

# Towards Interoperable Secondary Annotations in the E-Social Science Domain

Baden Hughes<sup>1</sup>, Desmond Schmidt<sup>2</sup>, and Andrew E. Smith<sup>2</sup>

<sup>1</sup> Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia

<sup>2</sup> Key Centre for Human Factors and Applied Cognitive Psychology, University of Queensland, Queensland 4072, Australia

Email address of corresponding author: [badenh@csse.unimelb.edu.au](mailto:badenh@csse.unimelb.edu.au)

**Abstract.** The sharing of data for secondary analysis has been very limited, especially in the social sciences. The reasons usually cited for this limited sharing are (1) strong privacy requirements on data and (2) lack of appropriate contextual knowledge by secondary investigators. We argue that a third reason, the lack of interoperability between software tools commonly used for data annotation and coding by social scientists, is critical even if the problems identified by earlier researchers are resolved. In this paper, we describe our work in attempting to address the data interoperability issue directly by developing standards for the syntactic expression of annotation, and core libraries which can be used to manipulate the annotations in their standard format, as well as the overall system architecture and examples of analytical applications which can be used for secondary coding and analysis.

## Introduction

Within the social sciences, the humanities and other fields such as information systems, the analysis of qualitative data is becoming an increasingly important research activity. Qualitative data can include spoken and written natural language, photographs, audio and video taken from: interviews, focus groups, ethnographic settings, open-ended survey answers, historical writings, observational notes and language corpora, to name a selection.

Qualitative data is particularly important in fields of research that deal with very complex social and cultural problems, and in emerging areas of inquiry. Fundamental problems for example in education, management and social policy are all highly suited to qualitative research. These and other complex areas of social, economic and cultural life benefit from the depth of insight provided by the ‘thick descriptions’: the nuanced, fine-grained and

penetrating observations that bring a deeper understanding of the complex dynamics that are emblematic of modern societies.

Given the potential for these data to contribute to deepening understandings of how societies and economies operate, there are many benefits in facilitating qualitative researchers' efforts to store and access data. Qualitative data sets are typically very large, for example, corpus linguistic analysis is frequently done on multi-million word data sets; ethnographies normally draw on large and multiple data sets of interview recordings and transcripts, field notes and documentary evidence; historians routinely draw on extensive documentary archives in state libraries and other places, and media and communication studies draw regularly on very large sets of video footage.

The growth of social science data archives throughout the world in recent years (e.g. Arces (Arces, n.d.) in Spain, Sophist in Russia (Sophist, n.d.), ZA-Information in Germany (ZA-Information, n.d.)) has made it clear that qualitative data is very much the poor cousin to quantitative data when it comes to archiving and interchange. Only a few archives (e.g. Qualidata in the UK (ESDS Qualidata, n.d.), the Murray Research Archive at Harvard (Henry A. Murray Research Archive, n.d.) and the Finnish Social Science Data Archive (Finnish Social Science Data Archive, n.d.)) currently store and distribute qualitative social science data. Even then, their holdings are limited to primary data sets, whereas in qualitative data analysis applications the secondary comments and coding are just as important.

An important point to be emphasised is that the value of qualitative data to the community would be amplified if these multiple types of data could be stored, searched, retrieved and shared more freely, and subject to secondary analysis by a larger group of researchers from a variety of disciplines, using a range of analytical methods. However, it is disappointing that the sharing of data for secondary analysis has been very limited, especially in the social sciences (Fielding, N., 2003, Corti, L., Witzel, A. and Bishop, L., 2005). The reasons usually cited for this limited sharing are (1) strong privacy requirements on data (Kuula, A., 2000a, Corti, L., 1995), and (2) lack of appropriate contextual knowledge by secondary investigators (Kuula, A., 2000b, Adams, A., 2005). We argue that a third reason, the lack of interoperability between software tools commonly used for data annotation and coding by social scientists, is critical even if the problems identified earlier are resolved.

We describe our work in attempting to address the data interoperability issue directly by developing standards for the syntactic expression of annotation, and core libraries which can be used to manipulate the annotations in their standard format. We also describe the overall system architecture and give examples of analytical applications which can be used for secondary coding and analysis. The structure of this paper is as follows. First we outline the background to, and motivation for our work, including the software survey. Next we describe the data interchange format in depth, followed by a description of our implementation experience and plans. Finally we identify a number of items for future work and draw conclusions.

