WP2 Report - Integration of NECEP Data

WP2 Note on the integration of Ethnological Data in the ECHO Infrastructure

Peter Wittenburg, Laurent Dousset, Daan Broeder
June 19, 2003

At 16/17.6. a work visit was organized in Nijmegen to speak about the details of the NECEP database and its integration into the ECHO metadata infrastructure.

1. Ethnology Databases
In the content area of ethnological data (WP3) it is the intention to build a database with descriptions of "societies" (NECEP) and to link this database with existing resources and resources to be setup at various ethnological places such as RMV[1]. Further, in WP2 the NECEP database itself (its structure, its web-front end) has to be built up and the content has to be integrated into the ECHO metadata framework.

2. NECEP Database
The first design of the NECEP database is ready and is visible on the web (www.ehess.fr/centres/logis/necep). Its design is not completely stable yet due to the ongoing discussions amongst the ethnology community about how to best characterize "societies" with the help of a number of selected characteristic elements. For communities that were not yet confronted with the request to describe "real world phenomena" these processes are normal and time intensive. Nevertheless, the database design is so far that conclusions about its integration can be drawn.

It exists of a complex structure that basically contains formal descriptions, prose descriptions (named articles) and media files for illustration purposes. The details are described in a report at the mentioned web site. The formal keyword-type of descriptions can be called "Descriptive Metadata" comparable to the IMDI set. The formal descriptions currently cover a flat list of about 140 descriptors. The articles partly describe the way in which the descriptors are interpreted and why they are valued in a certain way.

The database is implemented as a relational database covering a number of tables and offers two web-interfaces striving for simplicity: one to enter information and one to search. The database is accessed via the HTTP server that links to PHP scripts operating on the database and creating the html pages. Further, PHP scripts were built that create a set of XML files that contain all data. The XML files are organized in a clear path structure, so that it is easy to determine the access path.

.../necep/society/society_1.XML	the number is identical with the society ID, the societies contain references to article IDs

[1] RMV = Dutch Ethnology Museum; please, see report WP2-TR11-2003 for details about their DB

.../necep/articles/article_1.XML the number is identical with the article ID

It is intended within ECHO to create about 10 complete entries by various anthropologists[2].

3. Metadata - Semantic Mapping

The purpose of an integrated metadata infrastructure is to be able to combine as many repositories as possible to carry out joint searches and in doing so to link information. Bases for such an activity is the definition of a common semantics of the descriptors or of defining semantic mappings between the terms used.

The set of descriptors were checked and compared with IMDI descriptors and those 12 descriptors used by RMV[3]. The following mappings were identified as being suitable:

NECEP	RMV	IMDI
Usual anthropological designation (ethnic name)	cultural origin	language name
Designations in written sources (alternative ethnic names)	cultural origin	language name
continent	cultural origin	continent
countries of residence	cultural origin	country
ethnic region	cultural origin	region
language name		language name
countries of residence	geographical region	country
	content categories	content categories

In the RMV catalogue the category “cultural origin” covers location and society information in a structured form. For a suitable mapping the two different parts have to be separated. IMDI does not have a separate term for “society”, however, since society and language names are often identical or similar a mapping would make sense. RMV is using thesauri for geographical and language naming. It has to be checked in detail whether the hierarchical information can be exploited. In IMDI the content is described by several categories, while in the RMV catalogue prefers a description of the content according to a thesaurus. Here a separate detailed investigation has to be carried out to see in how far a more exhaustive mapping is possible. NECEP does not house content.
The analysis of the thesauri will be carried out in June.

Mapping has another dimension since we could be confronted with spelling variations etc. For a restricted set of categories tables with spelling variants will be provided. There will be no phonetic or fuzzy type of mapping at the beginning.

