Digital Futures: digitizing qnd publishing qualitative data

Louise Corti

Collections Development and Producer Support

IASSIST, Cologne

May 2013

UK Data Service

Digital futures project: rationale & aims

- Gap in the ESRC Methods & Infrastructure portfolio for qualitative data
- Provide enhanced access to key ESRC-funded qualitative data via online data browsing and exploration
- Offer a mechanism for reliably citing data located in the system
- Project includes large-scale digitisation of precious and undigitized materials
- Maximise the impact from existing research and resource investments

Our initial anticipated audiences

- Researchers: data re-use in activities that inform both research and policy. Innovative data–rich outputs
- Trainees: capacity building in social science data confrontation and analysis, enhancing data skills
- Appeal of the UK cohort studies and classic studies in social science to undergraduates, postgraduate and early career researchers

Users, data needs & enhanced outputs

USER TYPE	DISCIPLINE	DATA TYPES	CONTEXT	ENRICHED OUTPUTS
Academic Researcher	Hot topics	All	Rich	Enhanced Publication
Apprentice Researcher	Hot topics	Standard forms	Rich	Enhanced Dissertation
Teacher/Student	Tried & tested topics	Standard forms	Limited, specific	Enhanced teaching materials/Assessment
Citizen Researcher	Local issues/current affairs/trends	Sample data, simple data – real life issues/'facts'	Little	Own private research/Public visualisation

UK Data Service and its own needs

- We have one of the largest qualitative data collections– over 300 data collections
- A proportion of these have been digitised from older paper sources
- Currently users find and download these from our website
  - Not so easy to find, but study documentation good
  - No searching within collections
  - No file manifest shown until download
  - It can be a bit of guess work!
  - Cannot reliably cite parts of data

Finding & accessing qualitative data

- Search for “health”
- Retrieve catalogue record, e.g.
  - SN 4943: Mothers And Daughters: Accounts Of Health In The Grandmother Generation, 1945-1978
- View limited user guide
- Web download as RTF bundle (46 interview transcripts)

Clues to content: study level metadata

Morbidity and mortality - Population, vital statistics and censuses
Child development and child rearing - Social stratification and groupings
Elderly - Social stratification and groupings
Family life and marriage - Social stratification and groupings
Gender roles - Social stratification and groupings
Use and provision of specific social services - Social welfare policy and systems

Depositor(s):
Blaxter, M., University of East Anglia

Principal Investigator(s):
Blaxter, M., University of East Anglia

Data Collector(s):
Blaxter, M., University of East Anglia

Sponsor(s):
Economic and Social Research Council

Abstract:
This is an enhanced qualitative study.

The research looked at beliefs and attitudes to **health** and medical care, inter-generational relationships, and social history of members of a grandmother generation. The original study included interviews with daughters as well; this collection contains only the grandmother interviews.

Grandmothers are asked extensive questions about their own **health** and the **health** of other family members. Details are provided on episodes of illness and remedies used, both home and **health** services. Specific topics of accidents, nutrition, dental care, and immunisation are covered.

More generally, grandmothers are asked about their views of their personal doctors and institutional **health** services. They give opinions on the quality of **health** care before and after the introduction of the National **Health** Service.

Grandmother-daughter relationships are explored, especially around the subject of offering and taking of medical advice concerning care for the grandchildren.

The collection has been enhanced by: conversion from paper to searchable RTF format by OCR and extensive editing and formatting of all interview transcripts. This collection will also be made available through ESDS Qualidata Online.

Main Topics:
The interviews cover **health** and social history, beliefs and attitudes to medical care, and intergenerational relationships.

There are discussions of the grandmothers' backgrounds, including jobs they held and where they lived.

Many aspects of **health** are addressed: their own **health**, childhood diseases, the women's attitudes toward doctors and the National **Health** Service, including how medical care changed after the start of the NHS. The grandmothers are asked about causes of diseases, remedies and treatments used, and their thoughts on staying **health**y.

Relationships with daughters are covered, focusing on whether the grandmothers offer medical advice, and if they do, whether it is accepted by their daughters.

