Planning Digital Preservation for e-Health using Preservation Network Models

Michael WILSON[1], Esther CONWAY[1], Arif SHAON[1], Vasily BUNAKOV[1]
[1]Science and Technology Facilities Council, Rutherford Appleton Laboratory, Didcot, OX11 0QX, United Kingdom
Email: {michael.wilson, esther.conway, arif.shaon, vasily.bunakov}@stfc.ac.uk

Abstract: Health records and data can be preserved for the benefit of patients, their families and future medical research. The costs of storage itself are reducing, but the costs of ensuring the usability of the data to different audiences remain high. In order to make the preserved data usable, supplementary information to explain it and provide context in which to interpret it must also be preserved. The relationships between the target data and this supplementary information can be represented in Preservation Network Models (PNM). The utility of PNM are considered for planning the preservation of health data.

1. Introduction

e-Health generates large amounts of data in many different forms - e.g. body scans and tissue sample images, traces of monitors, patient records. There are both operational needs and regulatory requirements to preserve this data for the good of individual patients themselves in the future and for their potential use in future epidemiology and medical studies. Computing storage costs have halved every 18 months for several decades and are expected to continue to do so. However, digital preservation involves more than just storage, and the preservation process gives rise to costs which must be weighed against the benefits of preservation to justify the expenditure.

One approach that has been proposed to planning digital preservation (Conway et al, 2011) involves assessing the risks of future events that may limit the usefulness of the preserved digital information for its expected purpose, and selecting to preserve that digital information which will reduce the risks to an acceptable level given the costs. This approach uses preservation network models (PNM) of potential preservable digital information to propagate and aggregate these risks as a basis for making this decision. Technology to implement this PNM based preservation planning approach is being developed in several current EC funded projects to make the approach easier to use in a variety of domains.

This paper describes the approach to digital preservation planning using PNMs in the e-health domain and the technology which is being developed to support it in order to make preservation planning accessible and usable.

2. Objectives

The objective of the paper are: to explain how digital preservation is more than digital archiving, to show how the PNM based preservation planning process operates, describe how the technologies being developed will make it accessible in e-Health, and to encourage further adoption of digital preservation beyond data archiving, and preservation planning as a business activity.

3. Methodology

The preservation planning methodology used here is that proposed by Conway et al (2009), based on the ISO OAIS (2002) standard for preservation infrastructures. It divides the overall preservation process into the stages shown in Figure 1 below.

4. Technology Description

The main difference between digital backup or archiving, with which we are all familiar, and digital preservation occurs in the preservation actions. Once backed up or archived, data are normally left alone until they are wanted for use. In digital preservation, while the data are archived two checks are made. Firstly, fixity checks are regularly performed on them to ensure that the bits of data are being preserved and if there is any bit loss, then additional archived copies of data are restored to ensure the current copy is correct. Such checking and restoration are performed by all current digital preservation systems. Secondly, checks are made on the outside world to ensure that the expectations made during preservation planning are still correct. The scope of such checks, and the implementation of actions resulting from them vary between different commercial preservation systems and between research systems. All systems support some notification of potential obsolescence of a format in which data are stored (e.g. Microsoft announce that MS-Word V1 format will no longer be supported), which trigger an action to migrate data held in this format into one which is supported (e.g. migrate all documents held in MS-Word V1 format into the current version of PDF). Other forms of event monitoring or external watch processes will check for changes to policies or legislation relating to data security and encryption (e.g. to move data stored in an encryption algorithm which is no longer safe to a new one), changes to policies or regulations which set the duration of preservation (e.g. corporate decisions to delete e-mail after a period, or changes to court discovery regulations on patent cases extending the period of electronically stored information required to support a patent claim to 25 years). The useful range of watch events is currently an active research topic, so alternatives are being considered by different teams.

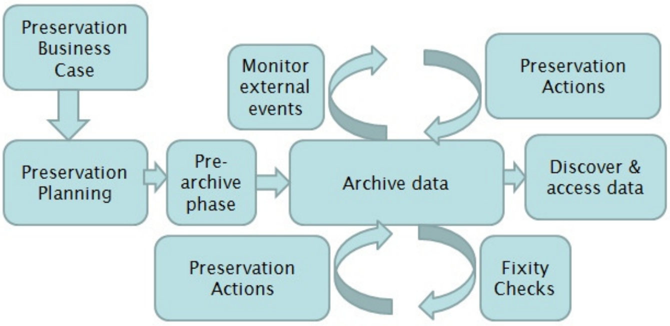


Figure 1: The overall preservation process

The overall preservation planning process is divided into two phases: The development of the general business case justifying the cost effectiveness of the digital preservation activity, and the generation of the detailed preservation plan at the operational level, giving all the details required to archive the data, monitor events and take the preservation actions required when those events occur.

