Technology Workshop Days
WP2-TR10-2003 Version 1

ECHO IT Days

18/19/20.September 2003
Lund[1]

Organizers: Sven Strömqvist, Marcus Uneson, Gerd Grasshoff, Peter Wittenburg

At 18/19/20. September 2003 WP2 will organize the ECHO IT Days. They will focus on two thematic topics relevant for the work in ECHO and give room for a meeting about all technical aspects within ECHO. The two selected special topics are “Annotation Systems” and “Interoperability Issues”.

The IT Days will be organized by University of Lund and Work Package 2 of ECHO. Both thematic areas will cover a whole day. They will have two parts each: (1) In the first part we will ask some speakers to describe a few important dimensions of the problem. (2) In the second part we will motivate other participants to contribute with short presentations and reserve time for intensive discussions. For the first part we will ask external speakers for presentations and use the expertise of the ECHO consortium where possible.

Annotation Systems Workshop

Motivation

ECHO is focusing on technologies that allow to bring cultural heritage content online, enrich the original material and to allow all sorts of users to interact about such content. We call all sorts of textual enrichments annotations independent whether they are complex descriptions of the linguistic structure, links created by scholars to combine content, keyword type of descriptions for discovery purposes or comments by arbitrary users. The original material that is subject of annotation can be texts, links, images, sounds, movies and 3D objects. Annotations can refer to the whole resource or part of it that can be identified by some formal means[2]. Annotations can be private or shared and being produced in isolation or emerge as a collaborative effort.

So, annotations are metadata that can appear in many different forms, can be associated with many different types of objects and can be created in different social circumstances. Therefore, we have very different annotation systems, i.e. formats, structures, ways of registration and linking methods are very different at this moment. We can refer to a couple of interesting initiatives to show the heterogeneity of the suggestions:

- In the CES[3] standard annotations are SGML/XML tags in a layered system of tiers with conventions for tags labels etc.
- In many traditional approaches annotations are specialized fields in a relational database according to the ER model.
- In the Annotation Graph[4] model from Liberman and Bird annotations are arcs in a directed acyclic graph. From this model they derive an API that allows them to manipulate annotations.

---

[1] Lund University kindly offers support for these activities and we are grateful to make use of the Lund facilities.
[2] No difference is made between references to a point/unit or a sequence or fragment, since a singular point is always seen as fragment of unitary length.
[3] Corpus Encoding Standard - a TEI compliant concretization for corpora; http://www.cs.vassar.edu/XCES/
[4] http://acl.ldc.upenn.edu/acl2001/STR/7-bird-et-al.pdf

- In the EUDICO/ELAN[5] project annotations are elements in an abstract corpus model that covers all known structural phenomena known in linguistics when dealing with multimedia language resources. From this model an XML format is derived allowing making the annotations persistent.
- In the GATE[6] system annotations are created automatically by Natural Language Processing components such as automatic parsers. These components need a well-defined formal framework to generate incrementally layers of annotations.
- In the Annotea/Annozilla[7] project annotations of web-documents or parts of them are described as sharable RDF assertions stored in distributed repositories. Here the annotations themselves are simple structures that can also be described by some keyword for easy discovery purposes.

This list could be extended by almost an infinite number of suggestions. Recently, the ISO TC37/SC4 subcommittee (chaired by Laurent Romary) took up this discussion for the area of language resources. Its aim is to find a generalization of the various useful suggestions. Within ECHO the scope is extended in so far that several disciplines are included. Nevertheless, it seems that there is much overlap in the requirements. Therefore, it is an important task to discuss the different suggestions, check their usefulness, compare their properties and investigate in how far they can be generalized as well.

Goal

The goal of the annotation workshop is to find possible generalizations of the currently used annotation systems for the ECHO future such that a unification process could be initiated. Such unification could lead to an ECHO standard compliant with the ISO suggestions, a reduction of the amount of work to build tools and a higher degree of interoperability.

Preliminary Agenda

8.30	Introduction and Overview	Peter Wittenburg	ECHO
9.00	Relational Linking of Documents as an Annotation Process	Hans Uszkoreit	DFKI
9.30	Annotating Web Resources	Dirk Wintergrün	ECHO
10.00	Automatic Annotation by NLP Components	Hamish Cunningham	U Sheffield
10.30	Break		
11.00	Web-based annotation of Images	Gerd Grasshoff	ECHO
11.30	Complex Annotations of Audio and Video Signals	Hennie Brugman	ECHO
12.00	Annotation - the ISO Perspective	Laurent Romary	LORIA
12.30	Lunch		
13.30		div	
15.00	Break		
15.30		div	
17.00	End		

The titles are working titles and will be subject of changes.

