LUND UNIVERSITY

Cross-Disciplinary Integration of Metadata Descriptions

Wittenburg, Peter; Gulrajani, Greg; Broeder, Daan; Uneson, Marcus

Published in:
Proceedings of LREC 2004

2004

Link to publication

Citation for published version (APA):
Wittenburg, P., Gulrajani, G., Broeder, D., & Uneson, M. (2004). Cross-Disciplinary Integration of Metadata Descriptions. In *Proceedings of LREC 2004*

Total number of authors:
4

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Cross-Disciplinary Integration of Metadata Descriptions

Peter Wittenburg, Greg Gulrajani, Daan Broeder, Marcus Uneson*

Max-Planck-Institute for Psycholinguistics, *Lund University
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
peter.wittenburg@mpi.nl

Abstract

Within the ECHO (European Cultural Heritage Online) project an integrated domain of metadata repositories was created covering data from 5 different humanities disciplines. The integration required intensive work on encoding, syntactical and especially the semantic level, since interoperability is still difficult to be achieved. An ontology was created and a search engine that makes use of the knowledge components. This work within ECHO is seen as one of the practical contributions on the way towards the Semantic Web.

Introduction

Metadata can be all types of data that is about other data. In the area of language resources metadata in this broad sense can be annotations of audio or video streams, annotations on annotations, lexicon data derived from corpora, sketch grammars describing grammatical structures found in transcribed texts and many others. In the context of this paper we would like to restrict the term "metadata" to the keyword type of description that can be used for example to discover language resources. Typical types of such metadata standards are defined by Dublin Core[1], IMDI[2] and OLAC[3].

In contrast to many other types of metadata this type of descriptions is based on a selected set of elements of which the semantics are more or less clearly described. So it seems that metadata lends itself perfectly well to create an interoperable domain that will allow for example searches across disciplines.

One of the goals of the ECHO[4] project was to create such an integrated domain of metadata descriptions. The ECHO project comprises 5 different disciplines: history of art (HoA), history of science (HoS), ethnology (E), linguistics (L) and philosophy (P). Within these disciplines we have different types of institutions involved such as research institutions, universities, museums and teams of researchers all coming from different backgrounds and requirements with respect to this type of metadata. For museums and other institutions that store large quantities of resources, metadata catalogues are a must to be able to manage the collection and to give users access to the material. Typical research groups may be interested in a more flexible notion of semantics that keyword type of metadata can provide.

In this paper we will report about the task of integration, the solution we found, the ontology that was created and the problems we came across.

Domain – Sub-domain	size	Type MD	Formal State	Harvesting Type	Comment
HoA - Fotothek	very large	MIDAS Iconclass	non validated	XML	export to XML from a database
HoA - Lineamenta	small	close to DC	non val	XML	export to XML from a database
HoA – Maps of Rome	small	self-defined	non val	XML	export to XML from a database
HoS – Berlin Collection	large	close to DC	validated	XML	export to XML from a database
HoS – IMSS	pot large	DC	non val	XML	export to XML from a database
E – Ethnology Museum Leiden RMV	very large	OMV OMV Thesaurus	validated	OAI	export to XML from a database
E – NECEP database	small	self defined	validated	XML	export to XML from a database
L – IMDI Domain	large	IMDI set	validated	XML/OAI	true XML domain
P – Collection of Texts	small	self defined	non val	XML	XML texts

Integration Task

In ECHO we are faced with metadata descriptions from 9 different providers/types. The table above gives an overview about the metadata types that we were confronted with. All providers are using their own metadata set, one is DC compliant, two produce descriptions that are close to DC, two provide true OAI compliance including delivering DC records, most of the data is extracted from relational databases of different types producing mostly non-validated XML and just one provider is using true XML metadata descriptions.

