WP2 Note on an ECHO Ontology

Peter Wittenburg
February 2, 2004

Essential part of the DORA[1] ECHO portal which was presented several times at meetings and discussed in detail with nearly all ECHO participants is the integration of ontological knowledge from several domains. This paper wants to document the knowledge components, their extraction processes and their relations. The resulting components will be available at the end of the ECHO project in well-documented formats.

This document can be seen as supplementary to the one that describes the DORA infrastructure, the selections made with respect to the semantics and the mapping choices. From several projects and initiatives we know that the mapping choices can be questioned, since two persons will not agree. But this is exactly the reason why we rely on practical ontologies that can easily be changed and amended by other persons such that the chosen mappings better reflect the intentions.

Despite many difficulties we can state that we were able to establish an ECHO ontology that covers the offered semantics of the participating disciplines and that is now base of the DORA machinery.

1. Provided Components

The following components were provided by the participants and external sources:

1. **Metadata Descriptions**
   XML repositories covering the metadata descriptions of the various data providers often without any form of validation. These were partially associated with
   a. the list of the metadata vocabularies of which some referred to Dublin Core concepts, others to proper definitions and others to verbal explanations[2];
   b. formal syntax descriptions (only in three cases).

2. **Content Thesauri**
   Two metadata sets are making use of thesauri to describe the content of the object.
   a. The RMV uses the OMV thesaurus that is derived from the AAT thesaurus[3].
   b. The Fotothek uses the IconClass[4] thesaurus which was available as an interactive CDROM.

[1] Digital Open Resource Area: see WP2-TR16-2003; web-site to come
[2] Metadata definitions will always include some tolerance in the usage due to the different interpretations of the definitions of the semantic scope. Non-existing definitions or unclear definitions lead to wider tolerances in usage of course.
[3] It should be noted here that Brik de Zwart supported the ECHO work by not only providing the only real OAI implementation, but also providing the OMV thesaurus in a structured form. Thanks a lot!!
[4] IconClass was bought from the KNAW Amsterdam.

c. Other metadata sets are using either unconstrained keyword elements or use a limited number of narrowly defined elements.

3. Geographic Information

a. The RMV is using a geographic thesaurus.
b. Other metadata sets are using either unconstrained elements or a limited number of more clearly and constraint elements such as continent and country.
c. It was noted that language and society names in many cases include geographical information.

2. Generated Components - Overview

From this basic information a number of essential components were extracted. Most of them are in XML, others are in a structured form that is easy to process, but will be transformed to XML until the end of ECHO. Yet RDF was not used to represent knowledge. Concept definitions can be done in XML and this is the way that is used by ISO groups such as TC37/SC4. For the mapping file that contains assertions about concepts RDF is the most suitable format. However, since there is no complete logic, since we have many fuzzy mapping relations and since we lack appropriate standard inference engines there is no immediate need to formulate the relations as RDF assertions. The mappings are embedded in XML so that they can be easily transformed to RDF.

1. Validated Metadata Sets

The metadata information was transformed into validated and machine readable formats. Structure and character encoding was standardized to XML and UNICODE.

2. ECHO Concepts

This XML file consists of all elements from the various metadata sets that were selected to be used in DORA, i.e. that are not too specialistic.
The current version is: ***echo-term-v6.xml***

3. ECHO Mappings

This XML file consists of an exhaustive mapping between all elements found in the concepts file. It is guided by the wish to do the access from different views.
The current version is: ***echo-mapping-v5.xml***

4. OVM-Geographic Thesaurus

This file contains the geographic thesaurus as used within the RMV descriptions. Where possible the OVM geographic thesaurus points to comparable entries in the MPI geographic thesaurus.
The current version is: ***ovm-geo-thesaurus-v3.xml***

5. MPI-Geographic Thesaurus

An analysis was carried out on all geographically oriented fields on all metadata records of all data providers except RMV to get a list of geographic concepts that are actually used. From these a "complete" geographic thesaurus[5] was created. Where possible the MPI geographic thesaurus points to comparable entries in the OVM geographical thesaurus.
The current version is: ***mpi-geo-thesaurus-v4.xml***

6. OVM Category Thesaurus

This thesaurus contains all values that are used in the RMV content description field and they are ordered in a hierarchical way. This thesaurus is based on the AAT thesaurus.
The current version is: ***ovm-category-thesaurus-v2.xml***

7. Iconclass Category Thesaurus

This thesaurus contains all values that are used in the Fotothek content description field (Iconography) and they are ordered in a hierarchical way.