## Background and Motivation

A survey of a number of commonly used qualitative data analysis tools reveals that moving annotated or coded data from one application to another is not usually supported (Fielding, N., 2003). We considered a range of popular tools including:

NVIVO (Richards, L., 2005)

ATLAS.ti (Muhr, T., 2000)

Leximancer (Smith, A.E. and Humphreys, M.S., 2006 to appear)

The Ethnograph (Seidel, J.V. and Clark, J.A., 1984)

Hyperresearch (ResearchWare Inc., n.d.)

Decision Explorer (Cropper, S. et al, 1995)

Maxqda (MaxQDA Verbi GmbH, n.d.)

Qualrus (Idea Works, n.d.)

Tatoo (Rostek, L. and Alexa, M., 1997)

All these tools were analysed from the perspective of data import/export, particularly examining how an annotation or coding scheme was able to be expressed internally to the host application and subsequently exposed to other applications. With two exceptions (ATLAS.ti and Tatoo) the data format and associated annotation or coding cannot be transferred from one analytical framework to the other. Hence from the perspective of data sharing and interoperability, it can be seen that the technological barrier is in fact considerable.

There are many methodologies associated with secondary markup of qualitative data. Perhaps the most well-known is Grounded Theory (Glaser, B.G., and Strauss, A.L., 1967); (Glaser, B.G., 1992), a methodology for allowing the emergence of categories from the data, and their gradual accretion into theories. Kelle has shown how modern qualitative data analysis tools have not in fact developed specifically from Grounded Theory, but have arisen from a variety of other approaches as well (Kelle, U., 1997). Against this background it would therefore be unwise to follow too strictly a single methodology in attempting to create a format for interchange of qualitative data. We have instead based it more loosely around simple annotation technologies, with some higher level constructs to enable hierarchical or overlapping relations among codes. The goal is to represent 80-90% of the features shared by the most commonly used qualitative data analysis programs. Any feature that can be meaningfully exchanged between at least two existing applications has potentially a place in the standard. Any program that does not understand a feature can then simply ignore it. The feature set described below is not yet complete, but is intended to serve as a basis for future discussion.

## Qualitative Data Interchange Format (QDIF)

We propose the Qualitative Data Interchange Format (QDIF)<sup>1</sup> for the exchange of annotated textual document collections, with special reference to qualitative social science data. The general data model was derived from a study of a number of qualitative data analysis tools mentioned above.

---

<sup>1</sup> <http://delacruz.csse.unimelb.edu.au/portcode/qdif>

QDIF is essentially an abstract standard that may be expressed in a variety of formal languages, although it has initially been described and tested in XML (QDIF Schema, n.d.). However, when it was recently presented at a workshop on the interchange of qualitative data (Smith, A.E. and Hughes, B., 2006), delegates expressed a desire that any such standard should not merely allow the conversion of data from one program to another, but also its dissemination on the Web. It is therefore planned to convert the coding to RDF (Beckett, D., 2004).

A QDIF document consists of components, which may be nested according to a formalism described below. The names of the components are indicated by italics and by an initial capital.

**Data Bundle** The top-level QDIF object for interchange is called the *Data Bundle*. A *Data Bundle* consists of:

- optional *Metadata* for the *Data Bundle* itself, as for other major components
- one or more *Primary Documents*
- zero or more *Comments*, i.e. free text annotations
- zero or more *Selections*, i.e. selected extents within *Primary Documents* or *Comments*
- zero or more *Codes* or concept tags
- zero or more *Relation Names* used in *Named Relations* within the *Data Bundle*
- zero or more *Named Relations* between *Codes* and *Codes* or between *Selections* and *Selections*, and unnamed *Relations* between *Codes* and *Selections* or *Comments* and *Selections*.

Apart from the *Primary Documents*, which are only referred to by the *Data Bundle* and exist independently, all the other components are contained within a single QDIF file. The components must follow this order to facilitate automated parsing.

