

Smart Qualitative Data: Methods and Community Tools for Data Mark-Up **SQUAD**

Louise Corti, UK Data Archive, Univ Essex
Maria Milosavljevic, School of Informatics, Univ. Edinburgh
Agenda Setting Workshop
28 April 2006



ESDS Qualidata

- specialist service of the ESDS led by the UK Data Archive (UKDA)
- focus is on acquiring digital data collections from purely qualitative and mixed methods contemporary research and from UK-based 'classic studies'
- facilitates the preservation of important large paper collections, and where appropriate, digitises samples of these collections.
- works closely with data creators (e.g academics) to ensure that high quality and well-documented qualitative data are produced
- offers user support and training to encourage professional researchers and research students alike to make full use of the rich sources of archived qualitative data

Access to our data

- access to some 150 qualitative research-based datasets available
- offers a resource discovery hub of some 4000 data collections
- has developed an online data browsing service for texts and is investigating both NLP and e-science applications to enable richer access to digital data banks
- wish to extend and share methods, standards and tools relating to this system

In search of standards and tools

- wish to enable more precise searching/browsing of archived qualitative data than catalogue records
- and to extend functionality of existing online data system to linking between sources (e.g. text, annotations, analysis, audio etc)
- for 5 years we have been developing a generic descriptive standard and format for data that is customised to social science research and which meets generic needs of varied data types

Applications of formats and standards

- mark-up data to an XML standard for data producers to publish data in multiple formats using style sheets/using web-based systems
 - E.g ESDS Qualidata Online
www.esds.ac.uk/qualidata/online/
- enable improved searching capability
- enable data exchange and data sharing across dispersed repositories (c.f. Nesstar)
- Enable the development of import/export functionality for CAQDAS software based on a common interoperable standard
- meet needs of researchers requesting a standard they can follow – much demand

How useful is digital data?

- dob: 1921
- Place: Oldham
- finaloccc: Oldham

- [Welham]

- U id='1' who='interviewer' Right, it starts with your grandparents. So give me the names and dates of birth of both. Do you remember those sets of grandparents?
- U id='2' who='subject' Yes.
- U id='3' who='interviewer' Well, we'll start with your mum's parents? Where did they live?
- U id='4' who='subject' They lived in Widness, Lancashire.
- U id='5' who='interviewer' How do you remember them?
- U id='6' who='subject' When we Mum used to take me to see them and me Grandma came to live with us in the end, didn't she?
- U id='7' who='Welham' Welham: Yes, when Granddad died - '48.
- U id='8' who='interviewer' So he died when he was 48?
- U id='9' who='Welham' Welham: No, he was 52. He died in 1948.
- U id='10' who='interviewer' But I remember it. How old would I be then?
- U id='11' who='Welham' Welham: Oh, you would have been little then.
- U id='12' who='subject' I remember him, he used to have whiskers. He used to put me on his knee and give me a kiss.
- ...

What are we interested in finding in data?

- Short term:
 - How can we exploit the contents of our data?
 - How can data be shared?
 - What is currently useful to mark-up?
- Long term
 - What might be useful in the future?
 - Who might want to use your data?
 - How might the data be linked to other data sets?

What features do we need to mark-up and why?

- Spoken interview texts provide the clearest—and most common—example of the kinds of encoding features needed
- Three basic groups of features
 - structural features representing basic format: utterance, specific turn taker, other speech tags e.g. defining idiosyncrasies
 - structural features representing links to other data types created in the course of the research process (e.g. audio or video referencing points, researcher annotations)
 - structural features representing identifying information such as real names, company names, place names, occupations, temporal information

Identifying elements

- Identify atomic elements of information in text
 - Person names
 - Company/Organisation names
 - Locations
 - Dates
 - Times
 - Percentages
 - Occupations
 - Monetary amounts
- Example:
 - Italy's business world was rocked by the announcement last Thursday that Mr. Verdi would leave his job as vice-president of Music Masters of Milan, Inc to become operations director of Arthur Anderson.

U id='1' who='interviewer' Right, it starts with your grandparents. So give me the names and dates of birth of both. Do you remember those sets of grandparents?

U id='2' who='subject' Yes.

U id='3' who='interviewer' Well, we'll start with your mum's parents? Where did they live?

