

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

IST-2001-37491

European Network of Centres of Excellence for Research and Education in Digital Culture

Work Package 6 – Deliverable 11b

Foundation for Distributed European Electronic Resource (DEER)

DEER and Digital Autonomous Cultural Objects:

A Demonstrator for ECultureNet.

Contact:

Kim H. Veltman

Maastricht McLuhan Institute

Universiteit Maastricht

Grote Gracht 82

6211 SZ Maastricht

The Netherlands

Phone: +31 43 3882699

Fax: +31 43 3252930

Mail to: k.veltman@mmi.unimaas.nl

deliverable11b.doc	Report	09 Juli 2003	1/37
--------------------	--------	--------------	------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

Project Number IST-2001-37491

Project Title European Network of Centres of Excellence for Research and Education in Digital Culture (E-Culture Net)

Document Type PU – Public usage of the result

Document Number D11b

Contractual date of delivery 01-06-2003

Actual date of delivery 13-06-2003

Title of Document Define Content Pilots

Contributing work package WP6

Nature of Document Report

Author(s) Torsten Schaßan and Manfred Thaller


Abstract The DEER is the concept presented by ECultureNet for the general framework in which digitised cultural resources shall be provided for Europe. As such it does not provide an actual technical mechanism how the interaction between different content providers can be accomplished. The present paper describes such a mechanism and is at the same time a blueprint for a demonstrator which has been built for ECultureNet.

Keyword list distributed access, digital libraries, ditributed autonomous cultural objects

deliverable11b.doc	Report	09 Juli 2003	2/37
--------------------	--------	--------------	------

Table of Contents

1. Purpose of this Paper	4
1.1 Structure of this Paper	4
2. Summary	5
2.1 Problem and Vision	5
2.2 Where do we go from here?	6
3. The Technical Concept: Digital Autonomous Cultural Objects	7
3.1 Available strategies	7
3.2 The concept: background material	9
4. Realising the Concept: A Protocol for the Exchange of DACOs	25
4.1 The ECNet – Protocol: General Design	25
4.2 The ECNet – Protocol: Currently Defined Requests	28
4.3 DACOs in the Semantic Web	35
5. The Demonstrator	36

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

DEER and Digital Autonomous Cultural Objects: A Demonstrator for ECultureNet.

Susanne Kurz, Torsten Schaßan and Manfred Thaller, University at Cologne

Report submitted as partial 11b of deliverable 11 of the ECultureNet project.


1. Purpose of this paper

The DEER is the concept presented by ECultureNet for the general framework in which digitised cultural resources shall be provided for Europe. As such it does not provide an actual technical mechanism how the interaction between different content providers can be accomplished. The present paper describes such a mechanism and is at the same time a blueprint for a demonstrator which has been built for ECultureNet.

1.1 Structure of this Paper

In section 2 below, we summarise the strategy followed and the expectations we have for the future usage of the work described here. Section 3 describes the basic concept for the exchange of cultural heritage material, the *Digital Autonomous Cultural Object* (DACO). Section 4 provides a functionally complete *protocol* for the exchange of such DACOs between arbitrary software systems. During the specification of the complete protocol, which has been derived from practical experience with the demonstrator and studies of other currently discussed protocols it became clear, that the way to access the DACOs should in future be changed slightly against the way in which it has been described in section 3. As section 3 describes a working system, where most details still apply to the future solution proposed, these slight differences have not been unified.

deliverable11b.doc	Report	09 Juli 2003	4/37
--------------------	--------	--------------	------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

2. Summary


2.1 Problem and Vision

Any technical solution for the provision of digital cultural content in a European perspective has to be aware of the character of the existing institutional framework which administers Europe's cultural heritage, being widely heterogeneous and not so much non-centralistic, but multi-centralistic. By this we mean, that in different sectors of the cultural heritage - libraries, archives and museums - national bodies exist which enforce strategies or standards, which are mandatory in their respective spheres of influence, not necessarily compatible with those enforced by other, similar bodies. It should also be noted, that heterogeneity is part of the institutional profile in a competitive environment: At the very least a cultural heritage institution will want to ensure, that its resources can easily be recognised at being provided for the users community by this institution and not any of the competing ones. Precisely the most significant institutions may beyond that insist on providing services which go beyond what has been generally agreed upon, to sharpen their profile within the perception of the user community.

This means, that some strategies for the implementation of a DEER framework are clearly less realistic than others. For the user, the digital resources offered must create the impression of a reasonably homogenous infoscape. The institutions providing the content must retain the greatest flexibility possible. If you avoid telling an institution, how it *must* solve its problems, however, you cannot create a shrink-wrapped software package to be handed over to that institution either. You can only provide a blueprint for a solution, which has to be applied by the contributing institution itself to its own software environment.

The vision presented here, therefore, assumes, that a *simple* set of rules exists, which allow a cultural heritage institution to offer its digital resources in such a way, that (a) these resources can easily be integrated into all kinds of integrated European Heritage information systems (b) the investment of the institution offering the resources is as small as possible, however. This notion of a cheap and simple way, in which everybody can prepare the own material with very little effort in such a way, that it becomes cheap and simple for others to use it, exists already in the realm of metadata: Here the Open Archive Initiative has presented a protocol, which can be implemented in typical libraries within one or two weeks, according to oral reports of participating library IT departments, and allows arbitrary other institutions to access these metadata to build their own information systems. On the OAI more below. The library committee of the German National Research Council (DFG) has recently, in a not yet published recommendation decided to recommend, that all German libraries shall in the future support the OAI protocol at least as an alternative to other access modes provided by them. One of the main reasons for this decision should be quoted here, as it clarifies a potentially non-obvious issue. While the protocol of the OAI has been created to provide for a *specific* harvesting service, the specifications simply describe a way in which *a* metadata harvester can access bibliographic metadata; its significance as a protocol goes therefore considerably beyond the immediate purpose of realising a specific vision for *one* harvester.

deliverable11b.doc	Report	09 Juli 2003	5/37
--------------------	--------	--------------	------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

The "DACOs" described in this document intend to provide the same vision for cultural heritage *data* which initiatives like the OAI provide for *metadata*. Provide a protocol, which any cultural heritage institution can follow easily and cheaply to offer digital versions of its content for anybody who wants to integrate these materials into systems with a larger scope.

The validity of this concept has been tested with servers, which have been created in projects with which the Cologne group responsible for this report is cooperating in other contexts as well. As such the DACOs / the protocol supporting them have been implemented within the context of only *one* software system, to demonstrate that the protocol proposed can actually be realised. The protocol claims, however, that is sufficiently easy to implement to be realised for other systems - Oracle, Informix, etc. etc. - to provide the same level of generality as the OAI in the area of metadata.

2.2 Where do we go from here?

The purpose of this work package within ECultureNet has been to show a general architecture for the exchange of cultural heritage data between arbitrary cultural heritage information systems within Europe and to demonstrate with a limited number of servers, that this concept actually works.

The future obviously depends on the institutional framework in which this initial work can be continued: If ECultureNet continues to be funded, we expect this blueprint to serve for the creation of a "demonstrator" which does provide unified access to a number of servers at least two orders of magnitude larger within ECultureNet. If it does not, the Cologne working group will use it for its own continued efforts at creating hybrid and distributed cultural heritage information systems.

deliverable11b.doc	Report	09 Juli 2003	6/37
--------------------	--------	--------------	------

3. The Technical Concept: Digital Autonomous Cultural Objects

3.1 Available strategies

The following technical strategies are possible to overcome the problem described initially:

1. Enforce adherence to a full set of standards.

In our opinion, this solution precludes itself, as it does not reflect the initially described political and administrative situation. For clarity's sake, we would like to add, that this does of course *not* doubt the advisability of adherence to industry standards: images made accessible by the DEER should be stored as TIFF or PNG and under general conditions offered over the WWW as PNG, JPEG or GIF. The metadata relating to them should be encoded with XML, not in any proprietary format. Unfortunately this does not say very much for the implementation of actual information systems. Standard DTDs exist, but they are: (a) Sometimes modelled so closely according to national idiosyncrasies, like the EAD¹, that there are very good reasons while many European providers of cultural heritage content will be willing to apply them only after considerable adaptations. (b) Representing only part of the trifold world of cultural heritage - archives, libraries, museums. (c) Sometimes so extremely general, that it is perfectly feasible to apply them in so many different ways, that from the point of view of an architect of an information system to be build upon them, a set of *n* information resources applying the very same standard might as well have been encoded using *n* different formats of metadata. (The TEI², e.g.)

2. Install a European cultural heritage broker accepting a set of different standards.

This would follow closely the solution proposed by the *Research Library Group*, which acknowledges the existence of a set of only partially compatible standards for different domains of the cultural heritage world and has therefore set up a central broker, which maps metadata formulated in these different standards to a common conceptual model, for which a implementation proprietary to the broker exists.

A perfect solution for a strong central institution. Almost by definition therefore, however, difficult to transfer to the European context, when the notion of strong cultural heritage institutions on the national level, under a purely subsidiary European layer, shall be kept up.