[5] The OVM geographical thesaurus is not complete and not appropriately structured. Different types of concepts appear at a certain depth. Therefore, we could not use it as master thesaurus. A conversion would have required manual work.

The current version is: ***iconclass-category-thesaurus-v2.xml***

8. **IconClass-to-OVM Mapping**
   This file contains a mapping between IconClass and OVM nodes where this is semantically feasible. It was clear that only a one-directional mapping would serve the needs.
   The current version is: ***iconclass2ovm-mapping-v3.xml***

9. **OVM-to-IconClass Mapping**
   This file contains a mapping between IconCLass and OVM nodes where this is semantically feasible. It was clear that only a one-directional mapping would serve the needs.
   The current version is: ***ovm2iconclass-mapping-v3.xml***

10. **MPI Content List**
    An analysis was made on all content type fields that can be found in all metadata records of all data providers except RMV and Fotothek. A mapping file was created that links these descriptors with those to be found in the OVM and the IconClass thesauri.
    The current version is: ***IMDI2iconclass-and-ovm-v1.xml***

11. **Other Components**
    There are a few other files that are used to facilitate the DORA searching machinery, but they don't contain essential knowledge representations.

3. Components in Detail

In this chapter we want to discuss some aspects in more detail.

3.1 ECHO Concepts

All concepts that were decided to be used for the DORA interface from the different metadata sets. So we choose a setup that seems now to be followed by many groups representing knowledge. Concept definitions are separated from any relational information except if a sub/superclass relation is an evident part of the concept definition. This gives everyone the possibility to relate concepts in the own way and nothing is prescribed. In ISO TC37/SC4 it is argued that equality and sub/superclass relations can be part of the definition of a concept. This is very dependent on the scope of the domain considered. According to the ISO 11179 model the domain description has to be part of the concept definition.

We have taken a strict role to separate definition and relation, since we don't have yet a sufficiently detailed view on the semantic scope of all terms. Each concept found is defined by a number of attributes which are indicated in the following XML fragment.

```
<terms>
  <term>
  <termID> 001 </termID>                                          unique identifier
    <term-name> title </term-name>                                concept name
    <xpath> dc.title </xpath>                                     how to find it
    <domain-name> DublinCore </domain-name>                       ECHO domain name
    <sub-domain-name> </sub-domain-name>                          ECHO sub-domain
    <description> name given to resource </description>           a prose definition
    <dedications>
        <fra> titre </fra>                                        French dedication
        <ger> Titel </ger>                                        German dedication
        <ita> titolo </ita>                                       Italian dedication
        <swe> titel </swe>                                        Swedish dedication
        <dut> titel </dut>                                        Dutch dedication
    </dedications>
  </term>
  <term>
  ....
  ....
```

```
  </term
</terms>
```

If there is enough time left in the ECHO project we will transform this into an ISO 11179, ISO 12620 compliant XML form so that it can be put openly on the web and used by others. However, in ECHO we will not introduce relational information into the document and will not eliminate equivalent concepts (synonyms etc). Mainly since the machinery is now developed such that it will use this normalized type of representation.

The file was generated only to a small extent automatically. All translations were created manually.

