Preservation Analysis from the Information Perspective: A Methodology to Support Long Term Reuse of Scientific Data Sets

Esther Conway

Science and Technology Facilities Council, Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, United Kingdom OX11 0QX

esther.conway@stfc.ac.uk

Brian Matthews

Science and Technology Facilities Council, Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, United Kingdom OX11 0QX

brian.matthews@stfc.ac.uk

David Giaretta

Science and Technology Facilities Council, Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, United Kingdom OX11 0QX

david.giaretta@stfc.ac.uk

Abstract

This Paper discusses an approach to preservation analysis which is driven by the need to meaningfully reuse scientific data. The development of the preservation analysis methodology presented in this paper was a response to that need. The methodology incorporates a number of analysis techniques and tools into an overall process capable of producing an actionable preservation plan for scientific data. The resultant preservation actions should ensure that future users are equipped with the necessary information to facilitate such re-use.

1. Introduction

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 Digital Curation Centre SCARP [1] and CASPAR [2] projects have a strong focus on the preservation and curation requirements for scientific data sets. These projects engaged with a number of archives based at the STFC [3] Rutherford Appleton Laboratory. In particular we carried out extensive analysis work to consider the preservation requirements of the British Atmospheric Data Centre [4], the World Data Centre [5] and the European Incoherent Scatter Scientific Association (EISCAT) [6]. During these studies it became clear that there was a need for a consistent preservation analysis methodology.

There are currently a number of tools available which have focus on digital preservation requirements. Drambora [7] provides audit/risk assessment and PLATTER [8] provides planning on the repository level but they do not provide an adequate analysis methodology for data set specific requirements. The Planets [9] planning tool Plato [10] deals with objects within a collection on an individual basis but doesn't examine the inclusion of additional digital information objects and how they interact to permit the meaningful re-use of data. Due to this gap it was deemed necessary to develop a new approach to preservation analysis.

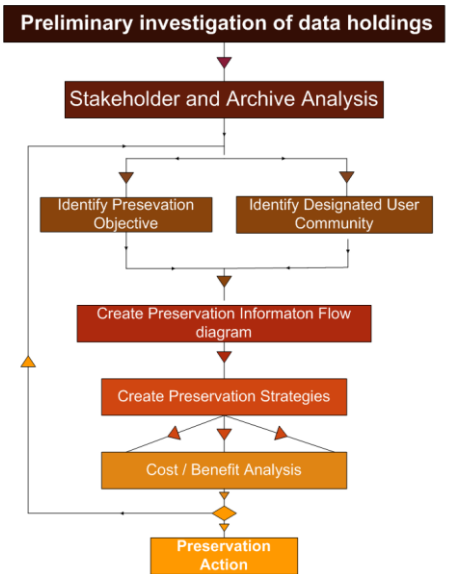

Figure 1. Preservation Analysis Workflow

In the resulting methodology we sought to incorporate a number of analysis techniques, tools and methods into an overall process capable of producing an actionable preservation plan for scientific data archives. Figure 1 illustrates the stages of this methodology. In the rest of this paper we shall discuss the stages in detail, illustrated with examples of work with the scientific archives.

2. Preliminary investigation of data holdings

The initial step is to undertake a preliminary investigation of the data holdings of the target archive. The CASPAR project developed a questionnaire [11] containing key questions which allowed the preservation analyst to initiate discussion with the archive. It critically allows the analyst to.

- Understand the information extracted by users from data
- Identify Preservation Description and Representation information
- Develop a clearer understanding of the data and what is necessary for is effective re-use
- Understand relationships between data files and what constitutes a digital object within the archive

While it is appreciated that this questionnaire is not an exhaustive list of questions which one may need to ask about a preservation target, it still provides sufficient information to commence the analysis process .The full questionnaire and results from the Ionosonde WDC holdings [12] can be obtained from the CASPAR website.

3. Stakeholder and Archive Analysis

After carrying out the questionnaire process for each of archive it is necessary to carry out a stakeholder analysis for these archives. This is because

- Stakeholders may hold different views of the knowledge a data set was capable of providing an end user
- Stakeholders can identify different end users whose skill sets and knowledge base vary
- Stakeholders may have produced or be custodians of information vital for re-use of data

3.1 Stakeholder Categories

The stakeholder analysis classifies stakeholders into a number of categories each with their own concerns. After inspecting a number of datasets the following categories of stakeholder were felt to be most useful.