3. Install a European cultural heritage broker accepting arbitrary metadata.


This solution has been realised in the German *Prometheus*³ project. There a central broker accepts contributions from arbitrary proprietary data bases and models information provided by them internally in such a way, that a generalised query system and the use of the selected material within a specific didactic user interface becomes possible.

This solution is *not* appropriate as blueprint for a effective DEER; on the one hand for the same reasons given against the RLG solution described and refuted in the preceding

¹<http://www.loc.gov/ead/>

²<http://www.tei-c.org/>

³<http://www.prometheus-bildarchiv.de/>

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

paragraph, but even more so, as a DEER broker would have to provide for an access which allows conceptually totally unrelated models to draw data from it.

4. Agree upon response models for totally unrelated servers.

Another strategy has recently been proposed by the Open Archives Initiative⁴. Here, no prescriptions whatsoever are made for individual servers participating in the initiative. A set of regulations exists, however, how a server agreeing to the principles of the initiative has to react, when queried in a given way via the standard http protocol. This (a) leaves at the same time maximum freedom for the strategies and policies of individual institutions, (b) defines an interface however, which can be accessed by arbitrary services and institutions to access materials to be presented within a unique interface build by such a service or institution. A problem of the OAI is, however, that it responds with metadata following the Dublin Core: while this standard is widely accepted, it has, by necessity, a rather restricted semantics. And: it is definitely no option, if beyond the metadata in the usual restricted sense of the term descriptions of the actual contents or the content of the described objects themselves shall be transported.

5. Agree upon behaviour of "digital cultural heritage objects".

Finally there is a strategy which has been advocated by the author for some time primarily in the library field, which is currently the platform from which planning for a national portal of certain classes of library materials to be created by the German National Research Council's library program starts. This strategy proposes to convert individual cultural heritage objects - books, manuscripts, but conceptually also castles or pieces of furniture - into objects which follow a common behavioural code. That behavioural code provides for the objects informing about themselves, when asked to do so by a request following the OAI approach, but it also involves the object wrapping itself in markup suitable for display in a browser, and integrating itself into a WWW interface, if called correctly. Objects following such a code of behaviour we call *Digital Autonomous Cultural Objects (DACOs)*.

This approach keeps the advantages of the OIA approach - i.e., it does *not* require a strong centralised institution, and leaves maximum autonomy to contributing institutions - carries the concept further to the delivery of actual content, not restricting itself to metadata.

⁴<http://www.openarchives.org/>

deliverable11b.doc	Report	09 Juli 2003	8/37
--------------------	--------	--------------	------

3.2. The concept: background material

We propose, therefore, that the DEER implements a strategy based upon such a behavioural code for the objects out of which Europe's cultural heritage consist. Before we go on to describe the demonstrator which is currently being built within ECultureNet, we would describe this concept of Digital Autonomous Cultural Objects further, focusing on two aspects: (a) The usage and usability of such objects within a concrete project and (b) a detailed description for a possible strategy for the implementation of a platform for a system of interconnected resource servers.

3.2.1 A current example: Medieval manuscripts as DACOs

The principle discussed here, has originally been developed in the context of a server for high resolution images of medieval manuscripts⁵. Here the guiding principle has been to understand a server of manuscripts not as a digital realisation of the metaphor "library", but rather as a generator of individually addressable quotations of manuscripts addressable on the page level, to allow for dynamic references to the resource presented analogously to the classical quotation in a scholarly note.

Quoting from a previous paper⁶, we would like to recapitulate the formal derivation of the concept here. Readers who are familiar with our concept of DACOs, as developed in this project, are advised to proceed immediately to section 4.2.

Digital libraries, particularly large ones, are still seen today as unique and significant projects. As a result, they are frequently constructed as self-contained systems, where the separation between the interface of the library and its contents is not as clear-cut as one would wish. This means that many digital libraries expect that a user will enter through the interface of the library. This is an example of when the implementation of a traditional metaphor is counterproductive.

In our reference project, we have experimentally created a functionally complete linkage interface that allows one to access the content of the library completely independent of its own user interface. While this specification is not yet fully stabilised and public, it is available, and the following ways of addressing are guaranteed to be as persistent as the floating discussion of persistent basic identifiers allows. A researcher who intends to refer to the content of the Cologne manuscript library will have a mechanism that allows him or her to address reliably and persistently the following:

1. A digital object that represents a conventional unit of reference within a given discipline. In our case it is a medieval codex.

⁵The CEEC project (Codices Electronici Ecclesiae Coloniensis), <http://www.ceec.uni-koeln.de>, which is the application side of the CEC (Codices Electronici Coloniensis) project for the definition of self contained manuscript representations in electronic media.

⁶<http://www.dlib.org/dlib/february01/thaller/02thaller.html>

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

2. A digital object that represents the same object at a finer level of granularity, reflecting the usage of a given discipline. In our case individual pages are at the finer level of granularity.

Note: We refer intentionally to "units of references" and "granular objects" instead of "codices" and "pages," not to introduce an additional level of complexity, but to prepare for the generalisation of such addressing schemes to other cultural heritage material. Ultimately codices can be seen as particularly simple cases, where only one level of subdivision exists and the granular objects are ordered linearly, as opposed to, e.g., museum objects, where a number of hierarchical levels for digitisation of details exist, and intuitive schemes for the naming of granular entities are considerably more complex. The basic problems remain the same, however.

The two types of reference above are necessary for two reasons:

1. From the end user's point of view, it is important to be able to include a reference to a digitally stored manuscript directly in a text. This will become much more important in the future when the results of research are themselves presented on digital media. In such cases it would be almost absurd to have an end user directed from a footnote to the search engine of a digital library, instead of the digital object itself, the address of which was obviously accessible to the author at the time of writing.

2. From the conserving institution's point of view, a clear tendency towards virtual libraries / archives / museums seems to exist. The most obvious way to construct such virtual collections is to envisage them as access platforms that hide from users the fact that the individual objects accessed are stored under different administrative and technical conditions. This is achieved most easily if an access machine can access individual digital objects in different holdings directly, that is, without a negotiation process with the access tools of the specific institution holding the object.

It would be highly impractical to rely on a central body, operating world wide, to create a new set of identifiers for all existing objects of cultural heritage. All existing collections of manuscripts, archives, museums, etc., would have to agree upon a common system of shelfmarking for their objects. This is not only impractical, but also directly damaging, as the reference systems within collections of cultural heritage material that have grown historically usually represent by themselves a specific intellectual view of that material.

We envisage, therefore, a solution that divides the general problem into three sub problems:

1. A persistent addressing scheme for collections, which by necessity must be organised nationally, with national (or regional) solutions being co-ordinated by appropriate international bodies.

2. A persistent addressing scheme for digital objects within individual collections that is under the control of the individual institution, but which guarantees a common functionality and interoperability of the different collections.

deliverable11b.doc	Report	09 Juli 2003	10/37
--------------------	--------	--------------	-------

3. A mapping scheme that allows referencing a granule of a digital object by a specific numbering scheme, which is then translated into the actual names of individual digital components, like page images. Such a mapping scheme is administered by the individual collection and should even exist if the names of the digital objects -- file names -- also reflect the traditional references directly. The order of access to granules of digital objects -- "next page", for example -- is a matter of interpretation. To allow operations like "virtual rebinding" of a digital codex, we strongly propose to differentiate clearly between this level and the preceding one.

An implementation of an addressing scheme for digital objects based on the preceding analysis would look as follows:

<collection-reference> <object-reference> <granule-reference>

where <granule-reference> is either a <direct-granule-reference> or a <mapped-granule-reference>.

CEC has a processing model for <object-reference> and <granule-reference>. For the discussion / definition of the concept of a <collection-reference> we seek the support of appropriate library institutions. CEEC has a working implementation for all three <object-reference>, <direct-granule-reference>, and <mapped-granule-reference>.

3.2.1.1 Discussion of individual access models

A complete Cologne codex can currently⁷ be reached via a WWW address like:

<http://www.ceec.uni-koeln.de/ceec-cgi/kleioc/0010KICEEC/exec/katk/%22kn28-0083ii%22>

Ignoring the "%22" (the CGI wrap-up for the quotation marks), this means our previous definitions are realised as follows:

<collection-reference> = <http://www.ceec.uni-koeln.de>

<object-reference> = [ceec-cgi/kleioc/0010KICEEC/exec/katk/%22kn28-0083ii%22](http://www.ceec.uni-koeln.de/ceec-cgi/kleioc/0010KICEEC/exec/katk/%22kn28-0083ii%22)

To access an individual page of a Cologne codex, a WWW address like the following can be used:

http://www.ceec.uni-koeln.de/ceec-cgi/kleioc/0010KICEEC/exec/pagedmed/%22kn28-0083ii_164.jpg%22

Here the following is applicable:

⁷See below under section 4.2 for a more general way, in which the objects can be accessed in the future.