Coverage:
Time Period Covered: 1945-1978
Dates of Fieldwork: 1977-1978
Index date: 1945-1978
Country: Scotland
Spatial Units: No Spatial Unit
Observation Units: Individuals
Kind of Data: Textual data; Semi-structured interview transcripts

Universe Sampled:
Location of Units of Observation: Subnational
Population: Women living in a city in Scotland who had a child between 1950-53, who were at that time in social class IV or V, who had a daughter who: had a child or children born in the same city; and were in the same social classes at the time of the birth; and still lived in the city; and were in touch with the grandmother.

Methodology:
Time Dimensions: Cross-sectional (one-time) study
Sampling Procedures: Purposive selection/case studies
Number of Units: 46
Method of Data Collection: Face-to-face interview; Transcription of existing materials; Audio recording
Weighting: Not applicable
Control Operations:

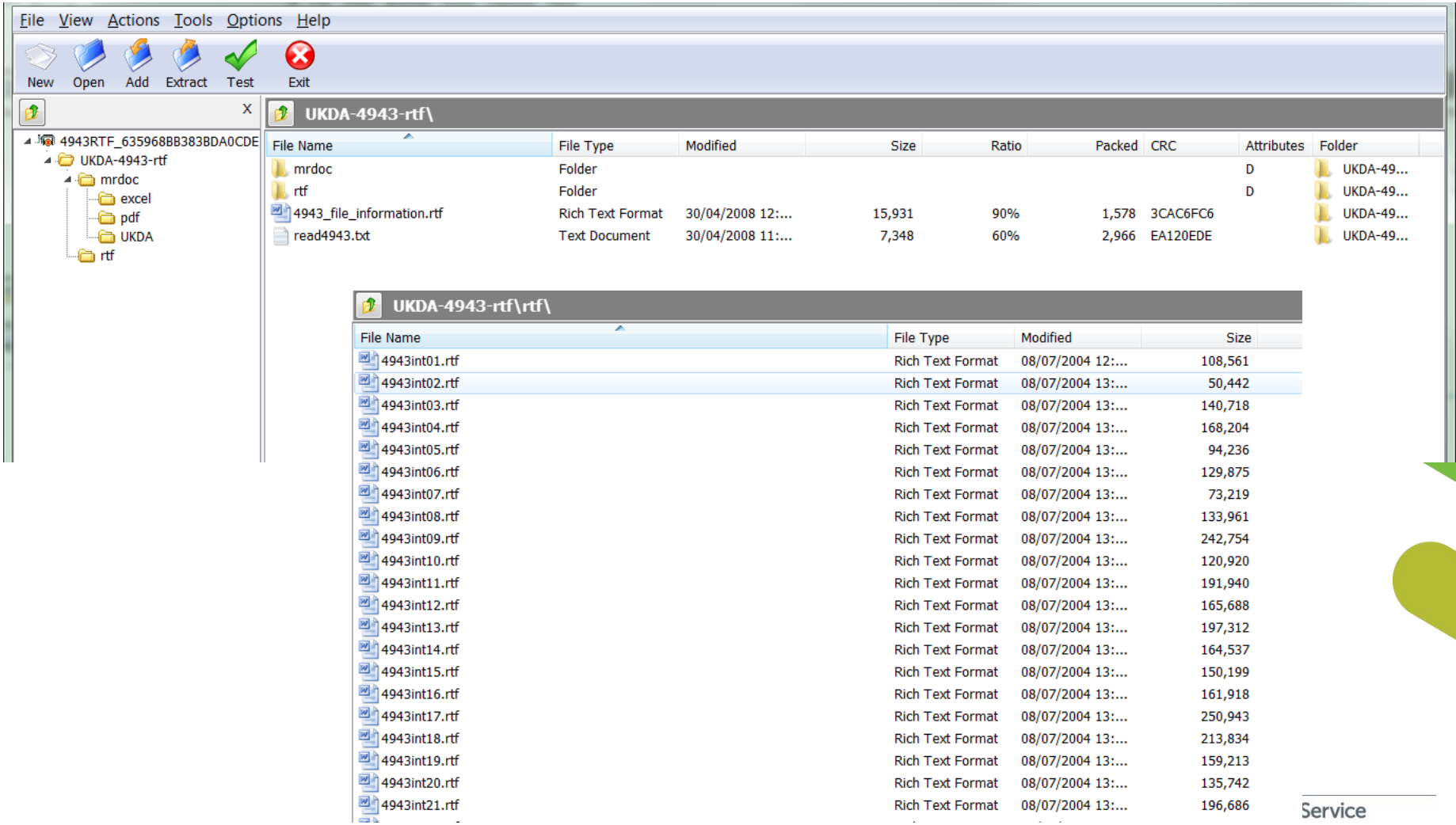
UK Data Service

Data listing

Study Number 6124
Being a Doctor: a Sociological Analysis, 2005-2006
Nettleton, S

