RESEARCH REPORT

1. Background

The ESRC Qualitative Archiving and Data Sharing Scheme (QUADS) aimed to support short-term research grants to develop new models of qualitative research archiving and data sharing which tackle in innovative ways the epistemological, ethical, methodological and practical challenges raised by the re-use and re-analysis of qualitative material.

In essence the SQUAD project sought to explore methodological and technical solutions surrounding: data context; systematic data descriptive standards, and information extraction and mark-up utilising language technology. These areas provide some of the key building blocks for enabling emerging innovations in qualitative methods, including increased, yet managed, access to data, linking data sources and data mining. The work builds on ten years of work in this field by ESDS Qualidata in enabling qualitative data sharing. But progress in data re-use has been somewhat hindered by preconceptions and sometimes less than innovative approaches to qualitative research (Corti and Thompson (2004)). However, a cultural shift is happening in a new willingness to share and utilise other research sources.

Fielding's (2003) scoping study examined issues for the role of qualitative data in e-social science and emphasised the need for 'tools that allow data to be published to the Web more easily and support online interrogation of data via standard Web browsers'. However, transforming the data into an acceptable web- or grid-exposable form is not straightforward, and a significant amount of manual effort is required which is both time-consuming and costly. This fundamental challenge was addressed by developing tools that support researchers and reduce the costs involved. Specifically, tools need to be developed for publishing marked-up enriched data and associated linked research materials (such as researcher observation or audio materials) to the web and for longer-term archiving.

SQUAD brought together research expertise and applications from social science research and methodology with computational linguistics as applied to qualitative data archiving and sharing. Evidence of successful bridges between these two disciplines is sparse and this project showed a practical contribution to interdisciplinary collaborative practice and innovation.

In creating tools the SQUAD project wished to provide user-friendly guidance that would help speed the process of adoption (by researcher communities and learners) of some of the methodological advances proposed in relation to qualitative research archiving, data sharing and re-use.

2. Objectives

The SQUAD project aimed to explore methodological and technical solutions for exposing digital qualitative data to make them fully shareable and exploitable. The three areas covered were XML standards and technologies for sharing qualitative data, contextualising research data and information extraction and anonymisation using automated tools.

The five main objectives were:

1. to specify and test commonly agreed 'open' standards for storing and 'marking-up' a wide range of qualitative data using universal (XML) standards and technologies.;

2. to investigate optimal requirements for contextualising research data (e.g. interview setting and interviewer characteristics), aiming to develop standards for data documentation and ways of capturing this information;

3. to develop and test user-friendly non-commercially-based tools for semi-automating (through the application of natural language processing technology) some of the very laborious processes already used to prepare qualitative data for both traditional digital archiving and more adventurous collaborative research and linking multiple data and information sources;

4. to research free, non-commercially based tools for online publishing and archiving marked-up data and associated linked research materials (Qualitative Data Mark-up Tools (QDMT);

5. to provide awareness-raising through the production of easy-to-follow guidelines and user-friendly step-by-step guides with exemplars centred on the use of these tools and the standards they utilise.

3. Methods and Results

The Methods and Results section have been merged and divided into three parts as the project investigated four quite distinct areas ranging from methodological to highly technical. The three areas are: 1. **XML standards**, covering metadata standards, including audiovisual archiving, data exchange standards and publishing tools; 2. **Capturing context; 3. natural language processing**, information extraction through named entity recognition, anonymisation and annotation.

3.1. XML Standards

One aim of the development work that was undertaken by ESDS Qualidata in the SQUAD project was to produce an application format that would enable sophisticated online searching of, and information retrieval from, digital materials. The data archiving community requires a standard and uniform format for richly encoding qualitative research that supports 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, newspaper articles, and so on). It is also essential that the application provide links between texts and associated audio and video materials, and indeed any other related object. 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. It must also support the development of common web-based publishing and search tools; and facilitates data interchange and comparison among datasets.

In defining the baseline schema for commonly collected and analysed qualitative data, EXtensble Mark-up Language (XML) was chosen. XML is useful because of its inter-operability and its extensibility. XML is widely used and recognises and new elements can be added easily. The R&D built on the work of ESDS Qualidata in enabling online access to multimedia qualitative data. SQUAD worked closely with the metadata consultants to extend the draft proposed ESDS Qualidata schema. It also consulted with CAQDAS software vendors to ascertain the level of buy in for a commonly agreed standards to support import and export between the commonly used software packages, such as Atlas-ti, Nvivo and so on.

