

Towards Interoperable Secondary Annotations in the E-Social Science Domain

Baden Hughes*, Desmond Schmidt** and Andrew Smith**

* Department of Computer Science and Software Engineering, The University of Melbourne

** Key Centre for Human Factors and Applied Cognitive Psychology, University of Queensland

Overview

- Introduction
- Background and Motivation
- Qualitative Data Interchange Format
- Implementations
- Future Work
- Conclusion
- Acknowledgements

Introduction

- In the social sciences, humanities and other areas the analysis of qualitative data is becoming more important.
- Fundamental problems in education, management and social policy are all highly suited to qualitative research.
- Although the number of social science data archives has recently grown, qualitative data is very much the poor cousin to quantitative data.
- The reasons for this are:
 1. Strong privacy requirements on data
 2. A lack of contextual knowledge by investigators
 3. A lack of interoperability between common software tools

Background and Motivation

- A survey of popular QDA tools from the perspective of data import/export reveals that:
 - Most do not support the import or export of primary data and secondary annotations
 - Qualitative data for each project is typically stored in proprietary binary form
- We identified a superset of 80-90% features shared by at least two programs as the initial basis for our standard

Qualitative Data Interchange Format (QDIF)

- Abstract standard which can be expressed in variety of formal languages
- Intended not just for interchange but for publication on the web
- Initial implementation in XML
- Moving to RDF for web-savvy applications

QDIF Internal Structure (1)

- Data Bundle
 - Primary QDIF object
 - Metadata, Primary Documents, Comments, Selections, Codes, Relation Names, Named Relations, Relations
- Metadata
 - Typical DC/DDI style description
 - Title, creator, description, language, source, date created, date modified, rights, rights holder

QDIF Internal Structure (2)

- Primary Documents
 - Primary source files, either local or remote
- Checksums
 - For verification of Data Bundle contents
- Comments
 - Short notes or memos
- Selections
 - Segment in Primary Document to which Comments are attached
 - Methods cover span, line, XML, graphic

QDIF Internal Structure (3)

- Codes
 - Concept codes, a la traditional coding of a source document
- Relation Names
 - Labels for Named Relations eg isPartOf, causeA
- Relations and Named Relations
 - Relations express the relationship between two components (subject and object)
 - Relation names

```
<?xml version="1.0"?>
<dataBundle xmlns="http://delacruz.csse.unimelb.edu.au/portcode/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/portcode/qdif
  http://delacruz.csse.unimelb.edu.au/portcode/qdif.xsd">
  <pdocs>
  <pdoc id="doc1">
  <metadata><dc:title>Alice's 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>
</dataBundle>
```

Implementations

- QDIF import / export support in two implementations
 - Leximancer (in progress)
 - Automated analysis of documents to extract codings based on concept occurrence
 - Another paper by Smith at ICESS2
- Annozilla (in progress)
 - Browser-based annotation tool based on open source framework for annotation
 - Annotations as native RDF
 - Client-server (i.e. shareable) or self-contained instantiation

Future Work

- Open issues to motivate future work
 - Complexity of metadata elements in QDIF and relation to external standards
 - Practical matters regarding storage and distribution of QDIF bundles
 - Alternate expressions (e.g. RDF) and compatibility with thematically related schemes
 - Audio and video support
 - Journaling (temporal versioning) support
 - Actively seeking input from other areas of the e-Social Science community on other requirements

Conclusion

- Positioned against background of data we see data interoperability as desirable in e-Social Science
- We are optimistic that data interchange can be realised in practice, QDIF is an early step
- Outstanding issues in policy area and social practice need to be addressed outside the technological sphere
- Project Website:
<http://delacruz.csse.unimelb.edu.au/portcode/>

Acknowledgements

- 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”