4. Search Modes

Four search modes were discussed to discover data: (1) Specialists of the individual domains use the specialized interfaces that are available. This mode does not have to be explained. (2) A simple IR-based full-text search on all integrated material where all structured data from the metadata descriptions is treated as being unstructured, i.e. this mode accepts that the recall and precision will be much less than for a real structured search. The advantages are that a term specified by the user who does not know the details of the structured sets can be discovered with a certain chance and that also the prose descriptions can be subject of search. (3) A modular search that supports those categories that are overlapping between the domains so that structured search is possible on the different domains. (4) A search that supports the 15 Dublin

[2] It was discussed in how far it would make sense to ask all groups that are active for example in Endangered Languages programs to also create descriptions for other societies, since often society areas are identical with language areas. This would increase the number of entries considerably. This has to be evaluated by the NECEP group.

[3] Other descriptor sets of ethnology partners have to be checked.

Core elements to allow the general user acquainted with the Dublin Core semantic could do structured search[4].

4.1 Full Text Search

In general users don't know per se the structure of the domain metadata sets, the categories used and also important the value ranges supported. Also strong categorization always will be associated with erroneous usage of the categories, i.e. of misplaced entries and much more. Further, the metadata sets from linguistics, ethnology (NECEP) and also from History of Arts have many general type of descriptions that are written in prose text, i.e. they cannot be subject of structured search although they contain utterly useful information.

A simple interface as is known from Google will be provided where the user may enter some terms. The search engine will go through the stored index and respond with the hits. Of course, unstructured search does not deliver so reliable information in terms of recall and precision, but due to the fact that structured (largely controlled) information is used the quality is not as bad as for search on general prose text.

To enable this all metadata information that can be harvested will be gathered and an index will be created at a central site (MPI and/or Lund). The search will make use of this index and the index has all necessary information to link to the source. The latter has to be sorted out in more detail in June/July.

4.2 Modular Search

A modular search uses the knowledge from the different disciplines, presents the discipline specific view, but nevertheless allows searching through discipline holdings connected. In the background a mapping (as described above) is carried out that translates between the domain specific terminologies.

Every metadata set represents a domain specific view of part of the resources reality, i.e. the metadata set includes domain specific structure and semantics. This also means that specialists from other domains can't make use of that view since they probably use a very different terminology. So the user in the modular scenario has to be confronted with his view. Nevertheless, he wants to get access to other holdings as well. This can be done by mapping his terminology to that of the other included disciplines. In strict terms every view has to support a unidirectional mapping to another discipline, in ECHO we will start with bidirectional mappings between two disciplines. Nevertheless, much effort is involved to create the many mappings.

It was decided that not all detailed metadata categories of the sets have to be provided at the interface to keep it simple. So the major categories of the metadata sets will be selected and provided. Each partner has to identify what his major categories are.

To enable this all metadata information that can be harvested will be gathered centrally and stored in an optimal form. Dependent on the view chosen the search engine will search on the discipline metadata repository with all search descriptors entered and will operate on the others by those terms that can be mapped including all available mapping information. The hits contain all information to link to the source. The latter has to be sorted out in more detail in June/July.

4.3 Dublin Core Search

Dublin Core search can be compared to the modular set except that a new view (the Dublin Core view) is introduced, i.e. for this view every domain metadata set has to be mapped to the DC set. Here it is important to not extend the semantics of the DC set.

5. Metadata Harvesting

Harvesting the NECEP metadata is very straightforward since the complete information is available in directly addressable XML files. They simply have to be downloaded.

[4] The DC search mode has the lowest priority and is in fact an extension of mode 3.

For the case of the RMV data and that of other potential sites detailed discussions have to take place with technical experts. For RMV this will be done in June/July. They support a shell called “Collection Connection” that seems to allow accessing resources.

6. User Interface

The user interface will have the following elements for searching

- a way to go to the discipline specific search environment
- a Google like field to enter terms
- several buttons to choose a specific view that can be seen as complex search options (one of these views is the DC view)

The output will be represented as a merged list where it is marked from where it comes. People should be able to click on specific search that will link them to the domain specific search engine and they should be able to immediately click on resources to get an impression about the document found. This has to be worked out in more detail in June/July, since it is not yet clear how to get access for example to media files in the RMV holding.

Appendix 1

Organization and Structure of XML Files
(this part will be adapted by Laurent Dousset dependent on the progress of NECEP)