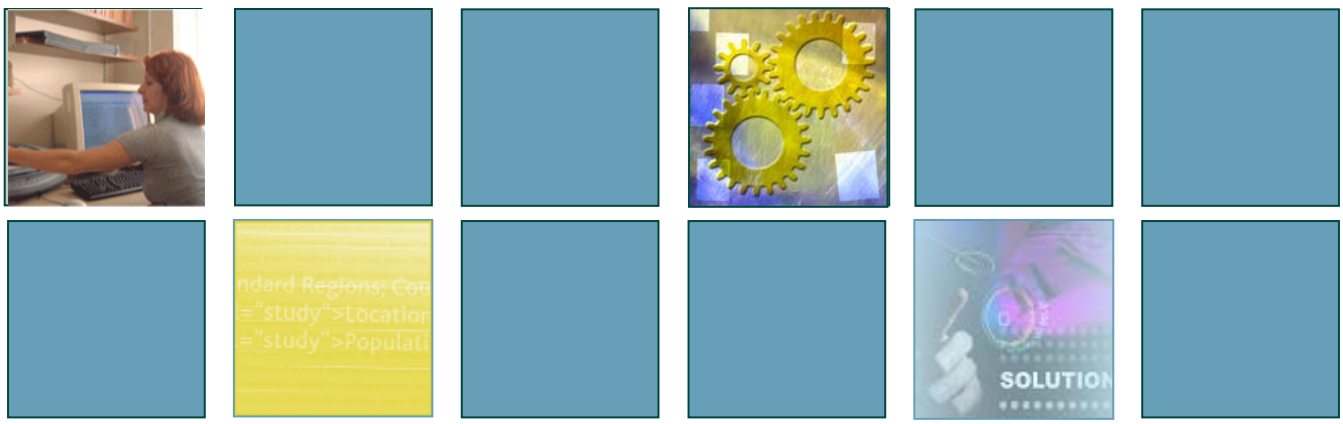


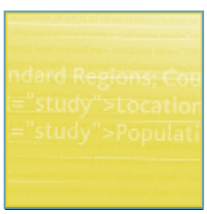
Searching and sharing qualitative data: the uses of XML

Smart Qualitative Data:

Methods and Community Tools for Data Mark-up (SQUAD)

quads.esds.ac.uk/squad

CONTENTS



Introduction

Qualitative research is a very large tent embracing diverse practices and epistemologies. Across that diversity, many researchers are grappling with the challenges of new technologies. It is never easy to sort useful tools and applications from spin and hype. Far too often, tools, software and gadgets become diversions from the inevitable need for meticulous data work, deep thought and rigorous analysis. But then, some tools are just unambiguously helpful. Even the staunchest critics usually acknowledge that word-processing has made the management of volumes of text far easier. But how does one begin to sort through the onslaught of newness to find what is really helpful?

For many qualitative researchers, XML (eXtensible Mark-up Language) may appear as one of these "new things", relegated to low priority as it seems either mystifying or distant from immediate work requirements.

This document is intended to demonstrate that XML is relevant to a broad range of very common practices in qualitative research. It will also show that XML enables new capabilities, and it will explain some of these applications of XML in non-technical language with practical examples.

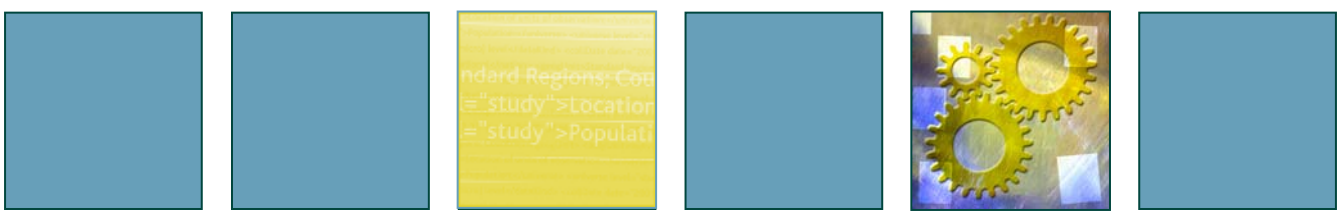
Typical research tasks made easier and better with XML

Searches

Text processors handle keyword searches fairly well, under certain conditions. However, if one is searching for a common name, it might show up too often for search results to be manageable. It would be handy to search in a more refined way: find all terms 'GM' only if they refer to the motor company, not genetic modification, or general managers. A document marked up with XML could easily identify occurrences of 'GM' that referred only to companies.