Every digital archive will have some form of funding body associated with it which provides the resources to collect and maintain the data. During its lifetime, the custody of a data set may pass through several bodies generating rich documentation which explains the scientific purpose of the dataset and how it has evolved over time. These documents can take the form of experimental proposals which will explain the original intent of the experiment/observation, institutional reports which state the intent of maintaining supply of the data to a scientific community, and reports which record scientific output.

Scientific organizations such as university departments or national and international institutes and laboratories, are frequently associated with datasets. They tend to work within a particular branch of science and can provide a great deal of detailed information on how a dataset can fulfill that particular area of scientific potential, providing for example software, support materials and field specific bibliographies.

Every dataset will have an individual scientist, or group of scientists responsible for its production. In addition to the scientific intent recorded in an experimental proposal, they may have made observations at the time of the data production which could can enhance use of the data or produce a new avenue of investigation. Theses could be associations of events with other phenomena for example lighting strikes with the ionization of a region of the atmosphere or identification of recurrent patterns which would merit further investigation.

Scientists in the Community are the most diverse and distributed. While they tend to be most difficult to assess this is nonetheless an important activity as they may generate and posses a lot of information critical to data reuse.

The data archivist is the group or individual who is the current custodian of the data. The extent to which they have interacted with other stakeholder groups and extracted knowledge requirement with its associated information will be highly dependent on the resources available to, the motivations, background and personal bias of the individual archivist

3.2 Archive Evolution and Management

In addition to identifying the stakeholders from the different categories it is also beneficial to understand how an archive has evolved and been managed. This can used to illuminate the different uses of data over time and the production of associated Representation Information. For example the following sorts of factors have influenced the use and re-use of data over time

- Birth and development of a science
- Events which influence data use such as the second world war or global warming
- Development of countries technologies and the emergence of global networks
- Publication of journals technical manuals, interpretative handbooks, conference proceeding, minutes of user group meetings, software etc.
- Emergence of branches of science and associated organisations
- Stewardship of data and the influence of different custodians

This is not an exhaustive list as many factors influencing data re-use are domain specific as is the categorization of the stakeholders. Naturally most of these can only be expected to be dealt with in the most cursory way in any practical study nevertheless even this can be extremely important in understanding the situation. As after this evaluation you should be in position to scope what types of reuse may be realistically achieved.

If we compare two archives that we examined as part of the SCARP project, that of a single site wind profiling instrument based in Wales and that of a global network of ionosondes which create ionization profiles of the atmosphere.

The Mesosphere Troposphere Stratosphere (MST) [13] data set is extremely well documented and tightly managed. Access to the data is restricted, with end users required to report back on how they have used the data. The Archivist is the key manager of these data for a number of reasons

- He is the project scientist involved in production of the data
- He is a field expert and practising scientist in close contact with relevant scientific organisations
- He provides support, runs and keeps records of user group meetings.

When we consider these factors we can see that that it is reasonable to try to capture information from current users which facilitate the re-use of data by future scientists. This is possible due to the archivist's domain knowledge and close connection to users.

By contrast the ionosonde data e data whilst being a skilled individual with some domain knowledge does not have the same strong connection with current users. The data currently comes from 252 geographically diverse locations and current users are simply required to provide an e-mail address to gain access. As a result it would be completely impractical to capture user generated information even if it might facilitate re-use.

The added value end users or the impact of the absence of such information must be considered in determining the value of the research asset to be created. If creation of such asset is deemed viable an archive may then begin to form preservation objectives and define user communities based on the information in scope.

4. Defining a preservation objective

The analysis carried out before this point may present one with a natural easily defined preservation objective or alternatively there may be a greater number of options which overlap and are more difficult to define. It is important to note that this type of analysis cannot advise you as to which preservation option to choose but merely clarifies the options available to you. Preservation objectives should be

- Specific well defined and clear to anyone with a basic knowledge of the domain
- Actionable the objective should be currently achievable.
- Measurable it is critical to be able to know when the objective has been attained in order to assess if any preservation strategy developed is adequate.
- Realistic based on findings from the previous stages of analysis

We shall now take an example preservation objective from the MST data. We set the preservation objective as follows. A user from a future designated community should be able to extract a specific set of 11 parameters from data files for a given time and altitude. These include typical measurements such as vertical wind shear and tropopause sharpness. We would also 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

This objective has the desired qualities of being specific, actionable, measureable and realistic. While it could be tempting to try and specify a replication of current use this may not be advisable. If we had set the preservation objective as the being the ability to study gravity waves or ozone layering occurring in the atmosphere above the MST site we would rapidly discover that this is too vague an objective. This opens too many avenues of investigation when determining the skill and knowledge base needed to correctly interpret or analyze the data for these purposes. The unfortunate consequence would have been a time consuming analysis process and a lack of certainty that this objective had been achieved for future users.

