Introducing the UK Data Service QualiBank

The UK Data Service QualiBank is an online tool for searching and browsing the contents of a selection of qualitative data collections held at the UK Data Service. The system complements the UK Data Service's Discover, where you can find over 5,700 data collections covering a wide range of social and economic data, spanning many disciplines and themes and including census data, as well as case studies and support guides.

What does the QualiBank do?

QualiBank is the UK Data Service's search and browse interface for qualitative data objects (text, audio and image files). You can find and retrieve extracts of textual data, audio files and images from a selection of qualitative data collections. The system allows searching of the content of text files, such as interviews, essays, open-ended questions and reports. It also allows searching of metadata attached to these objects, such as a description of a photo or of an audio recording, and it enables hyperlinking to related objects. A persistent citation can be made for selected extracts of data, such as an interview quote. Some collections are completely open while some are behind the UK Data Service access gate.

QualiBank uses faceted browsing or filters that allow you to refine your search or browse. Filters include:

● Collection title

● Access (showing access conditions) – to be implemented

● Resource type (type of object, such as interview transcript or image)

● Date (of coverage of the materials, not the dates of fieldwork)

● Sex

● Age group

● Socioeconomic status

● Region (to which the data refer)

Following a search or browse, the number of hits found is displayed at the top of the results list. A search may retrieve resources of different types and there is no limit to the maximum number of hits retrieved. Minimal metadata is shown below the title of the object (e.g. Interview with Mrs Omison). Where search terms appear in a text object (interview transcripts and so on) they are highlighted in the text in yellow.

Different types of resource are identified by a unique icon for text, audio and images placed next to the name of the retrieved hit.

To explore a hit, view simply click on the hyperlinked title, which will direct to a target page. Once on the target page, such as an interview transcript, the details are displayed, which consist of core metadata, the facility to open up and view additional metadata (click on +), and the object itself.

Related materials such as audio, images, study documentation and web resources are also shown, highlighted in red on the top left hand side of the page under clickable headings such as 'Collection documentation' and 'External resources'. Data collections can be downloaded using the 'Access this collection' from Discover.

What data are in QualiBank?

There are only a few collections in the online browsing system so far, and these were selected following a consultation process for prioritisation. Input was gained from academic users and archivists to choose a variety of important collections, some of which required digital enhancement. A dedicated project selectively digitised a number of paper-based materials including: original survey questionnaires; open-ended questions; transcribed interviews; essays and photographs. Detailed metadata was prepared for the collections.

The UK Data Service has over 350 records relating to qualitative data and mixed methods collections. These can be found using our Discover portal, which holds a detailed metadata record for each study/collection and, where possible, enables download of the whole data collection. Use the link within Qualibank to 'Access this collection from Discover'.

QualiBank is intended to provide access to two kinds of data: open data and safeguarded data. In the first beta phase, from April 2014, the system provides access to a limited number of open data collections only. Data files available as open data are licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Safeguarded data files will be made available under End User Licence conditions to users registered with the UK Data Service.

Currently the following five collections are openly available in the system:

British Oral Archive of Political and Administrative History, 1920-1980
British Oral Archive of Political and Administrative History (BOAPAH) was a pilot project conducted in 1979-1980 which aimed to collect a systematic oral archive of interviews with key figures from politics, the civil service and the armed forces. Interviewees selected were predominantly former Permanent Secretaries or former Cabinet Ministers, and were asked questions relating to the period of their official life, concentrating on the post-war years. There are 29 interviews in total, with 14 of these being made available through Qualibank.

Morale and Home Intelligence Reports, 1941-1949
The Morale and Home Intelligence Reports are government documents containing rich descriptions of attitudes and opinions of soldiers and the UK public during WWII. These reports include summaries of the morale of British personnel in post-war Germany and the Mediterranean Theatre. They focus on various topics including confidence in commanding officers, the government and the war efforts, postal censorship, leisure and entertainment, and personal finances and leave. In addition to these reports of soldiers abroad, observations and summaries of informal interviews with English civilians conducted by UK intelligence officers are also included.