This part of the SQUAD project produced several outputs:

3.1.1. Guide to searching and sharing qualitative data: the uses of XML

This document came through engagement with users (namely social scientists) who found the language we used in the project too technical. Through pictures the guide explains (in lay language) how XML is relevant to a broad range of very common practices in qualitative research and shows that XML enables new capabilities. It has been met with positive feedback.

3.1.2. XML schema for qualitative data

A core component of the XML application is a standard for *marking up* qualitative data. 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).

A simple DTD had already been developed by ESDS Qualidata for use in Qualidata Online that combined both file-level and content level metadata for social science resources. It incorporated a small set of basic metadata elements and established elements for the basic turn-taking structure typical of most interviews. In this project, the DDI (Data Documentation Initiative, DDI, 2007), already in use at UKDA, provides study and file-level metadata. In turn, the TEI (Text Encoding Initiative, TEI 2007) supplies content specific mark-up at the document level (e.g. g. single interview transcript). 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.

Using the DTD as a base, the SQUAD project hired an expert TEI consultant, James Cummings, at Oxford University Computing Centre, to further the development of the XML schema as it relates to TEI. He was charged with producing a schema for marking-up the most typical features of qualitative data (e.g., interview transcripts, still images, audio, video, and supporting research materials). The primary purpose of the schema is to enable web display of diverse forms of qualitative data.

The TEI elements cover a wide range of text formats and structures, set out in Annex 1. This version of the schema and documentation will be available on the ESDS Qualidata Online website. The SQUAD project worked with audio-visual data as case material for defining the metadata and data standards. The representation of audio-visual research data adds great power to textual output, but technical solutions are required to represent these materials in an efficient way. The ethical problems associated with sharing visual, can be counteracted by using XML as fine-grained and automated access control to data (e.g. if one wishes to let users see some data but not other parts of a collection) becomes possible.

3.1.3. Model transcript – layout and formatting

For all the tools and procedures used for processing and enhancing qualitative data, the initial format in which the data are received is critical. One outcome of the SQUAD project was to define a model transcript for both UKDA internal purposes and for others, usually prospective depositors, who want a consistent and standard format for their transcripts.

This model attempts to specify minimal requirements for preservation while not creating an unnecessarily complex template. There are three sections: a header that contains collection and interview information and appears on every page, a respondent details section that appears only on the first page, and the body of the transcription itself. This is shown in Annex 1.

In the header, key metadata about the collection from which the interview came is provided, such as the name of the collection, the depositor, and the interviewer and the date of the interview. In addition, any number assigned to the interview is here, as is the name (or pseudonym) of the respondent. Transcriber details and some version history (such as whether or not this transcript has been anonymised) are also provided. This information appears on every page because it is critical to identifying the respondent and connecting an individual interview to the collection from which it came.

The second set of metadata, respondent details such as date of birth and gender, are provided on only the first page of the document. The header includes the most critical metadata usually used to identify an interview, to be viewed as minimal. The respondent details are the ones most typically cited and are also provided (when available) in UKDA data listings.

This template is a guideline only and is expected to be adapted for specific projects, e.g. list of characteristics may vary, depending on the sample. For example, if children are a core research theme, then including the number of children might be appropriate here. The intent is not to have the transcript duplicate the full set of metadata, only to capture on the page what is most relevant and useful to those reading and analysing transcripts.

Regarding formatting content itself, there are very few guidelines. The intent was to use a format that was easily readable by humans and also conformed to the input guidelines for most qualitative data software applications. One requirement is to use a system of identifying speakers that is flexible enough to accommodate multiple respondents or focus group participants or interviewers. The use of the hyphen "-" instead of a colon makes this format compatible with the maximum number of qualitative software packages. Line spacing follows existing conventions of single spacing within a speaker's turn, and double spacing between speakers. As well as being a familiar and readable format, this spacing is also one required for some CAQDAS packages in order to automatically code questions and responses. Finally, text is added to indicate start and end points of each side of each audio tape.

Clearly, far more detailed specifications can be provided regarded transcription, such as the inclusion or not of non-verbal utterances, background noises, and conventions for transcribing accents, non-standard usage, grammatical errors, and so on. Such decisions are inextricably linked with the purposes of the research. The intent here has been to leave as much control over such matters to researchers, while still providing a flexible and useful model that meets minimum preservation requirements.