5. Defining a designated user community

The Designated Community is defined in OAIS [14] as **"**An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities"

An archive defines the Designated Community for which it is guaranteeing to preserve some digitally encoded information and must therefore create Archival Information Packages (AIP) with appropriate Representation Information.

The designated community will possess a skills and knowledge base which allow them to successfully interact with a set of information stored within an AIP in order to extract required knowledge or recreate the required performance or behavior. In common with the preservation objective the analysis up to this point may present one with a range of community groups which the archive may chose serve.

The definition of the skill set is vital as it limits to the amount of information which must necessarily be contained within an AIP in order to satisfy a preservation objective. In order to do this the definition of the designated community must be

- Clear with sufficient detail to permit meaningful decisions to made regarding information requirements for effective re-use of the data.
- Realistic and stable in so far as there is reasonable confidence in the persistence of the knowledge base and skill set.

While the need to define the designated user community is universal, the nature of a knowledge and skill set will tend to be domain specific. The following are typical examples from atmospheric science

- Ability of a community to successfully operate software i.e. knowledge of correct syntax to input commands into a UNIX command line.
- Ability to utilize correct analysis techniques with data to remove background noise or identify specific phenomena
- Comprehension of community vocabularies
- Appreciation of different scientific techniques employed during the production of data, their limitations and comparative success rates for picking up desired phenomena.
- Knowledge of atmospheric events or processes which may be affecting the atmospheric state being measured within a data set.

It is the appraisal of this knowledge skills base as permanent attribute of the designated user community which will determine whether it is necessary to preserve this information by inclusion in an AIP.

If we take an example from the ionospheric data set we can see how the designated community determines what needs to be included within an AIP. The image below is take from an html page which contains a structural description of ionospheric parameters which have been encoded within IIWG formatted files.

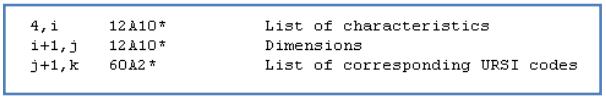

Figure 2. Structural description information

Upon inspection we can see that the current structural description contains FORTRAN notation. If knowledge of FORTRAN is not deemed to be a permanent stable attribute of the community this information must then be included directly within the AIP. This ensures the structural description can be interpreted correctly in the future.

6. Preservation information flows

Once the objective and community have been identified and described an analyst should be in position to determine the information required to achieve an objective for this community. An analyst proceeds by identifying risks which are to be addressed by preservation action. We advocate the creation of an OAIS preservation information flow diagram at this juncture.

An OAIS preservation information flow diagram is graphical representation and analysis tool which is a hybrid of an information flow diagram and the OAIS information model. It provides a convenient format to facilitate group discussions over preservation plans and strategies. A preservation information flow diagram we created for the MST data is shown below

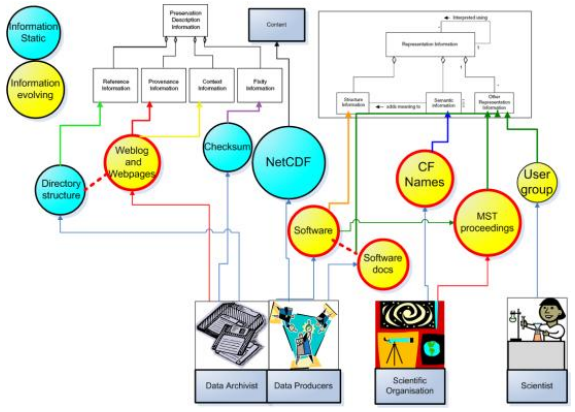

Figure 3. OAIS information flow diagram for the MST data set

The OAIS reference model specifies that within an archival system, a data item has a number of different information items associated with it, each performing a different role in the preservation process. The preservation objective for a designated community should be is satisfied when each item of the OAIS information model has been adequately populated with sufficient information. The information model provides a checklist which ensures that the preservation objective can be met. All information objects must be mapped to at least one of the element of the OAIS information model.

In addition to information objects and the standard OAIS information model the diagram contains a number of other components which we will now examine in turn.

