Report on ESF Nedimah workshop
‘Ontology based annotation’

Summary

The workshop on ‘Ontology based annotation’ was held on 17th July 2012 at the University of Hamburg. It formed part of a series of events organized by the NEDIMAH project at the same time as the large *Digital Humanities 2012* conference, which assured good participation and fruitful interchange of idea. This workshop was for WG 3, *Linked data and ontological methods*, chaired by Christian-Emil Ore (University of Oslo) and Sebastian Rahtz (University of Oxford). The aim of this workshop was to present and discuss current ontology-based annotation in text studies and to give participants an introduction and updated insight to the field.

The meeting was run over a morning, with approximately 15 minutes allowed for each paper to be presented, and then plenary discussion. There were 12 presenters for 9 papers, along with 3 organizers, and approximately another 20 people attending.

The meeting divided into two parts. The first set of four papers covered some of the underlying approaches and methodologies, and summarized the state of the relevant metadata standards. The second set of five papers were case studies and reflections on results achieved.

The workshop was positive and effective. It summarized the state of current practice, and demonstrated a common approach which is gaining acceptance across a wide range of humanities disciplines.

Scientific content of and discussion at the event

One of the workshop aims was to throw light on the consequences and experiences of the data-based approaches to computer assisted textual work, based on the developments over the last decade in text encoding as well as in ontological systems.

A half-century of computer-assisted work on texts has had two main traditions. One is that which includes corpus linguistics and the creation of digital scholarly editions, while the other strain is related to museum and archival texts. In the former tradition, texts are commonly seen as first class feasible objects of study, which can be examined by the reader using aesthetic, linguistic or similar methods. In the latter tradition, texts are seen mainly as a source for information; readings concentrate on the content of the texts, not the form of their writing. In the last decade the stand-off database approach has been reintroduced, this time in the form of ontologies (conceptual models) often expressed in the RDF formalism to enable its use in the linked data world, and the semantic web. A basic assumption is that reading a text includes a process of creating a model in the mind of the reader, which can manifested as an explicit ontology. By manipulating the computer-based model new things can be learned about the text in question, or it can be compared to other similarly-treated texts.

Sanderson and van Sompel discussed the *Open Annotation* data model, which describes a method of encoding annotations using RDF, following the best practice recommendations of the Linked Open Data effort. The model is in the process of standardization in the W3C, and is the merger of the Open Annotation Collaboration and the Annotation Ontology efforts. It is interesting because of its adoption in both the humanities and scientific scholarly domains, as well as in model tagging and annotation systems for the web in general.

Osborne, Gerber and Hunter continued the Open Annotation theme, discussing their work within the Australian Electronic Scholarly Editing (AustESE) framework. The aim of producing collaborative, open-ended electronic environments means making changes to the traditional cul-de-sac approach of edited texts, and needs changes to working methods to make use of multiple external, formal, annotations.

Brown, Goddard and Paredes-Olea described a Canadian project, the *Canadian Writing Research Collaboratory / Le Collaboratoire scientifique des écrits du Canada* (http://www.cwrc.ca), an online infrastructure project designed to facilitate the study of Canadian literature. One of the emphases is on large-scale data, and making consistent access across a variety of resources. The use of RDF and common ontologies was seen as the key to progress in this area.

Bradley and Pasin dealt with the same problem, concentrating on the software used by humanities researchers, and some of the user experiences learnt from their *Pliny* programme. This more VRE-like approach based on scholarly daily practice will be steered to integration with the OAC-like work coming from the scientific domains.

By contrast, Fielding looked in a more abstract way at what is meant by 'ontology', from his work on the Wittgenstein archive. He was able to cast doubt on whether the realist approach of the sciences really does apply to the humanities, given their amphasis on the particular rather than the universal.

Zöllner-Weber also cast some doubt on whether the derivation of open data from humanities resources would be as easy as sometimes imagined, given the problems arising from modeling vague content, ambiguous hierarchical structures, and changing content and spatial-temporal information.

