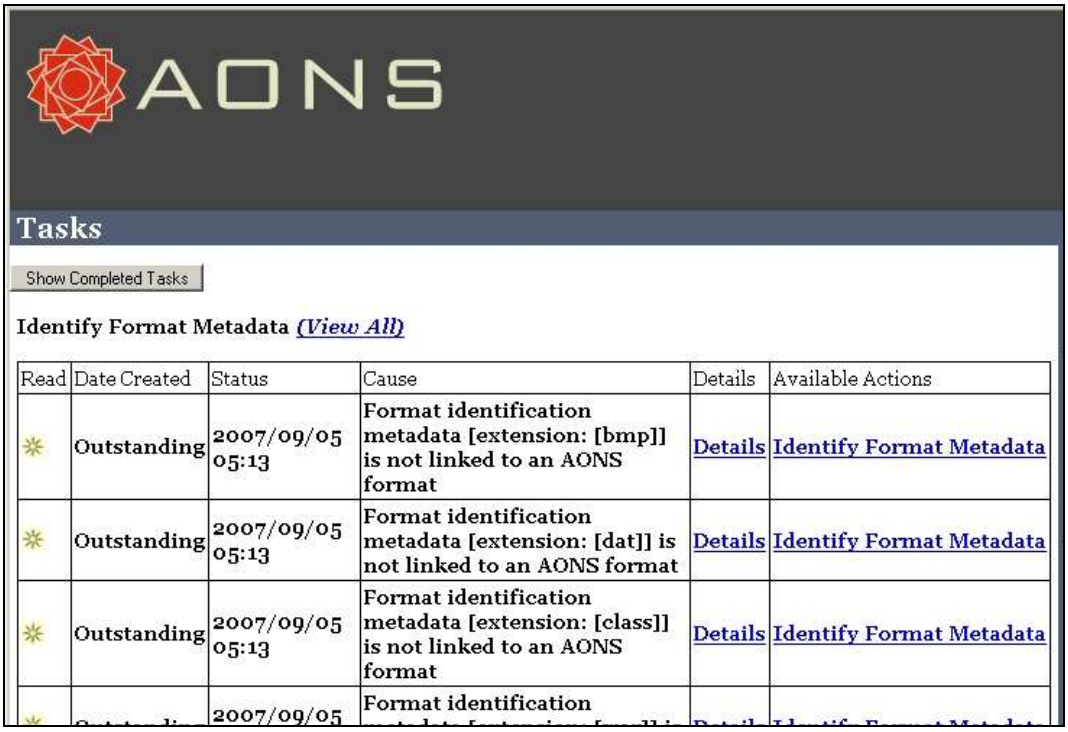


Fig. 3 Screen shot of the GUI notification screen. This notification view is informing the repository manager that there are a number of outstanding tasks; in this case a number of file formats need to be linked to an internal format identifier for each format found (*AONS II Beta 2007-09-05*).

Checking for Obsolescence Risk Information

Critically, AONS II software aims to help in assessing levels of obsolescence risk, with a view to informing decisions about the need for preservation action. That aim remains; however, it has been necessary to modify its interpretation in light of experience in developing the tool beyond its first prototype stages.

An initial business driver for the project was a perceived need for a tool which could automate much of the assessment process, using standardized 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 they might offer warnings about format obsolescence risks. Development of the project has involved close study of the information that known target registries offer, 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 not provide any format obsolescence risk metrics. One of them, PRONOM, has been declared by its owner institution, The National Archives (UK), to have a relevant long-term intention:

> "TNA intends to develop a holistic risk assessment methodology for electronic records that will enable us to identify risk factors at an early stage, predict their impact, and plan appropriate mitigation strategies" [23].

This functionality was not available during the 2007 development cycle.

Similarly, the current registries have not evolved to the stage where they are a good fit-for-purpose for a tool like AONS II. The data is not sufficiently structured to be useful in a system-automated context without considerable human intervention. Human intelligence is required to understand the content, and often little or no information is available.