**Metadata** *Metadata* may be specified for the *Data Bundle*, for *Primary Documents*, *Comments*, *Codes* and *Selections*. *Metadata* fields are based on standard Dublin Core Metadata terms (DCMI, 2005). All the fields are optional, as is the *Metadata* component itself. Some archives have already begun using the DDI metadata standard for describing qualitative data (Kuula, A., 2000b, ESDS Qualidata, n.d.). This decision, however, cannot be dissociated from their practice of archiving only primary documents, and they both admit that the DDI standard is in fact unsuited to qualitative data. Since the *Metadata* component is embedded in the *Data Bundle* at various points it makes more sense, at least in the case of XML or RDF, simply to refer to external definitions of individual fields as required on a case by case basis, rather than import an entire markup scheme. Although the following is not intended as an exhaustive list, currently defined fields thought to be useful in this context include:

- title – the name given to the resource e.g. a *Primary Document* or a *Data Bundle*

- creator – the entity primarily responsible for making the content of the resource
- description – an account of the content of the resource
- language – the language of the intellectual content of the resource. ISO639-3 is suggested as a standard, e.g. ‘eng’ represents English (ISO639-3, 2006)
- source – a reference to a resource from which the present resource is derived
- created – the date of creation of the resource
- modified – the date on which the resource was last changed
- rights – information about rights held in and over the resource
- rightsHolder – a person or organisation owning or managing rights over the resource

**Primary Documents** The *Primary Document* is a physical primary source file, and may reside on a local or remote file system or website. A *Primary Document* has a unique document-wide identifier, and consists of a URI, specifying its location, and a *Checksum*. Optionally a *Primary Document* may also contain a *Metadata* component and an indication of the encoding of its contents, e.g. UTF-8, USASCII etc.

**Checksums** *Checksums* are needed in order to verify that the content of a *Data Bundle* matches its externally referenced *Primary Documents*. A single *Checksum* consists of a number, which may be expressed in text form, for example, as a hexadecimal string. The format and length of the number is determined by the required *Checksum* type property. The only *Checksum* types currently defined are CRC32, Adler32 and MD5.

**Comments** A *Comment* is a short note or extended memo. No distinction is made between short *Comments* relating to a *Selection*, and extended secondary annotations, which may themselves contain *Selections* and be commented upon in their turn. A *Comment* must include a *Body* and may contain *Metadata*. It has a document-wide unique identifier.

**Selections** A *Selection* specifies a segment within a *Primary Document* or *Comment*, to which it may be related by its document-wide unique identifier. The selection mechanism varies according to the type of target document. *Selections* in text documents may be *SpanSelections*, *LineSelections* or *XMLSelections*. *Selections* in graphical documents are specified by *GraphicSelections*. Other types of *Selections*, intended to cover video and audio data, are envisaged for a future version of the QDIF standard.

- A *SpanSelection* is defined by two numbers giving its start and length. This method is suitable for plain text documents or *Comments*.
- A *LineSelection* is defined by four numbers specifying the start line and the offset within that line of the first character of the selection, and by an end line and end offset, which is just beyond the end of the selection. This method is designed to overcome the difficulty of importing or exporting selections to or from a proprietary viewer or editor, where offsets within the file do not always correspond to actual character positions.

- An *XMLSelection* is specified by an XPointer range (Grosso, P. et al (eds), 2003). This method is suitable for XML documents.
- A *GraphicSelection* consists of four numbers representing the top, left, bottom and right coordinates of points or pixels in a Cartesian coordinate system where 0,0 is the top left corner of the image.

**Codes** A Code or ‘tag’ is a short label for a concept, and is used to mark up a document. Each Code has a unique document-wide identifier and a name. It is used as a component in expressing ontological relationships between Codes and Selections or Codes with other Codes.

**Relation Names** A *Relation Name* defines a label for a *Named Relation*. The list of *Relation Names* includes all those names utilised in *Named Relations*, which do not refer to an external standard such as Dublin Core, e.g. ‘dcterms:isPartOf’ (DCMI, 2005). Examples of names might be: ‘isAssociatedWith’, ‘isPartOf’, ‘isA’, ‘causeA’, ‘contradictsA’ but new *Relation Names* can be freely invented for particular applications. A *Relation Name* may also have a description.