	IST-2001-37491	WP6	D11b
---	----------------	-----	------

<collection-reference> = <http://www.ceec.uni-koeln.de>

<object-reference> = [ceec-cgi/kleioc/0010KICEEC/exec/pagedmed/](http://www.ceec.uni-koeln.de/cgi/kleioc/0010KICEEC/exec/pagedmed/)

<granule-reference> = %22kn28-0083ii_164.jpg%22

<collection-reference>

In the example, <http://www.ceec.uni-koeln.de> is obviously a URL. This is where we seek the support of existing library institutions. Obviously a persistent identifier for the individual collections should replace the URL. It would be particularly helpful if the identifiers would incorporate existing schemes for the unambiguous reference to institutions. For example, it would be helpful if the identifier above could be replaced by something that contains a reference to "Kn28," the (within Germany) traditional unambiguous reference to the library in question.

Less formal than the rest of these proposals: Within the WWW the question of top-level domains is very much open to discussion, and with "*.museum", at least one type of cultural heritage institution has reached top-level status. Considering the fact that libraries in many ways are the nodes of the information network, when we consider the actual amount of information handled, it is fair to wonder if there are there any discussions underway that would lead to the creation of a library top-level domain and references like "www.kn28.de.lib". If not, why not? This could be a very good starting point for persistent implementations and, with the library community directly responsible for the administration of its domains, would do away with an entire level of problems.

This, of course, could alternatively be an important topic on the agenda for a DEER: To act as holder of a base address for accessing Europe's Cultural Heritage.

<object-reference>

Once the problem of the persistency of the basic identifier is resolved, we consider a robust technical solution reasonably simple.

We have implemented the following scheme:

<object-reference> = <interface> <access-mode> <resource-id>

with the following considerations:

<interface>

The <interface> of a CEC <object-reference> is a series of one or more identifiers separated by slashes. They represent a software system existing at a given point in time, in our example: [kleioc/0010KICEEC](http://www.ceec.uni-koeln.de/cgi/kleioc/0010KICEEC).

Notes:

The reference to a specific interface may be seen as directly opposed to locator persistency. It has been included based on the following assumptions, however:

1. The only thing about which we can be reasonably sure regarding the further development of net-oriented information access is that it will develop considerably beyond the current stage. It is

deliverable11b.doc	Report	09 Juli 2003	12/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

very likely, therefore, that future access systems to digital resources will make them accessible in new ways.

2. On the other hand, it is unlikely that a typical preserving institution will make fundamental changes to its software platform more regularly than, say every ten years.

We would therefore assume that a given institution, when exchanging a software platform 'x' with a software platform 'y', will provide scripts or their future equivalents that will direct all references to the software interface 'x' to methods provided by the new software which closely resemble the previous picture, while at the same time a reference to the new interface 'y' can be provided, making full use of any additional capabilities the new software provides.

As such changes will be infrequent, it is not unreasonable to ask an institution to provide the level of legacy support represented by such scripts.

<access-mode>

Almost all imaginable systems for the administration of digital objects will provide access to them according to different qualities, resolutions, access privileges and the like. The *<access-mode>* of a CEC *<object-reference>* provides a means to differentiate between different combinations of such properties. Like the interface, it is a series of one or more identifiers separated by slashes.

Notes:

1. It is important to differentiate the type of access granted to an individual object as cleanly as possible from the reference to that object itself.
2. Access-mode notation should, however, not be kept too simple. Relatively soon, standard qualities for digital objects will be developed. In this context, it is important that a mechanism exists that allows the combination of abstractly defined qualities, presumably by standardised names, with specific types of access provided by a library, modelling the peculiar properties of a certain object.


<resource-id>

Within the CEC mechanism, a resource-id is a string that allows a direct reference to a specific digital object. A functionally complete set of descriptive data exists so that this object can be accessed (and potentially transmitted) independently of the remainder of the collection. For reasons of better interoperability between resource-ids derived from different collections, we strongly propose that an abbreviated form of collection identifier is contained within the resource-id. This should make it possible to construct a complete reference to a digital object from the resource-id alone.

<granule-reference>

Within the CEC mechanism, a granule-reference is a string that allows a direct reference to the smallest division of digitised information within a digital object. Typically this will be the file containing a scanned page. For reasons of better interoperability between references derived from

deliverable11b.doc	Report	09 Juli 2003	13/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

different collections, we strongly propose that the complete resources id is contained within the granule-reference. This should make it possible to construct a complete reference to a digital object from that reference alone.

<direct-granule-reference>

A direct granule reference consists of a string that can be used directly to access digitised information on a specific server. It may be necessary to break the reference up into components that represent different levels of a storage hierarchy, and/or into components that map logical names unto physical storage addresses. It does not allow for any conceptual interpretation, however. A collection guarantee indicates that the *<direct-granule-reference>* of a digitised page or other atomic unit of digitisation will never change throughout its existence. In our example, kn28-0083ii_164.jpg is a direct-granule reference.

<mapped-granule-reference>

A mapped granule reference consists of a string that is separated by a dividing character. The CEEC implementation of the CEC concepts uses a vertical line "|". The first of the two parts is the identifier of a mechanism that allows the second part of the string to be mapped to a direct granule reference, according to a specific set of rules, which may be changed over the life span of the digital object or, indeed, be dropped as obsolete. If a mapped granule reference starts with the vertical line, it maps to a default mechanism that will exist for the complete life span of the object and is called a "canonical reference".


In our example: |kn28-0083ii_82r will map to the file which represents page 82 recto of the manuscript according to the canonical references given in the literature referring to it. Miller|kn28-0083ii_insertion4-3r may map to the file containing page 3 recto of the fourth insertion into a hypothesised original manuscript proposed to be assumed by researcher Miller. This interpretation may be adapted according to the researcher's progress or, indeed be dropped if he or she turns out to be mistaken.

3.2.2. A DEER built of DACOs

Building a strategy for the interconnection of resources for objects like these, we describe in the following a concept which has been developed in parallel with one for the creation of a unified national portal for digitised versions of printed books in German libraries. Germany being extremely non-centralistic in its cultural heritage funding, this seems to be a model which lends itself directly for the creation of a similar service on the European scale.

Collections of printed works normally consist of physical entities: -volumes, that allow a non-ambiguous identification of the underlying entity. The physical storage device "book" corresponds with a logical unit which is managed by a metadata record. Anomalies, like independent -writings bound together due to the accidental development of individual collections occur but rather as exceptions. Such anomalies are rather the rule, when one tries to expand to other areas of the cultural heritage: in the case of altars within a museum, loosely related objects - the individual painted tables making up a Gothic altar - being the norm here. In all fields of

deliverable11b.doc	Report	09 Juli 2003	14/37
--------------------	--------	--------------	-------

	IST-2001-37491	WP6	D11b
---	----------------	-----	------

cultural heritage, however, an abstract heritage object exists, which is the "obvious" building block for any repository of the respective class of material.

Thus heritage objects differ to great extent from digital representations, for which the "natural unit" is normally a single file while a "group of files", representing the individual pages of a physical unit described by metadata, is relatively rare. The problem is made worse as digital objects normally exist in more than one variety: One set of files representing the underlying book in one resolution, the other representing it in another, a third one providing a different type of compression etc.

We need, consequentially, first a naming convention for the lowest level of granularity of digital media -- the individual files -- which guarantees that all files belonging to the same underlying heritage object, the book, can be recognised automatically as components of one distinct object. This naming convention has to be resistant against migration faults.

It has to be ensured, e.g., that no necessary context information will get lost when the file systems are transferred from one physical medium to another. A typical problem in that context arises, when the structure of directories and their names bear semantic significance: the way in which files are distributed across several directories represents a consequence of a very specific stage of the development of computer technology. It can in no way be assumed, that currently optimal directory structures will remain to be so, when operating systems develop further. The naming convention chosen has, finally, to allow for the handling of files containing metadata as well as files providing the actual content of the digital object with *one* consistent convention covering both areas.

3.2.2.1 Convention for the naming of digital media units

As a basis for such a naming convention we propose:

`<media-ID>_<digital-ID>`

Basis of the name of all digitised objects is a non-ambiguous identification of the respective physical media unit (= *media-ID*). For the derivation of the *media-ID* from existing systems of identifiers used within a national context see section 4.2.2 below. For reasons given there one can expect more variation in the special characters making up this part of the of the name than in other parts. Therefore the *media-ID*, which reflects the original heritage object, is separated from the following *digital-ID*, which identifies an individual component of the set of digital files representing that object, by the character combination "_". The *media-ID* must not contain this character combination (as well as the "-_" to be introduced below). Apart from that it may contain any characters as long as there are no other restrictions inherent in current web-technologies.

Furthermore applies:

`<digital-ID> ::= <meta-ID>|<digital object-ID>`

where the *meta-ID* represents metadata describing the heritage object, the *digital object-ID* identifies one discrete unit of digital data representing the content of the heritage object.

deliverable11b.doc	Report	09 Juli 2003	15/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

3.2.2.1.1 The "meta file".

For every digital heritage object the following file is *obligatory*:

<meta-ID>_ _meta[.state-of-the-art-extension]

At the moment one has to assume that this file will be an XML-file. The file contains the metadata in the usual sense.

We propose to choose for this file a XML-binding of the Dublin Core Format. In this case the "state-of -the-art-extension" of the first generation of objects should be ".xml".

