Extended Abstract for Open Repositories 2008 Conference

Best of Both:
Connecting a institutional digital repository and a digital preservation service

Libby Bishop
University of Leeds, Senior Research Archivist
University of Essex, Research Liaison Officer
e.l.bishop@leeds.ac.uk
01206 872664

Institutional digital repositories are growing at a rapid rate. Diverse pressures such as research councils mandating deposit, funders hoping to capture intellectual capital, researchers' desire for greater control over output dissemination, and the growth of open access initiatives are all contributing to this successful expansion. With expansion, however, come new challenges. Three areas identified in the recent JISC digital repository review (add cite) include handling data and other file formats in addition to text and pdf research outputs, sustainability and long-term preservation, and better integrating repositories into researchers' work practices and the life cycle of research projects. A newly launched project, Timescapes, will create an archive of qualitative longitudinal data based at the University of Leeds and has been designed to develop and test solutions in these three areas.

Timescapes is a £4.5 million, five year ESRC-funded study designed to shed light on the dynamics of personal relationships over the life course, and the identities that flow from those relationships. A key objective of this initiative is methodological in nature: to establish a working archive of data derived from the empirical projects as a valuable resource for sharing within the social scientific community and for future historical use. This paper will first address the distinctive features of the Timescapes archive and its design process. Secondly it will discuss the implementation strategies used to date in the project. Finally, it will report on successes and challenges encountered so far. The project is still in its first year so all findings are tentative. However, it is hoped that by sharing our experience early, the project will benefit from others' experience and perhaps also be of some value to others.

Distinctive Features of the Timescapes Archive

The Timescapes Archive will have two components: there will be a digital repository at Leeds that will be the receiving facility for incoming content. This repository will be an extension of the existing MIDESS system at Leeds and explicitly designed to accommodate multi-media file formats. This repository will support data preparation, metadata enhancement, and data sharing, both within the Timescapes team and with other authorised users. The repository will send appropriately prepared (e.g., compliant with OAIS and DDI standards) to the UK Data Archive at the University of Essex for preservation. Dissemination versions of files (whether produced at Leeds or UKDA) will be managed

from the repository. Thus the repository will have primary responsibility for ingest and dissemination with the UKDA handling preservation.

The main reason for considering such an option was that the Timescapes archive needs a broad range of functions, not all of which have (typically) been provided by either a repository or an archive alone. Repositories don't usually focus on long term preservation. In contrast, however, preservation is a core competence for UKDA. However, the UKDA is not set up to allow depositors to have fine-grained dynamic control regarding sharing the content they deposit. It assumes that data are no longer in active use ("fixed") prior to deposit. (This matters because costs of preservation are much higher for continually changing content.) The ability of depositors to control access is a function the Timescapes facility requires.

Many other factors contributed to this design decision and will be elaborated in the full paper. Examples include a desire to avoid wasting resources by replicating preservation services and the desire to embed the Timescapes archive into the wider Library and IS infrastructures at Leeds.

A key element in this design choice was the fact that somewhat similar models already existed. A model exists (Sherpa DP2) for connecting one or more IRs to a single preservation service. This project has been funded by AHRB and has preserved cultural and historical materials.

There are a number of distinctive features of this project, both in the substance and content of the materials to be processed and also in the proposed design. Some of these will be listed and then discussed briefly here, with fuller explanations provided in the final paper.

- The primary file formats to be processed are data, including multi-media data (images, audio and video).
- The data are qualitative and longitudinal.
- The combination of sensitive content (family life) with longitudinal data creates a need for secure systems and complex access conditions and rights management.
- There is a strong domain focus for the initial data to be archived.
- The project is explicitly designed to integrate three strands of activity:
    - active research
    - preparation of data for preservation, and
    - reuse of the preserved data by the core team and by external users.
- The UKDA seeks to explore diverse ways of supporting repositories and creating a disaggregated preservation service for digital data is a key activity.