**Relations and Named Relations** *Relations* express an ontological relationship between two components of a *Data Bundle*, one component acting as the subject and the other as the object. Thus a *Relation* acts like a verb. *Relations* between *Codes* and *Selections* or *Comments* and *Selections*, being the most common, are treated as unnamed, although they can be conceived of as containing an implicit ‘is a’ or ‘is an instance of’ relation. A *Relation* always has at least two properties: a subject and an object. In the case of a *Named Relation*, which may be between two *Codes* or two *Selections*, an additional *RelationName* property is required. Both named and unnamed *Relations* may contain an optional comment.

The ability of *Codes* to relate to one another allows a *Data Bundle* to record a complex ontology or conceptual network, which in some applications may be visualised as a map. Other kinds of *Relation* other than the four basic types described here are not currently considered useful.

## Example QDIF Document

```
<?xml version="1.0"?>
<dataBundle xmlns="http://delacruz.csse.unimelb.edu.au/qdif"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xsi:schemaLocation="http://delacruz.csse.unimelb.edu.au/qdif
  http://delacruz.csse.unimelb.edu.au/portcode/qdif.xsd">
  <metadata>
    <dc:title>Example 6</dc:title>
  </metadata>
```

```

<pdocs>
  <pdoc id="doc1">
    <metadata>
      <dc:title>Alices Adventures in Wonderland</dc:title>
    </metadata>
    <checksum><value>F21B34C0</value><type>CRC32</type></checksum>
    <loc>
      http://www.gutenberg.org/dirs/etext91/alice30.txt
    </loc>
    <mimeType>text/plain</mimeType>
    <encoding>USASCII</encoding>
  </pdoc>
</pdocs>
<comments>
  <comment id="com1">
    <body>Alice considers her sister's book dull because it has no
      pictures or conversation</body>
  </comment>
</comments>
<selections>
  <spanSelection id="sel1" doc="com1">
    <!-- define the selection "considers her sister's book dull" -->
    <start>6</start>
    <length>32</length>
  </spanSelection>
  <lineSelection id="sel2" doc="doc1">
    <startLine>273</startLine>
    <startOffset>33</startOffset>
    <endLine>274</endLine>
    <endOffset>49</endOffset>
  </lineSelection>
  <lineSelection id="sel3" doc="doc1">
    <startLine>1513</startLine>
    <startOffset>6</startOffset>
    <endLine>1513</endLine>
    <endOffset>60</endOffset>
  </lineSelection>
</selections>
<codes>
  <code id="cod1"><name>boredom</name></code>
  <code id="cod2"><name>tiredness</name></code>
</codes>
<relationNames>
  <relationName name="isInstanceOf">
    <description>
      the subject is an example of the object
    </description>
  </relationName>

```

```
</relationNames>
<relations>
  <!-- relate "boredom" to "falls asleep" -->
  <codeSelRelation subject="cod1" object="sel1"/>
  <!-- relate comment about falling asleep to Alice doc -->
  <comSelRelation subject="com1" object="sel2"/>
  <!-- identify where Pigeon complains about being tired -->
  <comSelRelation subject="cod2" object="sel3"/>
  <!-- connect "tiredness" with "boredom" -->
  <codeCodeRelation subject="cod2" object="cod1"
    name="isInstanceOf"/>
</relations>
</dataBundle>
```

## Implementations

QDIF is currently based on a fully documented Java library incorporating all the features described above. Input and output of QDIF itself is currently unsupported while the physical format is still being refined. The only way of currently reading a *Data Bundle* is to import an ATLAS.ti project, but it is hoped to add other popular binary formats in the near future.

## Leximancer

The first analytical application is an enhanced version of Leximancer (Smith, A.E., 2000, Smith, A.E. and Humphreys, M.S., 2006 to appear). Leximancer is a tool that can be used to analyse automatically the content of document collections in a variety of common binary and textual formats. In contrast to other qualitative analysis software, it derives concepts from the source documents rather than requiring the user to code sections of the text manually or semi-automatically. The extracted information is displayed primarily via an explorable map, showing the strength of interrelationships between concepts. Our extensions to Leximancer focus on supporting the QDIF library, and allowing for the import and export of secondary annotations and encodings. Another key difference between this enhanced version of Leximancer and other qualitative data analysis software is that it will be implemented as a web application. All other QDA software is currently desktop-bound. This opens up possibilities for viewing pre-generated concept maps over the web and of conducting qualitative data analysis online.