The file "abc123+7a_ _meta.xml" provides, therefore, the basic DC metadata for the heritage object "abc123+7a".

3.2.2.1.2 The "meta-toc file".

For every digital media the following file is *recommended*:

<meta-ID>_ _meta-toc[.state-of-the-art-extension]

At the moment we assume that this file will be an XML-file. The file contains a list of all digital objects belonging to the respective heritage object in the sequence intended as default for their presentation to a user. This may be combined with the table of contents of the printed work.

The precise format for this file remains to be determined. The author considers the notion of a slight generalisation of the Ebind⁸ DTD, which is simple, straightforward and easily adaptable as a very pragmatic and useful way to go. Unfortunately in the meantime, the site presenting that DTD itself proclaims to be superseded by the METS⁹ standard, which "was designed to support all of the administrative, technical and structural metadata now known to be critical for effective and reliable long-term preservation of digital files". At least in the opinion of the author, METS has not achieved that goal, becoming a typical standard by committee, which is interlinked closely and inseparably with various other standards and contains a number or decisions which are highly debatable. As such recommending METS requires a much greater uniformity from participating institutions than requiring a simple structure preserving a table of content. And: while Ebind, by virtue of its great simplicity, can easily be a blueprint from which to generate slight variations, which administer different types of digital objects, METS is closely and inseparably connected to the notion of a printed book. In any case, a table of content following rules to be determined within the scopes of these models *should* be provided by a file with the name f"<media-ID>_ _meta-toc.xml".

⁸<http://sunsite.berkeley.edu/Ebind/>

⁹<http://www.loc.gov/standards/mets/>

deliverable11b.doc	Report	09 Juli 2003	16/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

If no such file representing the interrelationship between individual component files is provided, it is obligatory, that the individual files making up the heritage object are named in such a way, that their sequence according to the ISO 8859-1 collating sequence creates a meaningful default order for their representation. A naming convention for "associated objects" - single sheets and the like has to be prepared.

One has to assume that these two files - the basic metadata and the "table of contents" - will in many cases be created by automatic procedures which are independent from each other: in a setup, e.g., where the core metadata are created by an export from the local computer based catalogue of the holding institution, while the table of content is the result of a OCR process in the responsibility of an external service provider. In principal these two files should, therefore, be considered independent of each other. Nevertheless, a DTD has to be provided, which allows to include both types of information within the meta.xml file.

At the definition of the DTD for the meta-toc.xml components one has to attach importance to a deliberate division into a core which is as compact as it can possibly be and well defined interfaces towards local extensions. For local extensions one has to prepare possibilities for documentation, e.g. in RDF, that make their automatic processing easier.

3.2.2.1.3 The "meta-extension file".

For every digital media the following file is *allowed*:

<meta-ID>_ _meta-<local-extension>

Files of this kind contain information which offer content exploration beyond a) the metadata conventions of the type of cultural heritage institution in question and b) the table of digital objects / table of contents. Recommendations for the processing of those files will be given later.

3.2.2.1.4 Digital objects and their names.

The following rule applies for names of files providing digital content:

<digital object-ID> ::= <base-ID>[<segment-ID>][<version-ID>][<technical supplement>]

All four components of the <digital object-ID> it have to be constructed exclusively by characters from the pool of the "Latin core alphabet", "Arabian digits", "hyphen", "underscore" and "full stop". Beyond these assumptions, which may seem rigid, should allow the immediate usefulness of digital resources under the technical conditions of today, however, the following applies to the individual components of the *digital object-ID*:

deliverable11b.doc	Report	09 Juli 2003	17/37
--------------------	--------	--------------	-------

	IST-2001-37491	WP6	D11b
---	----------------	-----	------

The <base-ID> is *obligatory*.

It describes such a unit of the original physical object, as gives the digital object an intuitively suitable granularity. With printed works this normally will be the single page. It is recommended to represent the physical succession of the single digitised objects within the original physical media unit according to the sorting mechanisms of the ISO 8859-1 collating sequence in the <base-ID>. In case no <media-ID>_ _meta-toc.xml exists, this is *obligatory*.

The <segment-ID> is *optional*.

It will be used in cases where the technical state of the art recommends to represent a base unit of the physical media unit by more than one digitised object. Examples are the digitisation of historical maps in Tiled Image Systems or the digitisation of book pages with characteristics of colours that vary to great extent, for which the overall image will be supplemented by semi-independent images of details.

If a <segment-ID> is used, it has to be separated from the <base-ID> by a character combination used exclusively for that purpose within the individual heritage server.

The <version-ID> is *optional*.

It will be used in cases where a system offers different versions (solutions, levels of compression, etc.) of one and the same physical object.

If a <base-ID>[<segment-ID>] is supplemented by different <version-ID>s, it has to be ensured that all such versions refer to the same physical object.

If a <version-ID> is used it has to be separated from the <base-ID>[<segment-ID>] by a character combination used exclusively for that purpose within the individual heritage server.


The <technical supplement> is *optional*.

Normally it is a file extension which reflects information relevant for processing, e.g. the format of a graphic file.

One has to point out that newer software systems more and more do not deduce the type of data from the file extension, but from an analysis of the content of the file. As the absence of such inherently physical information from the name representing the shortest logical description of the content, which while shortest, is at the same time most intimately connected to that content, it is recommended to use the <technical supplement> only in those cases where this is necessary due to unavoidable technical restrictions.

If a <technical supplement> is used, it has to be separated from the <base-ID>[<segment-ID>][<version-ID>] by a character combination used exclusively for that purpose within the individual heritage server.

deliverable11b.doc	Report	09 Juli 2003	18/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

It is *recommended*, but not *obligatory*, to choose three different character combinations to separate the four parts of the <digital object-ID>.

3.2.2.2 Derivation of the <media-ID>

Formally we define:

<media-ID> ::= <heritage-object-ID>[-_<copy-ID>]

The <heritage-object-Id> specifies a non-ambiguous identification for a cultural heritage object. Here the greatest differences between the three classical types of heritage institutions exist: printed books can exist in functionally identical copies in several libraries. In the world of archives or museums - e.g. rather recent coinage - such a situation may also occur, will not be the rule, however. "Functionally identical" means that for the user it is irrelevant which of the copies he or she will access. Thus an integrated heritage server has the right to choose from any of the digital units if to the <heritage-object-Id> requested by a user no <copy-ID> has been added. The reason for a heritage server to access a specific digital media unit and not another one may be that it is most suitable for an optimal distribution of net traffic or a similar consideration.

The <heritage-object-Id> will be generated the following way:


- When a project digitises holdings for which according a widely accepted net-based cataloguing system exists, the possibly shortened- identification / shelf-mark used by this standard catalogue *must* be used as <heritage-object-Id>.
- When a project digitises holdings for which no such catalogue exists, but metadata are available from any less widely accepted net-based catalogue, the <heritage-object-Id> will be created by using an appropriate id for the respective institution supporting that catalogue, suffixed by the identification extracted from it as above.
- When a project digitises holdings for which no net-based catalogue exists a local computerised catalogue may be used. In this case the procedure given in the preceding item applies mutatis mutandis. The identification derive din that way *must* be persistent. Persistency can be assumed, if it is an identification being used in printed catalogues or if it exists in an electronic form, which has clear requirements for long term preservation, as e.g. non machine generated identifiers in library OPACs.
- If none of the above applies, a digitisation project has either to create a catalogue according to the criteria given above, or it is unsuitable for inclusion in a wider heritage network.

A <heritage-object-Id> can be supplemented by a <copy-ID>. The two of them will be connected by the character set "-_".

The <copy-ID> will be generated in the following way:

- When a project digitises holdings for which according a widely accepted net-based cataloguing system exists, the <copy-Id> *must* be derived form that system.

deliverable11b.doc	Report	09 Juli 2003	19/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

- In all other cases the <copy-ID> can be built from an unambiguous id for the the heritage institution responsible for the digitisation.

3.2.2.3 Generation of metadata for a DEER using DACOs

Metadata should be generated in such a way, that whenever possible already existing data can be used and the systems administering them at the moment shall automatically, with the smallest possible expenditure, become able to offer the digitised material representing the heritage object described by such metadata, without any change of software or procedures necessary for participating institutions.

In the following we describe of such process, which should be cheap and reliable, applicable to as many European heritage institutions as possible, without requiring a high investment into a central institution. We make a few assumptions which every system for the administration of metadata should be able to fulfil effortlessly, if its current IT infrastructure does not seriously fall short of the State of the Art of database technologies.

- There has to be a possibility to export the data of the system of a particular heritage institution / network of such in a non-binary form, i.e. the ASCII (UNICODE) character set.
- In doing so it has to be possible to generate within the system, which holds currently the metadata for the heritage object to be digitised, a persistent and non-ambiguous identification which becomes part of the exported data.
- Additionally it has to be possible to alter a data record of the metadata-creating system at a later point of time by an re-imported data set, which refers to the persistent non-ambiguous identifier created during the export.

Given the chronic overload of IT-departments of heritage institutions, in the following it will be assumed, that the above described import and export operations which can be organised as routine procedures will be the only operations that have to be undertaken to generate a connection between existing heritage information systems and emerging digital resources.

