The International Journal of Digital Curation
Volume 7, Issue 1 | 2012

# Managing Risks in the Preservation of Research Data with Preservation Networks

Esther Conway, Brian Matthews, David Giaretta, Simon Lambert and Michael Wilson,

Science and Technology Facilities Council

Nick Draper,

Tessella

## Abstract

Network modelling provides a framework for the systematic analysis of needs and options for preservation. A number of general strategies can be identified, characterised and applied to many situations; these strategies may be combined to produce robust preservation solutions tailored to the needs of the community and responsive to their environment. This paper provides an overview of this approach. We describe the components of a Preservation Network Model and go on to show how it may be used to plan preservation actions according to the requirements of the particular situation using illustrative examples from scientific archives.

 doi:10.2218/ijdc.v7i1.210

# Introduction

This paper outlines an OAIS (CCSDS, 2002) compliant approach to the curation, description, management and preservation of scientific research assets. We discuss how, after undergoing preservation analysis (Giaretta, 2011), it is possible to describe the resultant solution within a Preservation Network Model. We developed the concepts in this paper while considering preservation scenarios within the Science and Technology Facilities Council[1], including the ISIS facility[2], Solar Terrestrial Physics[3] and the British Atmospheric Data Centre[4] archives. The ISIS facility is a pulsed neutron and muon source at the Rutherford Appleton Laboratory in Oxfordshire, which has an archive of raw data dating back 25 years. The Ionosonde data holdings from the solar terrestrial data archive contains atmospheric ionization records from a global network of stations, some of which date back to the 1920's, while the Mesophere-Stratosphere-Troposphere (MST) Radar archive held by the British Atmospheric Data Centres contains over ten years of environmental data from a modern wind profiling instrument based in Wales.

This paper goes on to characterise and describe the main types of preservation action available to an archivist. We describe the effect each of these types of action has upon the network, risks accepted, modes of stabilisation, cost and benefits. We explore how more than one type of strategy can be employed as alternates in order to create the required balance of risk and usability in a preservation solution. We then examine positive feedback relationships between the strategy types in an evolving research asset. In conclusion, we will appraise how such a model will permit the correct configuration and evaluation of a solution in practice in the context of the ENSURE project[5].

# Preservation Networks

Preservation Network Models (PNMs) were originally developed within the CASPAR[6] project and are described in detail by Conway et al. (2009). A PNM is a formal model for conceptualising the relationships between resources within the scenario of a preservation objective. The preservation network model consists of two components: the digital objects and the relationships between them.

## Objects

Objects are uniquely identified digital entities capable of an independent existence which possess the following attributes:

[1] Science and Technology Facilities Council: www.stfc.ac.uk

[2] ISIS: www.isis.stfc.ac.uk

[3] UK Solar System Data Centre: www.ukssdc.ac.uk

[4] British Atmospheric Data Centre: http://badc.nerc.ac.uk/home/index.html

[5] ENSURE project: http://ensure-fp7-plone.fe.up.pt/site

[6] CASPAR project: www.casparpreserves.eu

- **Information:** a description of the key information contained by the digital object. This information should have been identified during preservation analysis as being the information required to satisfy the preservation objective for the designated user community.
- **Location information:** the information required by the end user to physically locate and retrieve the object. Archival Information Packages (AIPs) may be logical in construction with key digital objects being distributed and managed within different information systems. This tends to be the case when data is in active use with resources moving in a dynamic environment. Note the difference between being able to locate and retrieve information, and having control of it.
- **Physical state:** a description of the form of the digital object. It should contain sufficient information relating to the version, variant, instance or format.

A PNM terminates when a user requires no additional information or assistance to achieve the defined preservation objective, given that the accepted risks will not be imminently realised.

### Relationship

A relationship captures how two objects are related to one another in order to fulfill the specified preservation objective whilst being utilized by a member of the designated user community. Relationships can possess the following attributes:

- **Function:** In order to satisfy the preservation objective a digital object will perform a specific function, such as the delivery of textual information or the extraction and graphical visualisation of specific parameters.
- **Risks and Dependencies:** Most digital solutions will have inherent risks and a finite lifespan. Risks could include the interpretability of information, technical dependencies or loss of designated community skills. Risks should be recorded against the appropriate object so they can be monitored and the implication of them being realised assessed.
- **Tolerance:** Not every function is critical for the fulfilment of the preservation objective. Some objects are included as they enhance the quality of the solution or ease of use. Loss of this function is denoted in the model as a tolerance.
- **Quality Assurance and Testing:** The ability of an object to perform the specified function may have been subjected to quality assurance and testing, which may be recorded against the relationship.