U id='4' who='subject' They lived in Widness, Lancashire.

U id='5' who='interviewer' How do you remember them?

U id='6' who='subject' When we Mum used to take me to see them and me Grandma came to live with us in the end, didn't she?

U id='7' who='Welham' Welham: Yes, when Granddad died - 48.

U id='8' who='interviewer' So he died when he was 48?

U id='9' who='Welham' Welham: No, he was 52. He died in 1948.

U id='10' who='interviewer' But I remember it. How old would I be then?

U id='11' who='Welham' Welham: Oh, you would have been little then.

U id='12' who='subject' I remember him, he used to have whiskers. He used to put me on his knee and give me a kiss.

U id='13' who='Welham' Welham: Well, he'd have whiskers where he probably had not had a shave and he used to rub her cheek - literally.

U id='14' who='interviewer' What was his occupation?

U id='15' who='Welham' Welham: He worked at ICI, Widness.

U id='16' who='interviewer' Doing what?

U id='17' who='Welham' Welham: What did he do at ICI? Worked at ICI.

U id='18' who='interviewer' What was Grandmother's occupation?

U id='19' who='Welham' Welham: She had nought, because she had arthritis. She couldn't walk. That's why I used to go home every weekend to look after her. We had a Home Help until Friday and I used to go Saturday and stop until Monday when the Home Help came again.

U id='20' who='interviewer' Really? This was when you were married?

U id='21' who='Welham' Welham: Yes. I took me children with me. Me husband, give him his due, he let me go every Saturday, tea-time, until Monday and I used to take the children with me.

U id='22' who='interviewer' And then she came to live with you eventually?

U id='23' who='Welham' Welham: Well, me Dad died. Me sister was younger than me and she got TB and she went in Rufford Sanatorium but I still come every week and looked after me Mum on a Saturday night. Got up, got the dinner ready and got the bus from Widness to Rufford every Sunday to see me sister. Well, she was in for two years in Rufford and then they said they couldn't do nothing, so I brought her home and I nursed her for six weeks and she died and that's when we Mum came to me and we give the house up. There was nothing left. I got three brothers, but wives are not the same as their own daughter. So we all fell out over who was going to have her. So I said well, nobody's having her only me. I brought her to Altrincham and she lived two years. I made her comfortable in the kitchen with a single bed and looked after her. Gave me little job up and I looked after her and she just lived two years.

Key: organisation person location date time money percent geo-political vehicle facility weapon

How do we annotate our data?

- Human effort?
 - How long does one document take to mark up?
 - How much data do you want/need?
 - How many annotators do you have?
 - How well does a person do this job?
 - Accuracy
 - Novice/Expert in subject area
 - Boredom
 - Subjective opinions
 - What if we decide to add more categories for mark-up at a later date?
- Can we automate this?
 - The short answer: “it depends”
 - The long answer...

Automating content extraction using rules

- Why don't we just write rules?
 - Persons:
 - lists of common names, useful to a point
 - lists of pronouns (I, he, she, me, my, they, them, etc)
 - “me mum”; “them cats”, but which entities do pronouns refer to?
- Rules regarding typical surface cues:
 - CapitalisedWord
 - probably a name of some sort e.g “John found it interesting...”
 - first word of sentences is useless though e.g “Italy’s business world...”
 - Title CapitalisedWord
 - probably a person name, e.g “Mr. Smith” or “Mr. Average”
- How well does this work?
 - not too bad, but...
 - requires several months for a person to write these rules
 - Each new domain/entity type requires more time
 - Requires experienced experts (linguists, biologists, etc.)

What about more intelligent content extraction mechanisms?