5. Developments

The development of the preservation business case is based upon the selection of a preservation strategy. In OAIS terminology, the strategy will define what data are to be preserved in order to meet the *preservation objective* of the *designated community*. There are four main alternative strategies, in addition to combinations of them, to choose between, to reduce the risk of failing to preserve the data as required:

1) preserve the data in its current format and assume that it will be usable when required (archive strategy);

2) preserve the data in its current format and a description of the format which can be used by a generic emulation tool to use it when required, and assume that the emulation tool will be available when required (emulation strategy);

3) migrate the data to one of a few supported formats which can be migrated again later if potential format obsolescence is identified in order to ensure that the data is always in a format which can be used (migration strategy);

4) store the data and the tools to use them, and the libraries, operating systems, etc. which those tools depend on, and virtualise the platform to run the tools on the data when required (virtualisation strategy).

Each of these strategies has strong advocates and we are not promoting a single strategy either in general, or for specific types of e-Health data. The distinguishing feature of digital preservation explained above is that it monitors changes, and the environment in which preservation takes place is changing, so we must design any preservation planning system to manage change too. For example, a US court ruling defined migration as failing to preserve the data, thereby eliminating strategy 3 (WL 1952680 (Fed.Cl.) 2007). However, the scope of this decision is unclear and will be further clarified in future decisions, and it may be overruled by higher courts. The important point is that all four strategies are possible, and they each require the preservation of different sets of information along with the original data and consequently, each result in different risks of the preservation objective not being met, depending on what changes happen in the world, or more precisely, which events occur which will impact on the future users' ability to use the data to achieve their objective.

When choosing a preservation strategy there is a probability of different types of events happening (e.g. software obsolescence, data format obsolescence, policy or legislation changes) each of which will have an impact on potential user being able to perform their task on the data. Therefore there is a risk of failing to achieve the preservation objective associated with each potential event. Different events are associated with different objects being preserved to support different strategies. Therefore the risks associated with each strategy can be aggregated to provide the overall risk for that strategy. The risk of the strategy failing can be used in conjunction with the cost profile of the strategy and the return on the investment (ROI) in the preservation to determine which strategy to select. No attempt is being made to automate the strategy selection itself but merely to calculate the overall risk, the cost profile and the ROI for each strategy so that a human can select the appropriate strategy. The interaction of these concepts resulting in the strategy selection decision is shown in Figure 2 below.

The second phase of the preservation planning process following the generation of the business case, is the generation of the detailed preservation plan. The structure of this follows the same structure as the business case planning, but instead of calculating the risks and costs for a preservation strategy, they are calculated for the detailed implementation of a plan where the exact storage service, frequency of fixity checks, migration transformations etc... are all calculated. Once again, the user makes the choice of preservation plan based on the cost, ROI, risks associated with each plan. Only the best options for each and the best overall options are presented to the user in order to make the human task clearer.

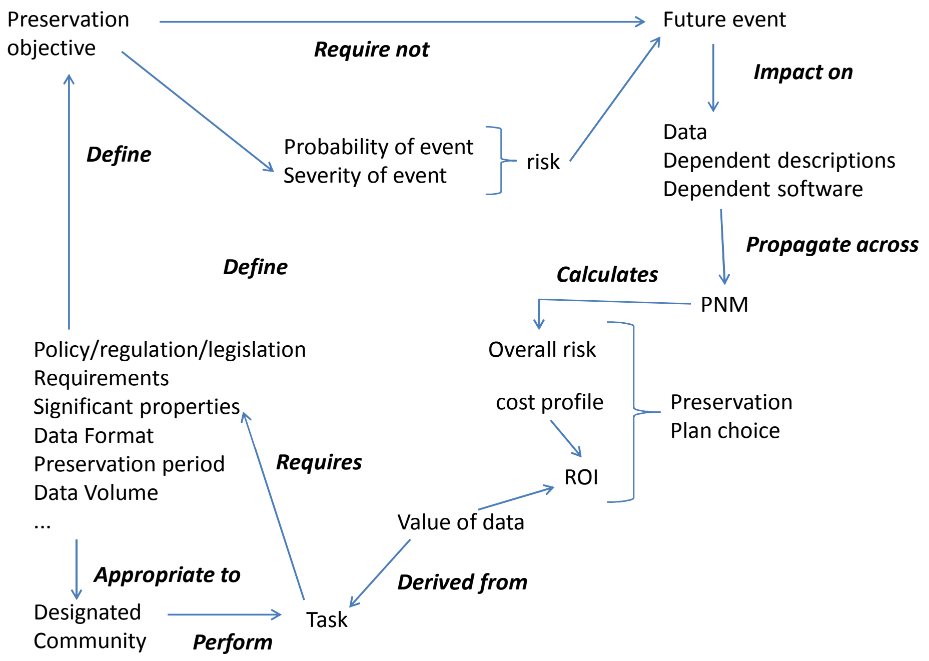
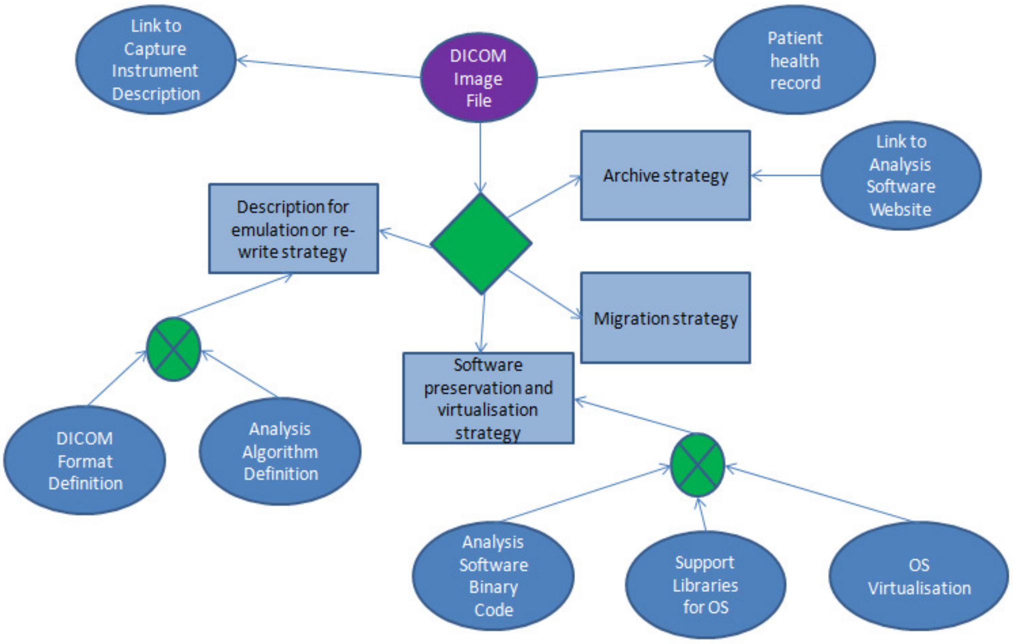