Relationships can be composed into Alternate and Composite relationships, which can be thought of as logical “Or” (denoted in diagrams by diamonds) or “And” (denoted in diagrams by circles with relationships emanating from target) relationships. The former is interpreted as all relationships must function in order to fulfill the required objective, and the latter is interpreted as only one relationship needs to function in order to fulfill the specified objective.

 doi:10.2218/ijdc.v7i1.210

# Preservation Actions on a Network

Networks are created and evolve through preservation actions which alter its structure. There are a limited number of types of action that can be taken which have distinct effect upon a network of information. These are detailed below.

## Risk Acceptance and Monitoring

Risk acceptance and monitoring is likely to the least expensive but most fragile type of preservation action. This action occurs when an archive makes the decision to actively monitor a dependency. By doing so, an archive must accept that important information, vital to the long term exploitation of the data, is outside the control of archive. If we take the example below of a data file which has been generated as a result of an experiment conducted at the ISIS facility. Reuse of the data is dependent upon a user's ability understand the context of the experiment and re-analyze it using particular methods. Figure 1 illustrates a preservation network model for a file in the "NeXus"[7] format. This network accepts that vital information for reuse exists outside the control or "active management" of the data archive. Rather than acquire this information, the archive records specific instructions about the external information sources, the nature of what needs to be monitored and the frequency at which this needs to occur. In this case three different instances of risk acceptance and monitoring are employed:

1. The GEM instrument site[8] is referenced along with a description of the relevant information in the sample environment/preparation methods;
2. The Mantid project site[9] is referenced along with specific details of the analysis methods which much must be supported;
3. Location of the ISIS office and the identifier for the experimental proposal.

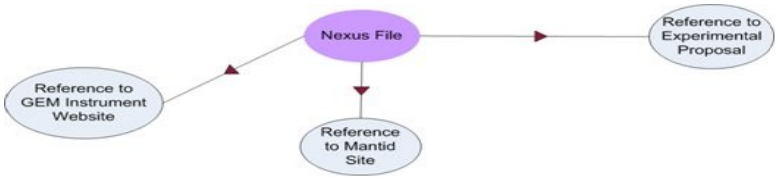


Figure 1. ISIS PNM which utilizes risk acceptance and monitoring.

The main advantage of this approach, in addition to the low initial investment required, is that it supports the evolution of an information asset. In many cases information about a scientific facilities and analysis methods improve over time. It then becomes advantageous to monitor information in the community in order to support such evolution. The main disadvantage of this approach is the level of risk the archive becomes exposed to. Information can become irrevocably corrupted or lost with little notice. Even if the destabilisation of community-based information has been

[7] NeXus Format: http://www.nexusformat.org/Main_Page

[8] GEM instrument websites: http://www.isis.stfc.ac.uk/instruments/gem/

[9] Mantid Project: http://www.mantidproject.org/Main_Page

detected in time, the costs of acquisition can rapidly escalate. Over time funding can be lost with key scientists and software development groups dispersing within the organisation or leaving completely. This can result in a sudden step change in acquisition costs. Long term preservation can only be confidently guaranteed if there is a reasonable expectation that sufficient funding will be available at such critical junctures to acquire all necessary information. An important distinction to draw is the difference between an object referenced within a logical Archival Information Package which the archive controls, and an externally referenced resource.

### Capture of Software and Extension Through the Stack

This type of preservation always involves the acquisition of digital objects which perform a function on the data. A scientist's ability to exploit data is normally dependant on software of some type, which processes and presents data to the end user in a usable form. The function of the software can be as simple as extracting parameters and rendering them in a usable format. However, the function of much software tends to be more sophisticated, potentially transforming the data in some way, which has serious implications for preservation. The software may take averages of parameters, apply some other form of statistical operation and fill in missing data from detectors using assumed values, or process data using models based on current scientific assumptions. Such assumptions could prove to be inaccurate in hindsight with advances in scientific knowledge, and therefore should be revised at a later date. Preserving software without full awareness of the operations/functions it performs effectively fossilizes a user's interaction with data. This permits no redress and may hinder positive progression of data exploitation. Figure 2 is a preservation network model for Ionosonde data, which contains an ionization profile of the atmosphere. The network consists of the SAO-explorer[10] executable file. The branch has then been extended through the software stack. In this case, we need to preserve the Java VM and software platform to maintain the executability of the SAO-explorer. By doing so we preserve the ability to process the data in the file, apply a standard analysis model and then render it in the form of an ionogram.[11]

