August 2014

**Changing Landscapes for the Third Sector: Enhancing Knowledge and Informing Practice**

# Report on the Timescapes Archive

*Bo Middleton, Library Head of Research Support Services*[1]

*John Beaman, Systems Team Officer, University Library Systems Team*

*Graham Blyth, Interim Technical Lead, Research Data Management Service*

*Kahryn Hughes, Senior Research Fellow, School of Sociology and Social Policy*

*Bren Neale, Professor of Life Course and Family Research, School of Sociology and Social Policy*

*Brenda Phillips, Research Data Officer, Research Data Management Service*

*Rachel Proudfoot, Interim Coordinator, Research Data Management Service*

*John Salter, Systems Team Officer, University Library Systems Team*

## Introduction

The broad aim of the Changing Landscapes project, as set out in the proposal to ESRC, was "to bring together a body of qualitative longitudinal and life course research on the third sector in order to exchange knowledge and data of relevance to the future development of the sector". Underpinning this aim is the Timescapes Archive – a specialist archive of Qualitative Longitudinal (QL) data for sharing and re-use. The long term strategy for the archive is to build collections of thematically related QL datasets, including non ESRC funded datasets, in order to facilitate data discovery and secondary analysis across a range of substantive topics. This is in a context where QL methodology is fast advancing and a growing number of projects are being funded. The archive originally contained a collection of 9 datasets (Changing Relationships and Identities through the Life Course - short hand title, Changing Relationships and Identities). Under the new funding, the aim was to develop a new collection of datasets (Changing Landscapes for the Third Sector). The specific objective was to "prepare data from two complementary datasets (NCVO and Birmingham) and ingest the data into the Timescapes Archive" (ESRC proposal).

This report describes the work undertaken in order to achieve the objective of adding two datasets to the Timescapes Archive, but also details the important development work undertaken to establish the Timescapes Archive on a new technical platform which will support the long term strategy for the Archive and ensure that the Archive is aligned with the University of Leeds institutional data management provision.

The development work carried out over the past year is detailed below, but has included the following key activities:

[1] Corresponding author

- Migrating the existing collection of 9 datasets (Changing Relationships and Identities) to the EPrints platform.
- Setting up a new search and browse function for the collection, to aid data discovery and facilitate re-use.
- Modifying EPrints to create an access control layer suited to QL data.
- Creating a new collection (Changing Landscapes for the Third Sector) comprising two datasets, for ingestion into the new platform.
- Provision of support and guidance to the depositors (NCVO and Birmingham) on data management and preparation of the datasets for archiving.
- Creating a new Guide to the Timescapes Archive for both depositors and users.
- Ongoing collaboration with the UKDA to ensure compatible, complementary systems are in place.

## Migration of data across platforms

The Timescapes Archive was originally built on a Digitool[2] platform. Although this platform was previously used to host the University of Leeds' digital image collections, over the last few years all digital collections have been migrated to an EPrints[3] platform and a decision was taken to decommission the Digitool platform. In addition, over the same time period, the University of Leeds has been running a project to identify options and implement a research data repository. In April 2013 the Changing Landscapes project was awarded ESRC funding; this was shortly after the University of Leeds had decided to implement a pilot research data repository on the EPrints platform. In order to secure the future of the Timescapes Archive, the Leeds Research Data Management Team[4] has worked with the Changing Landscapes Project academic team to migrate the Timescapes Archive to an EPrints platform. The datasets provided an excellent case study for working through the logistics of archiving data on EPrints and also the logistics of establishing a University of Leeds research data management archiving service. A decision was taken early on in the Changing Landscapes Project to ingest the additional data sets from NCVO and Birmingham directly into the new platform rather than ingesting them into the Digitool system and then migrating to the new platform.

[2] Digitool is commercial digital asset management software supplied by ExLibris, http://www.exlibrisgroup.com/category/DigiToolOverview

[3] EPrints is open source software commonly used as a platform for University repositories housing research publications, http://www.eprints.org/

[4] Leeds secured Jisc funding for a preliminary research data management project, RoaDMaP (Jan 2012 – June 2013); subsequent to this, the Research Data Management Team and infrastructure has been supported by the institution

## Relationship between the institutional data platform and Timescapes

## EPrints: pros and cons

During the investigative phase of the University's research data management project (RoaDMaP), a list of requirements for a research data repository was drawn up[5] indicating which repository functions were essential. The requirements list was matched against several potential platforms. It was quickly evident that none of the platforms was able to fulfil all requirements and a decision was taken to build a research data repository on EPrints, extending its functionality where appropriate. The Leeds requirements mapping indicated that the primary requirement not fully met by EPrints was 'access control'.