Poverty in the United Kingdom : a Survey of Household Resources and Standards of Living, 1967-1969
This study, conducted in 1968 by Peter Townsend, aimed to evaluate and assess relative poverty in the United Kingdom. The study collected comprehensive information on all forms of resources (including income and assets) and indicative information on deprivation

and style of living among a representative sample of the population of the United Kingdom. Although a quantitative study, these surveys include marginalia and paradata which have recently been reused in a qualitative project by Ros Edwards, 'Advancing Paradata'. A small sample of 100 surveys with marginalia has been digitised and is available for exploration through Qualibank.

Family Life and Work Experience Before 1918, 1870-1973
The Edwardians examined family life and work experience in the early twentieth century in Britain and contains over 450 oral histories of individuals from this time period. Paul Thompson, the investigator, recognised that there existed little direct evidence of life during this period from a working class point of view and that such material would be highly valuable. The interviews were undertaken in the late 1960s and early 1970s and formed the basis of the first national oral history project in the UK.

School Leavers Study, 1978
The School Leavers Study is a collection of 141 essays written by school children from the Isle of Sheppey in 1978. Ray Pahl, the principal investigator, asked teachers in a comprehensive school on the Isle Of Sheppey to set an essay to all those in English lessons about ten days before they were due to leave school. The students were asked to imagine that they were 60 and write a short, reflective account of their life. The data offers a unique insight into the hopes and aspirations of young people on the island at the time.

What were the key requirements for the QualiBank?

The key requirements for the system were that qualitative data should be easily discoverable, retrievable, and could be examined in context, and finally, cited in a robust way.

The ability to provide context requires that there is adequate rich descriptive metadata for files. This metadata should also capture relationships between objects in a collection, so that a system can display all the relevant information about an object. A typical example might be to:

- display the text of an interview transcript on a web page with links available to explore the associated audio or video

- examine the metadata about the interview, e.g. the speaker and various attributes

- be able to visit related information, such as external websites holding contextual documentation about the study or topic, e.g. maps

To prepare materials for the system, data must be highly structured and consistently marked-up. This system uses standardised metadata elements that, providing others use similar elements, enable federated catalogues across providers and borders.

Data collections used to pilot the system were chosen to represent a range of 'typical' outputs from a qualitative research project.

What standards does QualiBank use?

QualiBank is based on existing metadata schemas and known technologies. Destination formats for data in the system are: text files as XML; handwriting/photos/pictures images as JPEG; multipage non-XML document images as PDF; and audio as MP3.

Three metadata standards were used:

- DDI Codebook 2.5 for basic study-level catalogue metadata

- Qualitative Data Exchange (QuDEx) schema for rich file-level description, document coding and annotation and intra-collection relationships

- Text Encoding Initiative (TEI) schema for mark-up of textual data items

DDI CODEBOOK 2.5

The Data Documentation Initiative (DDI) is a specification for capturing metadata about social science data. It was originally created to capture the information found in survey codebooks, which remains the focus of earlier versions. Most survey organisations around the world use this for documenting survey data collections. At the UK Data Archive it is also successfully used for cataloguing qualitative data collections. It is used for:

- basic study-level catalogue metadata

- rich variable description for survey files

The standard is maintained by the Data Documentation Initiative Alliance (ddi.org), a membership-driven consortium including universities, data archives, and international organisations.

