WP2 Report on the ECHO IT Days

Peter Wittenburg, Sven Stromqvist, Marcus Uneson, Gerd Grasshoff
September 21, 2003

The Lund IT Days were organized by WP2 as part of the AGORA in the ECHO project. Everyone interested in technological matters was invited.

The first 9 months of the project have been devoted to interesting core developments with a great amount of domain specific impulses. This phase has been utterly important for the stabilization of ideas and frameworks. Within shortest time the ECHO partners have been able to realize interesting concepts that partly have benefited from already existing projects and collaborations. New types of services can be offered, such as easy ingesting of textual, image, sound and video collections, archival storage of cultural heritage material and rich exploitation frameworks.

Therefore, the IT Days were intended for exchange of information about the achieved state in the various activities, for discussions in detail of two important aspects that were raised again and again during the first phase developments and interactions (Annotation-Structures and Interoperability), for information about relevant work outside from ECHO from recognized specialists[1] and for bringing people together to discuss future plans.

The workshop was seen as extremely positive and raised many interesting perspectives. This report is meant to give a comprehensive picture of all activities and perspectives.

The program and the participants can be found at www.ling.lu.se/projects/echo/contributors

1. Annotation Topic

ECHO is focusing on technologies that allow bringing cultural heritage content online, enrich the original material and allow all sorts of users to interact about such content. We call all sorts of enrichments of raw documents annotations, independently of 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 to annotation can be texts, links, images, sounds, movies and 3D objects. Annotations can refer to the whole resource or a part of it that can be identified by some formal means[2]. Annotations can be private or shared and can be 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:

[1] The invited specialists are also members of the Technical Committee of ECHO.
[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.

- In the CES[3] standard, annotations are SGML/XML tags in a layered system of tiers with conventions for tag 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.
- 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, which allows 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 layers of annotations incrementally.
- 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 an almost 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 some overlap in the requirements. Therefore, it was seen as an important task to discuss the different suggestions, check their usefulness, compare their properties and investigate how far they can be generalized as well.

The original goal of the annotation workshop was to find possible generalizations of the currently used annotation systems for the ECHO future, such that exchanging can be facilitated. Experts from ECHO and from outside were invited to shed light on the annotation problem from different views[8]:

- Peter Wittenburg (ECHO-Nijmegen): introduction to the topic PW
- Dirk Wintergrün (ECHO-Berlin): annotation of (textual) web-resources DW
- Gerd Grasshoff (ECHO-Bern): annotation of images GG
- Hennie Brugman (ECHO-Nijmegen): annotation of sounds and movies HB
- Laurent Romary (LORIA-Nancy): annotations from an NLP and ISO perspective LR
- Hans Uszkoreit (DFKI-Saarbrücken): hyperlinking as an annotation process HU

1.1 Presentations

Peter Wittenburg

PW briefly introduced the topic by asking what annotations are and what the resources we want to annotate are. In fact, all electronic document types (texts, images, sounds, videos, time series, and links) can be subject of annotations. Annotations themselves can be subject of further annotations, i.e. we are faced with a recursive phenomenon. The different types of annotations can have the same range of document types, i.e. we could annotate a sound fragment by associating an image showing the position of the tongue with it. So he concluded that an “annotation is a document associated with another document or a fragment of that document that adds information”. This very general statement results in a number of questions that have to be addressed and are being discussed within ECHO, TEI, ISO and other organizations:

- What is the granularity of annotations?

[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
[5] http://www.mpi.nl/tools
[6] http://gate.ac.uk
[7] http://www.w3.org/2001/Annotea
[8] Originally Hamsih Cunningham was invited to speak about automatic annotations in NLP environments, but he had to cancel due to private reasons. Laurent Romary took over his part partially.

