Preservation Network Models: Creating Stable Networks of Information to Ensure the Long Term use of Scientific Data

Esther Conway [(1)], Matthew Dunckley [(1)], Brian McIlwrath [(1)] and David Giaretta [(1)]

[(1)] Science and Technology Facilities Council
Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, OX11 0QX, UK
Email: e.a.conway@rl.ac.uk

ABSTRACT

Meaningful preservation of scientific data is that which permits reuse. This is frequently dependent on a number of digital objects and sources of information which will have been subjected to preservation action such as format conversion or the addition of representation information. A future user may be required to interact with a number of unfamiliar digital objects in order to achieve meaningful reuse of data. As a result an archivist will be confronted with the task of designing an information network which a future data user can navigate and effectively engage with.

A preservation information network model is a representation of the digital objects, operations and relationships which allow a preservation objective to be met for a future designated community. The model provides a sharable, stable and organized structure for digital objects and their associated requirements. The model also directs the capture and description of digital objects which need to be packaged and stored within an OAIS compliant Archival Information Package.

In this paper we wish to present our approach to modelling these networks using illustrative examples from the CASPAR testbeds. We intend to demonstrate how preservation networks can be designed modelled and terminated for a variety of digital objects supporting the long term reuse of scientific data.

Keywords: Digital Preservation, Data Reuse, Information Modelling, Preservation Network Models

INTRODUCTION

This paper presents an overview of preservation network models developed on the CASPAR project [1]. We aim to show how preservation networks are an output put from preservation analysis where an OAIS [2] compliant Archival Information Package (AIP) has been designed and preservation strategies have been actioned. Using illustrative examples from the CASPAR testbeds we intend to show how these network models are a representation of the digital objects and relationships which allow preservation objectives to be met for a future designated community.

In this paper aim to demonstrate how preservation networks provide a sharable, stable and organized structure for digital objects and their associated functions. The model exposes the risks, dependencies and tolerances within an archival information package. This allows for the automation of event driven or the periodic review of archival holdings by knowledge management technologies.

The clear definition of relationships also facilitates the identification of reusable solutions which can be deposited within registry repositories of representation information, thus sharing preservation efforts within and across communities. We then conclude by summarizing how the model facilitates long term preservation of scientific data.

Preservation Network Models as an Output of Preservation Analysis

The challenge of digital preservation of scientific data lies in the need to preserve not only the dataset itself but also the ability it has to deliver knowledge to a future user community. This entails allowing future users to reanalyze the data within new contexts. Thus, in order to carry out meaningful preservation we need to ensure that future users are equipped with the necessary information to re-use the data. The analysis method facilitates modeling of information networks based on the archival information package solution.

If we take the following example from the SCARP project [3] case study Curating Atmospheric Data for long term use: Infrastructure and Preservation Issues for the Atmospheric Sciences community [4]. We set the preservation objective as follows. A user from a future designated community should be able to extract a specific set of parameters from data files for a given time and altitude. These include typical measurements such as vertical wind shear and tropopause sharpness. In addition we would want the data user to be able to correctly interpret the scientific parameter definitions and to be able access and read the following materials.

- Scientific output resulting from use of the data set
- The MST international workshop conference proceedings
- The MST user group meeting minutes

Resultant preservation action produced a collection of digital objects and relationships described by the diagram below.