EPrints offers several advantages:

- University of Leeds has extensive experience of working with EPrints.
- A number of UK HEIs plan to use EPrints to house their research data and thus have a common interest in developing the EPrints platform to support research data more effectively.
- EPrints is open an open source platform and so offers the potential to develop functionality locally or in collaboration with others.

The requirements list and the EPrints mapping were then matched against the requirements for the Timescapes archive in order to quickly identify where there may be unfulfilled requirements. Although there was a good match in some areas, it was clear that there were two broad areas where Timescapes requirements were different from University of Leeds/EPrints: slightly different access control functionality and more extensive functionality to support reuse of the Timescapes datasets. In both cases, it was agreed that the Timescapes Archive requirements should be modified slightly so that they aligned with the University of Leeds research data management requirements; this approach should support the sustainability of the Timescapes Archive as it is being built on a supported institutional platform and to almost identical standards as other institutionally hosted data.

## Access Control

The Timescapes Archive is now available on the new EPrints platform although access is currently restricted to a small group of staff while we complete work on the platform.

The University of Leeds has recently undertaken technical development work on EPrints in order to create access control functionality. Extensive consultation with the wider community including EPrints Services at the University of Southampton[6], highlighted

---

[5] See *Functional Requirements for a Data Repository* http://library.leeds.ac.uk/downloads/file/389/data_repository_platform_functional_requirements

[6] http://www.eprints.org/services/

significant interest in access control; we have been able to create functionality which may be widely adopted and included as a core component of EPrints in the near future[7].

The access control functionality has been informed by Timescapes' access control requirements. Some data need to be restricted for ethical reasons; where appropriate, restricted data content will be available only on request, in some cases at the discretion of the relevant PI, and registered users will be required to adhere to ethical conditions as part of their access, particularly in the re-presentation of these data. Data can be assigned to different access categories:

- Openly accessible to all
- Accessible by 'trusted' sources e.g. users logging in with a Shibboleth login and password; users accessing the archive from a specific IP address
- Closed access – metadata will be visible but access to the data will depend on agreement from the PI or other data owner

In essence, the access control development work enables application of data access 'rules' to determine which data can be accessed by whom and thus improves the granularity of access control offered by EPrints.

The work is nearing completion now (24/7/14) and will allow the University of Leeds' Research Data Management Team to implement necessary access control for the Timescapes Archive. Once these are in place, the new platform will be available for reuse and the Digitool platform will be discontinued. Until this work is complete the original Timescapes Archive on the Digitool platform remains available for reuse.

The access control work can be applied to other datasets at the University of Leeds but if, as we anticipate, the functionality is widely adopted, this work will have a significant and lasting impact.

## Timescapes Data Structure

EPrints was originally created to house publications and has a 'flat' structure. In other words, a repository 'record' will describe a published journal paper and hold a copy of that paper. Research data tends to be more heterogeneous than publications: there are different ways of grouping related materials and, in the case of Timescapes, data may grow over time as new waves are added. The Changing Landscapes and Leeds Research Data Management Teams explored how to organise the data within the EPrints platform so that it was searchable, browsable and had a logical structure. The hierarchical nature of the data was a particular challenge. We agreed the building block of the data was the 'case' and we needed to ensure cases were associated with the relevant collection and project to which they belong, and to the waves of data and varied data files that make up each case.

[7] For more technical information see http://wiki.eprints.org/w/Access_Control_Layer

## Browsing and searching data

The screenshots below show

(i) Browsing to a specific case (RF04) in the Changing Landscapes and Identities 'Pathways through Participation' dataset.

