Bringing qualitative data to life though online access and citation

Louise Corti (Director, Collections Development and Data Publishing) and Darren Bell (Repository Architect), UK Data Archive, University of Essex
Email: corti@essex.ac.uk
Tel: +44 (0)1206 872145

Introduction

The UK Data Service (UKDS) makes available collections of primary research data to social scientists. It has around a 1000 records relating to qualitative data and mixed methods collections that include transcribed interviews, original survey questionnaires, open-ended questions and essays and photographs. These can be found using the **Data Catalogue**, which holds a detailed metadata record for each study/collection and enables download of the whole data collection (Figure 1).

Figure 1: A UKDS Data Catalogue record for a qualitative data collection

While many datasets can be downloaded, content cannot be interrogated until the individual data has been downloaded. In 2014, the **Digital Futures** project built and launched **QualiBank**, an enhanced and user-friendly platform for publishing, presenting and searching the content of qualitative data. The QualiBank system offers a streamlined data publishing workflow for data files and metadata marked-up with known schemas.
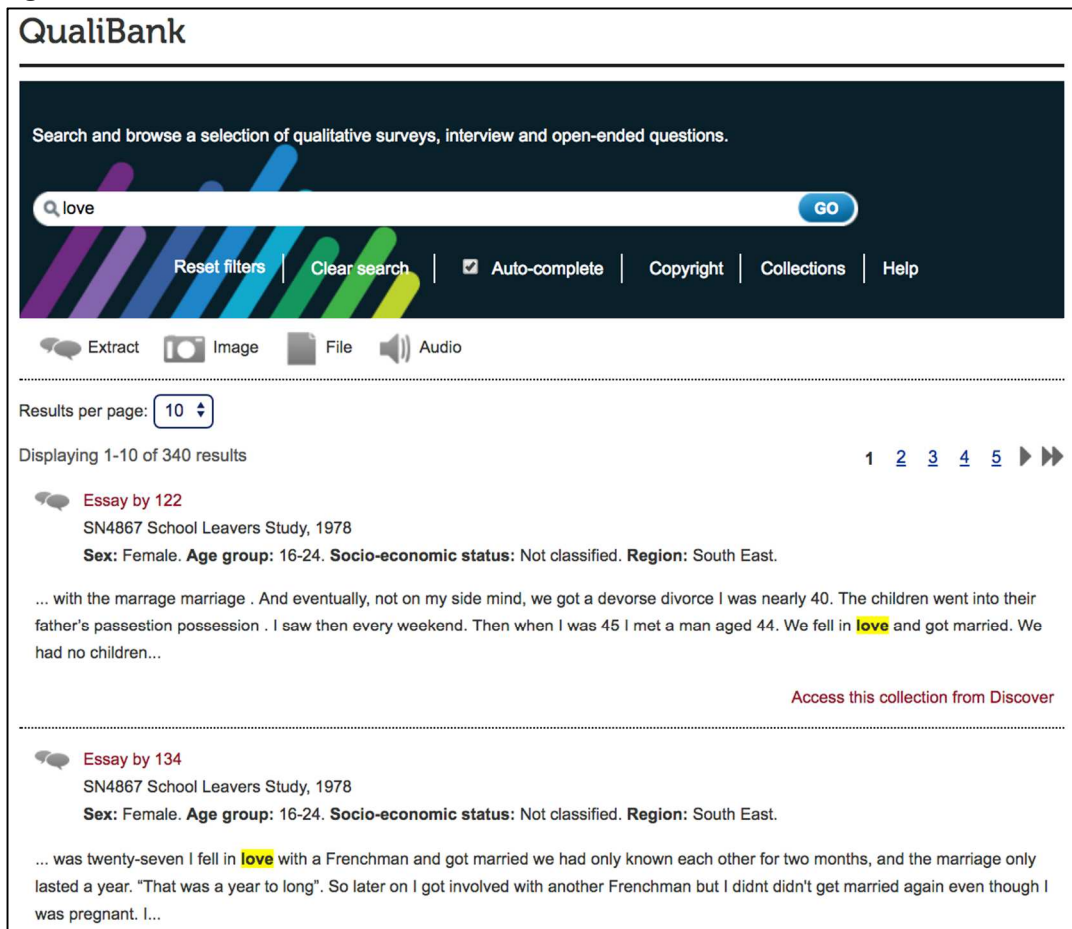
QualiBank functionality