So in ECHO we were confronted with interoperability problems at all levels: (1) The character encoding was not documented and showed problems. (2) The syntactic description was poor – only a few provided DTDs or Schemas. (3) Of course, the semantic interoperability had to be achieved in ECHO. Even worse was that in many cases the elements used were not well defined. This leads

[1] Dublin Core (DC): http://dublincore.org
[2] ISLE Metadata Initiative: http://www.mpi.nl/IMDI
[3] Open Language Archives Community: http://www.language-archives.org
[4] ECHO: http://echo.mpiwg-berlin.mpg.de/ECHO/home

to differences in usage by the metadata creators and difficulties during the integration phase.

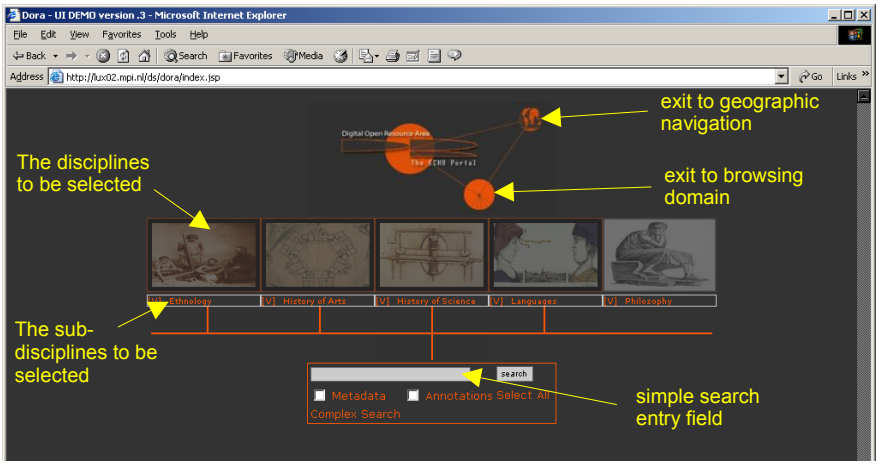
It has to be added that the way the content of resources is described differs substantially. In the Fotothek the IconClass thesaurus is used to categorize the content of photos and images. In the RMV catalogue the OVM thesaurus is used which is similar to the AAT thesaurus. Some use the subject field from the DC element set with all its weaknesses, others have an unconstrained keyword field and the elaborate IMDI set has a couple of elements that describe the content such as "task", "genre", "subgenre", "language" and "modalities".

Also for the indication of geographic regions a variety of description options is used. In the RMV case a geographic thesaurus is used, however, the thesaurus does not have a canonical structure, i.e., country names for example can occur at different levels of depth. In some instances language names have to be used to indicate the geographical region.

The task was to create one interoperable domain of metadata that would allow searches and that would generate hits from the various collections. It should then be possible to click on the hits to either go to the selected object or to link to the special web-site where further investigations can be started.

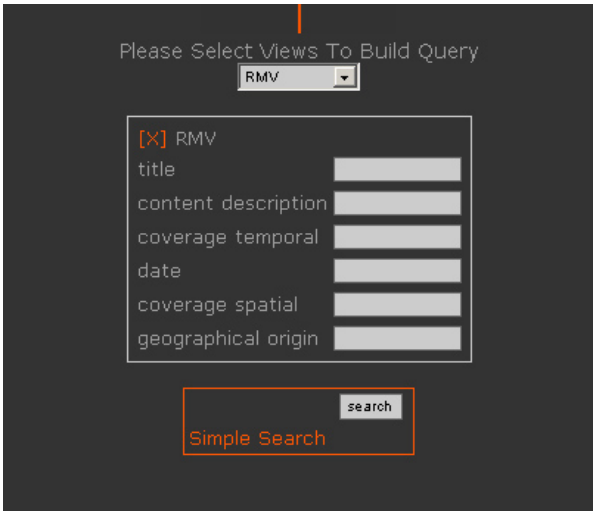
DORA – the ECHO Portal

The Digital Open Resource Area portal of ECHO is meant as entrance to the cross-disciplinary search domain. It offers a number of options. First it allows the user to select which of the 5 disciplines should be included in the search. When clicking on the discipline names the user can also select sub-domains from the menu. The programs are table-driven, i.e., when the participation will change the interface will be adapted automatically.