- What are the dimensions of fragment selection (space, time) for annotations?
- How exact do we have to be in the temporal and spatial domains?
- What are the necessary annotation structures and their internal complexity?
- How to refer to documents?
- How to create the annotations?
- In which formats do we have to store formats?
- How can we uniquely identify annotations?
- How to register, ingest, store, and discover annotations?
- How to visualize annotations?
- Can one speak about components that can be jointly developed?

He expressed his confidence in the choice that within ECHO it was necessary that the different approaches were primarily discipline-driven rather than standard-driven and that during the second phase, based on the achieved deeper understanding, one has to investigate in how far generalizations can be done.

Dirk Wintergrün
DW focused in his talk on annotations of (textual) web-resources where a resource can be identified by a URL or a unique web-resource identifier[9]. He made the interesting claim that annotations fit perfectly well with the RDF structure model, where an annotation is an identifiable web-resource that has a typed relation with another web-resource. This opens ways to seamless treating annotations of annotations etc.

Currently, their model includes “arcs” that are typed references associated for example with Dublin Core like metadata. Arcs are defined by a schema that includes references to the source and the target and a description of the relation type. He would like to see Arcs as a standard for ECHO.

Using the RDF syntax, he showed how arcs can be formulated in RDF. Further, he explained the kind of tools that are necessary to create arcs, such as text editors, Annota for images, an Annotea derivate for HTML, web-interfaces for commenting, exports form existing databases and others. These annotations have typically been stored on publicly available servers or on local file systems for private annotations. These annotations must be searchable along the dimensions target, source, relation type and metadata elements. Since not all annotations may be openly accessible, the need of a rights management system was introduced.

Further, efficient and user-friendly visualization is important, and in this connection DW referred to a recent suggestion from Gerd Grasshoff about developing a system that allows visualization of relation clusters within the whole network of annotations. Within texts, XML-tagged structure elements must be subject of annotations such as sentences, words, and pages. Mostly, annotating is a manual process in the humanities, therefore efficient annotation is dependent on good tools. The tools required are not available yet. For some lexicon-based annotations, there are automatic tools available at present.

Finally, DW explained what will be done within the coming months/years:

- Agreement on standards
- Setting up an annotation server
- Setting up an DRI Server for documents
- Standardizing the existing tools
- Completing feature lists for additional tools
- Putting some manpower into the development of manual annotation of fulltexts (XML, HTML)

Gerd Grasshoff
GG focused on the aspects that occur when annotating images. According to GG one has to define first what an annotation is. He describes an annotation as a relation between two objects. However,

[9] In ECHO we are working on a service that will resolve unique resource identifiers, i.e. at the end of the ECHO project, possibly all resources will be associated with an identifier that is unique for the ECHO namespace authority.

he would like to exclude simplistic resources such as pure references from being treated as annotations. He also assumes that annotated objects are typically more complex. In image annotation, one typically has annotations on parts that may even overlap and therefore could form complex inherent spatial relations. He then discussed a number of important aspects:

- Different anchoring methods have to be considered, this can go up to SVG graphs describing an object.
- The annotations can be of different types such as sounds, texts or web-resources.
- Annotations have to be associated with a minimal set of metadata elements such as Author and Date.
- The modes of access must support displaying, searching and also backlinking

In particular he stressed that handling access rights is very important, since annotations collections may be private at first. He gave a few examples that confirmed the need for at least temporal disclosure. There also will be copyright issues that have to be considered.

GG then gave an impression of the Alcatraz tool set that allows to operate on images in a very flexible way and therefore can be an excellent basis to produce annotations on images. Alcatraz will support browsing features, currently a privately generated taxonomy is used for this purpose. GG showed various types of typical images from the history of science area that demonstrated the need for a flexible image handling environment to anchor image annotations.

Finally, he also asked the participants to support the RDF model as basis for representing annotations and described a standard for an ECHO bookmark format.