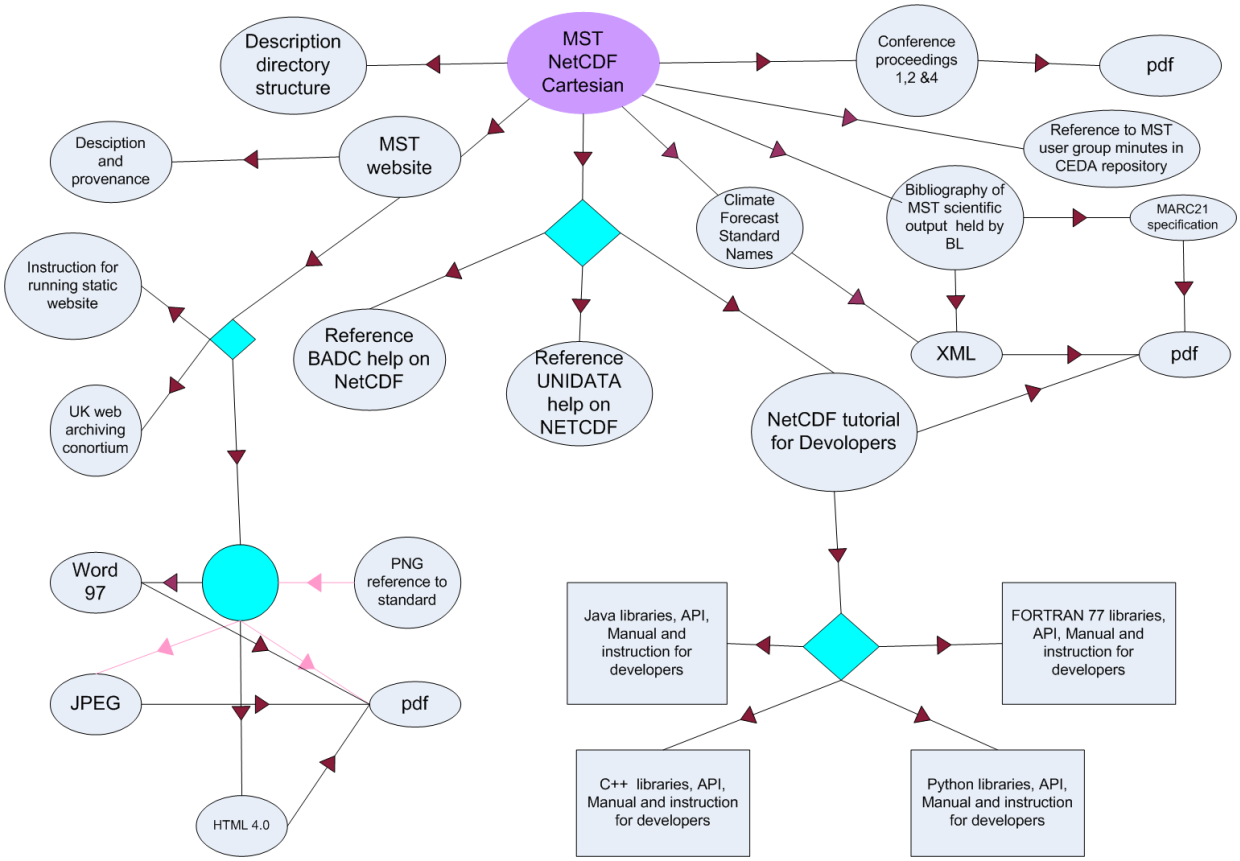

Figure 1.Preservation Network Model for MST data

Components of a Preservation Network Models

Preservation network modelling has many similarities to classic conceptual modeling approaches such as Entity-Relationship or Class diagrams, as it is based upon the idea of making statements about resources. The preservation network model consist of two components the digital objects and the relationships between them.

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

- **Information** is 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 is the information required by the end user to physically locate and retrieve the object. AIP's may be logical in construction with key digital object being distributed and managed within different information systems. This tends to be the case when data is in active use with resources evolving in dynamic environment.
- **Physical State** describes the form of the digital object. It should contain sufficient information relating to the version, variant, instance and dependencies.
- **Risks** most digital solutions will have inherent risks and a finite lifespan. Risks such interpretability of information, technical dependencies or loss designated community skill. Risks should be recorded against the appropriate object so they can be monitored and the implication of them being realised assessed.
- **Termination** of network occurs when a user requires no additional information or assistance to achieve, the defined preservation objective given the accepted risks will not be imminently realised.

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.

- **Function**, in order to satisfy the preservation objective a digital object will perform a specific function for example the delivery of textual information or the extraction and graphical visualisation of specific parameters
- **Tolerance,** not every function is critical for the fulfilment of the preservation objective with some digital objects included as they enhance the quality of the solution or ease of use. The 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.
- **Alternate and Composite relationships** can be thought of as logical "And" (denoted in diagrams by circle) or "Or" (denoted in diagrams by diamond) relationships. Where either all relationships must function in order to fulfill the required objective or in the case of the later only one relationship needs to function in order to fulfill the specified objective.