Represented as XML, different outputs can now be produced - create once, write many times. Figures 3a, 3b, 3c and 3d in Annex 3 show respectively the raw XML document, the .rtf version for download, a web page version in html and tagged items used in a web search.

3.1.4. Data exchange tools

An XML format that describes and represents a complex data collection is imperative. But, one that can enable value-added products (e.g. coding, annotation and analysis) to be imported and exported directly into and out of CAQDAS packages, avoiding the reliance on just a single product, and offering the opportunity to share analytic workings outside the confines of the particular software, would be even richer.

The project consulted with two set of collaborators to help guide work in this area. The first was a research group at Australian National University (ANU) who in 2006 proposed a standard called QDIF – Qualitative Data Interchange Format (Baden Hughes 2006). Based on an XML model this proposed a means of translating between the output of CAQDAS packages (e.g. Atlas-ti) and an open format. Meetings with the Australians took place but no testing of the standard was carried out, due to insufficient technical staffing in the SQUAD project (very limited programmer support). The other important communication was discussion via the qual-software JISC discussion list with CAQDAS software vendors, about the need for and interest in a data exchange format. When ESDS Qualidata began to promote its work on this idea, a flurry of emails was initiated between the markers leaders of CAQDAS packages. The biggest progress was that they agreed, in principal, that a common interchange translation format *was* needed. The Essex PI has been promoting this idea since 1996 through Qualidata and it is a welcomed breakthrough for them to break the silence and stalemate on this proposal.

While there was not time in the SQUAD project to focus more on the scoping and building of tools for exchange, Corti was successful in gained a follow-on JISC award to look at data exchange standards and tools, under the Repositories and Preservation Programme. The Data Exchange Tools and Conversion Utilities (DExT) funded from December 06 for a year project will develop, refine and test XML-based models for data exchange (for both survey data and qualitative research data) and will explore the development of tools for data import and export from some of the most popular social research software packages. Once a robust schema is in place it will easy to produce publishing tools to expose complex data collections to the web and archive for the longer term as a coherent 'bundle'.

3.2. Capturing data context

The SQUAD project aimed to come up with some measurable 'constructs' against which context could be assessed. The emphasis was on practical strategies with a positive outlook arguing that it is possible to capture and expose better and more systematically the context and the interrelationships among data and between data and other academic products, like analyses and write ups. This strand of work was approached by analyzing existing complex collections in the ESDS Qualidata catalogue to look at available context and discussing with users what areas of context would be useful to make their (re)-analysis easier. Sharing ideas with colleagues also enable some of the collation of base-line constructs to be drawn up. Two main outputs arose from the work

3.2.1 Workshop and special journal issue on defining and capturing context

A workshop was jointly organised by QUADS and SQUAD on context which fostered an opportunity for QUADS projects and other groups working with large collections of previously collected qualitative data, to share experiences from work in progress, with a remit of addressing context. An edited collection was secured by the Essex PI for the second issue of the new journal, *Methodological Innovations Online* of raw qualitative data. The editorial and papers are a testament to the analysis put into considering context as it relates to various kinds of collections.

3.2.2 Guidelines of collating context

A draft set of guidelines based on providing information about key aspects of the research project and resulting data has been drawn up for data creators and depositors. They should be viewed as mandatory elements for providing 'necessary' but maybe never 'sufficient' context. These are set out in the paper by Bishop (2006). The idea of *structuring* context has even been met with some criticism by the sceptics, as it apparently moves us away from the openness and complexity of qualitative data. However, a framework on which to hang study-specific context is useful. The SQUAD approach to researching context focused on the objectives of recreating context, concurring that original context cannot be recreated but can be 'recontextualised'. Multiple levels or layers of context were identified together with the processes of recontextualisation, from conversational context at the interview level to cultural context at the global level.

The SQUAD work offers practical advice on how to build up context information at the 'data unit' level (e.g. a single interview), such as descriptions of participants and interrelationships. The institutional/cultural level is also important and too rarely taken into account in the archiving process, – although ESDS Qualidata does provide chosen UK classic studies with as much published context as possible - newspaper clippings, article and book reviews and so on.