Sharing across software

Some researchers use software to help in qualitative data management and analysis, otherwise known as Computer-Assisted Qualitative Data Analysis Software or CAQDAS. In various situations, sharing parts of this work is necessary, perhaps among project team members collaborating on a coding scheme. In their current state, there is no standard import or export format for these software applications. XML, because it is a non-proprietary standard, enables researchers to code text and add annotations in ATLAS.ti, export coded data from ATLAS.ti, import it into MAXqda while retaining the coded data, code list and annotations within MAXqda.



Typical tasks – continued…

Anonymising data

Many researchers face the challenges of needing to anonymise data. This might arise when preparing data for publication, for analysis amongst a large team, or for self-storage or formal archiving. As with the search example, finding names is not terribly difficult, but it is both tedious and sometimes complicated to do replacement. Difficulties arise with multiple forms of names, for example: John Horatio Smith, Mr. Smith, Dr. Smith, John, John H. Smith etc. An efficient way to locate all such forms and replace them with suitable forms would save time. XML can enable this. ESDS Qualidata is developing a tool that will simplify the process of anonymising or pseudonymising data - a document that has real names marked up with XML tags for a given entity (e.g. name, place, company) can be imported into the system and pseudonyms applied across the co-references to that single entity.

Innovative publishing

When a qualitative research project is nearing completion, there are many ways to consider sharing it. Of course, printing and online journals are central means of disseminating findings, but some projects are increasingly suitable for web-based publication, especially those with significant image and audio-visual content. However, producing content for the web requires a certain format, in the same way that there are standards for titles, headings, text, for traditional publication outlets. XML enables such innovative publishing methods.

Long-term preservation

One kind of preservation involves putting unsorted materials into boxes in the home office or attic. While a common approach, it is increasingly seen as less than optimal – to say the least. Some kind of formalised preservation is far preferable, particularly when coupled with a way to share the data – and ensuring that confidentiality is protected where necessary. Whether data are self-archived or deposited in an authorised archive, certain features of the data and supporting materials have to be made clear to enable adequate cataloguing, identification and retrieval. XML makes these processes possible.

XML matters after all

In short, for many purposes - be it complex data searching, data sharing, anonymisation, publishing to the web, and data preservation - there is a need for a language that identifies structural features of text in a non-proprietary manner, and that language is XML. These are just a few of the practices and potential new capabilities that involve identifying the nature and structure of qualitative data. Though it might not be obvious at first, there are many good reasons to care about XML and its role in marking up data to enable searching and sharing.