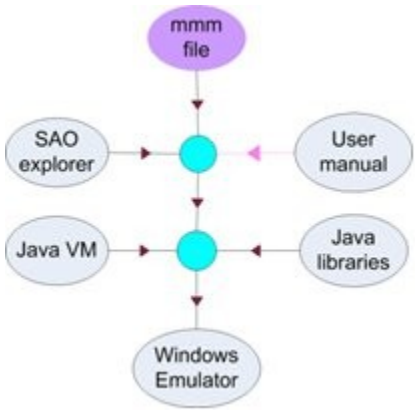


Figure 2. Illustrates stabilisation by extension through the stack (locally archived and managed digital objects).

[10] SAO Explorer: http://ulcar.uml.edu/SAO-X/SAO-X.html

[11] Wikipedia Definition of Ionogam: http://en.wikipedia.org/wiki/Ionogram

 doi:10.2218/ijdc.v7i1.210

Extension through the stack supports two of three methods described in the CEDARS project (2002) as Technical Preservation and Emulation. Matthews et al. (2009) applies the former in the context of software, as maintaining the original software (typically a *binary*), and hardware, of the original operating environment. In this approach otherwise obsolete hardware is maintained to keep vital software in operation or the original operating environment is recreated by programming future platforms and operating systems to reproduce that original environment so that software can be preserved in binary and run "as is" in on a new platform. In the emulation approach, a new item of software, an emulator, is added to simulate the original software platform, which can then be executed on a different hardware and operating system environment. These types of approach have been advocated by the PLANETS[12] project and are currently being developed on the KEEP[13] digital preservation projects. A discussion of this type of approach can be found in Suchodoletz & Hoeven (2009). In addition to stabilising a software-based solution by extending through the software stack (a set of software subsystems or components needed to deliver a fully functional solution), risk can also be spread through capture of different platform-based variants as alternates.

There is a third option for software preservation open when source code is available. For example, the Mantid software project[14] has publicly available source code that can added to the network as an alternate strategy to the archived executables. The inclusion of this type of strategy can be described as process-centric, CEDARS' third method. Matthews et al. describes this as transferring digital information to a new platform, and applies it to software as meaning recompiling and reconfiguring the software source code to generate new binaries, and applying it to a new software environment with updated operating system languages, libraries and so on. In other words, a software migration. This needs to be supported by validation tests to ensure that the migration successfully reconstructs the desired functionality. In Figure 2 alternate strategies are available for preserving Mantid depending on what operating system and source code compilers are available.

**Description**

Description strategies always involve the acquisition of information, but unlike software capture description is technology agnostic with no dependencies on the persistence of any form of technology apart from reading the ones and zeros from the media it was encoded on. This can done using verbose textual description or by using formal data description languages, such as those described by Rankin and Giaretta (2007). A properly executed preservation strategy should mean that a sufficiently large and motivated team of scientists, given enough time, could extract and process the encoded binary information into the desired form and derive any information specified by the preservation objective without ever needing to use any form of technology. This is, of course, the test of a true descriptive strategy. The reality of using a descriptive strategy is that it simply allows you to re-implement algorithms and processes using appropriate new technologies. Once a descriptive strategy has been successfully executed all dependencies and risks are associated with:

[12] PLANETS project: http://www.planets-project.eu/

[13] KEEP project: http://www.keep-project.eu/ezpub2/index.php

[14] Mantid Software: http://download.mantidproject.org/

- A community's ability to correctly interpret the descriptive information. A designated communities ability to do this should always be monitored with this type of strategy; and
- A reasonable expectation of resources/manpower to re-implement new solutions based on this information. This is an effort intensive, high cost strategy.