3.2.3 Teaching exercises on context

This advice has been made into a set of teaching exercises on context that have been used in post-graduate methods classes. The exercises, available from ESDS Qualidata, take the student through a reinterpretation with and without their associated contextual information. They have found to be rather insightful.

3.3 Information extraction and automated anonymisation

Most large document archives are simple repositories for documents with little thought given to improving the long-term search capabilities of those documents. Semantic Web initiatives aim to reverse this assumption by attaching machine-readable data to documents that will improve search precision. The SQUAD project investigated methodologies and technical solutions for *exposing the structured metadata* contained within digital qualitative data, to make them more shareable and exploitable.

3.3.1. Information Extraction Tools

. It developed mechanisms for using Information Extraction (IE) technology to provide user-friendly tools for semi-automating the process of preparing qualitative data in the social science domain for digital archiving, in order to archive enriched marked-up data. It also explored the possibilities of linking multiple data and information sources.

Identifying Useful Social Science Entities

Information Extraction (IE) is a sub-field of computational linguistics that aims to identify key pieces of information in unstructured texts using 'shallow' text analysis techniques. These techniques include a series of sub-tasks such as tokenisation and sentence boundary detection and Named Entity Recognition (NER). A typical IE system employs NER to identify, classify and mark-up particular kinds of proper names and terms. Examples of named entities include the names of people (individuals or groups), organisations, places (both physical locations and geo-political entities), occupations, dates, times and quantities such as sums of money or distances.

A second stage in a typical IE system will construct an information template by identifying and marking up particular facts relating to the entities that have been recognised (e.g. facts pertaining to an employment history such as employer's name, length of service, salary etc.) The goals of IE have been formalised largely in the context competitions such as the Message Understanding Conferences (MUC), the BioCreAtIvE competitions and the Automatic Content Extraction (ACE) program.

The two most common methods for identifying references to entities within a document are a rule-based approach in which entity recognition grammars are written and a machine learning approach, which includes the deployment and training of statistical taggers for entity recognition. In the rule-based approach, domain experts identify the contexts in which entities are found within documents, and write rules to locate these and classify the entity's type accordingly. Machine-learning techniques for named entity recognition rely on the use of training data constructed through the process of human annotation. Two or more annotators mark up the names of entities in a corpus of documents and the documents are fed into a statistical learning system, which automatically infers rules about the context and classification of the named entities. These models can also be informed by domain knowledge including gazetteers. In both cases, test data is used to calculate how well a system identifies and classifies named entities.

Improved Search

One of the aims in identifying and classifying named entities in qualitative data collections is to provide a framework for more precise and efficient web indexing and search. Full IE can be viewed as a means to annotate documents with semantic metadata, creating a machine-readable semantics for use in fully automated reasoning or highly sophisticated browsing and search. SQUAD used a range of XML-based language processing tools (Specifically, LT-XML , LT-TTT and LT-TTT2) that are exploited to reduce the manual efforts that are typically required to create marked-up data with shallow semantic information (including entities such as person names, company names, place names, and temporal information). Five types of entities were nominated which were considered to be broadly useful within the social science domain. These included the names of people, organisations, locations, occupations and dates selected to to enable: (i) linking between entities within and between documents and (ii) anonymisation of names or places (as described next). An example of an un-marked-up interview and a marked-up one is shown in Annex 2.

Automated Anonymisation

The SQUAD tools provided another potentially useful extension of IE research, that of data anonymisation. Effective editing of data, such as interview transcriptions, can involve using pseudonyms, abstract systems of coding or simply the crude removal of text. Manual anonymisation is time-consuming and labour-intensive. Providing user-friendly tools to semi-automate this process when preparing data for archives would be extremely beneficial in increasing the flow of web-enabled data. Currently in the international realm only a few new projects are utilising IE tools in this context (e.g. Poesio *et. al.*, 2006). Just identifying named entities is not enough as many texts include *co-references* (as when "John Smith" is later referred to as "Mr Smith"), and true anonymisation should consider this. This kind of reference resolution has attracted a great interest in the NLP community in recent years.