## Annozilla

The second client is a browser based annotation tool, extended from the open source Annozilla suite (Wilson, M., n.d.). Annozilla is designed to view and create annotations associated with a web page, using the W3C's Annotea standard (Kahan, J. et al, 2001). Annotations are stored as RDF on a server, using XPointer (or at least XPointer-like

constructs) to identify the region of the document being annotated. Annozilla leverages the native functionality of the Mozilla web browser to manipulate annotation data and its built-in RDF handling to parse and submit annotations. Our extensions to Annozilla involve extending the application level support for RDF embedded annotations, and for the library functions described earlier.

## Future Work

Our experience in instantiating the first version of QDIF has led to the identification of a number of open issues that motivate future work. This list is by no means exhaustive, and we actively solicit input from the e-Social Science community and from interested tool builders to refine the requirements further.

- Some work still needs to be done on deciding which metadata elements can be safely included in QDIF, and from which existing standards. Making them dependent on an external reference has the disadvantage that changes in the chosen external standard will also endanger the validity of archived QDIF documents.
- How will QDIF bundles be stored and distributed in practice? The most logical technique would seem to be to include all primary documents in one folder and also store there the QDIF document describing them. The primary document URIs would need to be altered to point within the bundle.
- The expression of QDIF as RDF needs to be checked for suitability and the details clarified.
- Audio and video selections need to be added to the format, including differently shaped regions for video.
- A journaling facility has been identified as a possible enhancement, to enable recording of the development of qualitative coding as a process. However, this feature is not currently supported by any of the programs examined, and might be better implemented via an external version control system.

## Conclusion

While our research and associated development is still relatively immature, we are optimistic that a technological solution to the data (and annotation) exchange problem is feasible in the short to medium term. What remains are the underlying issues in data privacy and contextual interpretation, which are largely policy and social issues. Our hope is that the availability of a technological solution for the exchange of annotated social science data will motivate researchers to engage with the difficult questions that need to be addressed in these other areas.

## References

- Adams, A. (2005): 'Grounded Theory: Case Studies and Methodological Issues.' in *Proceedings of the JCDL 2005 Workshop on Studying Digital Library Users in the Wild*. Association for Computing Machinery. [http://www.dlib.org/dlib/july05/khoo/01\\_adams.pdf](http://www.dlib.org/dlib/july05/khoo/01_adams.pdf)
- Arces (n.d.): Centro de Investigaciones Sociológicas. <http://217.140.16.67/cis/opencms/EN/index.html>
- Beckett, D. (2004): *RDF/XML Syntax Specification (Revised)*. <http://www.w3.org/TR/rdf-syntax-grammar/>
- Corti, L., Witzel, A. and Bishop, L. (2005): 'On the Potentials and Problems of Secondary Analysis. An Introduction to the FQS Special Issue on Secondary Analysis of Qualitative Data.' *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 6(1), Article 49. <http://www.qualitative-research.net/fqs-texte/1-05/05-1-49-e.htm>
- Corti, L. (1995): 'Qualidata Resource Centre.' *EURODATA Newsletter* 1995, No. 2, pp.14-16.
- Cropper, S. et al (1995): 'Keeping Sense of Accounts Using Computer-Based Cognitive Maps.' *Social Science Computer Review*, 1990(8), pp. 345-366.
- DCMI (2005): *Dublin Core Metadata Initiative Metadata Terms*. <http://dublincore.org/documents/dcmi-terms/>
- ESDS Qualidata (n.d.): *ESDS Qualidata*. <http://www.esds.ac.uk/qualidata/>
- Fielding, N. (2003): *Qualitative Research and E-Social Science: Appraising the Potential*. ESRC Consultative Study. Engineering and Social Research Council, UK.
- Finnish Social Science Data Archive (n.d.): *Finnish Social Science Data Archive (FSD)*. <http://www.fsd.uta.fi/english/>
- Glaser, B.G., and Strauss, A.L. (1967): *The discovery of grounded theory*. Chicago: Aldine.
- Glaser, B.G. (1992): *Basics of Grounded Theory Analysis: Emergence Vs. Forcing*. Mill Valley, Ca., Sociology Press.
- Grosso, P. et al (eds) (2003): *XPointer Framework*. World Wide Web Consortium (W3C). <http://www.w3.org/TR/2003/REC-xptr-framework-20030325/>
- Henry A. Murray Research Archive (n.d.): *Henry A. Murray Research Archive*. <http://www.murray.harvard.edu/mra/index.jsp>
- Idea Works (n.d.): *Qualrus*. <http://www.ideaworks.com/Qualrus.shtml>
- ISO639-3 (2006): *ISO/DIS 639-3 Codes for the representation of names of languages - Part 3: Alpha-3 code for comprehensive coverage of languages*. <http://www.sil.org/iso639-3/>
- Kahan, J. et al (2001): 'Annotea: An Open RDF Infrastructure for Shared Web Annotations.' *Proceedings of the 10th International World Wide Web Conference*. International World Wide Web Conference Committee.