Francesco Membrini described practical work on the Hellespont project, promoted by the German Archaeological Institute and Tufts University, bridging two of the largest publicly available online databases in the field of Classical Studies, namely Arachne and Perseus. These are already highly-structured 'data' resources which lend themselves to use in annotation. This project looked at annotating a sample of a text by Thucydides.

Wang returned to Ludwig Wittgenstein, and his *Lectures and Conversation on Aesthetics, Psychology and Religious Belief*, considering how to analyze the peculiarly precise vocabulary in the edited text.

Blondel and Segala completed the workshop by talking about how to manage the substantial digital resource around the scientist Ampère.

The themes in discussion may be summarized as follows:

1. the Open Annotation data model is gaining quite widespread traction and understanding, and looks to be a plausible basis on which some at least of the humanities disciplines can join the semantic web of open data;
2. there are large scale collections of digital data which are amenable to data modelling, and in a form from which more formal data can be inferred;
3. on the other hand, there remains quite a large gulf between the simple annotation of/linking to eg prosopographical or geographic data from textual material, and ontologies which can mirror the very fluid and particular understanding scholars have of their field of study;
4. humanities data remains imperfect and uncertain; the problems of annotation go beyond the purely technical.

Assessment of the results and impact of the event on the future direction of the field

An objective of the workshop was to throw light on consequences and experiences of the renewed database approach in computer assisted textual work, based on the development in text encoding over the last decade as well as in ontological systems.

As Zöllner-Weber aptly said, 'The challenge in Digital Humanities is to incorporate information that is often not precise, that is vague and interpretable, i.e. mostly the opposite of structured information, into an ontology'.

The workshop succeeeded in highlighting for the participants some key messages:

1. that the semantic web tools and and standards like OAC and CIDOC CRM provide a workable and powerful architecture upon which stand-off annotation can actually work;
2. that the humanities disciplines have no shortage of data with which to work;
3. that existing technologies like XML markup for text are not conflicting with the new ontological approach, with eg sensible mapping of TEI markup to CIDOC CRM RDF available;
4. that the problems of uncertainty, particularlity and interpretation remain in the forefront of our minds.

It is to be expected that the outcomes of the workshop will be seen in conferences ranging from Computer Applications in Archaeology to Digital Humanities.

No publications of material from the workshop is planned at this time.

Final programme of the meeting

July 17th 2012, University of Hanmburg, 0900–1230

1. *The Open Annotation Ontology: Applications in Textual Scholarship*: Robert Sanderson and Herbert Van de Sompel (Los Alamos National Laboratory, USA)
2. *Annotation Collaboration (OAC) Data Model*: Roger Osborne, Anna Gerber, Jane Hunter (The University of Queensland, Australia)
3. *RDF for a Dynamic Literary Studies Collaboratory: A Pragmatic and Incrementalist Approach*: Susan Brown (University of Alberta, University of Guelph), Lisa Goddard (University of Alberta, Memorial University of Newfoundland), Mariana Paredes-Olea (University of Alberta), Canada
4. *Annotation and Ontology in most Humanities research: accommodating a more informal interpretation context*: John Bradley, Michele Pasin (King's College London, UK)
5. *Ontology in the Age of Digital Reproduction; Using Philosophy to Improve the Coherence of Database Management in the Humanities*: James Matthew Fielding (Université Paris I, France)
6. *Ontologies in Digital Humanities: without Limitations?*: Amélie Zöllner-Weber (University of Bergen, Norway)
7. *The Hellespont Project – Integrating different sources for ancient History in a Virtual Research Environment*: Francesco Mambrini (University of Cologne, Germany), Agnes Thomas (University of Cologne), Matteo Romanello (Kings College, London, UK)
8. *Using Centrality-Analysis for Keyword-Graphs*: Joseph Wang (University Innsbruck, Austria)
9. *Towards a new kind of research in the history of science: annotation of Ampère's corpus*: Christine Blondel (CNRS, CRHST/Centre Alexandre-Koyré, Paris) and Marco Segala (University of L'Aquila, Italy; Centre Alexandre-Koyré, Paris), France