(ii) Finding the same case through a keyword search

### Browsing

Project -> Changing Landscapes: Pathways through Participation -> RF04

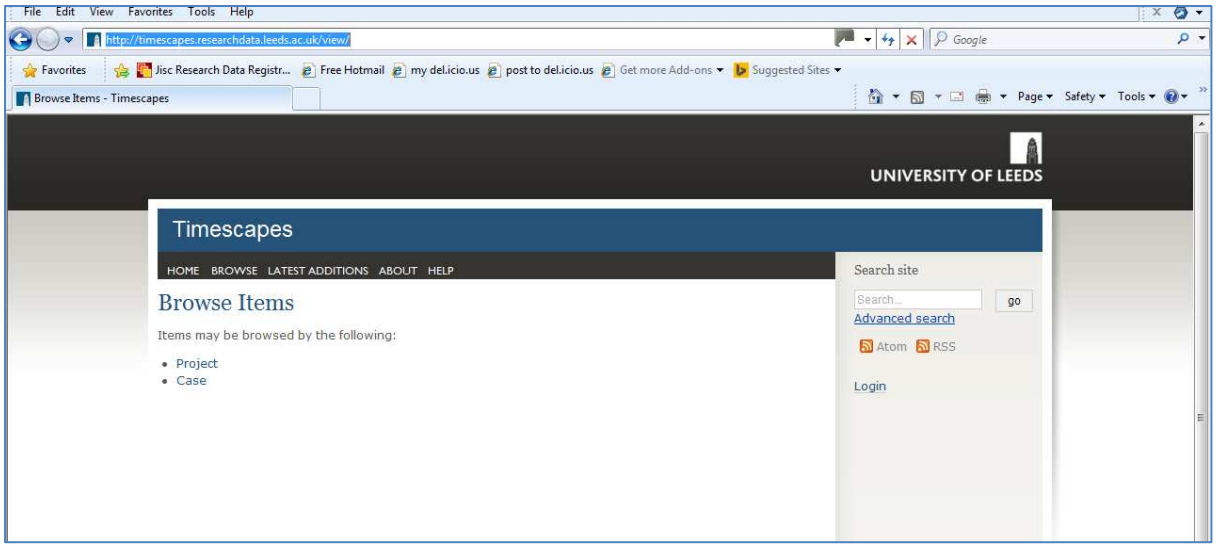


**Figure 1: Main browse options**

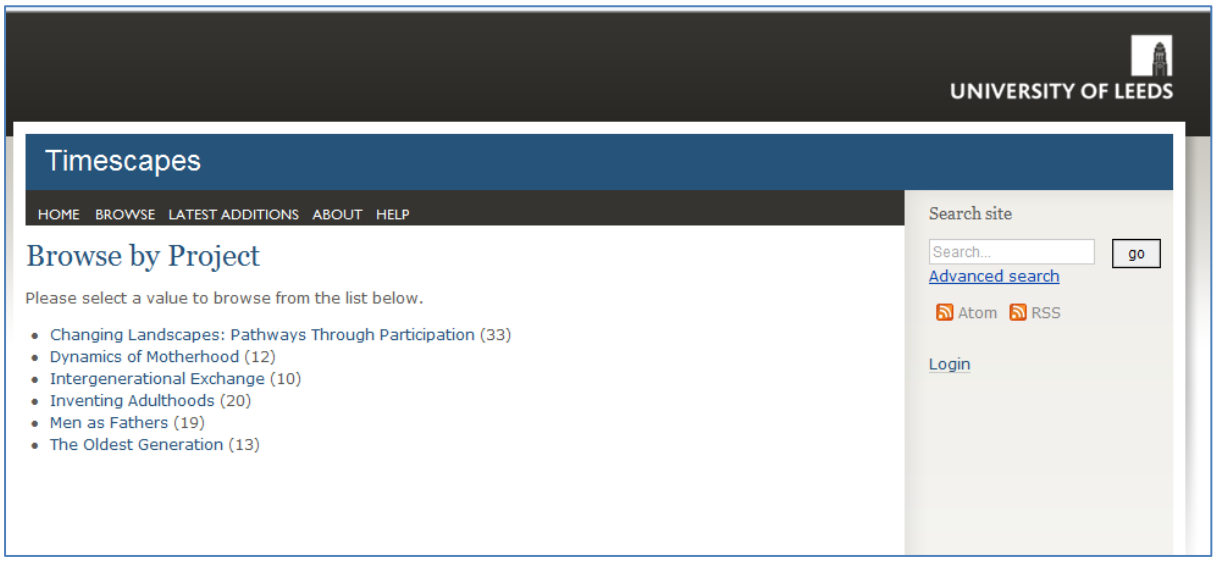


**Figure 2: List of Projects**

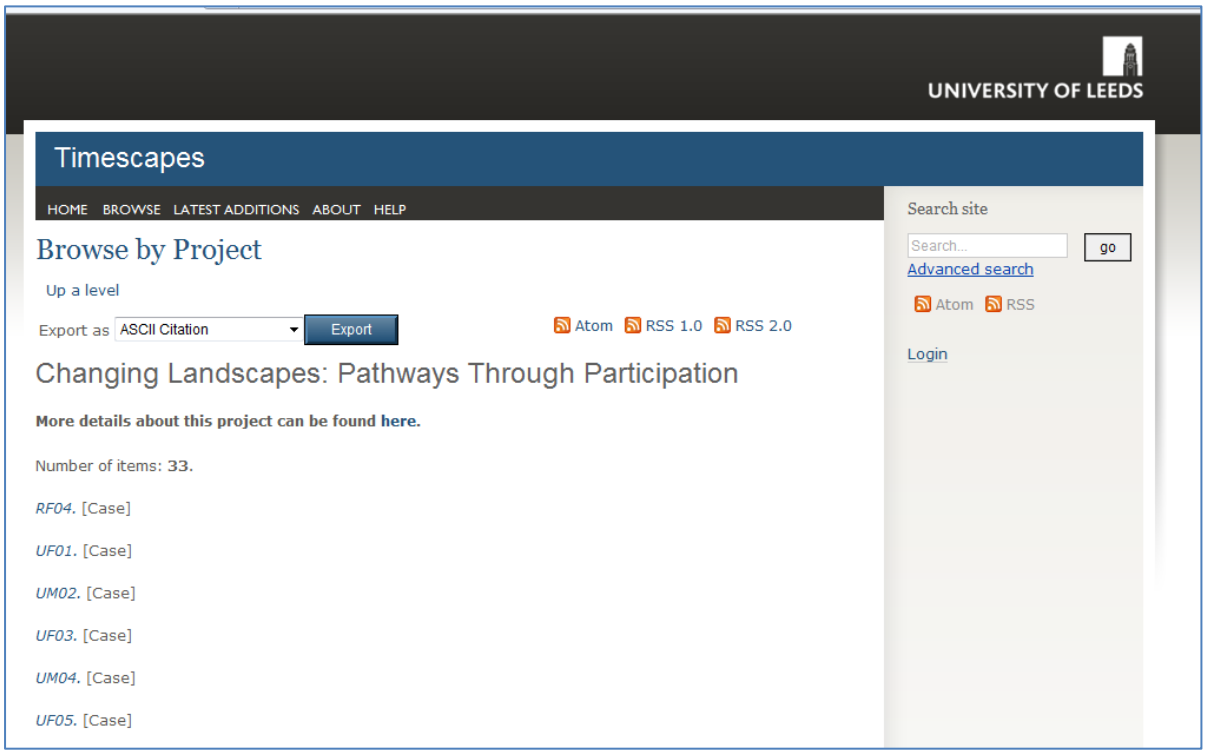


**Figure 3: List of cases**

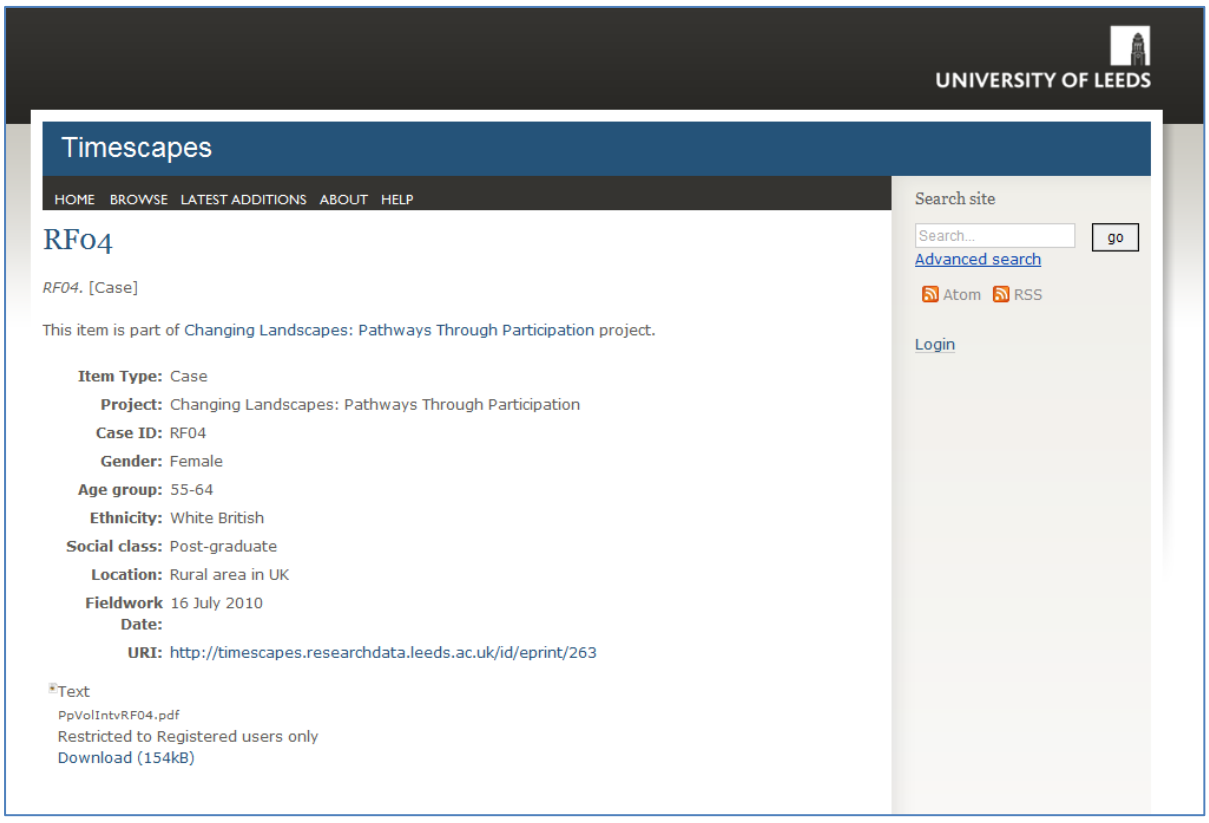


**Figure 4: Description of the case and access to the data for those with appropriate login credentials**

**Searching by topic**

Various search options are available which will search specific fields – like 'title' - or will search the full contents of the associated documents/data (if this is appropriate). The exact nature of some of the search options has been discussed and may change. For example, the option to search for either Male or Female may need more sophistication to encompass cases where, for example, the subject is male but there is case content from other individuals. We should also consider transgender or non-gender specific search options.

Individuals do not need to be logged in to execute the search in Figure 5, however when presented with Figure 6, only those with appropriate rights can access the content of the case files.