Interview ID	Gender	Description	Date of Interview	No of Pages	Text File Name
x001	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	07/09/2005	36	6124int001
x002	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	09/09/2005	41	6124int002
x003	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	09/09/2005	39	6124int003
x004	Female	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	13/09/2005	36	6124int004
x005	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	13/09/2005	34	6124int005
x006	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	14/09/2005	50	6124int006
x007	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	21/09/2005	31	6124int007
x008	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	21/09/2005	35	6124int008
x009	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	22/09/2005	33	6124int009
x010	Female	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	22/09/2005	23	6124int010
x011	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	22/09/2005	36	6124int011
x012	Female	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	27/09/2005	41	6124int012
x013	Female	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	27/09/2005	21	6124int013
x014	Female	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	30/09/2005	20	6124int014
x015	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	05/10/2005	19	6124int015
x016	Female	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	05/10/2005	27	6124int016
x017	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	07/10/2005	27	6124int017
x018	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	17/10/2005	11	6124int018
x019	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	19/10/2005	33	6124int019
x020	Female	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	07/11/2005	21	6124int020
z001	Male	Interview with Hospital Doctor in Northern Tourist City	19/07/2005	50	6124int021
z002	Male	Interview with Hospital Doctor in Northern Tourist City	10/08/2005	45	6124int022
z003	Male	Interview with Hospital Doctor in Northern Tourist City	17/08/2005	56	6124int023
z004	Male	Interview with Hospital Doctor in Northern Tourist City	07/11/2005	21	6124int024
z005	Female	Interview with Hospital Doctor in Northern Tourist City	14/11/2005	32	6124int025
z006	Male	Interview with Hospital Doctor in Northern Tourist City	15/11/2005	23	6124int026
z007	Male	Interview with Hospital Doctor in Northern Tourist City	16/11/2005	23	6124int027
z008	Female	Interview with Hospital Doctor in Northern Tourist City	17/11/2005	18	6124int028
z009	Male	Interview with Hospital Doctor in Northern Tourist City	18/11/2005	20	6124int029

Download Zip



Typical rft transcript template

Study Name:
Depositor:
Interviewer:

Interview number:
Interview ID: Firstname Lastname
Date of interview:

Information about interviewee
Date of birth:
Gender:
Marital status:
Occupation:
Geographic region:

I: Just one or two factual details first of all before we go on to your health and that.... how old are you?

FL: I'm 58 in June.

I: What schools did you go to? Can you remember that far back!

FL: Oh... the last school was at Longside,, aye, ken Longside?

I: No, where is that?

FL: Peterheid. Village... That was the last school.

I: Uh-huh, so you lived in Peterhead...

FL: No, Longside.

I: Longside. And, do you work at all? At the moment?

FL: Just look efter my grandchildren. Like that... well, my grand-daughter comes in at night,, well, her mither cleans the school... and I look efter my grandson whose mother works... aye, that kind of thing. I take him in an' keep an eye on him. Well, he's at the school but I give him his dinner an' look efter him at night till she comes an' picks him up.

I: Just one or two factual details first of all before we go on to your health and that.... how old are you?

FL: I'm 58 in June.

I: What schools did you go to? Can you remember that far back!

FL: Oh... the last school was at Longside,, aye, ken Longside?

I: No, where is that?

FL: Peterheid. Village... That was the last school.