Given that the target registries were not designed with tools like AONS II in mind, it is not surprising that there are some frustrations in automatically deriving 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.

Therefore, the AONS II project involved deriving a series of questions which it is believed provides an effective basis for judging the level of obsolescence risk for a file format at a particular time. At this time, the rule set has not been automated. As a consequence of having to cater for potentially thousands of possible file formats, the questions have to be generic and somewhat simplistic. However, the questions aim to allow a repository owner to build a risk profile of an individual file format. At this stage they are a series of questions with corresponding free-text entry fields (Fig. 4). Information from PRONOM, LCSDF as well as any other user-defined web sites can be made available for the operator to help answer these questions. At the completion of the assessment, based on the answers to the series of questions, the operator assigns a subjective risk level to each format. The results of all the format risk assessments are presented in the main format summary screen of the application. On a practical basis, there was a decision to wait on community feedback about the usefulness/appropriateness of the questions before hard coding workflows metrics into the software.

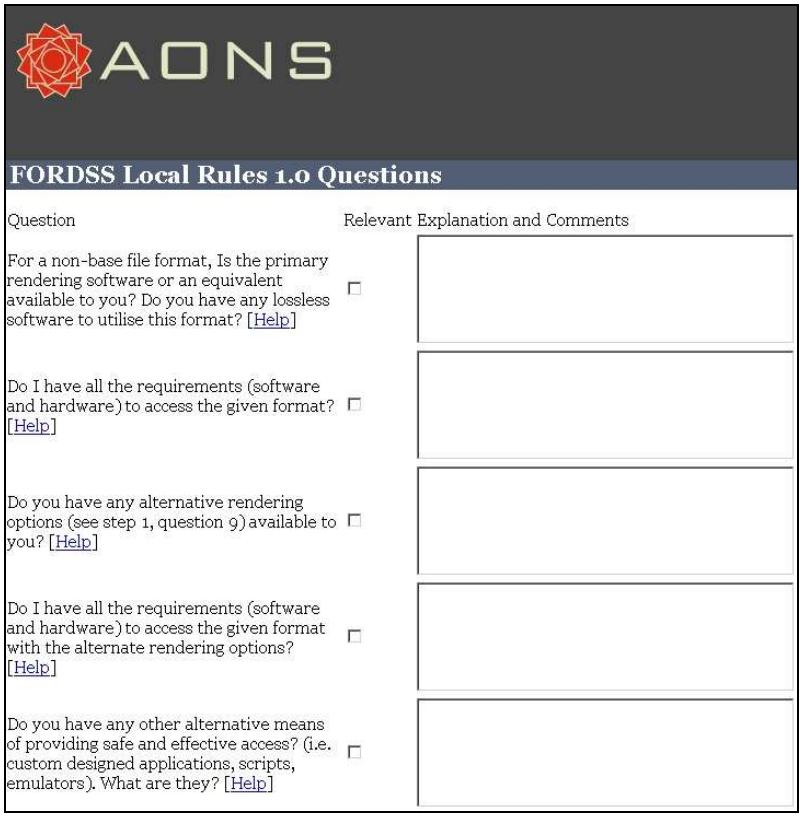


Fig. 4 Screen shot of the GUI file format risk assessment screen – Step 2 local risk assessment. Step 1 assesses community risk, while Step 2 assesses local risk. These questions should be answered for each file format in order to obtain a meaningful subjective risk metric (*AONS II Beta 2007-09-05*).

4. Future

The goal of a preservation manager is to sustainably preserve, manage and provide access to digital material as long as the business needs dictate. There are currently many open-source and proprietary tools which perform a single function towards this goal. However, AONS II has been purpose-built to manage the overall process of the identification of file format and associated risk. It builds upon the many other preservation community tools whether they are format identification tools, registries or useful websites, and attempts to obtain maximum value from them. These tools can potentially be added or subtracted with minimal effort. AONS II has also been designed in a modular fashion so that it, or parts of it, can be re-used in other preservation tools. Using information provided by AONS II could also be the enabler for many other preservation services.