- Machine learning
 - manually annotate texts with entities
 - 100,000 words can be done in 1-3 days depending on experience
 - the more data you have, the higher the accuracy
 - (the less annotated data you have, the poorer the results)
 - if the system hasn't seen it or hasn't seen anything that looks like it, then it can't tell what it is
 - garbage in, garbage out

State of the Art

- Use a mixture of rules and machine learning
- Use other sources (e.g. the web) to find out if something is an entity
 - number of hits indicates likelihood something is true
 - e.g. finding if Capitalised Word X is a country
 - search google for:
 - “Country X”; “The prime minister of X”
- New focus on relation and event extraction
 - **Mike Johnson** is now head of **the department of computing**. Today he announced new funding opportunities.
 - person(**Mike-Johnson**)
 - head-of(**the-department-of-computing**, **Mike-Johnson**)
 - announced(**Mike-Johnson**, new funding opportunities, today)

Data archive - NLP collaboration

- ESDS Qualidata making use of options for semi-automated mark-up of some components of its data bank using natural language processing and information extraction
- New partnerships created. New area of application for NLP to social science data

SQUAD Project Background

Smart Qualitative Data

Primary aim:

- to explore methodological and technical solutions for exposing digital qualitative data to make them fully shareable and exploitable
- Collaboration between
 - UK Data Archive, University of Essex (lead partner)
 - Language Technology Group, Human Communication Research Centre, School of Informatics, University of Edinburgh
- 18 months duration, 1 March 2005 – 31 August 2006

SQUAD: main objectives

- specify and test non-proprietary means of storing and 'marking-up' data using universal (XML) standards and technologies
- develop, implement and document user-friendly tools for semi-automating processes already used to prepare qualitative data for digital archiving and e-science type exploitation
- develop non-proprietary Qualitative Data Mark-up Tools (QDMT) for archiving and publishing XML marked-up data and associated research materials
- investigate requirements for contextualising research data (e.g. interview settings and dynamics and micro/macro factors)
- increase awareness and provide training with step-by-step guides and exemplars

Qualitative Data Mark-up Tools (QDMT)

1. systematic preparation of digital data: create formatted text documents ready for xml output
2. mark-up of data to capture basic structural features of textual data: e.g. speakers and selected demographic details
3. advanced annotation or mark-up of data
 - automated information extraction of basic semantic information: inserting tags for names, places, occupations and temporal information
 - automated anonymisation: replacing names with dummy forms, including co-references
 - geographic mark-up to enable data linking: identifying and applying geographic mark-up, and scoping researchers' needs for geo-linking

Qualitative Data Mark-up Tools (QDMT)

- 4 basic classification or thematic coding of textual data: will investigate linking into a domain ontology
 - e.g. UK and EU social science thesaurus
 - solution - 'stand-off annotation' approach whereby data and coding stored in different documents
- 5 contextual documentation to capture richness of the research methods, data collection and analytic interpretation and representation
 - looking at the interrelationships between complex intra-project data, annotations and context
- 6 exposure of annotated and contextualised qualitative data to the web
 - investigating publishing of above QDMT XML outputs to ESDS Qualidata Online, opportunities for exchange within CAQDAS tools, etc.

Progress

- have a draft DTD with mandatory elements
- have chosen an NLP annotation tool–NXT (NITE XML Toolkit) www.ltg.ed.ac.uk/NITE/) to automate the mark-up of qualitative data
- building a GUI – with step-by-step components for ‘data processing’
 - Data clean up tool
 - Named entity and annotation mark-up tool
 - Anonymise tool
 - Archiving tool – annotated data
 - Publishing tool – style sheet for ESDS Qualidata Online system
 - Demonstrator for July
- extending functionality of ESDS Qualidata Online system to include audio-visual material and linking to researcher products and mapping system

from summer, looking at key word extraction systems to help conceptually index qualitative data and exploring grid-enabling data

Metadata standards in use (LB)

- Study description
- Data file description
 - file contents; format; data checks; processing; software
- Other study related materials
- Document description
- Can also employ variable description:
 - for study survey data (mixed methods) or numeric outputs from qualitative data:
 - demographic profile of sample
 - other quantified responses to qualitative data (attributes or thematic classifications often assigned (coded) in CAQDAS software)
- For data content and data annotation: the Text Encoding Initiative
 - standard for text mark-up in humanities and social sciences



ESDS Qualidata XML Schema

“Reduced” set of TEI elements

- start with core tag set for transcription, then add:
- editorial changes <unclear>
- names, numbers, dates <name>
- links and cross references <ref>
- notes and annotations <note>
- text structure <div>
- unique to spoken texts <kinesic>
- linking, segmentation and alignment <anchor>
- advanced pointing, will use XPointer framework
- synchronisation
- contextual information (participants, setting, text)