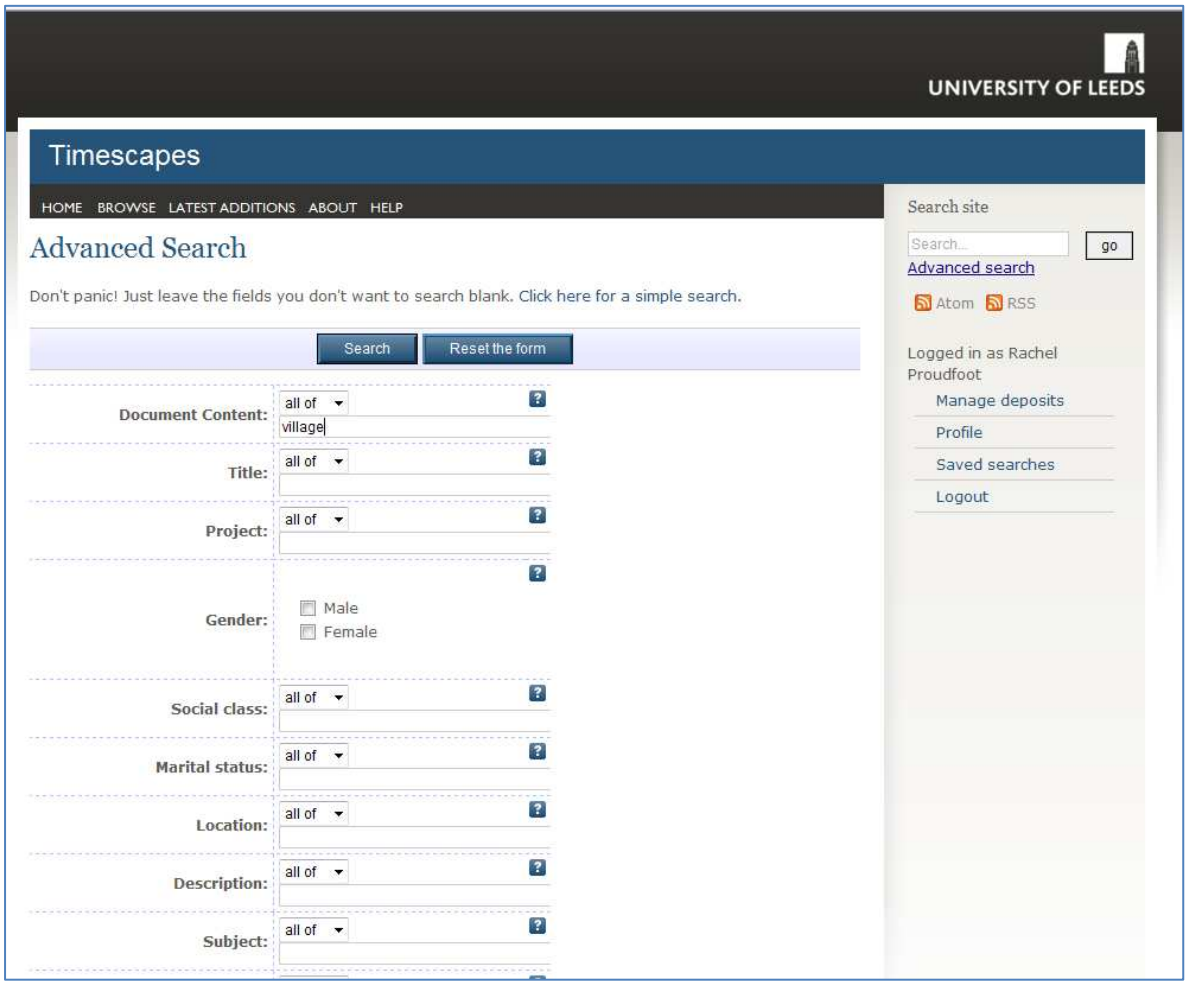


**Figure 5: Search all content for 'village'**

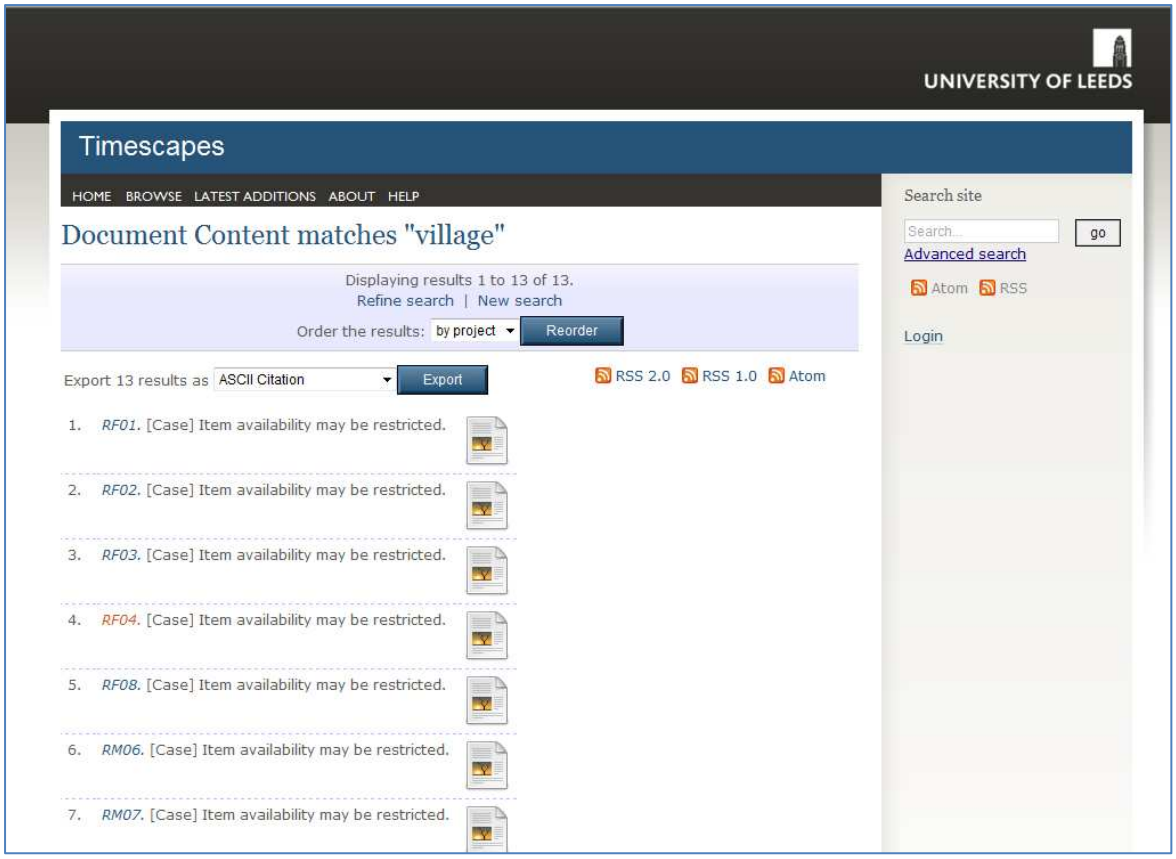


**Figure 6: Cases found using the search term 'village' – only those with appropriate rights can access the content of the files**

## Ingestion by script

There are scenarios where it will be beneficial to be able to bulk load data and metadata using an automated or scripted method. This may be because the research will generate a number of very similar data sets with much repeating metadata or where the data is coming from an instrument with metadata generated at source. A third scenario is where the data is already in a repository or has used templates for the collection of metadata – such as the two collections that are currently being ingested into the EPrints platform.

The data sets from Changing Relationship and Identities used XML files to hold the metadata to be used by Digitool for data ingestion. Scripts were written to convert the XML into a form that could be handled by EPrints. For the older datasets it was found that the XML files had been created using a different methodology such that either the script would have required significant additional development or the XML files would have needed major editing. It was decided that the metadata should be extracted from these XML files and modified within a spreadsheet. A second script was developed to read the content of the spreadsheet and convert it for EPrints use.

The use of scripts gives additional advantages. Should any problems be noted during or subsequent to ingestion it is less difficult to update the metadata files and repeat the ingestion process.

## Ingestion of datasets

The Changing Relationships and Identities datasets were grouped according to the methodology used in preparing the XML files. For the newer datasets the XML files were updated then used for ingestion using the script for XML. For the older datasets the metadata was extracted from the XML files and modified within a spreadsheet, then used for ingestion using the relevant script.