For designing the system, key requirements were that qualitative data should be easily discoverable, retrievable, could be examined in context, and cited in a robust way. The UK QualiBank (https://discover.ukdataservice.ac.uk/QualiBank/) offers a search and browse

interface allowing searching and retrieval of the content of textual files and of the metadata attached to these objects, such as a description of an image or of an audio recording. It enables hyperlinking to related objects and includes a persistent citation feature enabling users to select a whole text or an extract of the text.

Faceted browsing allows user to filer and refine their search or browse, with facets representing a minimal core of attributes for qualitative interviews. Summary metadata is displayed below the title of the object (e.g. in Figure 2, Interview with Mr Patridge). Where search terms appear in a text object (interview transcripts and so on) they are highlighted in the text in yellow.

Figure 2. QualiBank search results

Linked resources (e.g. text, audio and images) are identified by a unique icon. Target pages, such as an interview transcript, display core metadata and view additional metadata (+) and a link to open a non-textual file (Figures 3a and b).

Figure 3a Interview metadata from qualitative text in the UK QualiBank

Figure 3b Interview text from qualitative text in the UK QualiBank

Materials related to a text object such as audio, images, or web resources are also clearly displayed.

Citation functionality

Users can select an extract of text, based on one or more consecutive paragraphs and dynamically retrieve a citation, which can be pasted into any outputs, as in Figure 4.

Figure 4 On-the-fly citation created for an extract from qualitative text in UK QualiBank

The citation URI created takes the user back into the quoted extract (highlighted in green) in QualiBank, unless the data is Safeguarded and needs a login (Figure 5).

Figure 5 Citation of an interview extract in QualiBank

Metadata requirements

The ability to provide context for data objects requires sufficient descriptive metadata and metadata must 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 and other external links;
- examine the metadata about the interview, e.g. the speaker and various attributes;

To prepare materials for the system, data must be highly structured and consistently marked-up. 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. QualiBank exploits known technologies and utilised three standardised metadata schema:

1. **The Data Documentation Initiative (DDI)**[1] - an open international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences, used by all national social science data archives around the world. In the system, DDI Codebook 2.5 is used to capture basic collection or 'study'-level catalogue metadata';

2. **Qualitative Data Exchange (QuDEx)** [2] – an XML schema maintained by the author at the UKDS that preserves annotations of, and relationships between, data, parts and segments of data and other related objects in a qualitative data collection; and is a baseline data exchange model for the archiving and interchange of data and metadata. A simple MS SharePoint tool is used to enter metadata in a tabular format (Figure 6). In QualiBank it is used for rich file-level description and intra-collection relationships (Figure 7).
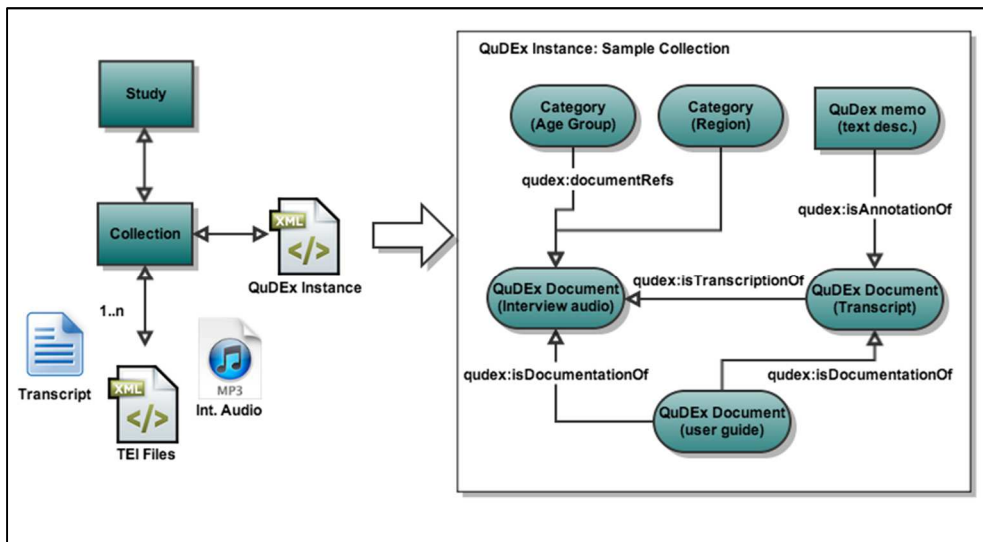
Figure 6. Simple MS Sharepoint metadata entry tool for QualiBank