Under these conditions the following procedure becomes possible for digitisation projects where the metadata already exist in electronic catalogues.

1. After the decision to digitise a specific object, the appropriate metadata will be obtained as ASCII file with least possible effort from an heritage institution / network, which has a metadata set for that object.
2. For the existing cataloguing systems of European heritage institutions modules for conversion have to be supplied, which create a <media-ID>_ _meta.xml file from these data. If such

deliverable11b.doc	Report	09 Juli 2003	20/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

modules are created by a closely monitored, software engineering driven project, the cost for such modules can remain very low.

3. When digitised objects become available, as the result of a digitisation project, the base URL of the server together with a list of the <media-ID>s will be passed on to comparable modules. These modules will generate ASCII files which serve as update instructions for the supplying catalogue system to create there a connection between the existing metadata and the digital version of a heritage object. (I.e., a URL to the digital content.)
4. After updating the system with the new files the according systems are connected to the digital documents.

This workflow on one hand ensures that immediately by the end of the digitisation process of an individual object the metadata are available which are necessary for the integration of digitised objects in a central portal, which can be, but does not need to be the role of an institutionalisation of a DEER; on the other hand a connection of the digitised objects to the already existing cataloguing systems is guaranteed.

3.2.2.4 Structure of access to the digitised objects

To be able to use the <media-ID>, the following structure of accessing digitised units will be supported. *How* a participating institution realises that functionality is completely left for that institution.

Every digitisation project is obliged to provide the following basic access mechanisms. (This does not preclude it to provide additional ones, offering additional services.)

1. Every digital collection of heritage objects has to provide a base-URL, to which any of the <media-ID>s supported by that institution can be added.
2. The server of the heritage institution addressed responds with a web page that is autonomous in that sense, that it (a) can be displayed in other sets of pages / frames without any attempt to influence the geometry management of the respective pages / frames and (b) provides a minimal set of instruments for navigation within the digital resource. (I.e., within the distinct digital units representing the material object.)
3. The optical design of the dynamically generated WWW pages is up to the respective institution. To create a common look and feel for the European cultural heritage, the following rules apply:
 - ✗ The major part of the page has to be reserved for the digital image. At least 90% of the area of the digital image has to be made available for displaying of the content.
 - ✗ The rest of the page has to contain the instruments for navigation and a unobtrusive logo / identifier of the institution providing the object. To generate an intuitive behaviour, this logo should function as a link to the general on-line services of the institution providing the digital object.
 - ✗ Every generated page may also contain a logo of the funding organisation responsible for funding the project.
 - ✗ If the European commission decides to provide funding for a general European heritage server as nexus of a DEER, which provides direct access to all European cultural heritage resources

deliverable11b.doc	Report	09 Juli 2003	21/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

following this behaviour, the pages should also contain a suitable logo provided by the commission acting as link to a general interface to all heritage resources available under that model.

4. The following forms of access to a specific <media-ID> have to be supported by any heritage server realising these recommendations, irrespective of the origin of a request for this type of access.

Requests for plain <media-ID>s

If a <media-ID> is not augmented further (i.e., if its is just a <heritage-object-ID> with or without a <copy-ID> being specified) the addressed heritage server has to respond with a page, which contains as a minimum the following items of information:

- ✕A complete set of metadata for this heritage object.
- ✕A navigational element, which leads to a "basic navigation page" for the digital objects representing the heritage object. (A table of content, basically.)
- ✕A navigational object, which provides a persistent URL for this page. (To provide for a possibility to "quote" the inspected heritage object, if a user has arrived at its location in a way which does not make the URL obvious.)

The minimal requirement, therefore, is a visual representation of the <media-ID>_ _ _meta.xml file.


If the navigational element, which leads to the "basic navigation page" is activated, this page is generated, containing minimally the following items:

- ✕for heritage objects, for which no <media-ID>_ _ _meat-toc.xml file exists, a survey about the available digital objects, allowing to address any of them by entering an id for it.
- ✕for heritage objects, for which such a file exists, a table of contents is created, which allows to select a digital component by browsing through a visualisation of the content of the <media-ID>_ _ _meat-toc.xml file. This page also allows to address a component object directly by entering its id.

The access to a page containing the basic metadata and to another "basic navigation page" are discussed separately above, as in some situations this separation will provide considerable gains of performance. (A good example would be the palace of Versailles: Its metadata can be kept to a reasonably short record; any meaningful "table of content" would be highly complex, so it should be generated - and transferred - to the user only, if it is absolutely clear that he or she wishes this to happen.) In many cases of servers offering heritage objects with a more simple structure, the servers will respond already to a <media-ID> by a table of content. This is legitimate.

Digital objects, which are accessed via a table of content produced as just discussed, are displayed in the form described below for access generated in response to a specification of <media-ID>_ _ _<base-ID>.

deliverable11b.doc	Report	09 Juli 2003	22/37
--------------------	--------	--------------	-------

	IST-2001-37491	WP6	D11b
---	----------------	-----	------

Note:

We discuss here the minimal functionality of a *single* digital library. The role of the <copy-ID> has not been discussed, therefore. If a heritage server receives a <heritage-object-ID> without a <copy-ID> added, it is expected to deliver the copy the institution considers to be most preferred. It has the right to ignore a <copy-ID> which is not supported by that server, replacing it with a copy available. If a individual cultural heritage server offers more than one copy of a heritage objects, it is free to ignore a request for a specific copy. <copy-ID>s are only intended for virtual servers, integrating services from many heritage servers, as a base for optimising overall server / network usage.

Requests for augmented <media-ID>s

If a request specifies the combination <cultural-heritage-ID>_ _ <base-ID> the server reacts with a page, which as a *minimum* contains the following information

- ✂ The digital object.
- ✂ A navigational instrument to go to each "logical direction" in which another digital object exists relative to the one addressed. (*Previous page / next page; overview of images displaying details; adjacent room on the left etc.*)
- ✂ A navigational instrument, to reach the basic navigational page for this heritage object (table of content).
- ✂ A navigational instrument, which generates a URL which allows to address this digital object persistently, to allow for mechanisms of quotation.

If for a <base-ID> more than one <segment-ID> is available:

- ✂ A heritage server is entitled to ignore - within the mandatory minimal responses - the <segment-ID>s.
- ✂ It has to guarantee, however, that in such a case a presentation of the heritage object is selected, which allows to view it completely.


If more than on <version-ID> exists:

- ✂ A participating heritage institution is entitled to make some of the versions available only under special restricted access modes. Does it receive a request for a <version-ID>, however, which is not available unrestrictedly, it is required, however, to respond with a valid page, which refers to a version available without restrictions.

Note: The two rules above shall encourage heritage institutions, which intend to make some of their holdings available only under specific schemes of payment, to participate within a DEER with other, free, versions of the same objects.

So far we have described a structure, by which different heritage servers are easy to refer to and to integrate by following a common behaviour. The two remaining items assume, that beyond this, the commission decides to provide startup funding for a European cultural heritage server, which provides an actual location for such DEER services as can be based on these behavioural rules.

deliverable11b.doc	Report	09 Juli 2003	23/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

5. Participating heritage server are required to provide access to copies of their <media-ID>_ _meta.xml and <media-ID>_ _meta-toc.xml files in such a way, that a European cultural heritage server can download them regularly to generate a unified access interface out of these data.

Each participating heritage institution provided, furthermore, a URL, which allows this heritage server to access an XML file, which contains a up-to-date list of all the heritage objects offered digitally by that institution at any one time.

6. A DEER server funded by the commission generates from this data:

- ✗ Access tools which integrate the metadata of these DC XML formatted files.
- ✗ For users of these access tools it remains transparent at *which* of the participating heritage servers the digital files reside.
- ✗ As the persistent URLs for the material accessed via this central interface are provided by the participating institutions independent later access to the resources, once identified, do *not* require activities or bandwidth of the DEER server.
- ✗ Depending on the further development of the topology, bandwidth and usage conditions of the European networks, scheduler can be integrated into a DEER server, which distributes access to particularly frequently requested <media-ID>s to different participating heritage servers, which offer these objects with different <copy-ID>s.

deliverable11b.doc	Report	09 Juli 2003	24/37
--------------------	--------	--------------	-------

4. Realising the Concept: A Protocol for the Exchange of DACOs

4.1 The ECNet- Protocol: General Design

The communication between the demonstrator, called henceforth the "ECNet-broker" and the individual servers of cultural heritage objects, called DACO-servers from now on, which the demonstrator accesses, is based on the DACO-protocol, which manages all exchanges between the two.

A DACO-Server is an arbitrary database, which fulfils the requirements of the DACO-protocol and implements a DACO-interface. Both the architecture and content of the database is beyond the scope of the DACO-protocol so that every abstract data model (RDM, OODM, ORDM, native-XML, ...), every vendor of a DBMS implementing one of these models (Oracle, Filemaker, MySQL, Tamino, Kleio, ...) and all types of database content (text, images, audio, video, ...) can be addressed by the ECNet-broker, as long as a server supports the DACO-protocol.