The successful development and use of the scripts led to the decision that the metadata from the new datasets from Changing Landscapes should be captured and ingested from spreadsheets.

**Changing Relationships and Identities**

Ingest from XML: The Oldest Generation, Intergenerational Exchange
Ingest from spreadsheet: Inventing Adulthoods, Dynamics of Motherhood, Work and Family Lives

The remaining Timescapes datasets will similarly be prepared for ingestion: Siblings and Friends Young Lives and Times (including new waves of data gathered under the ESRC Following Young Fathers study), Masculinities, Identities and Risk, and the DOH funded study Choice and Change.

**Changing Landscapes**

Ingest from spreadsheet: Big Lottery Fund Pathways through Participation (NCVO) and ESRC Real Times (Third Sector Research Centre, Birmingham).

## Metadata Visibility: how will data be discovered?

The Timescapes Archive holds metadata – or descriptive information – about the data sets and makes this metadata available online via several routes so that it is highly visible and readily discoverable. Metadata include project title, collection, case identifier, waves, fieldworker, keywords, and details of interviewees including gender, age group and social class; each project can also be associated with contextual documents, including a guide to the deposited data.

The Timescapes Archive is searchable in its own right – for example, by keyword and project title – and supports document full-text searching (e.g. Word documents, PDFs) (see Figures 1-6 above). In addition, metadata is indexed by Google, Google Scholar, Yahoo and other generic search engines; typically, institutional repository systems attract most of their visitors via the generic search engine route.

The Digital Curation Centre at the University of Edinburgh is currently developing a UK Research Data Registry (UKRDR) based on the well-established Australian National Data Service model. This model enables searching of multiple repositories of research data across the UK. The University of Leeds Research Data Management Team is working closely with the UKRDR to ensure Timescapes and other University of Leeds curated data is discoverable through this potentially important route.