3.2 ECHO Mappings

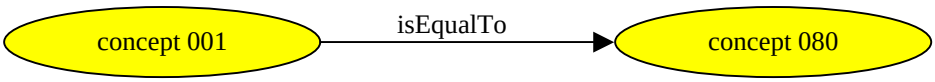
The mappings are done according to the Technical Report WP2-TR16-2003 about Mapping. They exist of references to the concept file, a relation type and a matching factor that currently is not used. Before using this information we first have to get more experience. The intention is to indicate the quality of the mapping, i.e. the amount of semantic overlap between the related concepts. The following XML fragment indicates how the file is structured. For easiness of reading a supplementary file was created that contains all concept information. However, this cannot be the basis for the DORA machinery, since the information would be stored at two places which is not acceptable from maintenance reasons.

```
<mappings>
  <mapping>
    <termID>001</termID>                          first concept reference
    <termID>080</termID>                          second concept reference
    <relation-type>isEqualTo</relation-type>      relation type
    <match-factor>1</match-factor>                matching factor
  </mapping>
  <mapping>
    <termID>002</termID>
    <termID>027</termID>
    <relation-type>mapsTo</relation-type>
    <match-factor>1</match-factor>
  </mapping>
  <mapping>
  ....
  ....
  </mapping>
</mappings>
```

It can easily be seen that the structure can be easily transformed into an RDF assertion. Let us take the example from the first fragment.

```
<termID>001</termID>
<termID>080</termID>
<relation-type>isEqualTo</relation-type>
```

This XML substructure would translate to the following RDF assertion.



The following semantic relations are used in the mapping file:

i**sEqualTo**	the two related terms are semantically equivalent Example: DC:Date isEqualTo IMDI:Date

isSubclassOf	the first concept is a hyperonym of the second one Example: DC:Creator is SubclassOf IMDI:Particpant
isSuperclassOf	the first concept is a hyponym of the second one Example: IMDI:Participant isSuperclassOf DC:Creator
MapsTo	the first concept is related with the second one this relation was chosen in many cases, but the semantic overlap cannot be specified in terms that can be exploited by strict logic; it represents a kind of fuzzy matching, i.e. only the move to some granular feature space would allow us to make the relation more specific and precise. Example: DC:Creator mapsTo RomeMaps:Editor

All relations were created based on manual inspection of the definitions and after having talked with the sub-domain experts. Currently, we start analyzing the usage of the fields which may lead to changes.

3.3 OVM-Geographic Thesaurus

This thesaurus was extracted semi-automatically from a web-representation. For reasons of simplicity we indicate the thesaurus in table form. It has three entries: (1) the OVM abbreviation that is used in the metadata records; (2) the geographic name used by OVM in Dutch and (3) a reference to the appropriate node in the so-called MPI geographic thesaurus.

OVM Abbreviation **OVM.AAA**	**OVM Geo-Name** **Geografische herkomst**	**MPI Geo-Name** **reference to mpi-geo-thesaurus**
OVM.AAA.AAA	Afrika	Africa
OVM.AAA.AAA.AAA	Afrikaanse eilanden	Island nations
OVM.AAA.AAA.AAA.AAA	Afrikaanse eilanden- Oost	
OVM.AAA.AAA.AAA.AAA.AAA	Comoren	Comoros
OVM.AAA.AAA.AAA.AAA.AAB	Madagascar	Madagascar
OVM.AAA.AAA.AAA.AAA.AAB.AAA	Antananarivo	
OVM.AAA.AAA.AAA.AAA.AAB.AAB	Betafo	
OVM.AAA.AAA.AAA.AAA.AAB.AAC	Nosy Bé	
OVM.AAA.AAA.AAA.AAA.AAC	Mauritius	Mauritius
OVM.AAA.AAA.AAA.AAA.AAD	Seychellen	Seychelles
OVM.AAA.AAA.AAA.AAB	Afrikaanse eilanden- West	
OVM.AAA.AAA.AAA.AAB.AAA	Canarische eilanden	
OVM.AAA.AAA.AAA.AAB.AAA.AAA	Tenerife	
OVM.AAA.AAA.AAA.AAB.AAB	St. Helena	
OVM.AAA.AAA.AAB	Centraal-Afrika	Central Africa
OVM.AAA.AAA.AAB.AAA	Angola	Angola
OVM.AAA.AAA.AAB.AAA.AAA	Angola:regionaal	
OVM.AAA.AAA.AAB.AAA.AAA.AAA	Angola- Noordwest	
OVM.AAA.AAA.AAB.AAA.AAB	Bengo	
OVM.AAA.AAA.AAB.AAA.AAC	Benguela	
OVM.AAA.AAA.AAB.AAA.AAC.AAA	Catumbela	
OVM.AAA.AAA.AAB.AAA.AAD	Bié	
OVM.AAA.AAA.AAB.AAA.AAD.AAA	Chinguar	
OVM.AAA.AAA.AAB.AAA.AAE	Cabinda	
OVM.AAA.AAA.AAB.AAA.AAE.AAA	Futila	
OVM.AAA.AAA.AAB.AAA.AAE.AAB	Loango	
OVM.AAA.AAA.AAB.AAA.AAF	Cuamato	
OVM.AAA.AAA.AAB.AAA.AAF.AAA	Forte Rocadas	
OVM.AAA.AAA.AAB.AAA.AAG	Cuanza	