6.1 Information Objects

An information object is a physical unit of information currently utilized in the use or re-use of the content data. This physical unit could be a document, compound object such as a website or indeed be a human being. An information object must have the following attributes

- Name
- Description of information contained by entity which is vital for the preservation objective e.g. a piece of software contains structural information and algorithms for the processing of data within its code
- Description of format i.e. website, PDF, database or software
- Assessment of preservation risks and dependencies

It is essential that the physical entity is identified so supply can be negotiated and risks identified. The analysis process asks you to separate the information which must persist from the physical form which can be change by preservation action in order to obviate risk.

We also found it useful to indicate on the diagram if the information was static or evolving. This allows the preservation plan to easily reviewed so additional information can be included at a later date, for example the inclusion of additional scientific output related to the use of the data which was published after the original AIP was created.

6.2 Stakeholder entities

A stakeholder entity is the named custodian of the required Information entity. It is the individual or organisation with who supply of the information object must be negotiated and confirmed.

6.3 Supply Relationship

The supply mechanism should simply be an indicator of any impediment to the current supply of an information entity such as an embargo or assertion of copyright. The attributes of a the supply relationship are

- Supply possible (Yes/No)
- Description of supply impediment

6.4 Supply Process

The supply process is any process carried out on information supplied by the stakeholder in order to produce the information object. Its attributes are

- Name
- Description of process e.g. dump of a database table into a csv file, archiving of public website or reformatting of data files

6.5 Packaging relationship

The only required attribute of the packaging relationship is that it links an Information entity to at least one standard OAIS reference model component of an AIP. This supports the checklist function which ensures the package a logical or physical AIP contains all the necessary information required to be OAIS compliant, thereby facilitating audit and certification processes.

6.6 Information object dependency relationships

The information object dependency relationship connects two information objects. If preservation action is carried out on one object this may impact on another object with a dependency. For example if a piece of software is identified as being at risk a number strategies present themselves. It may be deconstructed to a structural format and analysis algorithm descriptions recorded in ASCII text files for inclusion in the AIP in preference to an archived version of the software. The software user manual will be flagged up by the dependency relationship and may be removed on the basis that this information is now redundant.

7. Preservation strategies

Once the Information flow diagram has been created an archive must identify suitable preservation strategies in the following areas.

7.1 Strategies in response to a supply impediment

Where there is an impediment to the supply of a required information object strategy must be developed. One can overcome the impediment immediately or alternatively develop a mechanism that effectively references the external information object in tandem with a mechanism for monitoring the situation (preservation orchestration).

The international workshop on MST radar is held about every 2-3 years, and is a major event gathering together experts from all over the world, engaged in research and development of radar techniques to study the mesosphere, stratosphere and troposphere (MST). It was attended by young scientists, research students and new entrants to the field to facilitate close interactions with the experts on all technical and scientific aspects of MST radar techniques. It is this aspect which makes the proceedings an ideal resource to support future users who are new to the field.

Permanent access to these proceedings is again at risk with supply impeded by the distribution and failure to deposit proceedings in a single accessible institution. The MST 10 proceedings are available for download from the internet and from the British Library. Proceedings 3, 5-10 are also available from the British library, meeting 4 is only available from the Library of Congress and unfortunately the proceedings from meetings 1 and 2 have not been deposited in either institution.

A number of strategies present themselves. Copies of proceedings 1, 2 and 4 could be obtained from the still active community, digitised and incorporated into the AIP. The proceedings which are currently held by the British Library can be obtained, digitised and incorporated into the AIP. Alternatively bibliographic records which include the British Library as a location can be obtained and incorporated into the AIP as a reference. This is a satisfactory approach as there is a high to degree of confidence in the permanence of the holdings and the user communities' ability to access them.

7.2 In response to an identified information preservation risk

Information objects must be inspected on a case by case for their individual preservation risk based on dependencies which will be affected by the passage of time. Different strategies which effectively obviate these risks should be developed and evaluated.

I f we take another example from the MST data archives where an information object, the GNU plot software analysis programs is deemed to be at risk . This software extracts parameters and plots

Cartesian product of wind profiles from NetCDF data files.

Preservation risks have arisen due to the following user skill requirements and technical dependencies.

- The software requires a UNIX or Linux distribution the user community may lose access to or the ability to operate these systems.
- A future user may lose the ability to install different libraries and essential software packages python, the with python-dev module , numpy array package or pycdf
- GNU plot can no longer be installed
- The community may lose the technical ability to set environmental variables or run required python scripts through a UNIX command line
- The GNU plot template file to format plot output may no longer be accessible.

