Managing Preservation Networks: Issues of Scale for Scientific Research Assets

Esther Conway
STFC
Rutherford Appleton Laboratory
esther.conway@stfc.ac.uk

Simon Lambert
STFC
Rutherford Appleton Laboratory
simon.lambert@stfc.ac.uk

Brian Matthews
STFC
Rutherford Appleton Laboratory
brian.matthews@stfc.ac.uk

Arif Shaon
STFC
Rutherford Appleton Laboratory
arif.shaon@stfc.ac.uk

ABSTRACT

The preservation of science data requires consideration of a wide range of factors from file formats to analysis software. Previous work has reported on the development of Preservation Network Models that capture dependencies at multiple levels and allow reasoning about preservation planning and actions. However this is only one aspect of the development of a trusted preservation environment; there is also a need for quality assurance and relation to explicit policies on data preservation. This paper presents issues of scale for scientific research assets which will be explored further on the SCAPE project.

Keyword

Digital Preservation, Scientific Data, Preservation Network Models

1. Introduction

Preserving scientific research data has become increasingly recognized as necessary for the long term benefit of large scale data collection to be fully realized. However, the complex nature of science data and its dependencies on software, together with the scale of the data holdings involved make this a major challenge.

Given the sheer number of existing data files and the anticipated increase in production rates, scalability of any preservation action has become an important issue for the archive. Scale is absolutely critical in two respects for cost reduction through the re-use of preservation solutions and automation of preservation action. We consider these issues using Preservation Network Models (PNMs), a preservation analysis methodology which was originally developed within the CASPAR project.

2. Preservation Network Models

A PNM is a formal model for conceptualising the relationships between resources within the scenario of a preservation objective."The preservation network model consist of two components: the digital objects and the relationships between them possesing atrribute of (Information, Location ,Physical state) and (Function, Risks and Dependencies, Tolerance, Quality assurance/ testing) repectively

iPRES2011, Nov. 1–4, 2011, Singapore.

3. Preservation Action

Networks are created and evolve through preservation actions which are made on particular relationships or acknowledged dependencies within the network. Types of action are

- Risk acceptance and monitoring
- Software capture and extension through the stack
- Description
- Migration

4. Quality Assurance

The quality assurance of a preservation solution is provided by two mechanisms Trust in or Testing discussed below

4.1 Trust

Trust occurs when the archive appraises a solution as satisfactory for one of the following reasons

- Trust in a custodial organization; when the archive relies on an external organization to maintain the integrity and supply of important information. The accepted reputation of the organization supplies the required assurance.
- Trust in a standardization process; when an archive acquires descriptive information which has produced as a result of a standardization process such as ISO.
- Quality of Sources; this occurs when an external organization supplies an archive with an information object. Trust is based upon the belief that the supplier has delivered a quality preservation solution.

4.2 Testing

When an archive cannot fully trust a solution it must then employ testing to gain necessary assurance. We consider three testing scenarios.

- Passive testing; occurs when a preservation solution is exposed to an active user community with the expectation that they will report any deficiencies.
- Proactive testing; occurs when external experts are invited to test a preservation solution.
- Direct testing; occurs when the archive conducts testing itself.

5. Monitoring the Preservation Environment

No preservation solution is permanent and will always carry risks due to dependencies. Change is required for a number of reasons detailed below

When a preservation solution is longer valid this forces a reevaluation in terms of new information needs of the user community and the funding available to carry out preservation in a sustainable way.

Most changes are due to the realization of risk causing failures of the preservation solution that are within tolerance, partial or critical The explicit statement of technical dependencies within a network can be used to determine the types of things that need be monitored.

- Dependencies on external organizations risk acceptance which by definition inform watch services
- Dependencies on "software capture" strategies require the monitoring of libraries and operating systems
- Dependencies on a descriptive strategy involving community skill, support and resources

In addition to external triggers which invalidate the preservation solution "risk acceptance and monitoring" also has the capacity to support evolution of the scientific asset. It forms one of a number of positive feedback relationships when multiple strategy types are employed.