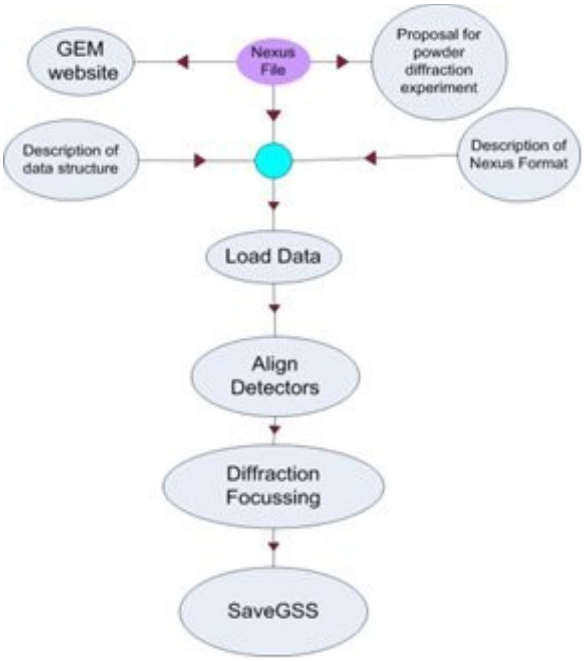


Figure 3. PNM employing descriptive strategy for ISIS data.

The example above permits a scientist in the future working with data from the ISIS GEM instrument to extract the necessary parameters and perform a simple powder diffraction analysis on the NeXus file. This uses descriptions of the NeXus format and four algorithms. If we look at the current algorithm description for Diffraction Focussing[15] we can see that all referenced information must be gathered together and more documentation needs to be carried out on how the data is grouped, how errors are calculated, how data points are normalised, and how the masked or partially masked bins are handled. Once this has been done an adequate description of the algorithm will have been created and, with sufficient expertise and effort, the software can be recreated.

**Migration (Transformation)**

Migration (strictly an information object migration, to distinguish from a software migration above) is different from the other strategy types in that it does not involve the acquisition of additional information objects but the transformation of an existing information object into another physical type of information objective. This may or may not involve the loss of information, but should always force the re-evaluation of the validity of any network branch below it.

[15] Mantid diffraction focussing algorithm: www.mantidproject.org/DiffractionFocussing

 doi:10.2218/ijdc.v7i1.210

The consequences of migration may be trivial, as in the example shown in Figure 4 where conference proceeding in Word document format have been migrated to PDF. While some of the formatting may have changed, no significant information has been lost and no major alterations to the network are necessary.

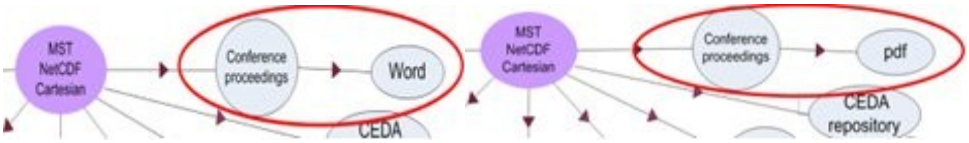


Figure 4. Trivial migration.

However, some types of migration have serious consequences and may involve a change in preservation objective. If we take the example in Figures 5 and 6, where an ionosonde "mmm" format file from an ionosonde instrument has been converted to an IIWG[16] formatted text file, we see that this transformation involves significant information loss. Where it was previously possible to extract a full ionogram, a scientist is now only able to extract a restricted list of parameters from the file.

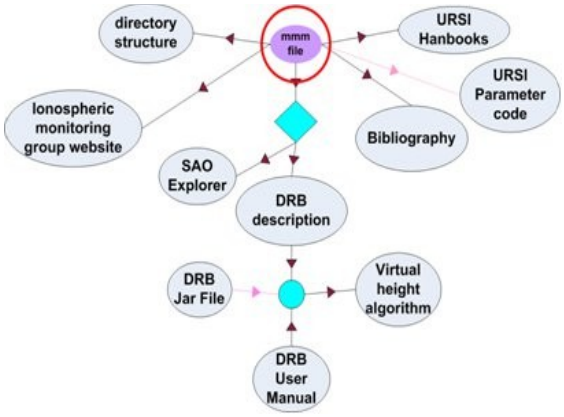


Figure 5. PNM before significant migration.

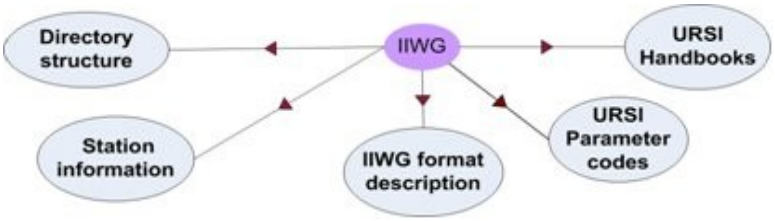


Figure 6. PNM after Significant Migration.

