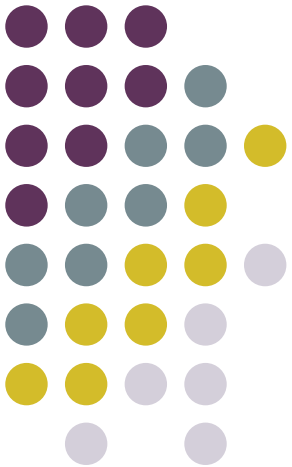
Two Data-Centric E-Research Projects at University of Melbourne

Baden Hughes

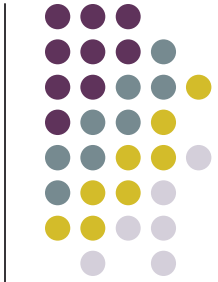
Department of Computer Science and Software Engineering

The University of Melbourne

badenh@csse.unimelb.edu.au



Overview

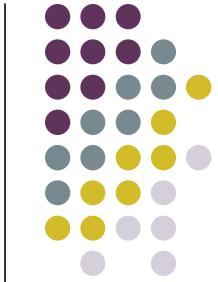


- Pervasive Issues and Themes
- An Intelligent Search Infrastructure for Language Resources on the Web (= LangSearch)
- Development of Tool Interfaces and Data Standards for Enabling Remote Secondary Analysis of Qualitative Data (= QualAnnot)
- Conclusions

Pervasive Issues and Themes

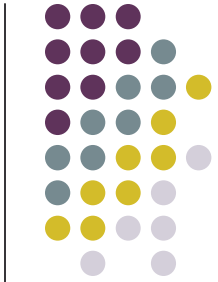
- Quantity of Data
- Discovery of Data
- Access to Data
- The 2 SRI projects discussed here address aspects of all 3 of these questions (albeit in different ways and for different audiences within the broader HCS community)

LangSearch SRI



- An Intelligent Search Infrastructure for Language Resources on the Web
- CIs: Tim Baldwin, Steven Bird, Baden Hughes (Melbourne)
- New infrastructure for accessing language resources, in form of a language-aware search engine (building on OLAC Search Engine)
- Advances in language identification to classify web content by language
- Large scale unsupervised content acquisition, classification, annotation, publication
- HPC infrastructure enabled (data/computation intensive)

Data Acquisition



- LingGator web content analysis for linguistic data
  - Seeded by language names and variants, lexical items, geospatial referents, encoding, character n-grams
  - Rank aggregation for results
  - Unsupervised execution
- Internet Archive's Heritrix crawler
  - State of the art web content acquisition
  - Seeded by URLs from LingGator
  - Crawl scoping (broad, narrow, link based)
- Highly parameterised and parallelised computation

Data Analysis and Annotation

- Dublin Core Metadata
- OLAC Metadata
  - Language Identification
  - Linguistic Data Type classification
  - Linguistic Subject classification
  - Discourse Type classification
  - Role classification
- OLAC Metadata Creation
  - Standoff model
    - Manual (XML, ORE, Kepler)
    - Semi-automated (OLAC-dot)
    - Automated (OLAC-X)

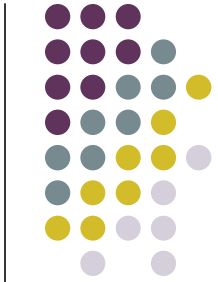
Data Discovery and Delivery

- Language data, tools advice described by OLAC metadata
- Data providers publish metadata catalogues to OLAC service providers via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
- End user discovery via OLAC, OAI via dedicated services
  http://www.language-archives.org/tools/search/
- End user discovery via broad coverage search engines such as Google via DP9 gateway

Partnerships

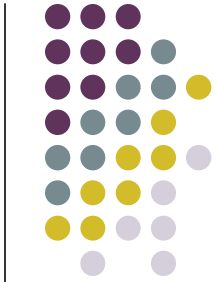
- SIL International
  - Ethnologue language classification
  - OLAC standards development
- Linguistic Data Consortium
  - Core OLAC infrastructure
- California State University Fresno
  - Online Database of Interlinear Text
- St Louis University
  - An Crudaban web crawler
  - Generated corpora

Issues and Themes Revisited



- Quantity of Data
  - Automated acquisition of large quantities of data
  - Variable data quality and formats
  - Dependent on web based publishing of some form of data
- Discovery of Data
  - Unsupervised data discovery from the web
  - End user discovery via domain specific or general purpose search engines
  - API access for programmatic interaction
- Access to Data
  - Assumes web based materials can minimally be indexed and accessed without restriction
  - OLAC does allow expression of rights metadata but not at the level of XACML style resource access formalism

QualAnnot SRI



- Development of Tool Interfaces and Data Standards for Enabling Remote Secondary Analysis of Qualitative Data
- CIs: Andrew Smith (UQ), Baden Hughes (Melbourne), David Rooney (UQ), Phil Graham (UWaterloo/UQ), Deborah Mitchell (ANU), Michael Humphreys (UQ), Cindy Gallois (UQ), Helen Chenery (UQ)
- Annotation interface standards for qualitative social science data (all natural language materials)
- Indexing of annotation materials and inline interpretation of annotations through library and applications
- Largely focused on middleware for collaborative research
- Leveraging existing ARC investment in qualitative data archiving

Data Analysis and Annotation

- Data
  - Variety of sources (surveys, interviews, cultural archives, creative endeavours, parliament etc)
    - From a linguistic perspective, of interest for information extraction, discourse analysis, semantic analysis
  - Variety of formats (Excel, Word, TEI, HTML, rich media)
  - Generally archived at domain specific repositories
- Annotation
  - Secondary to point of collection / curation
  - Required to be format independent, but durably linked to objects via URI
  - XML annotation expression for all formats
  - XInclude for merging XML data and XML annotations

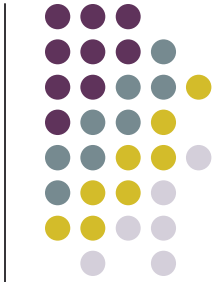
Data Discovery and Delivery

- Annotations stored as XML objects in Nesstar digital repository with persistent object identifier linkages
- Non-XML objects discovered and XML based annotations viewed inline by browsers (Annotea extensions)
- XML objects discovered and XML annotations merged by XInclude compliant web browser
- Both XML and non-XML objects discovered and viewed inline by specific analysis applications (Leximancer extensions)
- API (Java, Perl/Python)

Partnerships

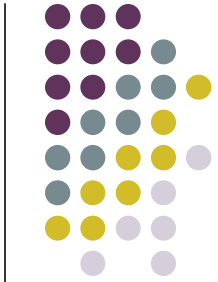
- Australian Creative Resources Online (ACRO)
- ACSPRI Australian Qualitative Data Archive (AQUA)
- UK ESDS Qualidata
- Canadian Centre for Cultural Innovation (CCCI) at CCAT

Issues and Themes Revisited



- Quantity of Data
  - Low quantity of data – currently dependent on archiving practice for individually funded projects with AQUA/ACRO etc
  - High quality data from significant data gathering efforts (eg surveys)
- Discovery of Data
  - User driven data discovery from curated collections
  - Programmatic discovery via Nesstar catalogue publication to OAI harvesters
- Access to Data
  - Assumes existing data distributors will handle access
  - New annotations may require access controls

Conclusions



- Both of these e-Research projects are data-centric and strongly motivated by interoperability concerns
- Following internationally accepted best practice for
  - data encoding
  - data analysis and description
  - data discovery and delivery
- Most infrastructure is existing open source software and can be extended to cover different domains