Figure 7. Use of QuDEx metadata in QualiBank

---

[1] Data Documentation Initiative (DDI): https://www.ddialliance.org/
[2] Qualitative Data Exchange Format (QuDEx): http://www.data-archive.ac.uk/curate/standards-tools/metadata?index=2

- **Text Encoding Initiative (TEI)**[3] - a standard for the representation of texts in digital form. QualiBank makes use of a limited bespoke set of TEI XML elements from a massive schema for structural mark-up of text: headers, speaker turn-taking, defining paragraphs and inline corrections and errors (see Figure 8).

Figure 8. QualiBank interview marked up in in TEI

Technologies in use

QualiBank uses a range of commonly used technologies for web interfaces that enables searching and retrieval of data and metadata, including:

- 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

The fine-gained within-system citation is made possible due to the use of randomly generated GUIDs to uniquely identify data collections, collection objects (an interview) and parts of objects (e.g. paragraphs).

The technologies used create a powerful interface that deals very well with large quantities of qualitative data. The system 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.

Copyright Acknowledgment form

I confirm that I have the legal right to allow the reproduction of the following image(s), and I hereby give permission for the Archives and Records Association to reproduce them in *ARC Magazine* for distribution both in print and digitally:

Filename	Caption	Acknowledgement
Figure 1	*A UKDS Data Catalogue record for a qualitative data collection*	UK Data Service, 2018
Figure 2	*QualiBank search results*	UK Data Service, 2018
Figure 3a	*Interview metadata from qualitative text in the UK QualiBank*	UK Data Service, 2018
Figure 3b	*Interview text from qualitative text in the UK QualiBank*	UK Data Service, 2018
4 QualiBank	*On-the-fly citation created for an extract from qualitative text in UK*	UK Data Service, 2018
Figure 5	*Citation of an interview extract in QualiBank*	UK Data Service, 2018
Figure 6	*Simple MS Sharepoint metadata entry tool for QualiBank*	UK Data Service, 2018
Figure 7	*Use of QuDEx metadata in QualiBank*	UK Data Service, 2018

Figure 8	*QualiBank interview marked up in in TEI*	UK Data Service, 2018

I also agree that these images can be used to promote *ARC Magazine* on the Archives and Records Association website (www.archives.org.uk): YES/NO*

Signed:

Full Name: Louise Corti

Organisation: UK Data Archive, University of

Essex

Full acknowledgement will be given in any Archives and Records Association publication in which this image appears.

Data Protection Disclaimer for Photographs

I give my explicit consent to my photograph being taken by a representative of the Archives and Records Association, for the purposes of publication in the Association's publications and the National Press, or National Journals.

I understand that my photograph will be stored in hard copy and that it may be used in the future for further publicity.

Signed:

Name (please print): Louise Corti

Date: 1 July 2018