This transformation involves significant information loss, so where it was previously possible to extract a full ionogram a scientist is now only able able to extract a restricted list of parameters from the file. As a direct result of the information loss the preservation objective must be altered, which has significant consequences and forces a completed re-analysis of the entire network. The re-analysis necessitates the use of other types of strategy, dramatically transforming the entire network.

[16] IIWG format help: www.ukssdc.ac.uk/wdcc1/ionosondes/iiwg_format.html

## Combining Strategy Types

As discussed above, each of the different types of strategy have unique benefits, risks and costs associated with them. In addition to considering individual types of strategy and the degree to which they should be stabilised/strengthened, the effect of combining strategy types should also be appraised. In the example shown in Figure 7, the combined usage of description, software capture and referencing means that the solution provides long term stability, immediate access to quality assured software and state of the art analysis techniques.

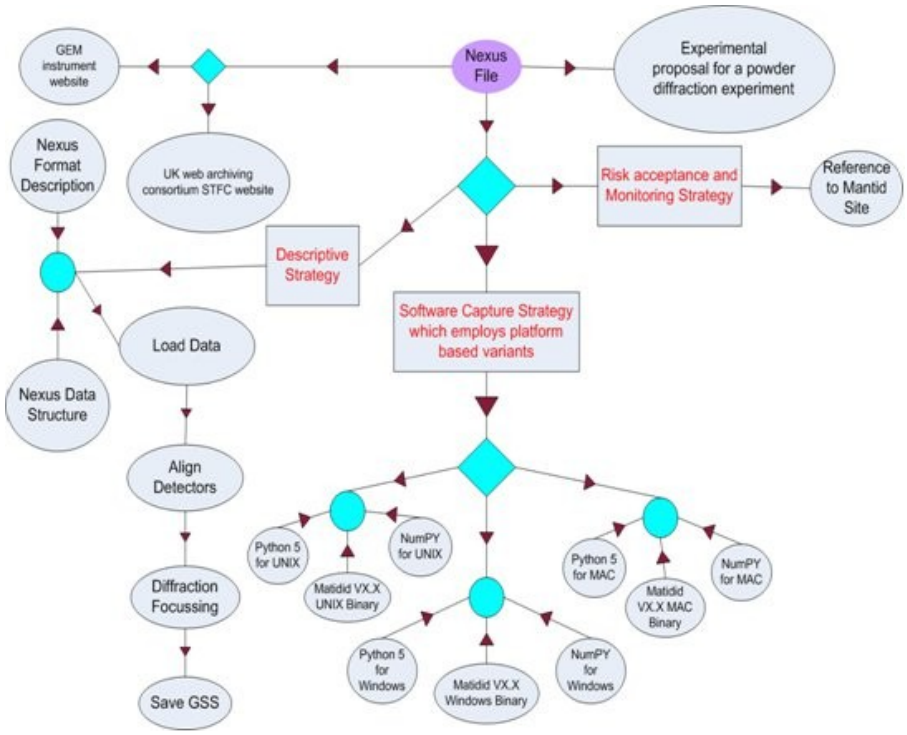


Figure 7. PNM employing combined strategy types.

If we take the position that no solution is permanent and that information networks will evolve over time, we can see that the benefit of using multiple solutions is the effect of more than just that of spreading the risk and becoming more than a "sum of its parts." The diagram in Figure 8 represents positive feedback between strategy types:

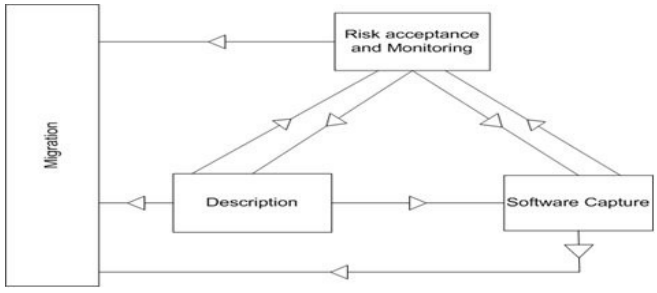


Figure 8. Positive feedback between strategy types.

