An institutional digital repository with a disaggregated preservation service:
3 January 2008
Version: incorporated comments from Michael Emly Nov 07

Remit: To create a facility for the Timescapes qualitative longitudinal dataset that enables rich data sharing and secondary analysis, through finely grained search facilities as part of a devolved resource, and also permits and supports preservation at the UK Data Archive.

Background and objectives:

Extract from original Timescapes bid:

> A key objective of this ESRC 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. The first step in this process is to address the practical, ethical, legal and epistemological tasks of archiving, representing and contextualising the Timescapes dataset.

There are a number of features of the Timescapes dataset that distinguish it from most UKDA qualitative data:

1. One important aspect of Timescapes is that fact that active research, sharing of "live" data, preparing for preservation, and conducting secondary analysis will happen nearly simultaneously, rather than in a sequential fashion. Therefore, the solution adopted should support both data sharing and preservation.
2. A second important feature concerns the requirement that data be "independently understandable". In addition to fully supporting this requirement, Timescapes is also committed to exploring the potential for enriching qualitative data, its context and its reuse by fostering relationships among primary researchers and between primary and secondary researchers.
3. Finally, the Timescapes programme includes seven projects and the resulting dataset will integrate content from all the projects. Although some activities (e.g., metadata assignment) may need to happen at the project level, plans for both data sharing and preservation must handle the dataset in its entirety as well as possibly supporting differential treatment for subsets of project-level data.

Working specification:

File types and formats

- Any type of qualitative longitudinal data must be supported, with an emphasis on research data, not only outputs.
- Items may include diverse file types: text, image, audio, video, etc. Materials might include raw data, analysed data, researcher-generated contextual

information (e.g. fieldwork notes), papers, analytical documents (e.g., coding schemes), files in CAQDAS application formats, etc.
- A size estimate for the initial Timescapes dataset is 5000 files (400 respondents, each will produce multiple files per session and across time; one 100 minute interview will produce a 1G .wav file); affiliated projects will contribute data as well; this estimate is conservative and certain to increase.
- Files can be viewed by collection and advanced search includes various criteria (project, case, date, type, theme etc.)

User types and activities

Different users will need to have different rights and that access may vary across the individual projects and the integrated dataset. These users and activities are examples only and will be modified once the "levels of access" document is made final.

- Guests (unauthenticated users) can search and browse archive-generated metadata (catalogue) and public items.
- Timescapes associates (authenticated users who are not depositors but who do have rights to use the dataset) can search, read and download data.
- Timescapes affiliates (authenticated users and depositors, with assigned rights to the dataset) can do above plus upload items, and assign metadata.
- Timescapes members (depositors) will have rights to the integrated dataset and the ability to fully manage (upload, share, assign rights) to their own project materials. Software is complex, so some of this may be via administrator, but rights are under Timescapes control.
- Depositors can be authenticated users on collections not their own (where they have been granted rights).
- IDR Administrator can add and remove users and groups of users and add and edit metadata.
- Metadata editing after initial upload requires Windows client, so will be restricted to the IDR Administrator.

Depositing and file sharing

These issues need to be reviewed in light of the need for controls at two levels: rights project leaders will need to manage data in their private restricted areas of the IDR and the creation and sharing of the Timescapes integrated dataset.

- Support for rights management:
    - Assign rights by individual file but not by folder,
    - Create groups for assigning of file rights.
- Simple web interface for file deposit, rights assignment, and initial metadata creation as part of initial upload.
- Interface should also enable assignment of metadata by Timescapes and IDR staff, though some operations will require Windows client.
- Batch upload of objects with web interface.

- Additional metadata will be required to conform to UKDA DDI requirements. DDI schema can be added to system.
- Depositors can add items to their collections; administrator can move items amongst collections.
- Each item should inherit appropriate collection-level metadata. Templates can be used to enable this requirement.
- Administrator (but not depositors) will be able to update content or metadata and remove objects.