Stability and Review

The preservation network model describes a preservation solution whereby a number of digital interact to fulfill a preservation objective for a designated community. The preservation solution consists of a number of digital objects and sources of information which will have been subjected to preservation action such as format conversion or the addition of representation information. A future user may be required to interact with a number of unfamiliar digital objects in order to achieve meaningful reuse of data. As a result an archivist will be confronted with the task of designing an information network which a future data user can navigate and effectively engage with. These solutions are also not permanent with dependencies and associated risks which must be monitored by an archive. These risks must be monitored and managed by an archive as the realization of these risks may result in a critical failure where the network cannot fulfill the defined objective. Realization of risk leads to the three different types of failure partial, within tolerance and critical. Critical failure requires immediate action as the solution will no longer be able to meet the preservation objective. With critical or partial failure the objective can still be met but realization risk often prove to be an appropriate juncture to review the overall solution. Review of the preservation network model provides stability and confidence that risks will be dealt with in a timely manner.

Partial Failure

The preservation network model below gives an example of a partial failure scenario. Where the British Atmospheric Data Centre [4] and UNIDATA [5] have withdrawn support for the NetCDF file format, the designated community has also lost the skill to write programs in C++, FORTRAN 77 and Python. As the community can still write a program to extract the required parameters the preservation objective can still be met. However withdrawal of the British Atmospheric Data support for the NetCDF format may prove to be an appropriate juncture to covert the file to a different format.