The interface of the ECNet-broker for the enduser is a set of html-pages which offer a number of search mechanisms. If the user issues a search request, the ECNet-broker generates on integrated list of responses form the responses received from all contributing DACO-servers. Every entry of this list contains a link to the original object, from which the searchable information has been generated, which refers to the address of that object in the DACO-server which offers its. By activating a link the ECNet-broker requests the object and displays it within its own display area.

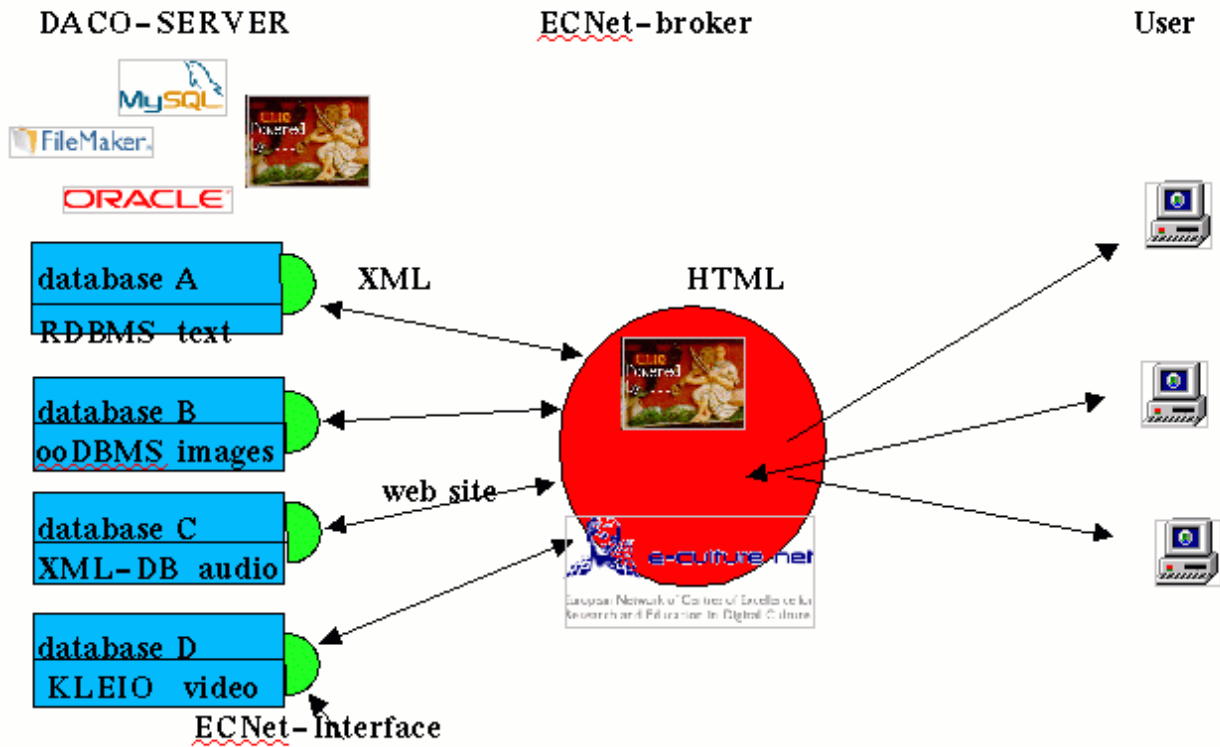
The experimental ECNet-broker is currently implemented as a Kleio-application¹⁰. The broker follows a general logic, which is closely related to that of a harvester. Metadata are, at regular intervals, requested form the DACO-servers and integrated into a system of common, centrally held index files. All user interfaces are self adapting: If, e.g., an participating DACO-server offers metadata for an additional central index at a particular stage, these data will be integrated into the interface without any human modification of the system. While at present all participating DACO-servers are taken from cultural heritage projects supported by the Cologne working group, other projects the working group is active in, notably the German *Prometheus* project quoted above, have already shown, that the identical strategy for such a broker, indeed the identical software, can also be used to integrate servers from different vendors.

This basic architecture of the demonstrator of the ECNet-broker is shown in graphic 1.

A DACO-server has to implement a DACO-interface (referred to as "ECNet-interface" in its current implementation within the demonstrator), that reacts in a defined way to a DACO-request sent to it by the broker. All communication is done via the http protocol, where all modes of

¹⁰Kleio is an open source data base and knowledge administration environment which supports a graph-oriented data model, which is a superset of all data structures that can be realised in XML. It is developed and maintained by the Cologne professorship for Humanities Computer Science responsible for this report. <http://www.hki.uni-koeln.de>.

communication are supported. In that reliance on an existing transport protocol the design of the system follows closely other more recent protocol, e.g. SOAP¹¹, and avoids the problems of earlier reliance on a proprietary transport layer, as, e.g. in Z39.50¹².



Graphic 1: The basic architecture of an ECNet-broker

A DACO-server can therefore be implemented using cheap standard components, as, e.g., the Apache web server which runs on many hardware platforms (UNIX, Linux, Windows NT, MacOS X, ...).

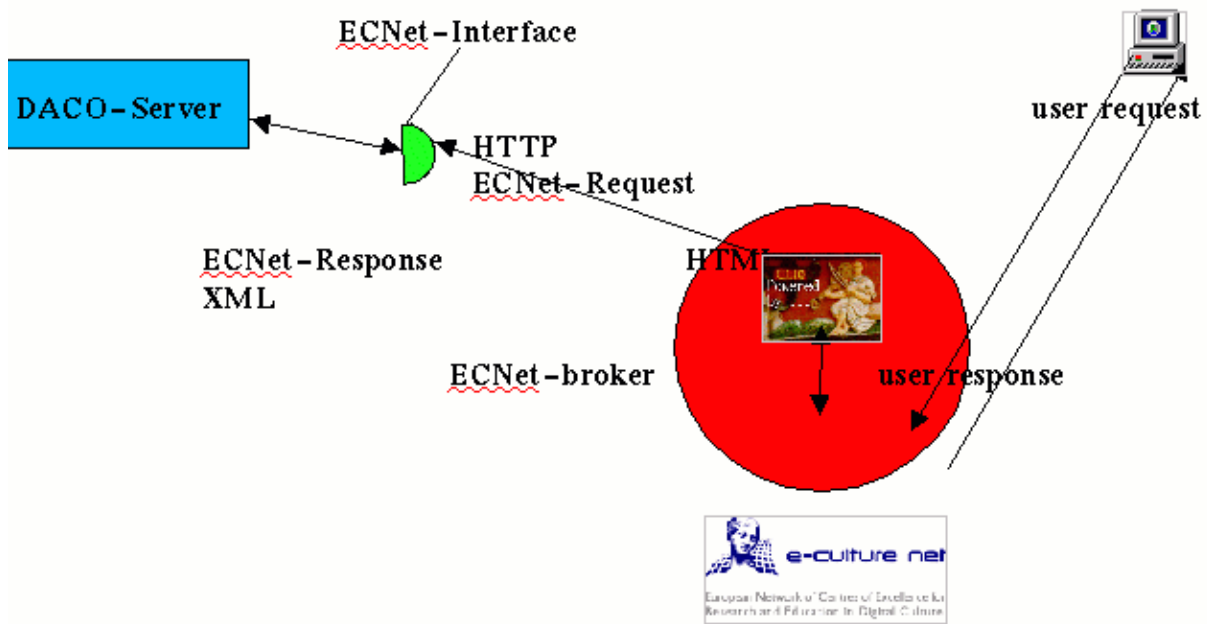
A DACO-server with an ECNet-interface communicates with the ECNet-broker by answering with a XML datastream, UTF-8 encoded, to a *request* (there are currently six request defined, see below) from the ECNet-broker transmitted via HTTP. Each request expects the answer as a message packaged within a specific XML format (see below) that can be processed by the ECNet-broker. A DACO-server has to generate well formed XML as response to a request. The protocol does not expect the responses to be validated against a specific DTD, to allow for greatest flexibility in the processing of responses generated by successive generations of DACO-servers.

¹¹<http://www.w3.org/2000/xp/Group/>

¹²<http://www.loc.gov/z3950/agency/>

The DACO protocol can indeed be seen as just a very shallow uppermost layer of defining the content of responses, which could be generated by many different mechanisms which define other, more general, XML based, message oriented protocols: e.g. by using SOAP (Simple Object Access Protocol), mentioned above, or the Java based Apache XML Project Cocoon¹³, that provides (among others) such an opportunity: it is an XML publishing framework, that interacts with most data sources, including filesystems, RDBMS, LDAP, native XML databases, and network-based data sources. One of its basic services is the generation of XML documents through generators.

This mechanism is shown in graphic 2.



Graphic2: Communication layers employed

¹³<http://cocoon.apache.org/2.0/>

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

4.2 The ECNet- Protocol: Currently Defined Requests

The protocol defines six different requests, which may be submitted via http - general typically using a CGI interface - by a broker (currently the ECNet-demonstrator) to the DACO-Servers. They reply to these requests with a specific XML encoded message, which informs the broker about the type of metadata and data provided by a specific server, and allows the ECNet-broker to build its own user interface from this information.