Figure 2: The flow of information in choosing a preservation plan



Link to Capture Instrument Description
DICOM Image File
Patient health record
Archive strategy
Link to Analysis Software Website
Description for emulation or re-write strategy
Migration strategy
Software preservation and virtualisation strategy
DICOM Format Definition
Analysis Algorithm Definition
Analysis Software Binary Code
Support Libraries for OS
OS Virtualisation

Figure 3: Preservation Network Model for a DICOM image file in the e-Health domain. DICOM is a format supported by the preservation service so the migration strategy is not an option. For simplicity the sub-networks under descriptions are not shown.

6. Results

Both the generic approach and specific tools are under development in several on-going projects, where the tools being developed to support this methodology include:

1) PNM editor to generate RDF representations of the preservation network model;

2) planning tool to generate alternative preservation plans from the PNM;

3) cost modelling tool to estimate the cost of each preservation plan based on its activity structure;

4) ROI tool to estimate the value of the data, and from that the return on investment;

5) Risk aggregation tool to infer the overall risk of a plan or preservation strategy;

6) Preservation plan selection tool, as a GUI to support the choice of strategy or plan.

7. Business Benefits

In order to meet the regulatory obligations in the health sector, e-health providers will have to preserve digital data. In order to justify particular preservation projects they will need to make suitable business cases and preservation plans that will enable their designated communities to meet their preservation objectives. To do this they need to adopt a method, supported by tools, which will manage the risks within an acceptable cost envelope. The preservation planning approach incorporating multiple preservation strategies and preservation network models presented here addresses these issues. Several current EC funded projects are implementing tools to support this approach, and once that tooling is available, the e-health community should be prepared to trial it and consider standardisation of their approach to this issue.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270000 to the ENSURE project, grant agreement n° 270137 to the SCAPE project and grant agreement n° 283401 to the SCIDIP-ES project.

References

[1] E Conway, MJ Dunckley, D Giaretta, B Mcilwrath (2009) Preservation Network Models: Creating Stable Networks of Information to Ensure the Long Term use of Scientific Data, In Proc. Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data (PV 2009), Villafranca del Castillo, Madrid, Spain, 01-03 Dec, URL: http://epubs.stfc.ac.uk/bitstream/4314/PV09_Conway_PNM.pdf

[2] E Conway; B Matthews; S Lambert; D Giaretta ; M Wilson; N Draper (2011) Managing risks in the preservation of research data with preservation networks. In Proc. 7th International Digital Curation Conference (IDCC2011), Bristol, UK, 05-07 Dec,
URL: http://epubs.stfc.ac.uk/bitstream/7171/Final_ManagingRisks_Conway_IDCC11.pdf

[3] OAIS (2002) The Reference Model for an Open Archival Information System (OAIS), Consultative Committee for Space Data Systems (CCSDS), URL: http://public.ccsds.org/publications/archive/650x0b1.PDF

[4] WL 1952680 (Fed.Cl.) 2007, United Med. Supply Co., Inc. v. United States, 73 Fed. Cl. 35,
URL: http://sos.mt.gov/Records/committees/erim_resources/G%20-%20United%20Medical%20Supply.pdf

[5] Stephan Strodl, Christoph Becker, Robert Neumayer, and Andreas Rauber. (2007). How to choose a digital preservation strategy: evaluating a preservation planning procedure. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (JCDL '07). ACM, New York, NY, USA, 29-38. DOI=10.1145/1255175.1255181