Sharing qualitative data for the longer term: technical issues

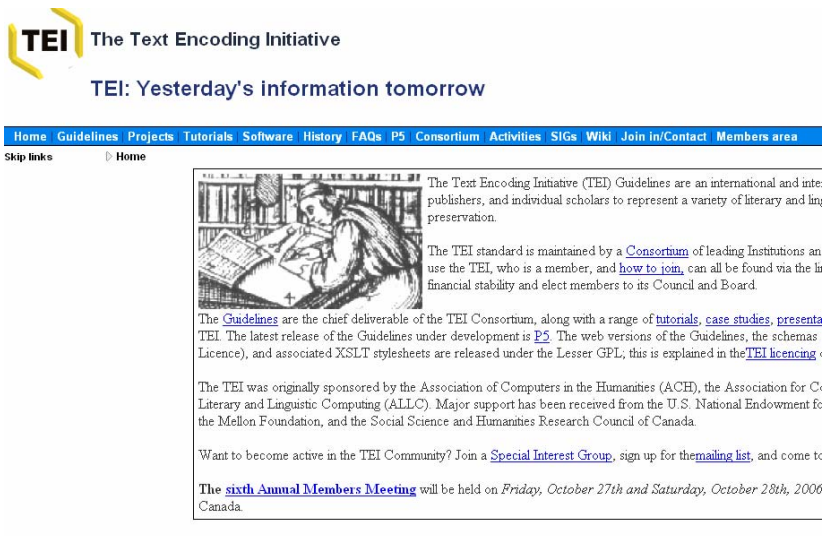
The aim of the development work being undertaken by ESDS Qualidata is to produce an application format that will enable sophisticated online searching of, and information retrieval from, digital materials. The data archiving community requires an application that will support the encoding of the content of various types of documents produced in qualitative research (e.g. interview transcriptions, research diaries, survey questionnaires) as well as contextual documentation (e.g. researchers' annotations and newspaper clippings). It is also essential that the application provide links between texts and associated audio and video materials. Finally, the application should be able to represent metadata (such as depositor's name or study title) at the individual file, or interview, level and for the entire collection.

What is XML?

XML stands for eXtensible Markup Language. It is a language for coding a text to define components of that text. It does this by embedding elements in the text. For example, one could use an element <occup> to mark every occurrence of an occupation. Elements consist of nested pairs of tags, so the marked-up text would look like this: <occup>teacher</occup>. Thus, mark-up language makes structural elements explicit in a document using a system of ordinary textual tags that are embedded in the text. The power of XML lies in the fact that it is extensible. That is, it allows for the creation of descriptive mark-up systems based upon a common vocabulary, but this vocabulary can be extended to accommodate the special requirements of a particular user or domain.

What is mark-up?

Basic annotation or mark-up of data is defined here as capturing the basic structural features of (primarily textual) data. This involves the basic layout of the transcript, such as the use of speaker tags (often initials) to indicate who is speaking, and double-spacing between speakers. At the study and file level, information about the study is routinely captured as metadata for catalogue records (e.g. study title, depositor name, and so on). At the document level (e.g. single interview transcript), mark-up will be based on the Text Encoding Initiative (TEI). The TEI was founded in 1987 to develop guidelines for encoding machine-readable texts of interest in the humanities and social sciences. It includes a large number of defined elements (e.g. <p> for paragraph) that are suitable for transcriptions and other social science data.



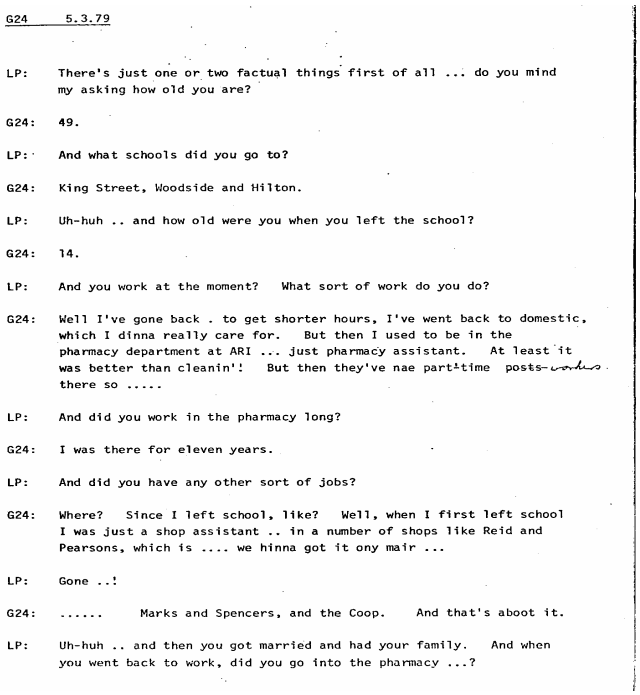
An example – searching on occupation

Mark-up of occupations, using native terms and standardised classifications, can enable data to be searched easily for specific information. Examples would be to search on a particular occupational classification, and retrieve all interview documents with that occurrence, or to browse documents by occupation. Another useful feature would be to view all occupations occurring in a single document to provide an overview. Finally, occupations occurring within text can be automatically extracted to provide classification metadata for each interview.