 doi:10.2218/ijdc.v7i1.210

Risk Acceptance and monitoring is a strategy which may be supported by, and potentially benefits from, both description and the capture of software. If we take the example of the Mantid software development project, long term funding is not guaranteed with control of this open source project potentially being transferred to different organisations. In the event of funding being withdrawn and re-started at later date or transferred to another organisation, a new development team can take advantage of well documented algorithms and properly archived software and code in its new development efforts. Even if the current organisation continues to support ongoing development, they still benefit from good quality documentation, which makes the natural evolution of the software more economically viable. In many ways this form of positive feedback between the strategy types becomes a form of preservation by cultivation in itself, as described in Software Preservation Benefits Framework (Hong et al., 2010).

The description strategy is itself supported by the act of monitoring a software development groups' activities. In the case of Mantid, the project deprecates old algorithms, replacing them with new algorithms endorsed by the scientific community. Monitoring the group's activities will alert an archivist to the creation of new, state of the art algorithms. This gives the archivist the opportunity to update the information contained as part of a description strategy, allowing the research asset to evolve in a positive way.

Software capture benefits from both risk acceptance/monitoring and description strategies, but in different ways. Again, by monitoring the Mantid project an archivist will see when new versions of software with different dependencies become available, giving the archivist the opportunity to add or replace existing executables and code in the networks. Description of the relevant algorithms benefits the software capture strategy by increasing its trustworthiness, as future users can scrutinise the algorithms to see if the analysis they perform is suitable for their line of scientific enquiry.

Migration benefits from all the other strategy types, as they may all potentially provide mechanisms for migration. In the ISIS example above, the Mantid development team have expertise and can assist in converting NeXus files to other popular current formats. The Mantid software can convert NeXus files in HDF5 format[17] and into XML[18] format, amongst others, and the NeXus specification will allow programmers in future to write tools that convert NeXus to whatever popular formats exist in the future.

## Added Value of Preservation Network Models

Preservation networks are about much more than simply visualising networks of information within diagrams. The full power or value of preservation network models is their ability to accommodate all four strategy types, which have inherently different risks, costs and benefits associated with them. The ability to employ multiple strategy types over time, adequately describing preservation of critical attributes, is of considerable benefit to an archive. It permits judicious appraisal of preservation strategies, selection of the optimal configurations and broader implication for the

[17] HDF5 format: http://www.hdfgroup.org/HDF5/

[18] An Introduction to XML: http://www.w3schools.com/xml/xml_whatis.asp

archive as a whole. In addition to aiding the initial design process, it permits the timely and considered evolution of research assets, allowing preservation action to be tailored to changing environments and advances in scientific understanding. The other major benefit of this approach is the ability to characterise data collections and apply generic preservation network models, allowing this approach to scale in a sustainable manner. We will now discuss each of the benefits in turn.

### Assess Exposure to Risk and Longevity of Solutions

As discussed above, the types of risk associated with the strategies differ in their nature and the likelihood of their realization. In addition to assessing whether certain individual risks are acceptable to an archive, it now becomes possible to statistically combine probabilities for all the composite and alternate relationships to produce time dependant risk profiles. This can be further enhanced, with consideration given to cost and benefits, to provide a rich and informative basis for decision making.

### Impact Analysis Based on Risk Realisation

Archives such as the British Atmospheric Data Centre hold many heterogeneous data sets and collections. As a result, they regularly survey their user communities to determine use of operating systems, programming languages and data formats. Once PNM's are expressed in a machine readable format, it then becomes possible to run archive wide queries to assess the impact of risk realization. This is because of PNM's ability to expose the risk an archive is subject to through their explicit statement. PNM's therefore provide the capacity to perform meaningful reporting on the broader implication of changes in the preservation environment.

### Tailoring Solutions that are Responsive to Evolving Environments

The preservation and stakeholder environment is a complex one, which is evolving all time. It is important to appreciate that situations can improve as well as degrade. Preservation technologies advance; scientific analysis techniques, models and understanding improve; costs reduce; and institutional infrastructure and skills bases develop. For example, it may be appropriate to have short term, weak solutions that employ referencing a temporary measure because superior analysis software is being developed and should be included in the PNM at a later date. Similarly, there is very little point in developing an expensive data description solution if it likely the data will be converted to a more popular format in the near future. The advantage provided by PNM's facilitate this by accommodating and clearly describing network branches of varying stability, ensuring research assets evolve in the most efficient way in response to their environment.