Hennie Brugman
HB focused on the annotation of time series such as sounds and videos, although the presented methods will also work on signals, e.g. from eye movements. He first described the complexity needed with the help of a UML diagram that is the basis for all his implementation work. In the linguistic domain, one typically has complex annotations that exist on several tiers. Tiers share annotations of a special type and have to adhere to a number of constraints such as being ordered and not overlapping.

At a high level of abstraction, in his Abstract Corpus Model (ACM) he distinguishes two types of annotations with respect to their relation: (1) Alignable annotations refer to sound or movie fragments and they do this by sharing time slots that point to the time axis (per annotation 2 slots). (2) Reference annotations are those that refer to other annotations. Transcriptions of speech signals and gesture annotations typically are linked to media fragments directly. Higher level linguistic concepts, such as morphemes, will be linked to words. He explained that linguistic annotations can become fairly complex in so far as one can find hierarchical dependencies between the tiers. In addition to these type-related dependencies one can find token-dependent dependencies that are typed relations between two or more elements on the same or on different tiers. To make ACM powerful enough, these co-references are themselves annotations, i.e. they have a type and can point to several elements on different tiers.

Of course, annotation tiers can be completely independent and therefore overlapping in time. This is the case if different channels of behaviour, such as speech and gesture, are encoded. Dependent tiers share time slots with the parent tier. This constraint allows the user to shift whole bundles of tiers with one operation.

HB further outlined the intended extensions of the model to allow 2D, 2+1D annotations. 2D annotations allow anchoring annotations to some form of contour to a still image that could also be a video frame or any other 2D representation. One cannot speak anymore about ordering, but other characteristics of ACM are still applicable. A further extension is the 2+1D case, where series of 2D annotations are linked to a time interval.

Finally, he briefly presented the ELAN annotation tool, with its components, that supports the features of ACM and that was extended based on the user requests, in particular in the Sign Language group. He presented ideas on how one could collaborate in WP2 when extending ELAN to the two-

dimensional case. Further, he outlined the intentions to prepare an ELAN version that allows collaboration of different users working at different locations.

Laurent Romary
LR gave a talk about annotation issues, based on his experience in the area of NLP and as a chairman of ISO TC37/SC4 about "Terminology and Management of Language Resources". He raised three questions that are currently in the focus of the discussion: (1) How can we share resources? (2) How can we share tools? (3) How can we assure meaning consistency between annotations? Annotations are secondary data that are added to primary data. He argued that now all resources should apply XML as the same underlying syntax. Also, one should adhere to the stand-off model as a general rule, in order not to touch the primary data. An annotation itself can be seen as primary data at that moment when someone wants to add annotations to the existing ones. Here as well, the stand-off principle is essential.

Then he explained the activities of ISO TC37/SC4 in the area of linguistic resources in order to draw attention to the problems associated with standardization, but also to the necessity to come to standards. ISO TC37/SC4 wants to give directions in the following areas: (1) Structuring primary resources, (2) Guidelines for establishing knowledge representations, (3) Establishing general models for lexicons, (4) Guidelines for the creation of annotations, and (5) Guidelines for metadata that can be used to discover the resources. Of course, TC37/SC4 will build upon of what has already been achieved by other standardization attempts in various areas such as TEI, EAGLES, ISLE, ISO TC37/SC3 and many others. He briefly explained that some initiatives within SC4 are close to proposing standards, while others have a longer way to go.

Further, he briefly indicated the current state of discussions with respect to annotations. A General Framework will be based on two components in particular: a metamodel representing the expressional power of possible annotation structures and data category repositories that provide semantic knowledge for re-usage. Using the example of morpho-syntactic encoding, he explained how to create easily more abstract representations. He introduced the tag "struct" that is used to label structural nodes in XML files and the tag "feat" to denote features. The specific linguistic information is added as a type of the given feature.