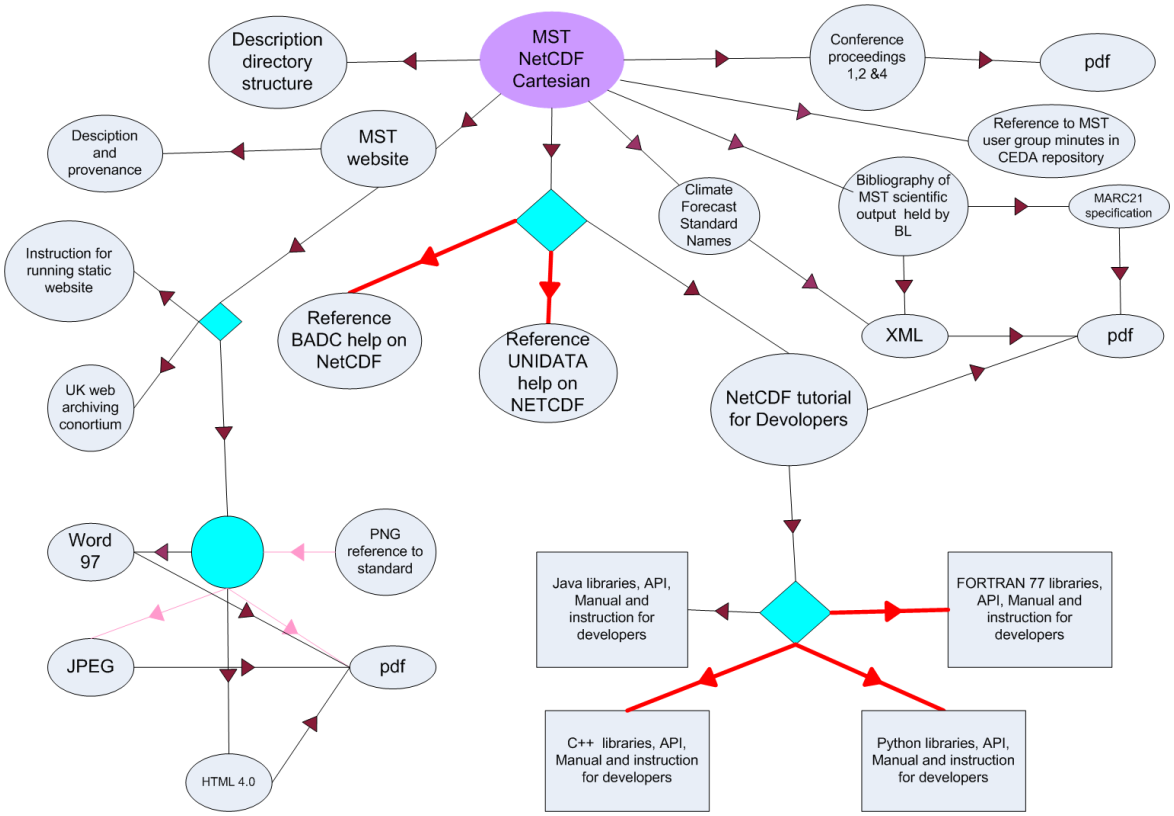

Figure 2. Partial Failure of MST data solution

Failure within Tolerances

The preservation network model section below gives an example scenario of failure within tolerances. The ionospheric monitoring group website contain vital provenance and context information relating to Ionosonde raw output files that are the target of our preservation efforts. The loss of jpeg images from the website can be tolerated as they do not contain any critical information.

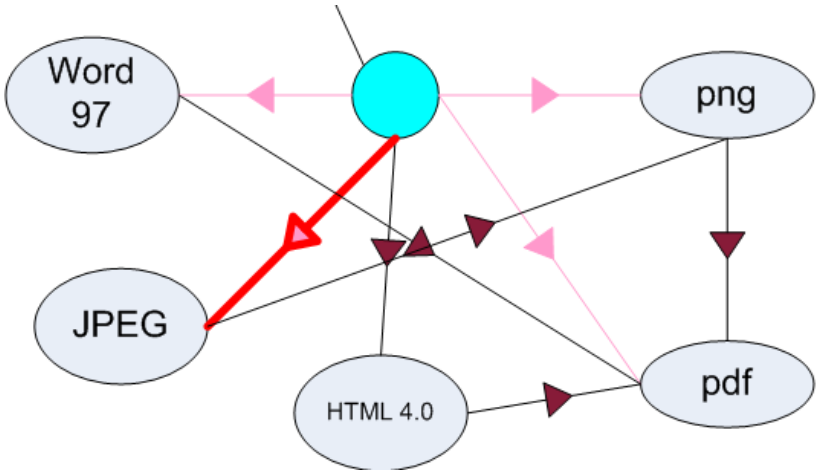

Figure 3. Failure of within tolerances for Ionospheric monitoring group website solution

Critical Failure

The preservation network model section below gives an example of critical failure. In this scenario failure of the communities ability to read xml documents would prevent them from reading the DEDSL dictionary which allows user to correctly interpret the parameter codes and therefore the contents of the data file causing critical failure of the solution.