Qualitative Data Exchange (QuDEx)

QuDEx is an XML metadata schema that enables discovery, location, retrieval and citation of complex qualitative data collections in context. It is maintained by the UK Data Archive, University of Essex and enables:

● highly structured and consistently marked-up data

● rich descriptive metadata for files e.g. interview characteristics, interview setting, type of object

● logical links between data objects: text to related audio, images, and other research outputs

● preservation of references to annotations performed on data

● common metadata elements that enable federated catalogues across providers and borders

The schema is maintained at:
www.data-archive.ac.uk/create-manage/projects/qudex?index=1
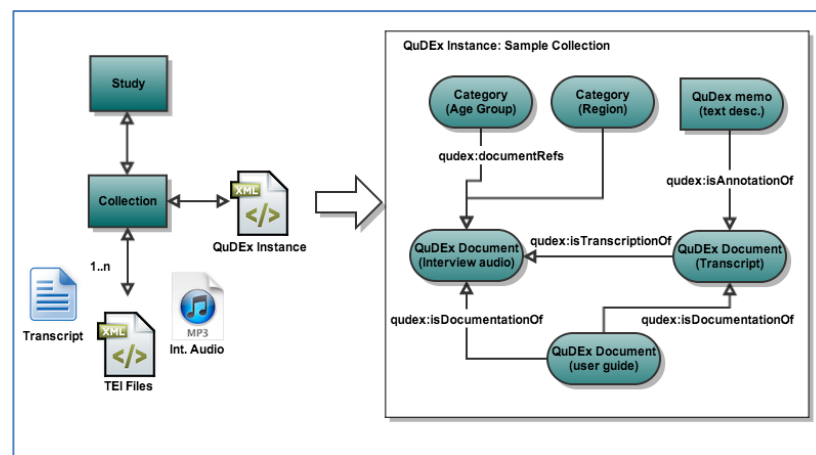
A simple MS SharePoint tool was used to enter metadata in a tabular format:

Text Encoding Initiative (TEI)

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. QualiBank uses a limited bespoke set of TEI elements from a massive profile. These are used primarily to denote structural mark-up and consist of:

● structural mark-up of textual qualitative data

● three mandatory TEI header elements

● body elements: turn takers, paragraphs, headers

● inline tags: corrections, errors

Example of QualiBank interview marked up in in TEI:

The TEI profile is available at: www.tei-c.org

What technologies does QualiBank use?

Aside from the metadata standards described above, QualiBank uses a range of commonly used technologies for web interfaces that enables searching and retrieval of data and metadata. Briefly, these include:

● BaseX for metadata and textual data storage and retrieval

● file server for non-XML docs

● simple QuDEx metadata data entry tool: Sharepoint

- C# scripts to process and validate against XML schemas

- Oxygen for manual mark-up

- Solr indexes used for faceted browsing and TEI text highlighting

- Xquery on BaseX for object metadata, text utterances and related materials

- GUID generator SQL database for QuDEx and TEI elements

- user interface in-house technologies, .NET and RESTful web services/APIs

Enabling citation of object and extracts in QualiBank

Fine-gained within-system citation is made possible due to the use of randomly generated GUIDs to uniquely identify data collections, collection objects and parts of objects, such as text paragraphs.

The interface allows users to select an extract of text, based on one or more consecutive paragraphs and dynamically retrieve a citation, which can be pasted into any outputs.

Using the URI in the citation takes the user back into the extract, highlighted in green, in QualiBank (that is the URI resolves directly to the data extract quoted).

The whole object can also be cited. We use the APA style for this and include the already

assigned DataCite DOI for the whole data collection. For open collections resolution is not an issue, but for closed collections, the user is prompted to login.

Reusing our standards

We have been spreading the word about our use of XML databases, the QuDEx and TEI metadata standards and the search platform, Solr, to create a powerful interface that deals very well with large quantities of qualitative data. QualiBank avoids the use of hard-coding of hyperlinks across data objects, so often found in multi-media archive and museum systems, making data publishing a much more straightforward process, and accommodating changes far more easily.

Please contact us if you would like to know more about our system or wish to reuse any of its components in any non-commercial applications.

Contact

digitalfutures@ukdataservice.ac.uk

Tel: +44 (0)1206 872145

UK Data Service