Interoperability Workshop

Motivation

ECHO covers several disciplines active in the area of cultural heritage such as History of Science, History of Arts, Languages, Ethnology and Philosophy. Covering several disciplines was done on purpose, since the ECHO partners wanted to study interaction and technology development processes that cross discipline boundaries and come to a framework that allows researchers to integrate resources from various disciplines to gain new insights. Mutual interest was shown from the participating disciplines in a cross-disciplinary approach. Until now, however, it seems that the disciplines are still operating in their scholarly domain of interest.

[5] http://www.mpi.nl/tools
[6] http://gate.ac.uk
[7] http://www.w3.org/2001/Annotea

Cross-disciplinary work requires interoperability at several levels. The most obvious level is that of discovering suitable resources that can be exploited in an interoperable way. The second level has to deal with accessing such resources, since they are stored in some containers and formats. The third level addresses the syntactical aspects, i.e. how can we identify the elements of the individual resources and extract relevant content. The last and most difficult level focuses on the semantic aspects[8], i.e. how can we exploit and combine the different contents.

There is yet another and even more fundamental level of interoperability that has to be addressed: Do scholars have a real interest in interdisciplinary work that exceeds the boundary of tailored projects? And if so do we understand each other's language and objectives sufficiently well?

Due to the AGORA approach of ECHO we have to address all levels of interoperability. WP2 already started discussing the first three levels and identified a couple of problems that are not at all trivial and need to be solved (see WP2 reports under http://www.mpi.nl/echo). With this workshop WP2 wants to continue the discussion about semantic interoperability. Within WP2 two institutes started already to discuss how their catalogue information could be mapped to allow joint queries[9].

Goal
The interoperability workshop is meant to get a better understanding of the problems, the solutions under way to close the "semantic gap" and to work out strategies for future ECHO work. The organizers don't want to give the impression that the current ECHO project can "solve" these problems, but we should at least make some tests such that we can derive perspectives for future work.

Preliminary Agenda

8.30	Interoperability in ECHO - an introduction	Sven Stromqvist	ECHO
9.00	Principle Difficulties for Interdisciplinary and Interoperability	Barbara Cassin	ECHO
9.45	Interdisciplinary work - hopes, expectations and illusions	??	Lund
10.30	Break		
11.00	Objects of and architectures for Interoperability	Peter Wittenburg	ECHO
11.45	Formal frameworks for interoperability	Frank van Harmelen	Amsterdam
12.30	Lunch		
13.30	Short Contributions	div	ECHO
15.00	Break		
15.30	Discussion	div	
17.00	End		

The titles are working titles and will be subject of changes. Lund University is currently looking for an excellent speaker.

Short Contributions

As already indicated we would like to promote the ECHO participants and our guests to suggest other contributions for the afternoon hours of the two workshop days. They should emerge from practical work and address concrete problems. It could also be the case that you raise specific questions where you want to have experts make statements.
Let us give an example: We have learned that we should make texts available by using XML as the standard exchange format. In addition, we are already confronted with RDF (Resource Description Framework) that is advocated by some as a container for metadata of all sorts. Many people don't know what the differences in expressional power between XML and RDF is and what the limitations are. This could be a useful contribution for an afternoon session.
The organizers are open for suggestions and will decide about the schedule about a week before the IT days start. So, please, specify your wishes and needs and we will try to find someone who can give explanations.

Send your contributions and wishes to Marcus Uneson (marcus.uneson@ling.lu.se) as soon as possible.

[8] On purpose we exclude aspects of pragmatics in this consideration.
[9] Bibliotheka Hertziana and MPI for Psycholinguistics started information exchange about this issue.

Technical Committee Workshop

Finally, at the third day (20.9.) we will discuss in an open framework all technical matters that were addressed in ECHO until now or that will be on the agenda. We will invite the members of our Technical Committee to participate and comment. This open discussion will also be used to draw conclusions form the work until now. For this part we will reserve the morning hours, i.e. we expect that the IT days will finish around 12 am.

Report

It is intended to write a special report about this workshop. To achieve this we want to gather all material about what was presented and work out conclusions that will be distributed for comments before distribution via the Web. For all presentations we expect to receive a summary of the major points (max. 2 pages) and the slides. At the workshop both should be ready. All will be made available via the ECHO WP2 web site.

Participation

All persons who want to participate have to register by sending an email to Marcus Uneson (marcus.uneson@ling.lu.se). Please, understand that we can only have a limited number of participants. The speakers listed above are registered automatically.