4.2.1 Basic form of a request.

Every DACO-server has a BaseURL under which it must respond to all DACO requests. These can be submitted by HTTP GET as well as POST.

The basic structure of a request, sent via HTTP GET, is as follows:

After the base BaseURL (e.g.: `http://a.server.somewhere/ecnetinterface`) follows at least one key-value-pair in the usual HTTP GET syntax with the key `ECNetRequest` and the code for the request as value. If additional specifications for a request are defined, they follow as additional key-value-pairs. A DACO-server can ignore the case of all symbolic names used, though the use of capitalisation to increase readability is encouraged. The order of the key-value pairs is arbitrary.

Example for a HTTP GET - `ECNetDeliverObject` - request:

```
http://a.server.somewhere/ecnetinterface?ECNetRequest=ecnetdeliverobjectbyid&ecnetobjectid=123456789
```

Basic form of a response:

A DACO-server replies to all six requests by returning a chunk of XML encoded data. Five of them inform about objects available via this DACO-server, return metadata, that is. The sixth request returns an the URL of an actual DACO. The chunk starts with the standard XML declaration, where we currently assume:

```
<?xml version="1.0" encoding="UTF-8" ?>
```

The remaining content is enclosed in a root element named `ECNetP`.

For all responses, the first two children of the root element are:

`ECNetResponseDate`: of type `UTCdatetime` indicating time and date when the response was sent.

`ECNetRequest`: indicating the protocol request that generated this response.

The third child of the root element is either:

deliverable11b.doc	Report	09 Juli 2003	28/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

an error element that must be used in case of an error or exception or an element with the same name as the value to 'ECNetRequest' of the request which has triggered this response.

Example of a response to a ECNetDeliverObject-request:

```
<?xml version="1.0" encoding="UTF-8"?>
<ECNetP>
<ECNetResponseDate>2002-05-01T19:20:30Z</ECNetResponseDate>
<ECNetRequest ECNetRequest="ECNetDeliverObject"
ECNetObjectID="|kn28-0083ii_87v"
</ECNetRequest>
<ECNetDeliverObject>
<ECNetDACO>http://www.ceec.uni-koeln.de/ceecgi/kleioc/0010KICEEC/exec/pagedmed/%22kn28-0083ii\_87v%22</ECNetDACO>
</ECNetDeliverObject>
</ECNetP>
```

4.2.1 Individual Requests.

4.2.1.1 ECNetDescribeYourself

arguments: none

description: this request will be used to acquire basic information about a DACO-Server

response: ECNetBaseProtocol, ECNetBaseURL, ECNetWrapup

note: A response to this request may contain more than one triplet of the response arguments indicated.


The specification of the base-protocol in the response is included, to provide for the possibility of using non-HTTP transport layers in the future. Currently the value of ECNetBaseProtocol is always "HTTP".

The return of ECNetBaseURL allows constructions, where a DACO-server, addressed via a known ECNetBaseURL, is used to act as logical gateway to a set of other DACO-servers.

ECNetWrapup is a specification about the encoding of the DACOs returned by that server. It consists of the pair: 'encoding1:encoding2'. The first segment (encoding1) relates to the encoding of metadata about the DACOs (to be processed by the broker), the second segment (encoding2) describes the encoding of the DACOs themselves to be displayed to the end-user.

The ECNetWrapup provides for future flexibility of the ECNet-brokers, as this allows for

deliverable11b.doc	Report	09 Juli 2003	29/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

different encoding schemes for the metadata as well as different ways to encode the DACOs to be presented to the end users. (Flash movies, VRML simulations, plain HTML pages.) At present metadata is always XML encoded and the data - the actual DACOs - are expected in HTML.

For example:

```
<?xml version="1.0" encoding="UTF-8"?>
<ECNetP>
<ECNetResponseDate>2002-02-08T12:00:01Z</ECNetResponseDate>
<ECNetRequest ECNetRequest ="ECNetDescribeYourself">http://www.hki.uni-
koeln.de/ecnet </ECNetRequest>
<ECNetDescribeYourself>
<ECNetBaseProtocol>HTTP</ECNetBaseProtocol>
<ECNetBaseURL>http://www.ceec.uni-koeln.de/ceec-
cgi/kleioc/0010KIDACO</baseURL>
<ECNetWrapup>XML:HTML</ECNetWrapup>
</ECNetDescribeYourself>
</ECNetP>
```

4.2.1.2 ECNetExportAccessVenues

arguments: none

description: The "access venues" of a DACO-server are essentially those sets of terms it exports to allow the broker to build an integrated central index. An "access venue" has to provide a mechanism, by which any term belonging to it can be found speedily and can act as a reference to a specific DACO in that server. In most cases, an "access venue" will therefore be implemented as a local index. To allow the broker to decide whether the information in two access venues from different servers can be meaningfully combine, we propose to describe the access venues by a CIDOC-CRM-based semantic description. Such descriptions are not defined at present; a description of the current stage of considerations of CRM use for that purpose is contained in section 4.3 below. In the current demonstrator simple attributes with mnemonic values are used, like the ones used in the examples below.

response: A list of ECNetAccessVenues available within a specific DACO-server and their semantic descriptions. The resulting names of access venues will be used as input to the ECNetExportAccessTerms, ECNetExportFormats and ECNetDeliverObjectByTerm requests described below.

For example:

```
<?xml version="1.0" encoding="UTF-8"?>
<ECNetP>
<ECNetResponseDate>2002-02-08T12:00:01Z</ECNetResponseDate>
```

deliverable11b.doc	Report	09 Juli 2003	30/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

```

<ECNetRequest ECNetRequest
="ECNetExportAccessVenues">http://a.server.somewhere</ECNetRequest>
<ECNetExportAccessVenues>
<ECNetAccessVenues semMean="people">access venue 1</ECNetAccessVenues>
<ECNetAccessVenues semMean="content words">access venue
2</ECNetAccessVenues>
<ECNetAccessVenues semMean="place names">access venue
3</ECNetAccessVenues>
<ECNetAccessVenues semMean="title words">access venue 4</ECNetAccessVenues>
<ECNetAccessVenues semMean="institutions">access venue
5</ECNetAccessVenues>
</ECNetExportAccessVenues>
</ECNetP>

```

4.2.1.3 ECNetExportAccessTerms

arguments: ECNetAccessVenue, ECNetBaseTerm, ECNetMaxTerm, ECNetTermSelection

description: This request is used to extract all - or a subset of the - ECNetTerms of a DACO-server which are available via a specific ECNetAccessVenue.


As there can be a great number of such terms it is necessary to provide a mechanism for splitting the response into segments to allow the broker to access the terms in packages of a size which fits into the local capacities and / or the bandwidth of a specific connection. The maximum allowable size for such a segment is given by the value of the argument ECNetMaxTerms and the first term required is specified as value of ECNetBaseTerm.

Additionally the value of the argument ECNetTermSelection {begin, end, middle} allows to choose a specific sector of the set of terms ordered in the collating sequence of the DACO-server. If begin is specified the ECNetBaseTerm and at most ECNetMaxTerms-1 additional terms *following* that term in the collating sequence are returned in the response; if end is specified the ECNetBaseTerm and at most ECNetMaxTerms-1 additional terms *preceding* that term in the collating sequence. The specification of middle returns a set of terms, which consists of the ECNetBaseTerm and sets of (ECNetMaxTerms / 2) -1 preceding and following that term.

If ECNetBaseTerm does not exist in ECNetAccessVenue it is replaced by the next greater term in the collating sequence; if none such exists, the preceding one is returned. If ECNetBaseTerm is missing, represents the null string or white space, the first term in the collating sequence is chosen. To access the last term in an access venue, a byte consisting of only ones - binary '11111111' - should be transmitted.

The response contains, besides the list of terms requested, the additional arguments ECNetPreviousTerm and ECNetNextTerm. The contain the term immediately preceding the first and following the last term of the segment of the access venue transmitted, allowing for successive requests retrieving the whole access venue in successive segments.

deliverable11b.doc	Report	09 Juli 2003	31/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

response: A list of ECNetTerms contained in the access venue, ECNetPreviousTerm and ECNetNextTerm

For example:

```
<?xml version="1.0" encoding="UTF-8"?>
<ECNetP>
<ECNetResponseDate>2002-02-08T12:00:01Z</ECNetResponseDate>
<ECNetRequest ECNetRequest ="ECNetExportAccessTerms"
ECNetExportAccessVenues="ECNetAccessVenues 4"
ECNetBaseTerm="ECNetBaseTerm 100" ECNetMaxTerm="100"
ECNetTermSelection="begin">http://a.server.somewhere</ECNetRequest>
<ECNetExportAccessTerms>
<ECNetTerm>term 100</ECNetTerm>
<ECNetTerm>term ...</ECNetTerm>
<ECNetTerm>term 200</ECNetTerm>
<ECNetPreviousTerm>term 100</ECNetPreviousTerm>
<ECNetNextTerm>term 201</ECNetNextTerm>
</ECNetExportAccessTerms>
</ECNetP>
```

4.2.1.4 ECNetDeliverObjectId

arguments: ECNetAccessVenue, ECNetTerm, ECNetFormat

description: The response to this request is a list containing all identifiers of objects which can be provided by a DACO-server as DACOs, which are connected with a specific ECNetTerm in a specific ECNetAccessVenue. A server is expected to administer such identifiers persistently; i.e.: even if a number of years passes between exporting an identifier to the outside world and a request to access the DACO addressed by it, it has to reference the same object as before.