6. Preservation Action in a Scalable Environment

In this section we e give the illustrative examples from the ISIS GEM powder diffraction instrument. We explore how each of the main types of preservation action are affected by issues of scale for the creation and maintenance of scientific research assets.

6.1 Monitoring Websites

The preservation network uses two different risk acceptance and monitoring strategies with different degrees of re-use. The archived website held by the UK web archiving consortium hold information which should be universally associated with all data files. However, the software which the Mantid website provides access to is not universally applicable. Data files from different beam lines and experiment types require different forms of analysis. Currently Mantid can support the minimal required analysis for approximately 60% of data holdings. The Mantid website needs therefore to be associated with files whose preservation objective requires the type of analysis supported by Mantid. In both cases automation is required to propagate necessary notifications and changes for example new URL's reference points when websites are migrated or failure of the solution through the networks, as thousands of file should be associated with both these externally managed information objects.

6.2 Capture of Mantid Software

As described above this type of strategy involves the acquisition and management of information objects. As with the risk acceptance and monitoring strategy re-use of this solution (network branch) is appropriate for around 60% of data holdings based on their preservation objective.

Again because of the number of files associated with a software capture solution, automated changes to large numbers of networks become desirable. Removal of platform dependent network branches when operating systems become obsolete or extension of the branches to include libraries and emulators is required in order to stabilize the solution. The automated addition of alternates involving new binaries or source code would also be advantageous. These can then be recompiled to work on different operating systems when communities begin using new technologies analysis techniques.

6.3 Description of Analysis Algorithms

The descriptive strategy a scientist to identify, extract and correctly interpret parameters and the relationship between them. A scientist can subsequently carry out a specified type of analysis by applying the described algorithms. The capture of specific algorithms mean the user is restricted to a particular analysis path which is a functional subset of both the software capture and risk acceptance strategies. As a result the degree to which this solution can be reused by different data files is much lower as experiment types have unique analysis requirements. As this type of preservation strategy is technology agnostic the only automation required is the ability to update the analysis path for multiple networks once the old have been deprecated and new algorithms gain community acceptance.

6.4 Conversion of Document formats

The need for automation becomes important when an archive needs to transform a large number of digital objects from one format to another. When the preservation network models are logical rather than physical there are variations in the numbers of actual objects which may require conversion. If we consider the example of an archive making a decision to convert all word documents to PDF. Automation is not necessary for the word documents describing the experimental environment and preparation methods within the instrument website. While the website is logically referenced by thousands of PNM's there exist only a couple of physical copies making manual conversion the most efficient option. However experimental proposals are unique to a discrete set of data files the ability to characterize the relationship and format of a digital object and automate its transformation is critical to the successful management of information on this scale.

7. Policy Formation

Dealing with large volumes of data with differing preservation objectives can place addition pressures on an archive. Based on our previous discussions we suggest areas where an archive may wish to develop policy to manage such large scale data holdings.

- Policy on number and types of dependencies and archive should be exposed to in order to maintain an acceptable risk burden.
- Policy can specify appropriate levels of vigilance and monitoring of the preservation environment.
- Policy can specify what is a trusted organization, institution or standards.
- Policy can mandate levels of testing required for any preservation solution to be deemed acceptable
- Policy can specify how much of the hardware/software environment should be captured or if the solution should be supplemented with source code.
- Policy can recommend the employment of multiple strategy types to lower risk burden and enhance long term usability
- Policy can also stipulate acceptable formats which an archive a can reasonably expect to support and monitor
- Policy can mandate descriptive preservation solution for non standard formats

8. Conclusions

The use of Preservation Network Models provides a basis not only for preservation planning and actions, but also for other preservation-related aspects such as quality assurance or trustworthiness. By conducting an analysis of dependencies, founded on specified preservation objectives, issues such as scalability can also be analyzed. Furthermore there is an interaction with preservation policies: Preservation Network Models can highlight areas where policies should be put in place, and help to guide their formulation. Thus these models are proving an invaluable framework for scientific data preservation at STFC facilities. Further exploration and trialing on part of the ISIS archive within the SCAPE project to fully address the issues of scale discussed in this paper is required.