EUROPEAN CULTURAL HERITAGE ONLINE
WP2 Infrastructure and Technology
WP2 TR05-2002 - Draft Version 2

WP2 Note on Architecture

Peter Wittenburg
3.1.2003

At our initial WP2 meeting during the kickoff meeting in Berlin a number of points were discussed. Based on this discussion, a few email exchanges and especially the information about the state of the contents this note about architectural concerns emerged. It starts with some statements to lay the ground. It should be stressed, that this note is only about short-term solutions defined by the concrete circumstances and meant to be implemented until September (at least so far that first demos can be realized). Long-term solutions have to be worked out together with the Technical Committee in the AGORA.

Statements

1. The general idea is that people can search & browse in a metadata domain and then when they located a (set of) resource(s) can immediately start special tools to operate on the real resources such as video, sound, photos and texts. For the history of science domain the situation is slightly different in so far that they speak about a certain archive such as Archimedes where the whole set of resources is seen as one holistic resource with strong internal interrelations. The need for describing the individual components is not seen as so relevant[1].
2. Nevertheless, where necessary or useful we have to assume that metadata records are available in different forms and formats.
3. WP2 will establish a mapping between the metadata sets on a modular level (not DC which is a pidginized form associated with loosing relevant data for the scientists). Disciplines who don't yet have an established MD set could serve DC records as a minimum. Application profiles for the different disciplines will be developed.
4. MD records can be in different containers, WP2 has to take care together with data providers that the containers are accessible to be able to harvest them such that they are accessible from the IMDI environment.
5. MD records have to provide references to the real resources. If the resources are bundled the MD records have to refer to all of them to be able to retrieve them.
6. The resources we have to deal with are related videos, sounds, photos and texts. Textual resources can be notes of all sort, complex annotations and lexica.
7. The resources are in different formats and containers as well. Some of the texts are in XML. Some or many of the photos in the project seem to be stored in the eDOC container[2].
8. WP2 has to take care together with the local people that there are mechanisms to retrieve resources or resource bundles from the containers or at least to support accessing them from a remote site given appropriate access rights.
9. Within ECHO there must be ways to uniquely identify each resource.
10. To operate on the resources themselves there are two well-described software packages (ELAN, DIGILIB) and some language processing modules/scripts that have to be integrated in a way that has to be worked out within WP2. There are a number of possible ways; dependent on the way to be chosen an information exchange protocol has to be defined.
    - Integration on functional level: Either ELAN or DIGILIB or both are taken as the basis to implement the full functionality, i.e. all features the programs have right now are integrated as methods in one or several of the existing tools.

[1] See WP2-TR3
[2] EDOC is based on a relational database system.

- Integration on complementary level: The programs remain as they are and strengthen the functionality they have right now and call each other to include the functionality missing.
- Integration on component level: The functionality of the two major programs is transformed into a component model such that the components can be integrated into one framework. The language processing modules have also to be realized as components and integrated as well.

11. Whatever solution will be chosen for the integration, WP2 has to speak about formats that are suitable for handling all types of texts in ECHO (raw texts and annotations) and narrow down the heterogeneity. Those XML-based suggestions being discussed right now in the ISO framework have to be considered seriously. The standoff principles for XML-based annotations have to be followed as well.
12. Whatever solution will be chosen in 10, the resulting tool(s) must be callable from the IMDI metadata browser and operate in a distributed manner, i.e. the resources such as photos, videos and texts may locate at other servers. The tools have to accept parameters when being called to know which resources to be fetched and which spatial and/or temporal fragment to work on or to visualize.

Architecture

In ECHO different disciplines with different traditions and solutions come together. Figure 2 describes a possible scenario including three different types of repositories as they can be seen right now (indicated by light-green color, there may be more). This scenario has to be taken into account for building the short-term solution.

We can distinguish between three different types of tools or services:

1. visualization, annotation and exploitation tools such as DIGILIB and ELAN
2. text tools/scripts such as provided by MPIWG Berlin that run offline to establish links
3. html generation scripts for special browsing such as used at MPIWG Berlin

There is some overlap between points 1 and 3, since they both are about visualizing resource data. While the two tools support specific views on the data (one could call them stereotypic views) the Berlin setup seems to be generative in some way[3].