It is expected that future implementations of DACO-servers will return descriptions of the objects identified using the CIDOC-CRM-based mechanism described below in section 4.3.

response: A list of DACO - identifiers.

For example:

```
<?xml version="1.0" encoding="UTF-8"?>
<ECNetP>
<ECNetResponseDate>2002-02-08T12:00:01Z</ECNetResponseDate>
<ECNetRequest ECNetRequest ="ECNetDeliverObjectId" ECNetAccessVenue="access
venue 4" ECNetTerm="term 2" ECNetFormat="format 1"
>http://a.server.somewhere</ECNetRequest>
<ECNetDeliverObjectByTerm>
<ECNetRef>DACO identifier 1
```

deliverable11b.doc	Report	09 Juli 2003	32/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

```

<optional-semantic-description/>
</ECNetRef>
<ECNetRef>DACO identifier 2
  <optional-semantic-description/>
</ECNetRef>
<ECNetRef>DACO identifier 3
  <optional-semantic-description/>
</ECNetRef>
<ECNetRef>DACO identifier 4
  <optional-semantic-description/>
</ECNetRef>
<ECNetRef>DACO identifier 5
  <optional-semantic-description/>
</ECNetRef>
</ECNetDeliverObjectByTerm>
</ECNetP>

```

4.2.1.5 ECNetDeliverObject

arguments: ECNetObjectID, accompanied by an optional format specification ECNetFormat.

description: This request, finally, accesses the URL of the actual DACO in the type of wrapup supported by this server / format.

response: URL of the DACO in the wrapup implied by the format selected. If no format is specified, the server will use one by default, which, however, may change without any broker being identified.

For example:

```

<?xml version="1.0" encoding="UTF-8"?>
<ECNetP>
<ECNetResponseDate>2002-05-01T19:20:30Z</ECNetResponseDate>
<ECNetRequest ECNetRequest="ECNetDeliverObject"
ECNetObjectID="|kn28-0083ii_87v"
</ECNetRequest>
<ECNetDeliverObject>
<ECNetDACO>http://www.ceec.uni-
koeln.de/ceecgi/kleioc/0010KICEEC/exec/pagemed/%22kn28-
0083ii_87v%22"</ECNETDACO>
</ECNetDeliverObject>
</ECNetP>

```

deliverable11b.doc	Report	09 Juli 2003	33/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

4.2.1.6 ECNetExportFormats

argument: ECNetAccessVenue

description: The response to this request is a list of all DACO-formats supported via the access venue in the argument, accompanied by semantic markup (see below), describing the properties of the format.

response: A list of DACO formats and their semantic meaning. (See also below, section 4.3.) The result will be used as argument for ECNetDeliverObject. Some or all of the formats specified may also specify which wrapup they are using with ECNetWrapup; this may deviate from the default wrapup, a server signals by the same element returned in response to the ECNetDescribeYourself request.

For example:

```
<?xml version="1.0" encoding="UTF-8"?>
<ECNetP>
<ECNetResponseDate>2002-02-08T12:00:01Z</ECNetResponseDate>
<ECNetRequest ECNetRequest ="ECNetExportFormats"
  ECNetAccessVenue="access venue 3">http://a.server.somewhere </ECNetRequest>
<ECNetExportFormats>
<ECNetForm semMean="thumbnail">DACO format 1
  <ECNetWrapup>XML:Flash</ECNetWrapup>
</ECNetForm>
<ECNetForm semMean="maximum resolution">ECNet format 2 </ECNetForm>
<ECNetForm semMean="TOC">ECNet format 3</ECNetForm>
</ECNetExportFormats>
</ECNetP>
```

deliverable11b.doc	Report	09 Juli 2003	34/37
--------------------	--------	--------------	-------

4.3. DACOs in the Semantic Web

The protocol presented above hints in a number of cases at the necessity, to include semantic descriptions of the type of data contained within the various servers and their components. To test the protocol, we have simply used the preliminary attribute "semMean", which uses a arbitrary subset of intuitive categories, as e.g. "people", "content vocabulary", "place name" to categorise the information contained within the access venues offered by a DACO server. This preliminary attribute will be replaced in later versions.

Two strategies are currently explored:

1) It would be conceivable to use an ontology language like OWL¹⁴ to describe the servers and components by a purpose built ontology. Within ECultureNet various work packages have touched upon the question of ontology; and the SEMKOS project is directly derived from considerations which have (also) been planned within EcultureNet. Employing any of these ontologies would be a logical extension.


2) At the moment, however, most of our considerations deal with the CIDOC Content Reference Model¹⁵ - CRM. We assume that an XML binding of the abstract CRM would:

Fit technically very well into the structure of the DACO protocol proposed here. (And be easily handled by the thesaurus administration system which is part of the software with which the broker has been realised, as mentioned in section 5 below.)

Provide sufficient flexibility for very detailed descriptions for those who want to provide them. Make it easily possible, however, to extract only those categories which are needed for the degree of conceptual differentiation a broker wants to achieve. (A very high level broker would presumably combine all ECNetAccessVenues which declare themselves to be related to E21 (person), while a more specialised one would only start at that level to decide which references to persons are appropriate for a very narrowly defined search tool.)

¹⁴<http://www.w3.org/2001/sw/WebOnt/>

¹⁵<http://cidoc.ics.forth.gr/>

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

5. The Demonstrator

To proof the viability of the concept of DACOs as base for a network of virtually integrated, but basically very heterogeneous servers of cultural heritage material, a demonstrator has been built. It is currently still available under <http://gilgamesch2.hki.uni-koeln.de/ECNet>, will during the summer be transferred to <http://lehre.hki.uni-koeln.de/ECNet>, however.

The server will be expanded after the end of the project to integrate at least the following resources:

The medieval manuscripts from the CEEC project quoted above (currently: <http://www.ceec.uni-koeln.de>; active in the demonstrator).

The medieval images of the Institut für Mittelalterliche Realienkunde of the Austrian Academy of Sciences (currently: <http://www.imareal.oeaw.ac.at/realonline/>)

The servers for early modern doctoral theses of law provided by the Max-Planck-Institut für Rechtsgeschichte, Frankfurt (currently: <http://dlib-diss.mpier.mpg.de/>; active in the demonstrator)

The server of the same institute offering a collection of printed books on legal history (currently: <http://dlib-pr.mpier.mpg.de/>; active in the demonstrator)

A server currently created within the context of the demonstrator project for the library of the Predigerseminar at Wittenberg. (Not yet accessible independently.)

A server currently being created within the context of the demonstrator project for the "Personenstandsarchiv Brühl" offering archival material for genealogy.

Possibly the archival server offered by the Stadtarchiv Duderstadt with late medieval / early modern documents (currently: <http://www.archive.geschichte.mpg.de/duderstadt/dud-e.htm>.)

Other institutions may be included.

Between them, the seven servers quoted above represent images (2), manuscripts (1) early prints(3, 5). modern prints (4) and archival material of different types (6, 7).

To prepare for integration of these servers offering actual digital sources finally, the data contained in the SUMS server implemented at MMI at Maastricht as part of the ECultureNet Project, have been transferred to Cologne and re-implemented (provisionally, not persistently!, available under <http://gilgamesch2.hki.uni-koeln.de/SUMS>; active in the demonstrator).

This server is offering metadata: it will be used as experimental host to study the possibility to integrate data coming from the heterogeneous primary servers via the technology proposed.

To study the possibility of teaching the methodologies required for the creation of such servers, the implementation of the servers 5 and 6 above have been integrated as teaching project into the regular teaching at the University at Cologne.

deliverable11b.doc	Report	09 Juli 2003	36/37
--------------------	--------	--------------	-------

 e-culture net	IST-2001-37491	WP6	D11b
--	----------------	-----	------

The demonstrator will, at the same time, include an implementation of a PURN16 name resolution service, currently being discussed with the "Die Deutsche Bibliothek" (roughly: The German National Library). This seems to be a very worthwhile approach to the the problem of finding digital identifiers with a reasonable longevity. One might considers this to be solved by a simple application of current European projects on the provision of digital identifiers¹⁷. In our opinion, this can not be done mechanically, however, in the Cultural Heritage domain, as the question of granularity of identifiers raises specific problems here. (As demonstrated in section 4.1 above with regard to the need to address individual pages.)

The software being implemented for the broker has currently stubs for the usage of thesauruses, which can provide thesaurus based "translation" from the language of a database into any intended object language. This works by simple token replacement, however, no linguistic analysis is performed. So far the technology would, therefore, only work with databases with controlled vocabularies.

¹⁶<http://www.ietf.org/html.charters/urn-charter.html>

¹⁷<http://www.doi.org/>

deliverable11b.doc	Report	09 Juli 2003	37/37
--------------------	--------	--------------	-------