Annotation and Evaluation

In evaluating the performance of the IE tools, standard practice was followed, by comparing the system output to that of a ‘gold-standard’. The gold-standard was developed and inter-annotator agreement scores calculated. Seven collections of interview transcripts were annotated with the names of people, locations, organisations, occupations and dates. The corpus contains thirty documents with an average of 12,019 words per document. Eleven of the documents were annotated by two people in order to calculate inter-annotator agreement scores. The agreement on person names and locations were as high as one might expect. However, the definitions for when to annotate the other entity types were obviously not as clear-cut. In particular, occupations are significantly lower than any other entity type. The reason for this was that annotators struggled to identify when an occupation was specific (e.g., “I worked as a cleaner”) or generic (e.g., “A teacher told me...”). The annotation guidelines focused on specific references to job titles held by people being discussed in the study. Generic statements that talk about people in particular positions (such as past teachers, doctors) with no references to names are important for the interview analysis, but are less important for search and anonymisation. However, specific and generic cases are not always easy to distinguish.

Named Entity Recognition Evaluation

The SQUAD system's performance on recognising the names of entities in social science data was evaluated. In general, both precision and recall are lower than what was hoped to be achieved. The names of entities are primarily extracted using NER models trained on newswire data. However, documents within the social science context are entirely different in both content and presentation to the annotated documents available for training machine learning models. These are set out in Table 1 in Appendix 2.

Geographic Referencing

The recognition of *location* entities warrants further discussion in its own right since location names in the text have the potential to be the basis for interfacing with a variety of GIS technologies. Associating accurate geo-spatial metadata with archived documents will allow for more accurate geography-based search of qualitative data, as well as enabling interfaces with mapping technologies.

The LTG has collaborated with Edina on the development of a geoparser as part of the JISC-funded GeoCrossWalk project. The geoparser is currently available as demonstrator which allows the user to upload a document or webpage which is then processed to discover the location names in the text. The locations are looked up against the GeoCrossWalk gazetteer and there is a map-drawing graphical user interface to allow the user to select the correct gazetteer entry so that each location name is paired with a unique geo-spatial footprint. LTG and Edina are currently continuing their collaboration in phase 5 of GeoCrossWalk where the focus will be on improved location name recognition paired with a component for the automatic disambiguation of location names. The aim is also to generalise the gazetteer component so as to permit look-up and disambiguation against other gazetteers because the coverage of the GeoCrossWalk gazetteer is limited to Great Britain.

The existing demonstrator version of the geoparser is a prototype was not mature enough to be easily integrated with the SQUAD processing model and user interface. In particular, it needs a map-drawing graphical user interface so that the user can disambiguate the location names (see Leidner 2007) and it would have been too costly to integrate this with the Squad interface. For this reason it was not possible to provide geo-referencing functionality in the Squad demonstrator, but it is hoped the next version will enable this.

Anonymiser and annotation user tools

The IE work described above all happens 'behind the scenes' in that data are parsed through processing 'pipes' on linux or unix machines. The average user does not get to see anything meaningful and certainly cannot interact with the system. The CME system was in charge of the *automatic* annotation (NLP) pipeline for mark-up of named entities) phase described above. Annotation tools such as the NITE XML Toolkit (Carletta *et al.*, 2003), WordFreak and MMAX2 allow users to *manually annotate* documents with semantic information. The project developed a solution in which the NITE XML Toolkit was integrated with the Edinburgh IE tools in order to semi-automate the process of annotating important entities in social science documents, called SQUADCoder.

.

SQUAD built a *graphic user interface* to the NLP tools to enable interview files to be loaded, and the marked-up entities checked for accuracy. This was called the SQUADRunner, which basically was a wrapper subsystem that integrated the two primary subsytems - CMe and SQUADCoder.

After the automated phase of annotation has been run and the SQUADCoder invoked, an initial window allows the user to indicate where the file is to annotated (the NXT metadata file). The main SQUADCoder window is then opened where new named entities can be tagged in the text. Another key feature is that the system enables *co-reference chains* identified in the NLP pipelines to be highlighted and then anonymised in one swoop, using the *anonymise* option. Some exemplary screen dumps for the SQUADCoder Tool for annotation and anonymiser are shown in Annex 3.

The NITE-NXT on which it is based system saves the original text file (with the named entity mark-up), creates a new anonymised version, saves a matrix of references - names to pseudonyms, and outputs the annotations (e.g. who worked on the file etc).

4. Activities

Promotion of the SQUAD work was made via the QUADS and ESDS proactive outreach work and to key centres and networks. The latter included the international network of social science data archives (IASSIST), National Centre for e-Social Science (NceSS), Association Of Survey Computing (ASC) and the Language Resources and Evaluation Conference (ALREC) international forums where papers were delivered at their international meetings in 2006.