Ultimately, it would seem a positive development to be able to share the results of risk assessments from individual AONS II instances with a central web service. Such a service could provide both a machine- and human-readable federated risk metric, based on an active community exporting individual risk metrics to this central service, and providing some form of voting system. Users could ascertain the community-derived metrics and the level of mitigation within individual repositories. This model would allow hosted registries to draw on the experiences and expertise of the contributing digital preservation community. As well as aiding risk assessment, a similar service could be hosted for file format recognition (e.g. a digital fingerprint) which could also improve the effectiveness of format recognition tools.

Only when there is a co-operating file format obsolescence community which includes registries, software tool creators and end users can we, as members of this community, hope to be able to effectively manage digital content in our care. It is the intention that this tool will continue to be developed based on community need, and thus become sustainable within our community. Until then, we believe AONS II is a step in the right direction, toward preservation software 'Nirvana'.

Acknowledgments. The author wishes to thank APSR and DEST for supporting this project. At the NLA the author would like to recognise the assistance given by Colin Webb, Gerard Clifton, David Levy and Matthew Walker for their work on AONS II. I am also indebted to Dr Jane Hunter for the PANIC model and for providing project assurance. I would also like to thank colleagues at the ANU Division of Information's Digital Resources Service, The Library Technology Service at University of Queensland Library and Sydney eScholarship of the University of Sydney Library for providing input and advice.

References

1. NLA (National Library of Australia), http://www.nla.gov.au/
2. APSR (Australian Partnership for Sustainable Repositories), http://www.apsr.edu.au/
3. Hunter, J., Choudhury, S.: A Semi-Automated Digital Preservation System based on Semantic Web Services. In: Joint Conference on Digital Libraries, JCDL. 2004, pp.269--278. Tucson, Arizona (2004)
4. Hunter, J., Choudhury, S.: 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 (2005)
5. Hunter, J., Choudhury, S.: PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. In: International Journal on Digital Libraries, Special Issue on Complex Digital Objects. vol.6, no.2, 174--183 (2006)
6. DEST (Department of Education, Science and Training), Systemic Infrastructure Initiative, http://www.dest.gov.au/sectors/higher_education/programmes_funding/programme_categories/research_related_opportunities/systemic_infrastructure_initiative/
7. APSR AONS I - Automatic Obsolescence Notification System I, http://www.apsr.edu.au/aons/
8. Curtis, J.: AONS System Documentation. Revision 169 2006-09-29. Australian National University, Canberra (2006)
9. Curtis, J., Koerbin, P., Raftos, P., Berriman, D., Hunter, J.: AONS – An Obsolescence Detection and Notification Service for Web Archives and Digital Repositories. In: Special issue on Web Archiving for the New Review on Hypermedia and Multimedia, JNRHM (accepted – in press)
10. PRONOM, http://www.nationalarchives.gov.uk/pronom/#
11. LCSDF (Library of Congress Sustainability of Digital Formats), http://www.digitalpreservation.gov/formats/
12. DSpace, http://www.dspace.org/
13. Fedora, http://www.fedora.info/
14. APSR AONS II - Home Page, http://www.apsr.edu.au/aons2/index.htm
15. APSR AONS II - Wiki, http://pilot.apsr.edu.au/wiki/index.php/AONS
16. AONS II - Development Blog, http://aons2dev.blogspot.com/
17. RESTful Definition, http://en.wikipedia.org/wiki/REST
18. AONS II – SourceForge Download, http://sourceforge.net/projects/aons/
19. DROID (Digital Record Object Identification), http://droid.sourceforge.net/wiki/index.php/Introduction
20. JHOVE (JSTOR/Harvard Object Validation Environment), http://hul.harvard.edu/jhove/
21. Hitchcock, S., Hey, J., Brody, T., Carr, L.: Laying the Foundations for Repository Preservation Services – Final Report from the PRESERV project. University of Southampton, Southampton (2007)
22. GDFR (Global Digital Format Registry), http://hul.harvard.edu/gdfr/
23. PRONOM, Future Development, http://www.nationalarchives.gov.uk/aboutapps/pronom/#future