A number of preservation strategies now present themselves,

One solution is preserving the software through emulation, using for example Dioscuri [15]. Current work with the PLANETS project should make Dioscuri capable of running operating systems such as Linux Ubuntu which should satisfy platform dependencies. With the capture of specified software packages/libraries and the provision of all necessary user instructions this is a potentially a viable strategy.

It is additionally possible to convert NetCDF files to another compatible format such as NASA AMES [16]. The conversion can be achieved using community developed software and the scripting language Python. This is a compatible self describing ASCII format, information would still be accessible and easily understood as long as ASCII encoded text can still be read. There would be however reluctance to do this as NASA AMES files are not as easily manipulated making it more cumbersome to analyse data in the desired manner.

Preservation by addition of Representation information strategy is an alternate strategy. Capturing NetCDF documentation and libraries from Unidata [17] means that if future user community still have skills in FORTRAN, C, C++ or Java they will be able to easily write software to access the required parameters.

7.3 As a secondary response to a preservation strategy

Where a dependency between information objects has been identified a secondary preservation strategy may need to be developed for the associated object.

As multiple strategies can be developed a number of competing preservation plans are available. A preservation plan should consists of a unique

- Set of information objects
- Set of supply relationships
- Set of preservation strategies

Each plan will allow an archive to carry out a series of clear preservation actions in order to create an AIP. The archive should now be in a position to take a number of plans to the cost/benefit/risk analysis stage where they can be evaluated and a preferred option chosen.

8. Cost/Benefit/Risk Analysis

The final stage of the workflow is where plan options can then be assessed according to

- Costs to the archive directly as well as the resources knowledge and time of archive staff
- Benefits to future users which ease and facilitate re-use of data
- Risks – what are the risks inherent the preservation strategies and are they acceptable to the archive.

Once this analysis is complete the optimal plan can be selected and progressed to preservation action. If no plans are deemed suitable then the process must begin again with an adjustment to the preservation objective and/or the designated community to be served.

9. Conclusions

We believe that this approach is successful at delivering preservation analysis which permits planning at the data set level. It critically allows an archive to establish a process which is comprehensive and aware of all elements required for the re-use of data in the long term.

Wider application, trialling and further development of the preservation analysis methodology would be desirable as would testing its

validity in a broader range of disciplines and organisational settings. In addition the production of training materials and support for archivists who wish to adopt the approach articulated in this paper would be an important next step in promoting its use in the community.

Although the methodology was created due the additional information requirements involved in the reuse of scientific data we believe that this approach could be seen as complimentary to Drambora, PLATTER and Plato. Where the results of analysis could be fed into the audit, risk analysis process and inform repository planning. There is also the potential of integrating the methodology with Plato at the preservation strategy selection stage and it was felt this was an area that would benefit from further research.

10. 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, Matthew Dunckley, Stephen Rankin and Brian McIlwrath.

11. References

[1] Digital Curation Centre SCARP project http://www.dcc.ac.uk/scarp/

[2] CASPAR Project http://www.casparpreserves.eu/

[3] Science and Technology Facilities Council http://www.stfc.ac.uk/

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

[5] World Data Centre for Solar Terrestrial Physics www.ukssdc.ac.uk/wdcc1/wdc_menu.html

[6] European Incoherent Scatter Radar http://www.eiscat.rl.ac.uk

[7] Drambora http://www.repositoryaudit.eu/

[8] Planets http://www.planets-project.eu/

[9] PLATTER http://www.digitalpreservationeurope.eu/platter/

[10] Plato http://www.ifs.tuwien.ac.at/dp/plato/intro.html

[11]CASPAR Questionnaire http://www.casparpreserves.eu/Members/cclrc/ReferenceDocuments/caspar-test-case-questionnaire/

[12] CASPAR Ionosonde case study http://www.casparpreserves.eu/other-caspar-products/other-caspar-products/ionosonde-case-study.pdf

[13] The Natural Environment Research Council (NERC) Mesosphere-Stratosphere-Troposphere (MST) Radar at Aberystwyth http://mst.nerc.ac.uk/

[14] Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1. Jan 2002 http://public.ccsds.org/publications/archive/650x0b1.pdf

[15] Dioscuri on Sourceforge http://dioscuri.sourceforge.net/

[16] NASA Ames http://badc.nerc.ac.uk/help/formats/NASA-Ames/

[17] NetCDF document library http://www.unidata.ucar.edu/software/NetCDF/docs/