	A	B	C	D	E
1	DRAFT DTD FOR QUALITATIVE DATA				
2	Partial List of TEI elements for use in ESDS Qualidata XML DTD for transcribed interviews and other qualitative research materials				
3	Citation information: Draft DTD for qualitative data. ESDS Qualidata. Colchester, Essex: UK Data Archive. March 2005.				
4	Revised: 28 June 05 lb				
5	Ch. Of TEI Guidelines (p5)	Element Name	Typical Use	Attribute	Typical use or other comments
16	6.3 Highlight and quotation	<q>	quotation		
17		<emph>	marks words or phrases which are stressed or emphasized for linguistic or rhetorical effect.		
18	6.4 Editorial changes				
19		<add>	text added to the document by author, transcriber, etc	place	indicates where the addition has been added by author, scribe, etc. (inline, left[margin], etc.); useful for "end of side n" notations
20				resp	person responsible for addition
21		<unclear>	any text which cannot be transcribed with certainty	reason	illegible, inaudible source, and so on
22	6.5 Names, numbers, dates+				
23		<rs>	contains a general purpose name or referring string.		
24		<name>	contains a proper noun or noun phrase		this can take attribute to handle subcategories: person, place, street, address, etc.
25		<num>	any number		
26	6.6 Links and cross references				
27		<ref>	reference to another location in the current document	target	can use URI to point to other docs locally or on other systems
28				type	
29		<ptr>	pointer to another location in the current document	targType	
30	10.2 Elements unique to spoken texts				

Metadata for model transcript output

Study Name <titlStmt><titl>Mothers and Daughters</titl></titlStmt>
Depositor <distStmt><depositr>Mildred
 Blaxter</depositr></distStmt>

Interview number <intNum>4943int01</intNum>
Date of interview <intDate>3 May 1979</intDate>
Interview ID <persName>g24</persName>
Date of birth <birth>1930</birth>
Gender <gender>Female</gender>
Occupation <occupation>pharmacy assistant</occupation>
Geo region <geoRegion>Scotland</geoRegion>
Marital status <marStat>Married</marStat>

Transcript with recommended XML mark-up

```
<?xml>
<titlStmt><titl>Mothers and Daughters</titl></titlStmt>
<distStmt><depositr>Mildred Blaxter</depositr></distStmt>
<intNum>4943int021</intNum>
<intDate>3 May 1979</intDate>
<persName>g24</persName>
<birth>1930</birth>
<gender>Female</gender>
<occupation>Pharmacy assistant</occupation>
<geoRegion>Scotland</geoRegion>
<marStat>Married</marStat>
<!--comment -->
<u id="1" who="interviewer"> There's just one or two factual things first of all do you mind my asking how old you are?</u>
<u id="2" who="subject"> 49.</u>
<u id="3" who="interviewer"> And what schools did you go to?</u>
<u id="4" who="subject"> <orgName>King Street</orgName>, <orgName>Woodside</orgName> and <orgName>Hilton</orgName>.</u>
<u id="5" who="interviewer"> Uh-huh .. and how old were you when you left the school?</u>
<u id="6" who="subject"> 14.</u>
<u id="7" who="interviewer"> And you work at the moment? What sort of work do you do?</u>
<u id="8" who="subject"> Well I've gone back to get shorter hours, I've went back to domestic, which I dinna really care for. But then I used to be in the pharmacy department at <orgName>ARI</orgName> ... just <occupation>pharmacy assistant</occupation>. At least it was better than cleanin'! But then they've nae part-time workers there so..</u>
<u id="9" who="interviewer"> And did you work in the pharmacy long?</u>
<u id="10" who="subject"> I was there for eleven years.</u>
<u id="11" who="interviewer"> And did you have any other sort of jobs?</u>
<u id="12" who="subject"> Where? Since I left school, like? Well, when I first left school I was just a <occupation>shop assistant</occupation> in a number of shops like <orgName>Reid</orgName> and <orgName>Pearsons</orgName>, which is.. we hinna got it ony mair.</u>
<u id="13" who="interviewer"> Gone ..!</u>
<u id="14" who="subject"> <orgName>Marks and Spencers</orgName>, and the <orgName>Coop</orgName>. And that's about it.</u>
<u id="15" who="interviewer"> Uh-huh .. and then you got married and had your family. And when you went back to work, did you go into the pharmacy?</u>
```

XML is source for .rtf download

Study Name:: Mothers and Daughters
Depositor: Mildred Blaxter
Interviewer: Liz Patterson

Interview number: 4943int021.doc
Interviewer ID: g24
Date of interview: 3 May 1979

Information about interviewee

Date of birth: 1930
Gender: female
Marital status: married
Occupation: pharmacy assistant
Geographic region: Scotland

LP: There's just one or two factual things first of all do you mind my asking how old you are?