The OVM geographic thesaurus does not have a canonical hierarchical structure that could look like:

```
<continent>
  <sub-continent>
    <country>
      <region>
        <place>
          ...
```

It leaves out nodes where nothing suitable could be filled in, i.e. countries can appear at different levels of depth. This makes it difficult to automatically transform this thesaurus into a canonical structure and it is too large to do a manual transformation within ECHO. Therefore, the resulting XML structure can only use arbitrary <struct> tags. This does not harm searching, since the nodes represent super-classes that can be exploited. The link to a node in the MPI geographic thesaurus can also be exploited.

The figure indicates the partial match between the two geographic thesauri. Partial matching in the geographical domain means in the far most cases that complete sub-trees can be matched. Only in few cases at the regional level the classifications may be unclear.

3.4 MPI-Geographic Thesaurus

Due to the non-canonical form of the OVM-geographic thesaurus it was decided to add another canonical thesaurus and enter all geographically oriented names that can be found in one of the metadata records (except OVM) into this one. An analysis of all other metadata records revealed that there were not too many different names. For example in the large Fotothek repository only a few names are re-occurring. Also in the large language domain mostly the categorization is done systematically until the country level. Some used the region element, but in total there were not too many different ones. So it was an easy job to add all names into a canonical structure that was extracted semi-automatically from an official and reliable web-site.

```
<continents>
  <continent>
    <cnt-name> Africa" </cnt-name>
    <dedications>
      <ger> Afrika </ger>
    </dedications>
    <ovm-code> OVM.AAA.AAA </ovm-code>
    <sub-continents>
      <sub-continent>
        <sc-name> North Africa </sc-name>
        <ovm-code> OVM.AAA.AAA.AAC" </ovm-code>
        <countries>
            <country>
              <cou-name> Algeria </cou-name>
              <ovm-code> OVM.AAA.AAA.AAC.AAA <ovm-code>
            </country>
            <country>
              <cou-name> Egypt </cou-name>
                <dedications>
                  <ger>Ägypten </ger>
                </dedications>
              <ovm-code> OVM.AAA.AAA.AAC.AAB </ovm-code>
            <country>
              <cou-name> Libya </cou-name>
              <ovm-code> OVM.AAA.AAA.AAC.AAC </ovm-code>
            </country>
            <country>
              <cou-name> Morocco </cou-name>
              <ovm-code> OVM.AAA.AAA.AAC.AAD </ovm-code>
            </country>
            <country>
              <cou-name> Sudan </cou-name>
              <ovm-code> OVM.AAA.AAA.AAC.AAF </ovm-code>
            </country>
            <country>
              <cou-name> Tunisia </cou-name>
              <ovm-code> OVM.AAA.AAA.AAC.AAG.AAX </ovm-code>
              <places>
                <place>
                  <pl-name> Tunis </pl-name>
                  <ovm-code> OVM.AAA.AAA.AAC.AAG.AAY </ovm-code>
```

```
                    </place>
                    ...
                  </places>
                <country>
                  ...
                </country>
              </countries>
              ...
            </sub-continent>
            ...
          </sub-continents>
        </continent>
        ...
      <continents>
```

Yet the links in the OVM geographical thesaurus are not XML path expressions. This has to be generated to make it a fully XML compliant version that can easily be re-used by others. For the DORA machinery it is not of relevance since optimal index structures are generated anyhow for fast processing.