As first important option a Google like search field is offered. This simple search can include metadata and/or annotations – the latter is not yet implemented. An index is created for all values found in the metadata records such that quick searches can be carried out. In the simple search yet no mechanisms are integrated that make use of the available ontological information. It offers unstructured search on structured metadata which would not be a convincing strategy if there is just one discipline involved. In a cross-disciplinary approach such a field is more interesting since structured search supposes that people know the semantics of the elements. The current ECHO project was too short to also extend the simple search option.

More interesting is the so-called complex search option. On request it allows the user to select a view which is a semantic perspective the user can chose. In the normal case this will be the perspective he is most familiar with. Figure 2 indicates the RMV perspective offering a terminology and a selection which RMV visitors will understand.



For every sub-domain a view is provided and there is also one supporting the Dublin Core view. To generate these views we selected those elements that show a certain amount of semantic overlap with other elements. If the query types are so special that they include elements without any semantic overlap the users should better go to the domain-specific sites.

The user can now enter criteria for his search in these fields. During search the query is now expanded to the other domains making use of the ECHO ontology which is described below. To generate these mappings all elements of all sets were analyzed carefully. For the selected fields that showed a potential for semantic overlap one-directional mapping tables were generated, i.e. for each view there is a description of how its elements can be mapped to the elements of the other domains. In addition to these mappings the search machine also makes use of the available thesaurus information.

The DORA user interface also has two symbolic elements that allow to browse in an integrated browsable domain,

however, the IMDI language domain is the only one that allows browsing at this moment.

There is also an exit to browsing in the geographical domain which actually means defining the geographical region by selecting areas on maps. Yet it is not integrated and again: this ECHO project may be too short to offer all these options. To efficiently setup such a geographic mapping one would need metadata descriptions that have exact GIS coordinates such that spots can be generated automatically on the maps. Yet there is no such information, i.e., all spots would have to be entered manually.

The backbone of the DORA search engine is a set of configuration files. One specifies where metadata can be harvested, which method is used, how frequently the harvesting should occur, what kind of conversions have to be carried out etc. The other one defines the paths where to find the relevant metadata fields within the delivered files.

ECHO Ontology

The core of DORA is the ontology developed within the ECHO project. It can only be a first version since dependent of the usage of metadata elements and the experiences a second version will be necessary.

The following data structures were developed:

- **Validated Metadata Sets**
  The harvested metadata records were transformed into validated and machine readable formats based on proper XML and UNICODE.
- **ECHO Concepts**
  A file was generated that contains all concepts that occur in the ECHO metadata domain. It is a structured XML file that specifies the concept's ID, its name, its normalized path, its domain and sub-domain, a description and the translations in French, German, Italian, Swedish, and Dutch.
- **ECHO Mappings**
  This XML file contains all mappings between the different elements in an exhaustive, non-optimized form. It basically contains RDF-like assertions where two concepts identified by its ID are linked by a predicate (relation type).
- **OVM Geographic Thesaurus**
  This file contains the geographic thesaurus as it is used by the ethnology museum in Leiden. This XML file contains a unique code indicating also the thesaurus position and hierarchy, the geographic name and where possible a link to the so-called MPI Geographic Thesaurus. This XML file was extracted from tables that did not present geographic items in a canonical form, i.e. structure elements such as "country" occur at different hierarchy levels.
- **MPI Geographic Thesaurus**
  For all elements in all metadata sets that have geographically exploitable information a list of all geographic terms was generated. These were then integrated into a canonical thesaurus that is complete up to the country level. Added to this XML file was a tag that contains mappings to the OVM thesaurus where possible.
- **OVM Category Thesaurus**
  This thesaurus contains the values that are used to describe the content of museum objects in a hierarchical order. It is similar to the AAT thesaurus. The generated XML structure contains three information types: a code indicating the position and the hierarchy of the value, its label in Dutch and English.
- **IconClass Category Thesaurus**
  This thesaurus is used by the Fotothek collection to describe the content of the photos and images. The generated XML structure contains a code indicating the position and the hierarchy of the value and its English label.
- **IconClass-to-OVM Mapping**
  This XML-structured file contains all IconClass nodes that map to OVM nodes. For all these nodes the possible mappings to OVM concepts are given. These mappings result in partial mappings between the two thesauri.
- **OVM-to-IconClass Mapping**
  This file contains all mappings from OVM nodes to IconClass nodes which is also often a 1:N mapping.
- **MPI-to-IC-and-OVM Mapping**
  From all metadata sets except Fotothek and RMV we extracted all values of the elements that describe the content of the resources. The resulting list was transformed into a structure that contains for all values found a mapping to the appropriate Iconclass and OVM nodes.