G24: 49.

LP: And what schools did you go to?

G24: King Street, Woodside and Hilton.

LP: Uh-huh .. and how old were you when you left the school?

Metadata used to display search results

ID: g24

Born: 1930 Female

Occupational Class: Sales and customer service **Geographic Region:** Scotland

... and Hilton. Uh-huh. and how old were you when you left the **school**? 14. And you work at the moment? What sort of work do you ...

ID: g25

Born: 1932 Female

Occupational Class: Elementary **Geographic Region:** Scotland

... No, no, no. I was a shop assistant when I left **school**. And then we had our own shop down in Ashgrove there. my sister an ...

ID: g31

Born: 1929 Female

Occupational Class: Unemployed **Geographic Region:** Scotland

... , what sort of work did you do? When I first left the **school**? An upholstery sewer. I worked in Roberts in Union Street for 16 years, ...

ID: g32

Born: 1921 Female

Occupational Class: Unemployed **Geographic Region:** Scotland

... Gordon's College. And a shop assistant before that. And when you left **school**? Oh, when I left **school** I went to Broadford. When I left **school** ...

ID: g6

Born: 1921 Female

Occupational Class: Elementary **Geographic Region:** Scotland

... . I'm 57. And what schools did you go to? Old Aberdeen **School** for a start, and then Sunnybank. And how old were you when you left ...

ID: g7

Born: 1916 Female

Occupational Class: Unemployed **Geographic Region:** Scotland

... what schools did you go to yourself? Well, Porthill and the Middle **School**. Aye, Porthill's the Primary and the Middle's the Secondary. Yes. And how ...

XML+XSL enables online publishing



The screenshot shows the ESDS Qualidata Online website. The header features the title "ESDS Qualidata Online" in a blue serif font, with the ESDS logo (a bar chart) and the text "Economic and Social Data Service" to the right. A navigation bar below the header contains links for "About", "Data Collections", and "Explore Multiple Collections". On the left side, there is a vertical menu with several categories: "Introduction", "Edwardians", "Mothers and Daughters (SN 4943)", and "100 Families (SN 4938)". The main content area is titled "Transcript" and contains the following text:

To view the interview summary, click the title. (Only available for Edwardians interviews at the moment).

g24

There's just one or two factual things first of all do you mind my asking how old you are?

49.

And what schools did you go to?

King Street, Woodside and Hilton.

Uh-huh .. and how old were you when you left the school?

14.

And you work at the moment? What sort of work do you do?

Well I've gone back to get shorter hours, I've went back to domestic, which I dinna really care for. But then I used to be in the pharmacy department at ARI ... just pharmacy assistant. At least it was better than cleanin'! But then they've nae part-time workers there so..

And did you work in the pharmacy long?

I was there for eleven years.

And did you have any other sort of jobs?

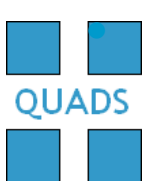
Information

- See ESDS Qualidata site:
www.esds.ac.uk/qualidata/online/
- and SQUAD website:
quads.esds.ac.uk/projects/squad.asp

We would like collaboration and testers!

Some questions to resolve

- What hierarchical elements should we use for collections of interview transcripts? Corpus, group/text, text/div?
- What is the best XPointer scheme (or schemes) to handle linking and pointing to external resources?
- Are there preferred standards for linking to, and synchronising with, audio and video?
- We have some text requiring non-hierarchical coding and need to determine which of the schemes for multiple hierarchies best suits our texts
- How can we best use TEI metadata to incorporate several DDI elements used by the UKDA for cataloguing?



We are seeking advice on some issues arising from the integration of TEI and NXT xml models