### Scalability Through Characterising Data Collections

Scalability is rapidly becoming an important issue, given the sheer number of existing data files and the increase in production rate most archives are anticipating. However, data files with common preservation requirements can be defined as a data collection. In the case of ISIS data, this could be files from the same instrument with a common experiment type, community and preservation objective. This data collection is

 doi:10.2218/ijdc.v7i1.210

supported by a generic PNM. This allows cost reduction through the re-use of preservation solutions and automation of preservation action.

## Conclusions

This paper has made the risks addressed in preserving research data explicit by showing a representation of preservation networks, which can be used to capture the dependencies of data, software and other types of digital object that give rise to those risks. Once the risks are explicit then a preservation strategy can be chosen to balance the risk aversion of the preserver, benefits to the archive or data user, and the cost of preservation. This paper has also characterised and described the different types of actions for the preservation of scientific data. The advantages and disadvantages of these approaches are discussed in terms of risks, stability, benefit and cost. We examined how these can be combined in the most effective ways. This allows the optimal balance between cost, benefit and acceptable risk burden to be sought when a preservation solution is designed, as well as supporting the sustainable evolution of the research asset.

The process described here has been designed to address the risks of scientific research data. However, the underlying approach that it has revealed would also be appropriate to the preservation of other digital resources. This is because we believe that the four types of preservation action described in this paper are potentially applicable to all types of data object and can be effectively captured for long term management within a Preservation Network Model. We feel that possible resistance within archives to employing the full range of strategy types may be due to historical reasons rather than full consideration of all possible solutions. However, more research into the domain/organisational applicability and the design of preservation solutions is required to fully test this.

Future work on the ENSURE project will study the costs of preservation actions and the cost benefit tradeoffs required to choose the appropriate preservation strategy. Network modelling will be investigated in relation to preservation policies: both the derivation of lower-level policies from the models themselves, for example about frequency of monitoring, and the use of high-level policies in conjunction with models to determine and constrain preservation actions.

## Acknowledgements

The research reported in this paper was partially funded by the EC through awards to the projects CASPAR (contract number, 033572) under EU-FP6 – ICT Programme. and ENSURE (contract number, 270000) under EU-FP7 – ICT Programme.

## References

Cedars Project. (2002). *The Cedars guide to digital preservation strategies*. Retrieved from http://www.webarchive.org.uk/wayback/archive/20050410120000/http://www.leeds.ac.uk/cedars/pubconf/pubconf.html

Consultative Committee for Space Data Systems. (2002). Reference model for an Open Archival Information System (OAIS). *Recommendation for Space Data Systems Standard.* CCSDS Blue Book. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf

Conway, E., Dunkley, M., Giaretta, D., & McIwrath, B. (2009). Preservation network models: Creating stable networks of information to ensure the long term use of scientific data. Paper presented at Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data conference, Madrid, Spain. Retrieved from http://epubs.cclrc.ac.uk/work-details?w=51437

Giaretta, D. (2011). Advanced digital preservation (1st ed.). Springer-Verlag, Heidelberg.

Hong, N., Crouch, S., Hettrick, S., Pakinson, T., & Sheeve, M. (2010). Software preservation benefits framework. Report of the Software Sustainability Institute. Retrieved from http://www.software.ac.uk/attach/SoftwarePreservationBenefitsFramework.pdf

Matthews, B., Shaon, A., Bicaregiu, J., Woodcock, J., Jones, C., & Conway, E. (2009). Towards a methodology for software preservation. Paper presented at the Sixth International Conference on Preservation of Digital Objects (iPres 2009), San Francisco, USA. Retrieved from http://epubs.cclrc.ac.uk/work-details?w=50801

Rankin, S., & Giaretta, D. (2007). Requirements for OAIS structure representation information ensuring, long-term preservation and adding value to scientific and technical data. Paper presented at the PV 2009 conference, Madrid, Spain. Retrieved from http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/papers/7_Rankin_OAISStructure.pdf

Suchodoletz, von D., & Hoeven, van der J. (2009). Emulation: From digital artefact to remotely rendered environments. *International Journal of Digital Curation 4*(3). Retrieved from http://www.planets-project.eu/docs/papers/SuchodoletzVanderhoevenIJDC.pdf