Access (user rights and authorisation procedures)

- Depositors potentially located any where in the world.
- A user can search metadata and public items without logging in.
- All further activity requires logging in.
- The login will require authorisation, such as Shibboleth.
- First time users will be required to register with name, organisation and email address. A web registration process is acceptable.
- Administrator can create collections, or grant access to existing collections where authorised to do so by the owners of those collections.

Search capabilities

- Search of metadata fields available to public without log in.
- Search enabled at file (item) level (e.g., find all interviews with 12 year olds)
- Simple search interface with Boolean operators (e.g., find male 12 year olds).
- Initial search categories defined by metadata.
- Expandable to additional search fields (e.g., if themes such as "fatherhood" are assigned to each interview, it should be possible to search by such themes).
- Browse features – need to be specified but existing functionality is adequate.
- Thumbnails of JPEG images (and sampler for audio) in search and browse results.

Relationship with and file transfer to UK Data Archive

This section concerns the handling of the integrated Timescapes dataset. Additional provisions may need to be made for individual project level data. Agreed that these processes can be developed or undertaken manually – need to clarify how metadata and object link can be exported. See note at end.

- The IDR administrator will indicate when a version of the dataset is ready for preservation.
- The dataset can then be transported to UKDA for preservation.
- The transport must support UKDA defined standards for data security and provision of adequate metadata.
- Standard UKDA procedures will apply for verifying data.

- Updates and new versions of data will need to be sent to the UKDA (possibly two editions in five years, one at mid point (2.5 years) and one at end of the five years).

Dissemination (downloading data)

- All preserved materials will be available through the UKDA catalogue.
- A "Leeds/Timescapes" branded interface also needs to be created. Frames may be an alternative solution.
- This dissemination interface will need to be a part of the IDR at Leeds.
- Authorisations, user agreements, etc. must be established and consistent with UKDA standards.
- Authorised users should be able to select files for download based on search results (e.g., download all files from Project 1; or all interviews with male 12 year olds, or all files marked up with a keyword such as "fatherhood"). Need to clarify download functionality – see note at end.

Metadata standards

- Support for OAI-PMH to expose metadata for external harvesting.
- Seamless integration with UKDA input programme requirements (DDI, Dublin Core, others?).
- Support for Z39.50
- Support for OAIS and persistent URLs; need to clarify best mechanism to do this.
- Support for METS, though METS may need more detailed specificity to be useful.

Backup, security and management statistics

- Full data security provided at the IDR. Planned upgrade will provide redundancy.
- Backup and recovery enabled.
- Statistics are available to measure activity of the repository (downloads, page hits, unique visitors, OAI requests, most popular items and collections, etc.).

Other requirements

- The IDR must have Timescapes and Leeds branding especially for deposit and dissemination functions.
- Resource discovery, deposit and initial metadata upload to be enabled through a user-friendly web interface. Metadata editing, file relocation, and rights assignments will require the Windows client and be done by the Administrator. Frames may be an alternative solution.
- A substantive part of the solution must be physically located at Leeds.

- For any proposed solution, full, long-term costs (especially costs linked to quantity of data deposited and numbers of users) and staffing implications must be considered.
- Timeline and targets:
    - Demonstration of DigiTool, including procedures for depositing and accessing data, ready by 31 Jan 2008 (Timescapes launch).
    - Service ready to support Timescapes project by 31 May 2008.
    - Link to UKDA and preservation completed for test data by Dec 2008.
    - System complete, including multiple channels for dissemination by June 2009.

Notes from Michael

In discussion with Libby, there seemed to be 3 key issues:

1. Ability to download the data stream. For file formats like zip, download would be the only delivery option. For some file types, an application would be invoked which offers a save facility (Excel, Word, JPEG). However for audio and video, this is more problematic. I am awaiting further information from Ex-Libris.
2. Exporting records/data. OAI-PMH harvesting only gives limited metadata. I am awaiting further information from Ex-Libris.
3. Metadata schemas and search of individual metadata elements. Additional schemas can be added, though this is not a completely straightforward process: it is particularly important to specify the rules mapping elements onto the various indexes. If this is done, then it is possible to change the search options according to the collection searched, via JavaScript or by providing specific search boxes outside the main DigiTool interface (e.g. via frames)