The QUADS and scheme and projects were promoted at around 45 different events over the 18 months, in which the SQUAD work often featured. The team were invited to present on all aspects of the work from context to text mining. SQUAD related presentations are shown in Annex 4.

SQUAD helped organise the workshop on context described in section 3.2.1, and contributed papers to the QUADS Online Resources Day in November 2005 and the NCRM Summer School in September 2006.

The SQUAD project also made collaborations with CAQDAS vendors and an Australian group in an agreement to held define a data exchange format for qualitative data (which ahs been taken forward).

5. Outputs

Some of the less technical outputs from the SQUAD project have been uploaded to the Society Today website. Some neat QUADS branded promotional materials, including a 2 sider sheet and a poster to summarise the SQUAD findings were produced. They are available at http://quads.esds.ac.uk/news/showcase.asp. A basic web presence was established but the content is spread across sites due to the natures of the strands. The metadata work sits on the ESDS Qualidata Online site as it follows on from and is being followed by key R&D in this area. quads.esds.ac.uk/projects/squad.asp and http://www.esds.ac.uk/qualidata/online/about/dtd.asp

Technical documentation for the named entity mark-up and SQUAD anonymiser tool were created. The draft metadata schema is also on the technical side but available as TEI DTD.

Three other less technical substantive outputs should be highlighted. The first is the Guide to searching and sharing qualitative data: the uses of XML described in section 3 which helps to demystify XML and has had great feedback. The second is the edited collection of papers collated form the second QUADS workshop on 'Defining Context for Qualitative Data'. Seven papers plus a long introduction were included in a 2006 special edition of the Methodological Innovations Online (Corti 2006). The SQUAD contributions were the Corti editorial and the Bishop article on capturing context. Finally, the work on context was made into a set of teaching exercises for use in post-graduate methods classes. The exercises, available from ESDS Qualidata, take the student through a reinterpretation with and without their associated contextual information.

A few articles have been published that cover some of the SQUAD input into the work on context to some of the NLP work (See Annex 4).

6. Impacts

The biggest impact has been the interest shown by the CAQDAS software vendors for the data exchange standards, which is set out in Section 3.1.4 and which has spun out into a funded award under JISC (PI: Corti 2006). The text mining and anonymisation applications have also gained some credibility, for example with the National Centre for Text Mining (NACTEM). Finally, the work on context and XML model transcript is being taken up by some keen depositors wishing to share data and by the Leeds based ESRC Timescapes project, who are hoping to adhere to the standard for their satellite archive of ESDS.

7. Future Research Priorities

All areas of the SQUAD work will be continued in some form, under the auspices and management of ESDS Qualidata. The data exchange work is already funded and proceeding.

The NLP tools developed here are very much proof of concept. Ideally they would form part of a suite of tools that would enable all step of processing a text for archival dissemination. There is a number of interesting future directions that could be explored based on the NLP work. Automatically identifying the relationships between documents on the basis of their entities is a particularly interesting area. This is important for automatic cross-document anonymisation in order to ensure that common pseudonyms are used for the same entities across documents. Moving on from named entity recognition to activities even closer to the ultimate goals of the Semantic Web initiative, such as key word extraction based on chosen ontologies or folksonomies, is also a direction the UKDA would like to explore. The automatic summarisation of texts that can also be employed to reduce a piece of text to its key attributes, for example a condensed interview. This might also prove to be useful for efficient automatic comparison of documents for a qualitative researcher (see Milosavljevic, 1997; 2003). A bid for looking further into information extraction and automated indexing of qualitative data using text mining tools has already been submitted to ESRC.

The SquadCoder tools were very much a proof of concept, and a bit rough and ready. They would really benefit from more work to make them bug free and trial them in Data Archiving working practices for preparing data. This would be very useful and feed into ESDS practice. This work could also fall under the e-science type banner looking at workflow or collaborative tools and virtual research environments, but this is more tenuous. However, there is ongoing dialogue between Essex , Edinburgh and Manchester NLP groups to ensure the collaboration does not completely stop.

Finally, the findings on context are being edited into a Best Practice Guide, but little more will be done on that front, as it has been well covered in the dedicated journal issue.