Only for some entries language dedications are specified. It would be too much work to create names in the different languages for all entries except that we will find reliable multilingual geographic lexicons.

3.5 OVM Category Thesaurus

The categories and the Dutch labels of this thesaurus were extracted semi-automatically from a web-representation. For reasons of simplicity we indicate the thesaurus in table form. It has three entries: (1) the OVM abbreviation that is used in the metadata records; (2) the English category naming and (3) the original Dutch category naming.

OVM indeling/categories	English	Dutch
OVM.AAC	OVM Category	OVM Categorie
OVM.AAC.AAA	"hunting, fishery, food gathering"	"jacht, visserij, voedselgaring"
OVM.AAC.AAA.AAA	hunting	jacht
OVM.AAC.AAA.AAA.AAA	hunting without tools	jacht zonder handwerktuigen
OVM.AAC.AAA.AAA.AAB	hunting with lures	jacht met lokmiddelen
OVM.AAC.AAA.AAA.AAC	hunting with traps and snares	jacht met vallen en strikken
OVM.AAC.AAA.AAA.AAE	hunting with weapons	jacht met wapens (inclusief accessoires)
OVM.AAC.AAA.AAA.AAE.AAA	hunting with fist weapons	jacht met handwapens
OVM.AAC.AAA.AAA.AAE.AAB	hunting with projectiles	jacht met projectielen
OVM.AAC.AAA.AAB	fishery	visserij
OVM.AAC.AAA.AAB.AAA	fishery without tools	visserij zonder handwerktuigen
OVM.AAC.AAA.AAB.AAB	fishery with lures	visserij met lokmiddelen
OVM.AAC.AAA.AAB.AAC	fishery with traps and nets	visserij met vallen en netten
OVM.AAC.AAA.AAB.AAE	fishery with weapons	visserij met wapens (inclusief accessoires)
OVM.AAC.AAA.AAB.AAE.AAA	fishery with fist weapons	visserij met handwapens
OVM.AAC.AAA.AAB.AAE.AAB	fishery with projectiles	visserij met projectielen
OVM.AAC.AAA.AAC	gathering food	voedsel verzamelen
OVM.AAC.AAB	"weapons, warfare, war"	"wapens, strijd en oorlog"
OVM.AAC.AAB.AAA	fist weapons and accessories	handwapens en accessoires

Since the IconClass thesaurus uses English labeling and since at the user interface at least English labeling should be used all entries were translated into English labels as well. It would be too much work within ECHO to generate other language dedications. This should be done semi-automatically by using appropriate technology.

An XML version is being created currently which will be made public at the end of the ECHO project.

3.6 Iconclass Category Thesaurus

The categories of this thesaurus were extracted semi-automatically from a CDROM. Again, for reasons of simplicity we indicate the thesaurus in table form. It has two entries: (1) the IC abbreviation that is used in the metadata records and (2) the English category labeling.

1 Religion and Magic

10	(symbolic) representations ~ creation, cosmos, cosmogony, universe, and life (in the broadest sense)
11	Christian religion
11A	Deity, God (in general) ~ Christian religion
11A1	God the Creator
11A11	God measuring the Universe (with compasses)
11A2	Divine Nature
11A21	Divinity, 'Divinità' (Ripa)
11A22	symbols ~ Divine Nature
11A221	circle symbolizing God's perfectness
11A23	God's perfections
11A3	God's wrath
11A31	'Flagello di Dio' (Ripa)
11B	the Holy Trinity, 'Trinitas coelestis'; Father, Son and Holy Ghost ~ Christian religion
11B1	Trinity represented by tripartite symbols
11B11	symbols of the Trinity ~ circular and/or triangular forms or arrangements
11B114	three animals, geometrically arranged within a circle or triangle
11B12	Trinity represented as a person with three heads
11B13	Trinity represented by three animals sharing one head
11B14	other tripartite symbols of the Trinity
11B2	Trinity in which each of the Persons (God, Christ, Holy Ghost) is represented either by an object or by an animal
11B21	representation of the Trinity: hand (Father), lamb (Son), and dove (Holy Ghost)
11B22	representation of the Trinity: hand, cross and dove
11B23	representation of the Trinity: hand, chalice and dove
11B3	Holy Trinity in which one, two or all figures are represented in human shape
11B31	Trinity as three persons
11B32	Trinity in which God the Father and Christ are represented as persons, the Holy Ghost as dove
11B321	God the Father seated, holding the youthful Christ (Emmanuel) in his lap
11B322	God the Father and Christ enthroned
11B3231	God the Father holding the crucifix, 'Gnadenstuhl', Mercy-Seat, Throne of Grace
11B3232	God the Father standing or seated, holding the body of Christ, 'Pitié-de-Notre-
11B33	representations of the Trinity