There are some particular benefits of using the IDR for ingest, especially for the kinds of data that Timescapes will produce. Qualitative data makes heavy demands during the pre-ingest and early processing phases. Because it is *data*, it needs more extensive metadata and contextual material to render it "independently understandable" than textual research outputs). Because it is also *qualitative* (and longitudinal), there are complex confidentiality requirements and access provisions. In such situations, the repository manager alone can not adequately prepare the data; there must be extensive and positive (not minimal) engagement

with researchers if the data are to be prepared in a way to make preservation and reuse optimal. Timescapes will create a setting where IDR staff can work with researchers "at the coalface" to gather and process IP and metadata.

Another area of distinctiveness is even more innovative. Timescapes intends to have primary researchers play a very active role in selecting which data get preserved and in promoting its reuse. Because processing qualitative data is so labour intensive, it is very expensive to waste resources preserving materials that don't get used. Estimating potential for reuse is more art than science, but both archivists and researchers have essential contributions to make to the discussion. Moreover, the IDR could also broker relationships between primary researchers and potential reuses by advising on collections, pointing out rich, unmined data, or connecting research teams. Qualitative data tends to benefit from this kind of extra promotion, in part because the culture of reuse is not as widespread as it is for quantitative data.

Implementation experience and strategies

This section of the paper will review some of the key implementation steps of the Timescapes archive. Key points will include benefits of the vast range of resources available (JISC, ESRC, IASSIST, personal networks) but also the difficulties in quickly navigating such a vast terrain.

Other keys issues will be highlighted, such as the need to choose between a partially developed software platform or "building from scratch". This overlapped with the need to choose between proprietary software (Ex Libris Digitool) and building a new Fedora or D-space based open repository. Although the MIDESS repository is using commercial software, in all other respects, we intend to pursue an open strategy regarding, for example, complying with open metadata standards.

Success and challenges to date

The Timescapes project formally started in January 2007, however, there were delays and the archivist was not in post until late summer. The formal launch is not until January 2008. The intent of the paper, therefore, will not be to present conclusions, but to share work in progress. Some insights, none novel, are apparent already and will be elaborated and deepened in the final paper.

1. Complex institutional collaborations may have benefits (especially for sustainability) but are very time-consuming to develop.
2. For IDR staff to gain active participation from researchers, they must have credibility with those researchers.
3. There are deep, if not irreconcilable, tensions, between some practices of qualitative research and system infrastructure design requirements. A trivial but illustrative example is that for some forms of research, it is not possible to say how much data will be collected. The variation can be high when large audio and video files are at issue. Estimating repository costs (where unit pricing is in place) or even simple storage requirements, requires diplomacy in interacting with both researchers and IS staff.

References (partial list and incomplete citations):

Allen, J. Interdisciplinary Differences in Attitudes towards deposit in institutional repositories.

Bishop, L. Archiving for the future: the archivist as researcher. 2007.

Gibbs, H. DISC-UK Data Share: State of the art review, Sept. 2007.

Green, A. Connecting the dots among digital repositories, data services, and social science researchers, ICPSR OR Meeting, Oct. 2007.

Green, A. and Gutmann, M. Building partnerships among social science researchers, institution-based repositories and domain specific data archives. 2007.

Heery, R. and Anderson, S. Digital Repositories Review, Feb. 2005.

Johnson. http://www.dlib.org/dlib/november02/johnson/11johnson.html

JISC "Digital Repositories: dealing with the digital deluge". June 2007.

Knight, G. Sherpa-DP OAIS report: An OAIS compliant model for disaggregated services. Aug. 2005.

Lavoie, B. The open archival information system reference model: introductory guide. Jan. 2004.

Lyon, L. Dealing with Data: roles, rights, responsibilities and relationships. June 2007.

MIDESS WP2 – Functional and Technical Requirements Specification.

MIDESS WP5 – Digital Preservation Requirements Specification.

Sherpa-dp wp2 report (various Sherpa DP and DP2 outputs)

Conference details:

Submission Deadline: Friday 7th December 2007
Notification of Acceptance: Monday January 21st 2008
Conference: April 1-4, 2008. University of Southampton, UK.
Program Committee Chair (e.lyon@ukoln.ac.uk) or General Chair (lac@ecs.soton.ac.uk)