Complex data collections

- SN 5801: Concepts of Healthy Eating Food Research: Phases I and II, 1992-1996
- 293 interview transcripts; 73 diaries; 6 observation field notes
- Not represented well at all in a DDI 2.X catalogue

Geography: Lewisham; Newport; Greater London; Pembrokeshire
Spatial Units: Unitary Authorities (England);Unitary Authorities (Wales)
Observation Units: Individuals; Families/households
Kind of Data: Textual data; Individual (micro) level; open-ended, semi-structured interviews; diaries

Universe Sampled:
*Location of Units of Observation:*Subnational
*Population:*Men and women of all ages, black or white British, English and/or Welsh speaking, middle-and working-class backgrounds; Retail, catering and health professionals; Tourists in the Newport area.

Methodology:
Time Dimensions: Cross-sectional (one-time) study
Sampling Procedures: Convenience sample
Number of Units: 293 interview transcripts; 73 diaries; 6 observation field notes
Method of Data Collection: Face-to-face interview; Observation; Diaries
Weighting: No weighting used

Satisfying modern-day users

- 21st Century system for delivering qualitative data
- Web-based data exploration system
- Interface should lower the barriers to non-academic users
- Explore data through a data journey
  - Find relevant extract, examine in context, cite
  - Link data to still and moving images, and other related research outputs

Metadata demands for the system

- Demands highly structured and consistently marked-up data

- Qualitative data requires object (file-level) descriptive metadata, e.g. interviews, audio-visual files, images

- Common metadata elements enable federated catalogues across providers and borders

Richer metadata = richer discovery

- Work on enabling rich description
- Combining DDI 2.5 and QuDex schema
- QuDEx allows identification of data objects:
  - Interview transcript or audio recording etc.
  - Relationship to another data object or part of data
  - Descriptive categories at the object level, e.g. mime type, interview characteristics, interview setting
  - Capacity to capture rich annotation of parts of data
- Qudex model in use (Schema at: www.data-archive.ac.uk/create-manage/projects/qudex/)
- Object-level description = a lot of manual work!

System expectations

- Search/browse for data
  - Search /faceted browse of data - text in XML or PDF
- Browse
  - Faceted browse by categories: study-level topics; data types; interview/object facets, e.g gender, geography
- Display no. hits; data in lists
  - Word in paragraph; thumbnail image/pdf; AV link
  - Context: other related objects, in system or outside
- Access full object
  - View data, key metadata and related links
  - Get citation for part of data

System assumptions – phase I

- BaseX for metadata storage; Java loading; Solr search
- Data must be fully prepared on loading/publishing to the system. Data not ‘managed’ within the system
  - Mark-up, metadata, relationships all pre-defined
  - Pre-defined GUIDs to be used for citation (DOI + drilldown)
- Cannot search audio-visual data content

- Simple metadata mark-up tool created
- Technologies for user interface will be .net

- No download of data collection/subset - route to the UK Data Service
- No user-annotation possible in this phase, e.g. adding coding or comments

Digitisation of key data sources

Selectively digitize paper-based materials:

- original survey questionnaires
- interviews
- handwritten field notes, essays
- diagrams
- photographs

- Wide consultation to choose sources: user-driven, not archival
- Large digitization budget

Digitising textual data

Create image file

- used for poor typeface, handwritten, text with tables and graphs
- anonymise a copy with black marker prior to scan
- scan and save as TIFF image file

Create searchable PDF

- collate TIFFs and convert to PDF (Adobe Acrobat)
- bookmark PDF file for navigation, with contents page and headings

Create rich text: optical character recognition (OCR)

- convert TIFF to RTF format
- requires detailed checking and proofreading
- XML mark-up using TEI

Transcription

- Selective transcription method
- create consistent structure; templates
- XML mark-up using TEI

Timelines and deliverables: 18 months

Nov 12 – March 13

- Requirements gathering and technical specification work, making use of project consultants and users
- Mark-up of test data collections in Qudex
- Mark-up of XML example text in TEI

April – June 13

- Metadata and data (XML) mark-up defined and signed off
- Metadata mark-up tool developed
- Tech spec for system finalised
- Collections scoped and selected for digitisation

Timelines and deliverables

June – December 13

- Data digitization and mark-up of 10-15 collections
- Data loading system and user interface build
- Investigate methodology for permanent citation of data extracts

October 13 – Feb 14

- Working with key users to create outputs; teaching and research
- Create an enhanced publication linking to the live system
- Launch system with data – open and closed

Questions

Louise Corti

corti@essex.ac.uk