Finally, LR described what data categories are and how they can be used. It is an elementary descriptor used in a linguistic description or annotation scheme. If such descriptors are part of open term repositories, the community can re-use them and thereby increase the semantic interoperability. Data categories have to be defined according standards such as ISO 11179, that describes a fairly comprehensive data model to define categories, or ISO 16642, having emerged in the terminology area and extending the definition to the multilingual dimension. A scenario is described in which term repositories with rich content are maintained as namespaces that will be re-used and combined. ISO TC37/SC4 will actively promote such repositories.

Hans Uszkoreit
HU introduced the term "digital memory" and described the essentials behind this idea. Obviously, his ideas describe the future perspectives of semantic technologies that are currently emerging and being tested out in the labs. The "digital memory" (DM) makes broad use of the stored information and the inherent semantic relations between documents. These can be exploited in various ways and therefore are more than pure references, as those known from the current web, for example. It allows the user to easily switch between semantic layers of documents to get new insights. It is also the basis for inferencing and learning, since agents will exploit the available relations.

HU envisions a scenario where automatic hyperlinking will be applied to cultural content, yielding a rich information density facilitating the scholarly work. Language Technology will play a very important role, since it can perform many steps such as recognition of named entities, morphology and syntax tolerant processing, synonym recognition, exploration of thesauri and ontologies, and recognition of syntactic functions and thematic roles. Polysemy, ambiguity and aspects still make the interpretation of sentences a hard task.

HU's goal is a densely hyperlinked text where any meaningful unit carries typed relational hyperlinks. There are many applications where such linked structures would be very useful. As one possible application, he briefly mentioned the LT World site, where all relevant information about authors, papers, projects and others in the area of Language Technology is hyperlinked automatically. This is

now a great tool for quick look up. Hand-crafted generic ontologies and specialized (personalized) dynamic ontologies are seen as necessary prerequisites.

Finally, he gave an introduction to the “Deep Thought” project. Based on the knowledge that 99% of the creative processes today consist of retrieval activities, a new approach for the creation of new knowledge by the exploitation of existing knowledge is made. When a text is entered, it is immediately interlinked to set it into its rich context and allow the user to make use of this additional information. Associative memories constructed in such a way are the natural step beyond digital content. Building such associative memories requires advanced language and knowledge technologies.

1.2 Discussion

The discussion was centered around a couple of main issues.

Annotations Structures

It was clear that there are different approaches to what an annotation is and how annotations should be represented. Only those enrichments that bear some content can be seen as “annotations”. Only here it makes sense to describe them by metadata, store them in repositories and make them searchable. Pure references are not annotations. While in some disciplines it makes sense to model annotations with the RDF model, other domains, in particular the language area, use more complex structures that can best be modelled with XML structures including Xlinks. Therefore, and that is coherent with the discussions within ISO TC37/SC4, it was concluded that it does not make sense to define a “unified standard for annotation structures” for ECHO. XML is the agreed syntactic basis, while RDF is not the primary choice for complex structured hierarchical annotations. For the relational type of annotations, RDF is a very promising model. XML will make it possible to easily transform a given file into another structure by using XSLT technology.

In the future, web-services will be used to access repositories. With UDDI a web-service will describe its type of service, with WSDL one specifies the functions and data structures that are available, and SOAP, finally, realizes the XML-based data exchange.

It was argued that for the simple commenting that are foreseen in many areas of cultural heritage, only a minimal standard for metadata descriptions is needed, in order to reduce the load. No agreement could be achieved, since there is no experience yet as to how these metadata descriptions will be used for discovery. The content of the annotations and their context could be sufficient enough.

A need was seen to combine metadata and content search. Metadata can be used as a filter for the resources to be included and content search could then operate on the detailed annotations. All developers in ECHO are working on implementation concepts. It will be a challenge to create interoperability, i.e. to extend searches on annotations that come from different disciplines. Due to its short life-time, ECHO is not the right framework to tackle such issues.

Semantic Level Aspects

A discussion revealed that the principal line as proposed by Romary was accepted. Term repositories will contain the definitions of concepts that occur in the humanities and dedications to the various languages. Other repositories will contain relations that are drawn between the concepts. This way of representing ontological knowledge offers maximal flexibility.