DIGILIB as well as ELAN can directly access the raw resource files and associated annotations[4]. Except when these resources are stored in a container (i.e. they are not directly addressable by a URL) some script has to be available to extract them. Both programs should use the same mechanisms if possible such as using the same XML-based annotation format or use the metadata record as an information source where all the associated resources can be found (see figure 1). In case that several resources are to be analyzed or viewed (several photos, several video streams, ...) the tools must be able to accept that information as well. It is up to the tool how they present the information.

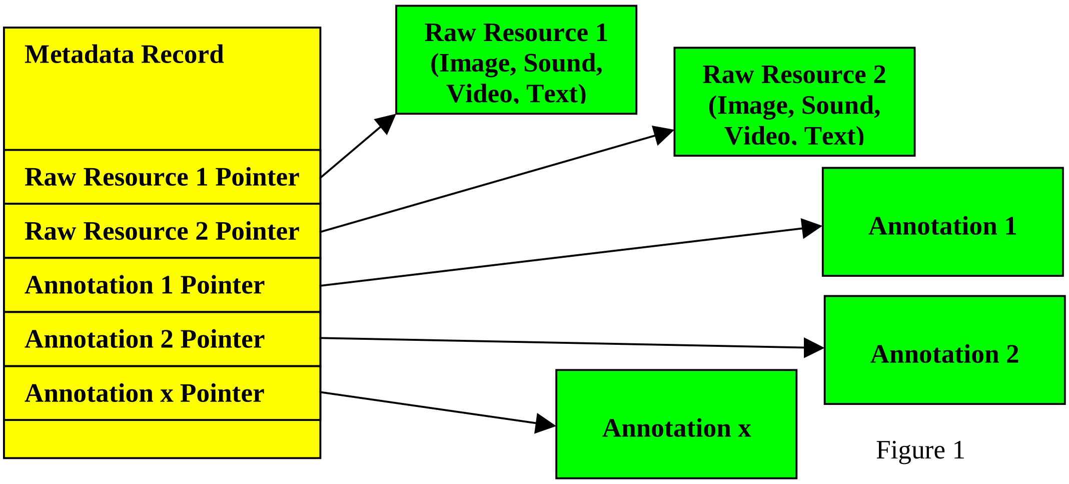
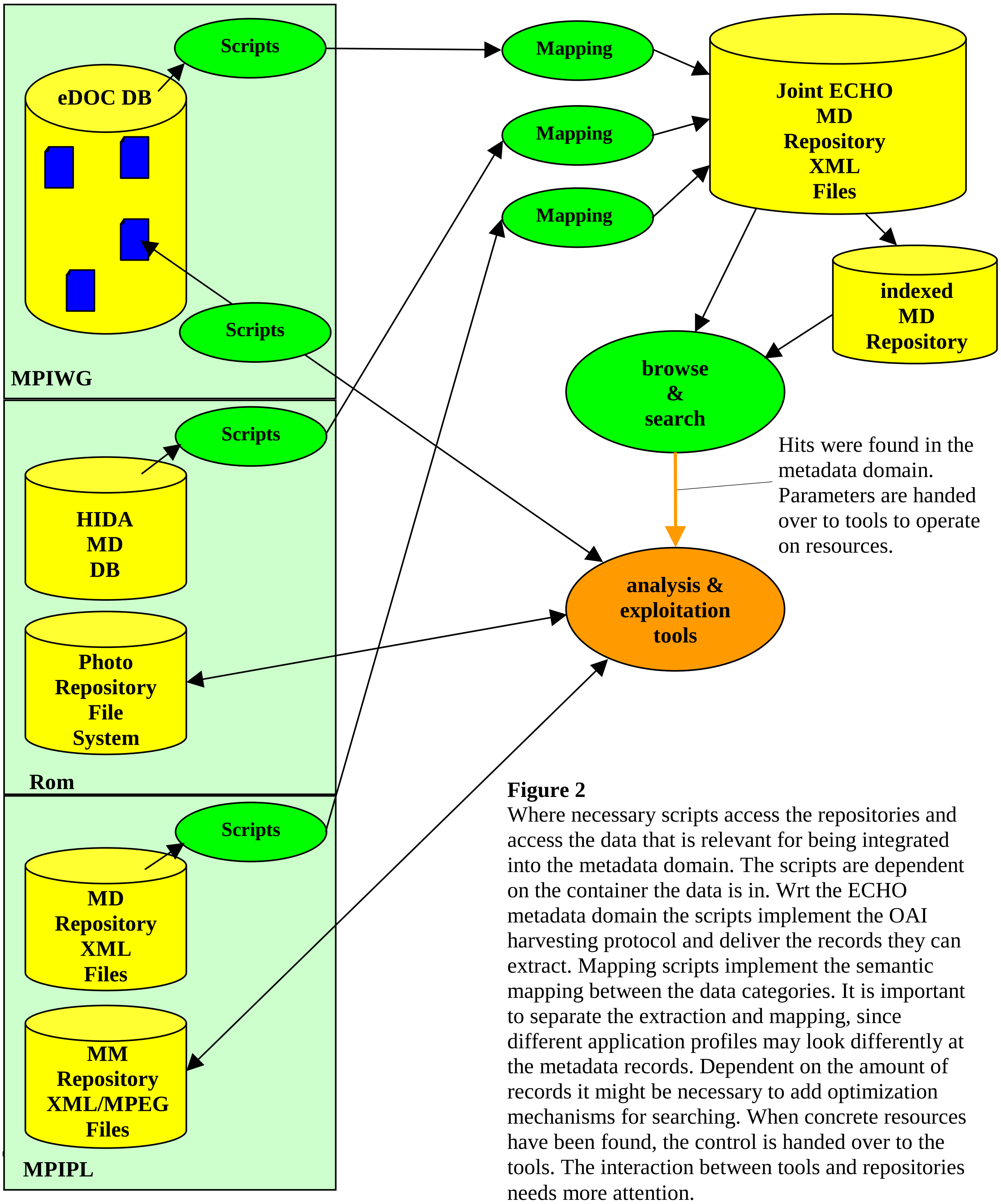