Original document – TIFF image produced by scanning original paper

This page is a direct scan from one of the Mothers and Daughters interviews deposited by Mildred Blaxter. ESDS Qualidata received this collection in paper format. Although the format is generally good (e.g. it has speaker tags and line breaks between speakers), there are areas for improvement as there is little identifying information about the interview or interviewee. In addition, though not visible in this image, the document produced from scanning can not always be searched as effectively as one that is fully digitised.



G24 5.3.79

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?

G24: 14.

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

G24: 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 posts there so .....

LP: And did you work in the pharmacy long?

G24: I was there for eleven years.

LP: And did you have any other sort of jobs?

G24: Where? Since I left school, like? Well, when I first left school I was just a shop assistant .. in a number of shops like Reid and Pearsons, which is .... we hinna got it ony mair ...

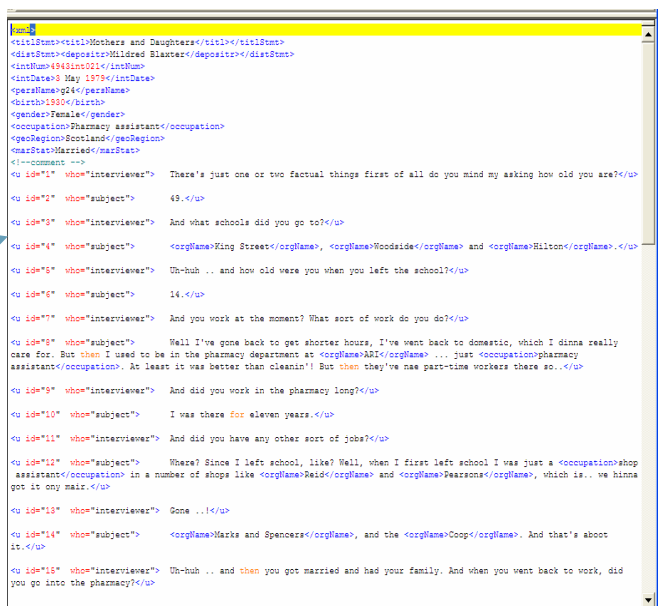
LP: Gone ..!

G24: ..... Marks and Spencers, and the Coop. And that's aboot it.

LP: Uh-huh .. and then you got married and had your family. And when you went back to work, did you go into the pharmacy ...?

Interview text with XML tags embedded

This shows the text with XML mark-up. 'Mark-up' simply refers to the tags, such as <occupation>, that are used to define text within the document. Several additions are visible. For example, metadata have been provided at the top of the file, such as study name, depositor name, interviewee name (or pseudonym), gender, and so on.



```
<?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 aboot
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>
```

Interview text with additional XML tags embedded

In addition, it is now possible to easily identify certain types of text. For example, organisational names are labelled <orgName>. This allows software (being developed at ESDS Qualidata) to partially automate the process of anonymising transcripts. The same procedure will be used to substitute pseudonyms for personal names.

```
<u who="#interviewer" xml:id="u1">There's just one or two factual things first of all do you mind my asking how old you are?</u>
<u who="#subject" xml:id="u2">49.</u>
<u who="#interviewer" xml:id="u3">And what schools did you go to?</u>
-<u who="#subject" xml:id="u4">
<orgName>King Street</orgName> , <orgName>Woodside</orgName> and <orgName>Hilton</orgName> .
</u>
<u who="#interviewer" xml:id="u5">Uh-huh .. and how old were you when you left the school?</u>
<u who="#subject" xml:id="u6">14.</u>
<u who="#interviewer" xml:id="u7">And you work at the moment? What sort of work do you do?</u>
<u who="#subject" xml:id="u8">
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
<seg type="occupation">pharmacy assistant</seg>
```

XML: enabling a standardised format for interview transcripts

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?
G24:14.
LP:And you work at the moment? What sort of work do you do?
G24: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..
LP:And did you work in the pharmacy long?

