CLARIN
Common Language Resources and Technology Infrastructure

Resource Discovery and the Oxford Text Archive: recent history and the current state of play

Jens Stegmann
Institut für Deutsche Sprache
&
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
jens.stegmann@gmail.com

Martin Wynne
Head of the Oxford Text Archive
Oxford e-Research Centre,
Oxford University Computing Services, &
Faculty of Linguistics, Philology and Phonetics
University of Oxford
martin.wynne@oucs.ox.ac.uk

A little bit of history

1249 University of Oxford

...

1958 computing at Oxford

1976 Oxford Text Archive

1990 Humanities Computing Unit

1996 Arts and Humanities Data Service

2008 CLARIN

Early collaborations

- Isolated, often independent scholars
- Other new text archives (e.g. Virginia)
- Publishers
- OCR and text analysis services
- Oxford libraries
- Emerging standards initiatives (TEI, COCOA, SGML/XML)

Collaborations in the Internet age

- **Relationships with funders and grant schemes**
- **More work on standards**
- **More systematic intervention throughout the digital life-cycle**
- **Shared resource discovery services**

Arts and Humanities Data Service

National service for the support of the creation, preservation and re-use of digital resources in the Arts and Humanities in Higher Education in the UK.

Five data centres:

- **Oxford Text Archive**
- **History Data Service**
- **Archaeology Data Service**
- **Visual Arts Data Service**
- **Performing Arts Data Service**

plus the AHDS Executive at King's College, London

AHDS: collaborations

Common procedures for:

- advisory services
- ingest procedure
- collections descriptive metadata
- administrative metadata
- preservation policies
- resource discovery
- access services

Post-AHDS collaborations (UK)

- Network of Expert Centres in the Digital Humanities in Britain and Ireland
- E-Research South (Oxford, Reading, Southampton, King's coming soon)
- CLARINET - corpus linguistics centres (e.g. Glasgow, Newcastle, Lancaster, Birmingham, Nottingham)
- Text Encoding Initiative
- Digital Humanities at Oxford

OLAC

Open Language Archives Community, with a shared resource discovery service using an extended Dublin Core metadata set, and with more ambitious aims...?

In 2003, while AHDS CMF was pulling in one direction, TEI was pulling in another, OLAC in yet another, while IMDI was another option...

Now we have various different communities with whom we want to share metadata.

Please note that some of our resources do not appear in this table. Reasons for this might include that they were deposited with us for preservation only, and not dissemination, or that our legal right to disseminate them has been questioned.

ID	Availability	Title	Language	Author
2541	free	VU Amsterdam Metaphor Corpus	English	Gerard J Steen; Aletta G Dorst; J Berenike Herrmann; Anna A Kaal; Tina Krennmayr
2540	free	Speech, Thought and Writing Presentation Corpus (STWP)	English	Culpeper, Jonathon; Semino, Elena; Short, Mick; Wynne, Martin
2539	restricted	British Academic Written English Corpus	English	Nesi, Hilary; Gardner, Sheena; Thompson, Paul; Wickens, Paul
2537	free	GerManC. A Historical Corpus of German Newspapers 1650-1800	German	Durrell, Martin; Ensslin, Astrid; Bennett, Paul (ed.)
2531	free	The Lancaster Newsbooks Corpus	English	Thomason, George, d. 1666
2530	restricted	Language convergence and grammatical borrowing database	English	
2529	restricted	The Workdiaries of Robert Boyle	English	Hunter, Michael; Centre for Editing Lives and Letters
2528	free	Demetrios Database of Septuagint Greek	English	
2527	restricted	Chambers-Le Baron Corpus of Research Articles in French	French	Chambers, Angela; Le Baron, Florence
2525	restricted	British Academic Spoken English corpus	English	Nesi, Hilary; Thompson, Paul
2524	free	Discourse on the Origin and the Foundations of Inequality Among Men	English	
2523	free	The Birth of Tragedy	English	
2522	free	On the Use and Abuse of History for Life	English	
2521	free	Universal Natural History and Theory of Heaven	English	
2520	free	Cognition, Biology and Idealist Philosophy	English	Randrup, Axel
2518	restricted	The Electronic Text Corpus of Sumerian Literature. Revised edition.	English	Cunningham, Graham; Ebeling, Jarle; Black, Jeremy (deceased); Flückiger-Hawker, Esther; Robson, Eleanor; Taylor, Jon; Zólyomi, Gábor (ed.)
2517	restricted	Angloromani (sample)	English	Matras, Yaron
2516	free	Discourse context and the processing of contrastive focus in silent reading (SPSS data files)	English	Paterson, Kevin

Discovering Babel: Enhanced Language Resource Discovery

"The digital literary and linguistic resources in the Oxford Text Archive and in the British National Corpus have been available to researchers throughout the world for several decades. Technical enhancements to the resource discovery infrastructure will allow wider dissemination of open metadata, will facilitate interaction with research infrastructures, and the knowledge and expertise achieved will be shared with the community."

http://www.jisc.ac.uk/whatwedo/programmes/inf11/infrastructureforresourcediscovery/discoveringbabel.aspx

Discovering Babel

- c. 1400 metadata records;
- c. 1400 electronic literary and linguistic datasets:
  - Electronic texts
  - Text corpora
  - Lexicons
  - Databases of linguistic information
  - Audio data
- British National Corpus
- TRACTOR archive of central and East European language resources

Discovering Babel

Resource discovery metadata:

- Text Encoding Initiative (TEI) XML headers
- Dublin Core;
- Open Language Archives Community (OLAC) format (extended DC);
- CLARIN Metadata Initiative (CMDI)
- RDF linked data;

- OAI-PMH target
- Harvested by OLAC, CLARIN, etc.

Key challenges

- Establishing sensible and standards-conformant architecture for resource file locations;
- Conformance to semantics of various target metadata schemas;
- Expressing quality assurance metadata for legacy data;
- Expressing information for web services processing;
- Mapping licence restrictions to CLARIN 'laundry symbols';
- Establishing procedures for ensuring persistence and high availability of services.

Thank you for your attention

CLARIN has received funding from
the European Community's Seventh Framework Programme
under grant agreement n° 212230