# *Technology in ECHO*

*(working paper)*

**Peter Wittenburg (MPIPL), Dirk Wintergrün (MPIWG), Robert Casties (MPIWG)**

**This document gives an overview about the key points of technology development in ECHO. It is based on the Technical Annex, discussions with the content providing teams and on the discussions of the first developer meeting. The key points were chosen such that first concrete results can be demonstrated in October 2003 and that the final versions will be delivered at the end of the project in June 2004. The realization of the ambitious program can only be managed when the participating institutes bring in own resources and when they are willing to participate actively in the AGORA.**

## *Embedding of Technology in ECHO*

The key objective of ECHO was and is to develop a Common Technological Framework to bring online as much content as possible during a limited time span, supplemented by tools which allow collaborative work on this content. The challenge of this task can only be met by a scenario where active participation of a significant number of archives and scholarly experts is guaranteed. This ECHO project is seen as a test bed for methods of interaction and requirements driven technology development to achieve the key objective.

Technology is dealt with at various layers: (1) It is a natural part of the AGORA discussions, in particular between specialists from humanity disciplines and from technology; (2) It is part of the content provision work in so far as content providers use tools, specify their practical needs and integrate their resources into a browsable and searchable domain; (3) Dependent on the requirements of content providers and users the infrastructure will be set up and tools will be developed.

In the following the key points of technology development will be briefly presented. For more details we refer to the web site “www.mpi.nl/echo” that will be updated continuously. It is evident that all technology development will build on existing components and standards where possible. It is also agreed that all software to be developed in ECHO will be open and freely usable.

### Integration of new data

Much expert knowledge about efficient workflow procedures to digitize texts, photos, sounds and movies is available in the participating institutes. The internally available workflow schemes have to be elaborated and simplified such that other institutes can understand and apply them. Some of the institutes are ready to offer a hosting service for institutions that do not have sufficient resources. Especially the Center of Information Management of the Max Planck Society will offer its potential.

### Resource Repositories

It will be worked out which type of repositories are needed within ECHO, what kind of functions they have to fulfill and how they will be organized such that persistent storage can be guaranteed to a certain extent. Broad experience from Max-Planck-Institutes will be re-used. The type of holdings differs largely between the involved disciplines. Appropriate harvesting and access methods will be defined for the different types of resources to be offered within ECHO.

### Document Identification

It is understood that the ECHO project needs a clear scheme to associate a unique identifier with each of the resources in the ECHO repositories. A registration authority and a distributed service will be designed and setup to guarantee proper maintenance and access to this crucial information. Metadata Layers will use this identifier to point to the resources. Knowledge gathered in ISO 11179 will be used here.

### Metadata Layer (Catalogue Metadata)

An integrated browsable and searchable metadata layer will be implemented retaining as much information as possible from the individual disciplines. The expert knowledge involved in ECHO will be used to develop suitable metadata schemas for the exchange of information within ECHO. Discipline interoperability is a non-trivial issue if a pidginization of formats and metadata is not acceptable. ECHO will study new methods that are in line with the trends in the Semantic Web age. Two types of user interfaces will be offered: (1) one very simple similar to Google and (2) one giving access to the complex descriptions. This work will be based on the rich experience of one of the partners and the work in ISO TC37/SC4.

### Text Technology

To operate on texts is crucial for all scholarly work and for interlinking of resources. Text forms the glue for the different types of sources, therefore an effective use of text is a central condition for forming an integrated ECHO corpus. It is intended to enable text interlinking including morphological normalization and lexicon integration. This work can be based on the experience in particular of one of the partners gained in several international projects making cultural heritage content available.

### Image Manipulation Environment

Images are a central resource for scientific work in the humanities, e.g high-resolution images of manuscripts or drawings enhance the comparative and collaborative work on these documents. Based on the existing DIGILIB software that already has a number of very attractive functions such as zooming, defining sub-images and associating comments with spatial points, a set of additional advanced components will be developed to create a toolbox covering most of the functionality required in the humanities.

### Multimedia Environment

Another important resource are sound and movie recordings although their historical appearance can only be traced back for about 150 years. Researchers working at various locations want to easily annotate such resources in any form of

complexity (speech, gestures, signs, eye movements) in a collaborative way across the Internet. They also want to do searches and analysis on such resources for various scientific purposes and immediately access the relevant media fragments. Based on the state-of-the-art ELAN tool extensions will be programmed to make it an advanced tool to satisfy these needs.

### Annotation Structures

It was understood from the developers that at various instances flexible annotation structures are necessary to safely store the texts, the annotations, the comments made by someone and the relations created to link various elements in text structures. Annotations to distributed resources have to be supported as a starting point for interactive interlinking. A group will work on this topic and will make use of the discussions and experience in ISO TC37/SC4.