The extraction of a clean, complete and well-structured file was not trivial and partially manual work had to be carried out. The thesaurus had to be complete since many mappings were found between OVM and IconClass nodes.

An XML version is being created currently which will be made public at the end of the ECHO project, if there are no IPR restrictions involved. This has to be discussed with KNAW.

3.7 IconClass-to-OVM Mapping

This mapping file is a result of a careful one-directional comparison. This comparison could only be done manually, since any formal comparison based on pure linguistic knowledge could lead to misleading results. The context had to be considered to do the right interpretations.

```
<mappings>
  <mapping>
    <ic-code> 1 </ic-code>
    <ic-label> Religion and Magic </ic-label>
    <ovm-mapping>
      <ovm-code> OVM.AAC.AAN.AAC </ovm-code>
      <ovm-label> altars, sanctuaries and their interior decoration and furniture </ovm-label>
    </ovm-mapping>
    <ovm-mapping>
      <ovm-code> OVM.AAC.AAN.AAD </ovm-code>
      <ovm-label> sacrifices </ovm-label>
    </ovm-mapping
    <ovm-mapping>
      <ovm-code> OVM.AAC.AAN.AAF </ovm-code>
      <ovm-label> ritual appliances </ovm-label>
    </ovm-mapping>
    <ovm-mapping>
      <ovm-code> OVM.AAC.AAN.AAG </ovm-code>
      <ovm-label> symbols of religious status </ovm-label>
    </ovm-mapping>
  </mapping>
  <mapping>
    <ic-code> 10 </ic-code>
    <ic-label> Religion and Magic </ic-label>
```

```
    <ovm-mapping>
      <ovm-code> OVM.AAC.AAN.AAC </ovm-code>
      <ovm-label> (symbolic) representations, creation, cosmos, cosmogony, universe, life </ovm-label>
    </ovm-mapping>
  </mapping>
  <mapping>
    <ic-code> 13 </ic-code>
    <ic-label> magic, supernaturalism, occultism </ic-label>
    <ovm-mapping>
      <ovm-code> OVM.AAC.AAN.AAB </ovm-code>
      <ovm-label> cult objects and other holy objects </ovm-label>
    </ovm-mapping>
  </mapping>
  <mapping>
    <ic-code> 13C3 </ic-code>
    <ic-label> magic objects, apotropaia </ic-label>
    <ovm-mapping>
      <ovm-code> OVM.AAC.AAN.AAE </ovm-code>
      <ovm-label> magical protection and defence </ovm-label>
    </ovm-mapping>
  </mapping>
  ...
</mappings
```

In contrast to the geographic mapping described above a mapping between two nodes often does not mean that complete sub-trees would map. For ECHO it would be too much to do a complete analysis. This has to be left over to other projects.

As indicated above there will be much debate about particular mappings. Therefore it is even more true that individuals or groups should be able to influence inferencing by being able to modify the mappings easily. This requires open definitions as they are envisaged for example in ISOTC37/SC4 based on ISO 11179 and ISO 12620 and suitable tools, but in the area of cultural heritage we are far away from such a situation.