To achieve a higher degree of semantic interoperability it would be excellent if all descriptors used would be entered into open term repositories. This holds for tag labels in annotations as well as for metadata elements. It would be an investment for the future.

For the representation of semantic relations between terms within ECHO, it was assumed that probably RDF(S) will be sufficient. The discussion was postponed to the next day.

Servers and Services

ECHO will need servers that can store annotations that are made according to the scheme as presented by Wintergrün. In this scenario, annotations have to be stored either on the private

notebook or in open repositories. Users who want to annotate have to have the choice where to store their annotations, since there will probably be several such servers.

Participants argued that often annotations will be private from various reasons. These could be very personal notes, they could be preliminary etc. This would require an access rights management system and an ingest procedure, since at a certain moment a set of annotations may become open. It was referred to Annotea, that also supports a password mechanism. There was a debate in how far access rights mechanisms can play a role in ECHO. It was agreed that ECHO can make use of systems that may be developed in other projects. Developing a full-fledged system would also go beyond the scope of the ECHO project.

It was explained which kind of unique resource ID (URID) resolving system was installed and is being tested. The Handle System was seen as the best choice of the available alternatives. The service set up at the MPI for Psycholinguistics should be available soon for concrete testing and usage by the partners. Since it is not acceptable that URID resolving is dependent on just one server, it was already discussed with another institution to set up a mirror server.

1.3 Future Perspectives

A couple of concrete activities was discussed to improve the ECHO scenario:

- A small working group will speak about possible generalization of the concept of annotations.
- It will be checked how far a component can be developed (or modified from an existing solution) that allows the generation of polygons to denote shapes and to associate annotations with such shapes.
- A requirements document will describe the function of an annotation server and its usage. When such a specification is available, concrete steps can be taken. A discussion about peer-to-peer and centralized server concepts made more clear what ECHO is aiming at.
- A requirements document will describe the characteristics of an access rights management system that can have a function for several ECHO partners.
- The term definition framework being worked out within ISO TC37/SC4 will be made available for ECHO and be used.

Future techniques such as automatic hyperlinking were seen as not yet ready to be used outside of specialized labs, since they require powerful infrastructures in language technology and in knowledge management.

1.4 Summary

The presentations and discussion contributions of the first day contributed to a much better understanding of the various approaches and of the mind sets within the different disciplines. Developing a common language is not at all a trivial enterprise. Therefore, the early decision to first build on the strengths of the existing initiatives was the right one to guarantee a quick start. It was also agreed that the ECHO collaboration until now was a big source of mutual fertilization across the disciplines.

Concrete joint activities were discussed, but it was also clear that for many problems ECHO as a framework is too short-term. The participants were convinced that the ECHO developments have already achieved a high standard after a relatively short period of time. It was also reported that the interactions with the users were very satisfying so far.

2. Interoperability Topic

The aim of the creation of the Internet in the late 1960s was to interconnect a small number of main frames, in order to share computational resources more efficiently. The creation of the World Wide Web in the early 1990s was initiated at the CERN laboratory in order for the research community in high-energy physics to be able to search and find information in a large number of documents across different computers and networks. Both endeavours were driven by economical reasoning – to make better usage of fragmented resources – and they both proved to be extremely good investments.

In the ECHO state-of-art report (D1.1) it is argued that part of the explanation for the lack of accessible digital content in the humanities so far has to do with lack of metadata, infrastructure and interoperability, as well as with insufficient attention to user perspectives. The topic of interoperability is, therefore, crucial to the success of ECHO.

The topic of interoperability has several aspects. Technical aspects include storage, accessibility, and analysis tools. But interoperabilty issues go far beyond technical aspects. Without a genuine interest in crossdisciplinary cooperation, the technical machinery stands without users. It is therefore crucial to ask questions about goals and motivational forces driving the scientific and educational communities. Further, political aspects are important to consider: the resolution of organizational problems in order to promote crossdisciplinary cooperation, questions concerning consequences for society at large, etc. The consequences of interoperability also affects our collective memory.