Once the XML tags are embedded in the document, they work in a way similar to a database. Different parts of text can be identified and placed within another document. In this case, some of the tags indicate metadata, such as name of interviewee and date of interview.

This template is ESDS Qualidata's 'model transcript' format. We recommend and promote this format as a research standard. It will aid interoperability and facilitate sharing.

From the XML version of the file, we produce a Rich Text Format (.rtf) file which is the version users receive when they download a collection of interviews from the ESDS Qualidata website.

XML: enabling web-enabled display, search and browse

Using a search query on the term ‘school’ in the Mothers and Daughters collection produces a list of interviews that contain the term. The interview identifier, g24 in this case, is highlighted. Clicking on it takes the user directly to the transcript - via Athens authentication as a layer of security protecting the data.

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 **...**

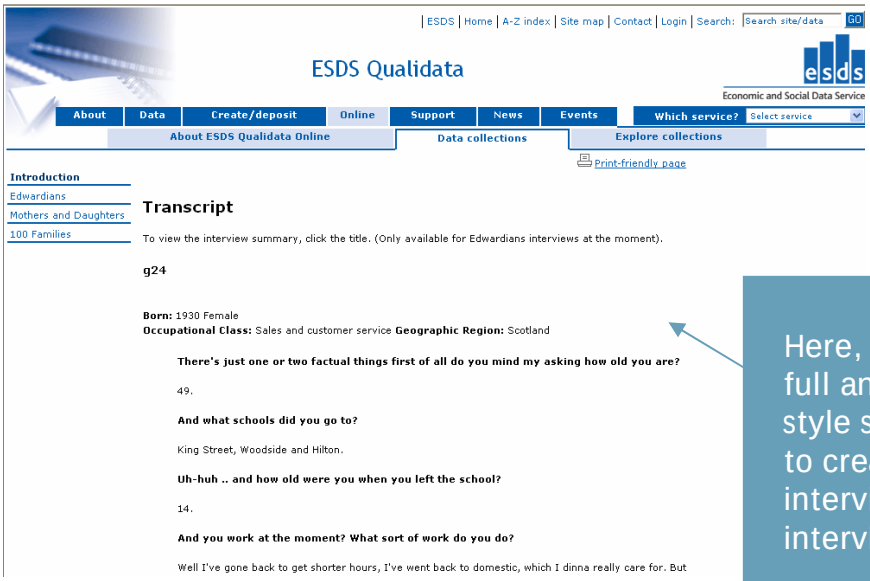
ID: g31
Born: 1929 Female
Occupational Class: Unemployed **Geographic Region:** Scotland
... , what sort of work did you do? When I first left the **school**? An upholstery sewer. I worked in Roberts in Union Street for 16 years, **...**

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 **...**

You are here : ESDS Qualidata Online > Explore > IndexZoom > Interview transcripts: search results



ESDS | Home | A-Z index | Site map | Contact | Login | Search: Search site/data GO

ESDS Qualidata

Economic and Social Data Service

About | Data | Create/deposit | Online | Support | News | Events | Which service? Select service

About ESDS Qualidata Online | Data collections | Explore collections

Print-friendly page

Introduction
Edwardians
Mothers and Daughters
100 Families

Transcript

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

g24

Born: 1930 Female
Occupational Class: Sales and customer service **Geographic Region:** Scotland

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

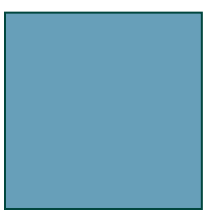
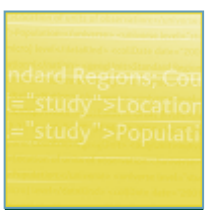
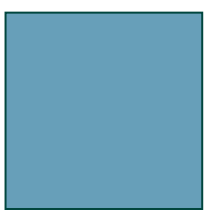
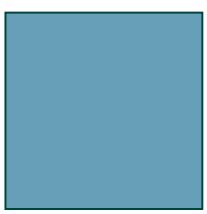
Here, the interview text is displayed in full and is available for browsing. A style sheet (a set of formatting guidelines to create web layout) displays the interviewer questions in bold and the interviewee responses in plain text.

