Ensuring profitability of commercial long term digital preservation

by Stephan Kiefer (Fraunhofer IBMT) and Michael Wilson (STFC RAL)

Financial and health records are often stored in record management systems which ensure that any changes are auditably recorded to justify that the retrieved information represents that deposited, within the constraints of regulatory requirements. When the regulated retention period expires, the records are normally deleted. In contrast, data from publically funded science is becoming preserved for the long term in repositories, to be open to discovery and access for future uses for which it was not originally collected. The experience gained from preserving scientific data over the long term is being transferred to the preservation of commercial data.

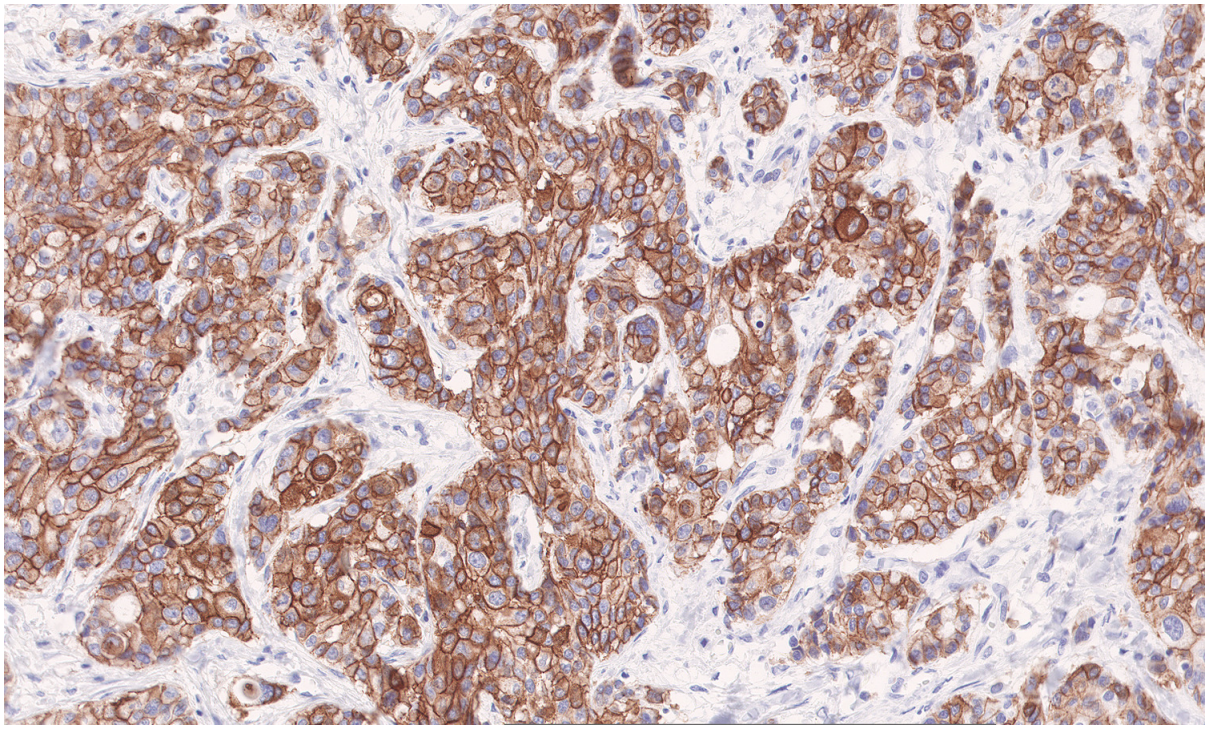
Data preservation goes beyond backup or archiving in order to manage change: to the data itself resulting from corruption, in the technology which will be used on the data, and in policies applied to the data. These changes are respectively managed by: performing fixity checks on the data and replacing corrupt data, using emulators of the original technology or transforming data to formats which new technology will support, and changing the preserved data or its environment to conform to new policies.

In applying preservation technology to commercial data, it is necessary to ensure firstly, that its preservation over the long term for new uses justifies the investment required [1], and secondly that the data can be discovered and accessed to meet those new uses. The EU funded ENSURE project has demonstrated a solution to the long term digital preservation of commercial data which addresses both of these issues for financial records, health records and clinical trial data. Given the current technology environment, ENSURE software supports clouds for the data storage and for the computation needed by the preservation functionality.

The approach taken to justifying profitability is to calculate the return on investment by subtracting the value of data from the cost of preservation. Cost of preservation is calculated from a hierarchical decomposition of the activities required for preservation on clouds, which can be costed individually at a low level, and then the costs can be aggregated to provide an overall cost. Private clouds have larger initial costs than commercial clouds, so the investment profiles differ accordingly.

Calculating the future value of data is a more contentious issue. Data valuation is normally based on a combination of the cost of collecting the data, the cost of not having the data, and estimates of the potential business revenue that the data could yield. The last of these can only be determined by considering potential uses of the data by a business. For scientific data there are several examples of datasets collected several hundred years ago for other purposes which are now proving invaluable in modelling and predicting climate change - a purpose which was not considered when the data were collected. Similarly, in predicting future uses of commercial data such low probability high impact possibilities must be included. For example, health records can be preserved for the benefit of patients during their lifetime, and for the benefit of their descendents in the future diagnosis of inherited illness, but also for use in future epidemiological studies and future medical research. Such future uses may have low probabilities, but their high potential impact contributes to the potential value of the data, and thereby to the return on investment calculations.

Predictions of future uses of data can only come true if the data will be discovered when it is required. Data that is indexed by its current role alone will not be easy to discover when required to meet new roles. ENSURE represents metadata for the data objects in terms of formal ontologies [2]. This supports the modelling of preservation knowledge as well as domain specific object formats and concepts effectively, in an application oriented way. The ontologies contain concepts describing general features of data objects as well as domain specific information. The metadata for data objects represent instances of the ontologies which are stored by the ENSURE software as an index. In order to ensure accessibility of structured and unstructured data by future users according to business oriented search criteria, a semantic search and query mechanism forms part of the access component. It leverages the semantic index created by the ingest component according to the ontologies in an ontology framework. The ontology framework includes an ontology registry populated by a set of initial preservation related ontologies. The ontology framework offers the flexibility required to serve future data retrieval needs. It also provides a platform to research how the evolution of ontologies over time can be exploited, both to trigger data transformations to ensure the long-term usability of the data, and to provide the basis for updating predictions of future data value to ensure the profitability of the preservation activity.



Caption: DICOM format image of a pathology slide from a patient's digital health record preserved by ENSURE. Credit: Philips Healthcare.

References

[1] E Conway et al. (2012) Managing Risks in the Preservation of Research Data with Preservation Networks, International Journal of Digital Curation **7** (1) 3-15

[2] S Kiefer et al. (2011) An Ontology-Driven Search Module for Accessing Chronic Pathology Literature. In On the Move to Meaningful Internet Systems, LNCS 7046, 382-391, Springer: Berlin, Heidelberg

Useful Link

ENSURE project website http://ensure-fp7-plone.fe.up.pt

Please contact

Michael Wilson, STFC. UK

Tel: +44 1235 446619

E-mail: Michael.Wilson@stfc.ac.uk