In order to shed light on this complex situation, experts from ECHO and outside were invited to discuss the topic of interoperability from different points of view:

- Sven Strömqvist (ECHO-Lund): Interoperability in ECHO – an introduction
- Barbara Cassin (Paris, Sorbonne): Difficulties for interdisciplinarity and interoperability
- Hans Andersson (Lund University): Interdisciplinary work – hopes and illusions
- Peter Wittenburg (ECHO-Nijmegen): Objects of and architectures for Interoperability
- Frank van Harmelen (LORIA-Nancy): Formal frameworks for interoperability

2.1 Presentations

Sven Strömqvist
Sven Strömqvist approached the issue of interoperabilty from a communication perspective, asking *What does successful communication presuppose?* Strömqvist pointed at three factors:

- Shared background knowledge
- Shared purpose/goal
- Shared attention

Shared background knowledge includes things like concepts and languages, attitudes (such as curiosity, or feelings of identity), and personal experience. In the case of ECHO, the dimension of shared purpose/goal includes things like counterbalancing the fragmentation of knowledge, promoting interdisciplinary cooperation, making use of resources which are otherwise waisted or underexplored, and creating a better basis for working with culturage heritage in the future. Shared attention has to do with, among other things, the need for those who cooperate to focus on the same thing.

In building external relations and attracting participation in the ECHO project, it is imperative to consider the above dimensions and to ask the question*Who sees the added value of interoperability?* An individual researcher? A research institute or university department? A university? A large cultural institution (such as a major museum)? A national research council? A ministry?

Barbara Cassin
Barbara Cassin presented a philosopher's angle to languages, concepts and the role of translation in crossdisciplinary understanding. Differences in terminological traditions have consequences for interdisciplinary cooperation and can easily lead to misunderstandings, due to translation problems. Cassin gave several examples of this, such as the translation of *philosophy of mind* (eng.) – *philosophie de l'ésprit* (fr.) – *phenomenologie des Geistes* (ger.). The semantic non-equivalence of the participating terms *mind* (eng.), *ésprit* (fr.) and *Geist* (ger.) is conducive to misunderstandings, when translated, and may interact with differences in how the disciplines have been pursued in the three countries in question.

Anchored in a discussion ranging from Aristotle via Leibniz to present day philopsophy, Cassin argued that there are two major ways of approaching the translation problem. One is to find the best word or to invent a good word for a given concept or phenomenon, following the priniciple "one thing

– one essence – one word-meaning". This results in a consistently ontology-driven terminology, such as Leibniz' Characteristica Universalis, where the expressions of the symbolism or language constructed should mirror the structure of the world. The other is to appreciate the force of words to represent subjective perspectives on reality. The diversity of languages represents a plurality of viewpoints. Applied to differences in terminological traditions between academic disciplines, this plurality presents a problem for translation, but it also presents a resource of perspectives for concepualizing a problem or phenomenon.

Hans Andersson

Hans Andersson pointed to three prerequisites for successful interdisciplinary cooperation: enough time, a common meeting place, and tools which can be used with sufficient ease.

Andersson shared aspects of his experience of coordinating one of Sweden's largests projects in archeology, trying to make archeologists, historians, natural and cultural geographers work together as a creative and efficient team. A key factor is a sufficient degree of common understanding across the participating experts, and this is sometimes harder to achieve than what you may think at first.Whereas there are many visible and open aspects of the scientific frame belonging to a given discipline, there are also hidden aspects which require more time to discover, make visible and integrate with the common understanding.

Further, the force of coming together – often in an improvised fashion - to share professional as well as personal experiences should not be underestimated as a source of getting acquainted and growing into a team. This requires a place to meet.