Once searchers have landed on Timescapes data via one of the discovery routes, it will be straightforward to register in order to access most of the data. Restricted data are discoverable through the standard search options within the repository and guidance provided on how to apply to access these data.

Because datasets held in the Timescapes Archive have been generated through qualitative longitudinal research methods, there are likely to be several 'waves' of data collection within each project, collected at different times throughout the study. The repository platform will aid navigation of the data, including by research project and by wave, by offering browse and search options.

## Reuse, citation and DOIs

Data require sufficient contextual information and accompanying metadata to be understood outside the originating team of researchers. The Timescapes Archive offers both structured and free text metadata so that data can be contextualised and more readily discovered, re-analysed or re-used.

The Research Data Management Team is planning to create Digital Object Identifiers for the Timescapes data. DOIs provide a permanent identifier which always points to information about the data, even if the data is moved to a new digital location. DOIs are important in tracking citations of data in scholarly outputs. Again, the Timescapes data has provided a useful test case for DOIs, prompting discussion of the level of granularity at which DOIs are created – to begin with, a DOI will be assigned per project. In the future we will have to consider whether there is the need to also assign DOIs to individual cases within a project. Each DOI must be associated with a 'landing page' which contains mandatory information about the data, at a bare minimum:

- the creators of the data
- title
- publication date
- publisher
- doi

Although it is straightforward to present these fields in EPrints at the individual EPrint level – in this example, the Timescapes 'case', it is less straightforward at other levels of granularity – for instance, at the project level or for individual documents. The options and

required functionality within EPrints are under active consideration and the Timescapes/Changing Landscapes data provided an excellent springboard for discussion of the issues.

## Relationship with UKDA

Some of the Timescapes data is deposited with the UK Data Archive. Staff involved in the two Archives have worked closely together to scope the relationship between the UKDA as a national, ESRC funded service and Timescapes Archive as an institutionally hosted service holding data arising from ESRC funded projects alongside other related data sets from varied sources and funders. We are currently working on the assumption that the Timescapes Archive ingests and manages access to the data, with the UKDA providing a preservation function. Potentially, the two archive systems complement each other without duplication of function; the Timescapes Archive is potentially evolving to be a specialist 'node' of the UKDA holding QL datasets. The relationship between data centres like the UKDA and data managed in institutional repositories poses many technical, policy and funding issues which are still to be clarified; the Timescapes Archive offers an important test case in this respect.

In terms of the prestige or 'trustedness' of the system hosting QL data, interviews with Timescapes researchers suggested most were agnostic about whether data are managed in a subject specific data centre or an institutional based repository, so long as the data are findable, securely held, well organised and appropriate safeguards of confidentiality are in place.[8] Most were keen to understand what their data would 'look like' in the host system; this would help them plan how to structure their data for presentation to re-users.

## Conclusion and Lessons Learned

- The datasets associated with Changing Landscapes provided an excellent driver and test case for technical developments.
- It is very important – and extremely useful – for researchers and technical staff to work together to agree a shared vision and approach; sometimes this means negotiation and compromise but this in itself is a valuable process.
- The interaction between researchers and technical staff facilitated innovative solutions and proved fertile ground for exchange of skills and experience.
- Direct and iterative input from researchers ensures that the technical solution does not stray too far away from scholarly requirements.
- Granular access control to data may be necessary to accommodate consent conditions offered to research subjects and/or to secure buy-in from researchers; we need further work on how to balance controlled access to data with the increasing

[8] Proudfoot, R. (2013) RoaDMaP Project: Timescapes Case Study Report. http://library.leeds.ac.uk/downloads/file/476/timescapes_case_study

expectations from research funders that data should be openly available where possible.

- It is possible to collect structured metadata – for example, via a spreadsheet – which can be manipulated and imported into a repository; this is attractive because researchers are being asked to work with familiar software and do not need to interact directly with the repository platform.
- Do not underestimate the time required to decide how to structure a dataset to enable navigation; there may be several viable options and choices will be influenced by (i) the internal logic of the dataset (ii) the capabilities of the delivery platform (iii) the sophistication of the metadata adopted and the extent to which this can encode relationships between data entities.