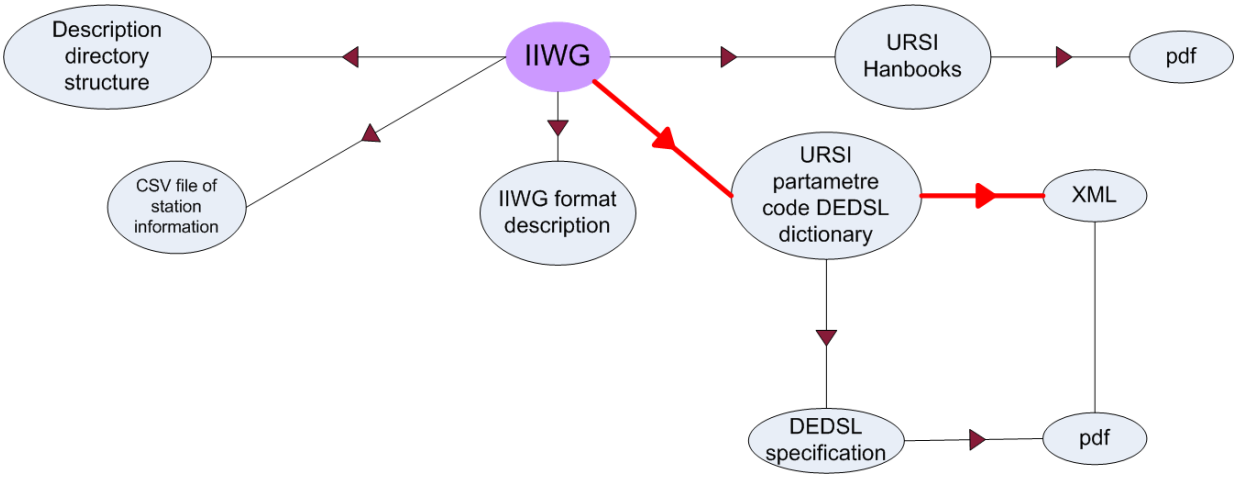

Figure 4 Critical Failure for Ionospheric data preservation solution

Re-usable Solutions and Registry Repositories of Representation Information

The Registry Repository of Representation Information [6] is a CASPAR/DCC component which allows centralised and persistent storage and retrieval of OAIS Representation Information (RepInfo) (including Preservation Description Information (PDI)) in a centralised Registry/Repository. It also contains maintenance tools for user interaction with the Registry for:-

- Manual RepInfo ingest
- Creation and maintenance of the XML structures (RepInfoLabels) which connect related RepInfo in the Registry into an OAIS network (using the defined categories Semantic, Structure and Other)
- Other RepInfo maintenance

Within CASPAR and the DCC the Registry component has the following responsibilities

- Ingest RepInfo into Registry - with appropriate name, description and classification
- Extract RepInfo from Registry reliably.
- Search Registry for RepInfo matching appropriate (wild carded) criteria (a combination of name, description or classification)

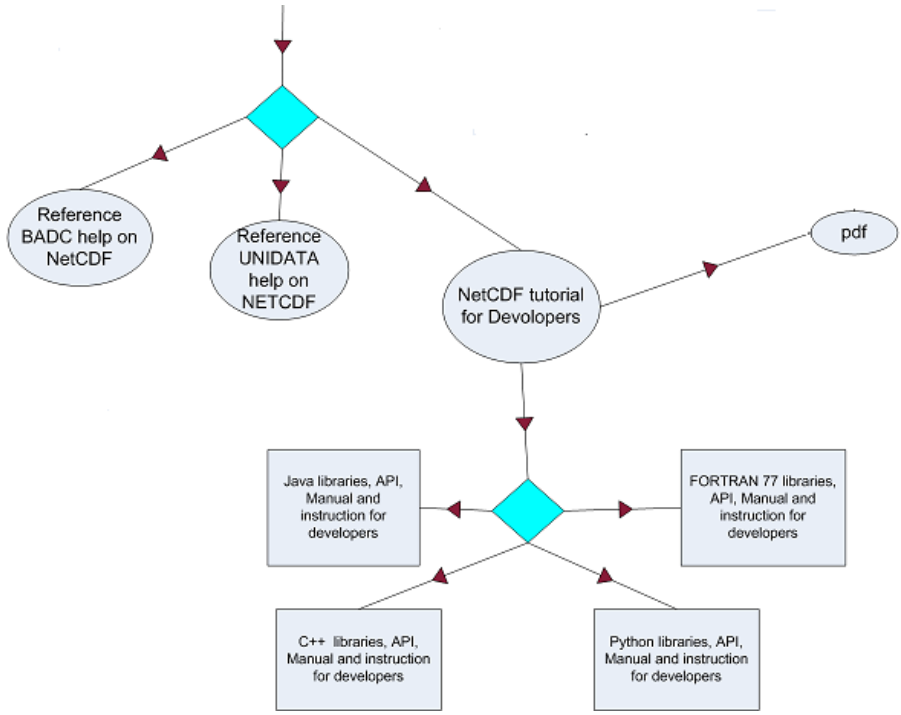

Figure 5 Preservation network model of a NetCDF reusable solution

If we look at the network section above, the objects and relationships described in this section allow a user to extract the desired parameters from NetCDF formatted files. There are eight different strategies a user can employ all of which must fail before there is a critical failure of the solution. As this section of the network has a specific well defined function which is to allow a user to extract parameters form NetCDF formatted files, the solution can be deposited within the repository of representation information. It can then be reused as part of wider solution for different atmospheric data sets which utilize the format.