Finally, common tools for solving research tasks is a powerful resource both for the individual researcher, but also for builing a team. In order to establish a tool as an attractive common resource, it is imperative that usability can be demonstrated. It is a great drawback if people refrain from using them, simply because they are perceived of as too difficult to handle.

Hans Andersson concluded with a general plea for crossdisciplinary endeavours: it is not always new theories or methods which accomplish breakthroughs; breakthroughs are often effected by putting old things together in new ways. And interoperability plays a key role in that process. So far, this aspect of scientific work has received little attention, and it is symptomatic that we do not talk about an interdisciplinary research front. Let us change that situation!

Peter Wittenburg

Peter Wittenburg gave a report about the work in progress with ECHO's technological response to the interoperability challenge, the Digital Open Resource Area (DORA).
Here, the task is to create a searchable and browsable metadata domain of all available resources in ECHO and then to provide access to the resources where possible.

The navigation modes of DORA include full-text search across all metadata (also prose descriptions), complex structured search supporting domain views, support of browsing were possible, geographic browsing were possible (three layers), support of selections when searching, escape to specific domain interfaces, in complex search support for semantic mapping and for all hits immediate jump to resources (MD, texts, media) if possible (executing a certain tool)

The content domains of DORA are at present History of Arts (Fotothek, Lineamenta, Ancient Maps of Rome), History of Science (various repositories via Bern & Berlin, IMSS collection), Ethnology (NECEP society database, RMV collection, DOGON collection), Languages (various repositories via Nijmegen & Lund, special profiles such as for the Sign language group) and Philosophy.

He pointed out that the vocabularies that are used for the metadata descriptions even for the collections within the disciplines are very different in several respects. Therefore there is no unified or common view (the vocabulary to formulate a query) such as Dublin Core may suggest. Dublin Core can only be seen as yet another view that may be used by some web-users. Consequently, in ECHO we have to investigate the structure and the semantics very carefully together with the specialists to come to suitable mappings. First, individual mapping schemes were described which formed the

basis for mapping schemes per view. The most problematic mapping occurs where the content of the resources is described. Some content descriptions are based on thesauri of different origin and the dimensions of describing the content are different. Here a mapping based on values has to be carried out.

At first instance a hard-coded version of the semantic mapping will be generated. At a second phase an RDF mapping should be applied where all terms are defined according to standards such as ISO 11179 and where relations are specified as well in open repositories. This allows people to modify them and use them in more flexible way.

Peter Wittenburg showed a brief demo of DORA and continued to discuss properties of the metadata sets of the content domains as well as various types of mapping problems.

Frank van Harmelen

Frank van Harmelen, who is active in defining web-based interoperability standards such as RDF and OWL gave a presentation divided in three main parts:

- Why we need "*formal* frameworks for interoperability"
- Leading open standards for such formal frameworks: RDF & OWL
- Interoperability as "ontology matching"

He introduced the basic concepts of RDF and RDF(S) and their relationship to the Ontology Web Language (OWL). According to him RDF is a mature infrastructure that can be used now. Also for the light OWL version tools are increasingly available.

Van Harmelen stressed that ontology matching remains the most important and difficult open problem, and discussed three approaches in greater detail:

- Shared vocabulary
- Upper-level ontology
- Instance-based matching

Van Harmelen concluded with a discussion of some forecasted developments with regard to the semantic web. It is obvious that to make the semantic web working we have to have methods to solve the ontology matching problem automatically or at least semi-automatically. He reported briefly about some examples of successful work.

2.1 Discussion

The interoperability talks were supplemented by a presentation of NECEP by Laurent Dousset, whereupon followed a general discussion. The remainder of the workshop was spent in sub-groups, in order to follow-up on issues and plans discussed and outlined during the two topical seminars. In addition, Jean Maroldt gave a much appreciated presentation of Marie Curie grants as a means for exchanging young researchers and promoting cooperation between the partners of ECHO.