Figure 1

[3] Actually I need a good description and design from MPIWG.
[4] Not yet sure whether I fully understand how DIGILIB handles annotations.

Figure 1 indicates a typical scenario where a set of related raw resources (could be two photos of one object from different perspectives, a sound and a video recorded at the same event, two videos taken from different views, a photo and a text associated with that photo, etc) will be annotated in an open manner by different researchers. Hereby some of the annotations will be stored in different files due to standoff or privacy reasons. The metadata record acts as the glue to describe the formal relation. This relation is different to those where an element in a text is semantically related to another element in a certain text (this is scholarly metadata (using the Berlin terminology) based on semantic relations).

{*The following has to be described by MPIWG, since I am not sure whether I understand how Berlin does it. So the following is more a debate then a good statement. For this reason the Berlin scripts are not integrated in this document*}. For the programs/scripts that Berlin uses to create the semantic relations between for example words, things may look a bit different. Here one has to specify a couple of resources such as texts and lexica. The selection of texts has to be done in certain ways (perhaps by finding resources on base of metadata such as sharing the same language) and also the lexica to be involved have to be found based on similar criteria. The interlinking scripts then create annotation layers to the original texts. According to the stand off model annotations of material should be kept separately and not modify the raw text. I guess that at this moment this is done differently due to practical reasons. So here is a topic for clarification and discussion.



Hits were found in the metadata domain. Parameters are handed over to tools to operate on resources.

Figure 2
Where necessary scripts access the repositories and access the data that is relevant for being integrated into the metadata domain. The scripts are dependent on the container the data is in. Wrt the ECHO metadata domain the scripts implement the OAI harvesting protocol and deliver the records they can extract. Mapping scripts implement the semantic mapping between the data categories. It is important to separate the extraction and mapping, since different application profiles may look differently at the metadata records. Dependent on the amount of records it might be necessary to add optimization mechanisms for searching. When concrete resources have been found, the control is handed over to the tools. The interaction between tools and repositories needs more attention.