3.8 OVM-to-IconClass Mapping

This mapping file is complementary to the one-directional comparison described above. For the same reasons also this comparison could only be done manually.

```
<mappings>
  <mapping>
    <ovm-code> OVM.AAC.AAA.AAA.AAA </ovm-code>
    <ic-label> hunting without tools </ic-label>
    <ovm-mapping>
      <ovm-code> 43C111 </ovm-code>
      <ovm-label> game, hunted animals, hunt, bird hunting </ovm-label>
    </ovm-mapping>
  </mapping>
  <mapping>
    <ovm-code> OVM.AAC.AAA.AAA.AAB </ovm-code>
    <ic-label> hunting with lures </ic-label>
    <ovm-mapping>
      <ovm-code> 43C132 </ovm-code>
      <ovm-label> duck decoy </ovm-label>
    </ovm-mapping>
    <ovm-mapping>
      <ovm-code> 43C1(+43)</ovm-code>
      <ovm-label> lures (hunting)</ovm-label>
    </ovm-mapping>
  </mapping>
  <mapping>
    <ovm-code> OVM.AAC.AAA.AAA.AAC </ovm-code>
    <ic-label> hunting with traps and snares </ic-label>
```

```
      <ovm-mapping>
        <ovm-code> 43C131</ovm-code>
        <ovm-label> finch trap, finchery </ovm-label>
      </ovm-mapping>
    </mapping>
    ...
  </mappings>
```

For some comments see above.

3.7 MPI Content List

To achieve content mappings were possible it is important to try to map all content describing elements from all metadata sets with the thesauri used by RMV and Fotothek and to find of course links between them. We extracted the list of all values we found so far and are currently comparing the entries. This all can only be done manually.

```
<mappings>
  <mapping>
    <mpi-label> Speech </mpi-label>
    <ic-code> 31B6235 </ic-code>
    <ic-label> speaking </ic-label>
  </mapping>
  <mapping>
    <mpi-label> writing </mpi-label>
    <ic-code>49L11</ic-code>
    <ic-label> handwriting, writing as activity </ic-label>
    <ovm-code> OVM.AAC.AAK.AAB </ovm-code>
    <ovm-label> script </ovm-label>
  </mapping>
  <mapping>
    <mpi-label> Speech, some gesture </mpi-label>
    <ic-code>31B6235</ic-code>
    <ic-label> speaking </ic-label>
    <ic-code>31A25</ic-code>
    <ic-label> postures and gestures of arms and hands </ic-label>
  </mapping>
```

4. ECHO Knowledge Repositories

In chapter 3 we made some comments about the need for flexible knowledge representation infrastructures for the area of cultural heritage. This mainly is due to the fact that people will not agree about definitions - so it should be possible to add new definitions. Even more problematic are the mappings, since only in a few cases one can speak about a perfect match.

In the case of the thesaurus mappings we yet did not use relation-types. It is beyond the scope of the ECHO project to sort out how the inherent semantics can be modeled more precisely to be able to exploit the mappings in a more fine-grained way. Currently, all mappings between the thesaurus nodes are of the type “mapsTo” which implement a fuzzy mapping indicating some form of overlap without being more precise.

To come to a more open and flexible knowledge representation infrastructure we will set up an ISO TC37/SC4 compliant repository and start defining the DORA categories with the help of this framework. For the mapping files appropriate open repositories will be offered at the MPI web-address including all schemas[6]. RDF seems to be a primary candidate for the representation in teh Semantic Web era. Currently, however, XML is seen as being sufficient. This could allow everyone to modify aspects of the mapping and use it in their machinery.

[6] Before doing this at the end of the ECHO project we have to check the IPR situation.

We see this start of an open knowledge representation infrastructure as one of the outcomes of ECHO. The current DORA machinery will not make use of this open infrastructure, since it would cost too much effort to rewrite all programs and scripts.

5. Exploitation

Within ECHO we have created a practical ontology covering a number of knowledge components. From careful inspection of certain representations such as the thesauri we could identify many useful mappings that can be exploited by the DORA machinery. However, we yet cannot say enough about the usage of the various metadata categories by those people who generate the metadata descriptions. From experience we know that there is some semantic spreading, yet we cannot make any quantifying statements.

When DORA uses the full set of components described here[7], we have to start investigations how effective the mappings are in exploiting possible relations between the different domains and sub-domains. Here we are at the beginning. Partly this has also to do with the fact that only few repositories have a large size (Fotothek, RMV, Languages).

[7] The machinery is constantly extended with the goal to be ready end of April 2004.