- Kelle, U. (1997): 'Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data.' *Sociological Research Online*, 2(2). <http://www.socresonline.org.uk/2/2/1.html>
- Kuula, A. (2000a): 'Laadullisen aineiston jljill.' *FSD Bulletin Newsletter Archive*. No. 3. pp. 16-20. English translation: <http://www.fsd.uta.fi/tietoarkistolehti/english/03/quali.html>
- Kuula, A. (2000b): 'Making Qualitative Research Material Reusable: Case in Finland.' *IASSIST Quarterly* 24(2). pp.14-17.
- MaxQDA Verbi GmbH (n.d.): *MaxQDA*. <http://www.maxqda.com/>
- Muhr, T. (2000): 'Increasing the Reusability of Qualitative Data with XML.' *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 1(3). <http://www.qualitative-research.net/fqs-texte/3-00/3-00muhr-e.htm>
- QDIF Schema (n.d.): *Qualitative Data Interchange Format (QDIF) Schema*. <http://delacruz.csse.unimelb.edu.au/portcode/qdif.xsd>
- ResearchWare Inc. (n.d.): *HyperRESEARCH*. ResearchWare Inc. <http://www.researchware.com/hr/index.html>
- Richards, L. (2005): *Handling Qualitative Data: a practical guide*. London: Sage.
- Rostek, L. and Alexa, M. (1997): 'Marking up in TATOE and exporting to SGML.' *Computers and the Humanities* 31(4), 1997. pp.311-326.
- Seidel, J.V. and Clark, J.A. (1984): 'THE ETHNOGRAPH: A Computer Program for the Analysis of Qualitative Data.' *Qualitative Sociology*, vol. 7, pp.110-125.
- Smith, A.E. (2000): 'Machine Learning of Well-Defined Thesaurus Concepts.' *Proceedings of the International Workshop on Text and Web Mining at PRICAI 2000*. LNCS 2112. Springer Verlag. pp.72-79.
- Smith, A.E. and Hughes, B. (2006): *Workshop on Standards for Storage and Interchange of Coding/Annotation Information for Qualitative Data*. 27-28th April 2006, The University of Queensland, Australia. <http://delacruz.csse.unimelb.edu.au/portcode>
- Smith, A.E. and Humphreys, M.S. (2006 to appear): 'Evaluation of Unsupervised Semantic Mapping of Natural Language with Leximancer Concept Mapping.' To appear in *Behavior Research Methods*. Accepted for publication on 29 March, 2005.
- Sophist (n.d.): *Sophist Information Bulletin*. Independent Institute for Social Policy. <http://www.socpol.ru/eng/archives/sofist.shtml>
- Wilson, M. (n.d.): *Annozilla: Annotea on Mozilla*. <http://annozilla.mozdev.org/>
- ZA-Information (n.d.): *Central Archive for Empirical Social Research*. University of Cologne / German Social Science Infrastructure Services. <http://www.esis.org/en/za/index.htm>

## Acknowledgements

This research is supported by the Australian Research Council through Special Research Initiatives - E-Research SR0567263 "Development of Tool Interfaces and Data Standards for Enabling Remote Secondary Analysis of Qualitative Data".