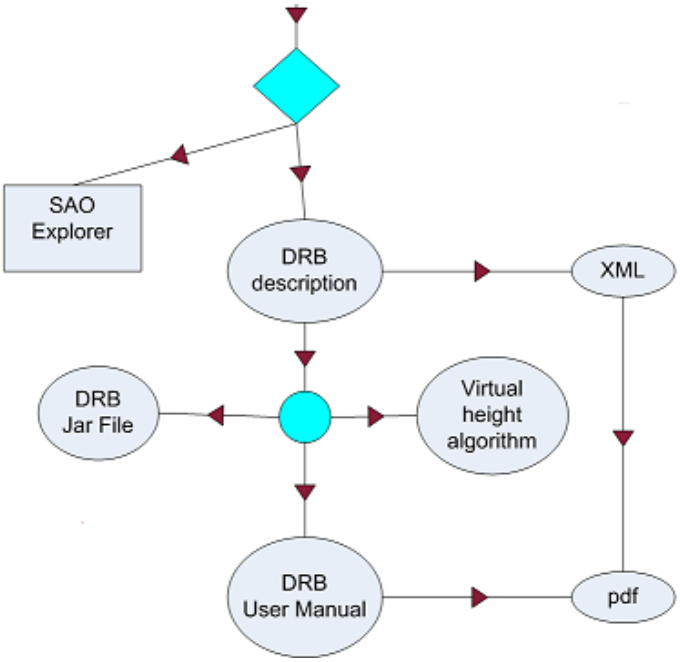

Figure 5 Preservation network model of a mmm file reusable solution

A further validation of the approach took place within a preservation exercise for solar-terrestrial physics data. This study considered raw data which could be analysed to extract an Ionogram – a graph showing ionization layers in the atmosphere. Current scientists use a software product called SAO explorer to extract ionograms from the data. This software was archived in accordance with the methodology described in Matthews iPRES09 [7]. As a result then archived software could with confidence be integrated into an larger OAIS compliant solution for the preservation of mmm data files implemented on CASPAR which permits the long term study of specified atmospheric phenomena from this geographic location. The archived SAO explorer solution could also then be deposited in the DCC registry repository of representation information thereby providing a solution which can be re-used by hundreds of ionosphere monitoring station which are active globally.

Conclusions

This paper presented an overview of a reservation network models which were developed on the CASPAR projects. We believe they support the long term preservation of scientific data by providing a sharable, stable and organized structure for digital objects and their associated requirements. This permits management of risk and promote reuse of solutions allowing the cost of digital preservation to be shared across communities. Wider application, trialling the preservation network modelling outlined here would be desirable to test its validity in a broader range of disciplines and organisational settings. As would the development of software tools to capture te necessary information and automate the risk management.

Acknowledgements

Work partially supported by European Community under the Information Society Technologies (IST) program of the 6th FP for RTD - project CASPAR and the Joint Information Systems Committee (JISC) for the Digital Curation Centre SCARP project.
We would also like to thank our colleagues at STFC David Hooper, Sam Pepler, Matthew Wild, Steve Crothers, Chris Davis, Rita Blake, Ruth Bamford, Simon Lambert, Stephen Rankin and Brian McIlwrath.

References

[1] CASPAR Project http://www.casparpreserves.eu/
[2] ISO 2002. Reference Model for an Open Archival Information System (OAIS). *Recommendation for Space Data Systems Standard, CCSDS Blue Book*.
http://public.ccsds.org/publications/archive/650x0b1.pdf
[3] Conway, E.; Curating Atmospheric Data for long term use: Infrastructure and Preservation Issues for the Atmospheric Sciences community, 2 June 2009 http://www.dcc.ac.uk/docs/publications/case-studies/SCARP_B4832_Atmospheric.pdf
[4] British Atmospheric Data Centre http://badc.nerc.ac.uk/home/index.html
[5]UNIDATA http://www.unidata.ucar.edu/
[6] CASPAR/DCC Representation Information Registry
http://registry.dcc.ac.uk:8080/RegistryWeb/Registry/
[7] Matthews, B; Shaon. A; Bicarregui, J; Jones, C; Woodcock, J; Conway E; Towards a methodology for software preservation In *proceedings of iPres 2009, The 6th International Conference on Preservation of Digital Objects*