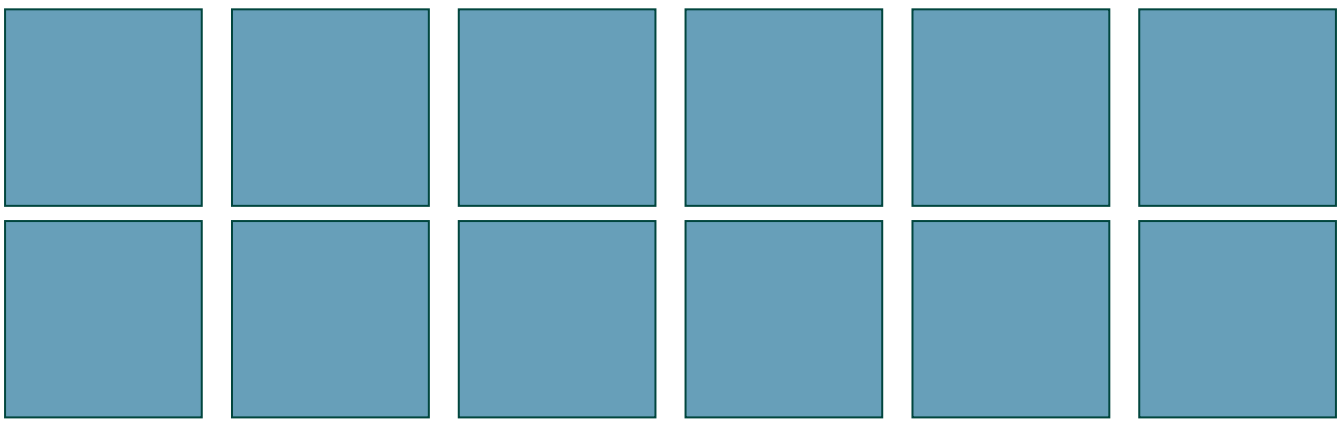
Useful URLs

ESDS Qualidata	www.esds.ac.uk/qualidata
ESDS Qualidata Online	www.esds.ac.uk/qualidata/online
QUADS	quads.esds.ac.uk
SQUAD project	quads.esds.ac.uk/projects/squad.asp
TEI	www.tei-c.org
DDI	www.icpsr.umich.edu/DDI/
Online QDA	onlineqda.hud.ac.uk/

Useful acronyms

ASCII	American Standard Code for Information Interchange
CAQDAS	Computer-Assisted Qualitative Data Analysis Software
DDI	Data Documentation Initiative
DTD	Document Type Definition
ESDS	Economic and Social Data Service
ESRC	Economic and Social Research Council
HTML	HyperText Markup Language
NLP	Natural Language Processing
PDF	Portable Document Format
QUADS	Qualitative Archiving and data Sharing Scheme
RTF	Rich Text Format
SQUAD	Smart Qualitative Data; ESD Qualidata's QUADS project
TEI	Text Encoding Initiative
TIFF	Tagged Image File Format
XHTML	Extensible HyperText Markup Language
XML	eXtensible Markup Language
XSL	Extensible Stylesheet Language
XSLT	XSL Transformations

QUADS
UK Data Archive
University of Essex
Wivenhoe Park
Colchester
Essex CO4 3SQ

Email: quads@essex.ac.uk
Tel: +44 (0)1206 872145
Fax: +44 (0)1206 872003
quads.esds.ac.uk