Without having had the chance to carefully analyze the usage of all metadata fields and without having had the time to carry out optimizations we see the current ECHO ontology as a first version that can be used by others. The ECHO ontology is described in more detail in (Wittenburg, 2004b).

Problems Encountered

While building the DORA search machinery and constructing the ontology we encountered many problems. We will report about some essential ones.

Harvesting of metadata poses for many repositories a major obstacle. In general they are not prepared to offer their catalogue to others in a machine-readable form. The files that often are extracted from relational databases do not provide validated XML, much post-processing as to be carried out before being able to integrate the data into a smoothly running machinery. Many archives have heard about the OAI metadata harvesting protocol[5] and are interested, but are not capable to provide a registered interface that provides reliable data. Therefore, mostly XML files had to be harvested via HTTP.

Important knowledge sources such as the content thesauri are not available as open and well-structured XML resources. They have to be extracted semi-automatically from web-sites or from binary files. This cannot be accepted in the long run, since they are so crucial for

[5] Open Archives Initiative: http://www.openarchives.org

achieving an interoperable domain. They should be open resources on the web in a standard format. Here the emerging Data Category Registry of ISO TC37/SC4[6] which is compliant with standards such as ISO 11179 and ISO 12620 could be a good example.

Many of the resources are only available in one language, in some cases not in English. It is time to generate multilingual extensions to the most important ones, however, such work is beyond the scope of the ECHO project. We provided translations of all metadata categories used in the complex search for 5 languages.

All semantic evaluation had to be done manually, various adhoc scripts were used to simplify the task. As a consequence all relations that are established are a result of manual inspection. Often the labels are slightly different, such that only human inspection can identify the types of semantic overlap. Two examples may indicate this:

Example 1:
- IMDI: matching game
- IconClass: games of calculation and chance
- OVM: recreation, sports, games

Example 2:
- IMDI: route direction
- IconClass: means of determining location
- IconClass: direction, orientation
- OVM: orientation in time and space
- OVM: route and appliances

In the first example the semantic overlap is given by the term "game". If used consistently one can imagine that this relationship between the three content description types could be exploited during search. Here a simple script looking for the same word stems can yield useful results. In the second example the semantic overlap between "route direction" and "means of determining location" is evident although the words used are different. The overlap between "route direction" and "route and appliances" is less evident although both include the word "route". Only very smart programs using world knowledge could discover these differences.

For the mapping we have identified only 4 useful types of relations. "isEqualTo" defines semantic equivalence, "isSubclassOf" defines a hyponymy relation, "isSuperclassOf" defines the inverse and "mapsTo" is used to express a semantic overlap. In most cases the "mapsTo" relation type was used – a relation that can only be dealt with by fuzzy logic. Therefore such a relation type does not appear in RDF(S)[7] or OWL[8]. The "mapsTo" relation may only be interpreted in one direction. This knowledge is implemented in the DORA search engine.

We have chosen XML as representation format for the ontology components. This makes the structure explicit, allows validity checking with existing tools and thus facilitates the re-usage of the resulting files. For the definition of concepts this seems to be the way that is chosen by several communities. For example the Data Category Registry developed within ISO TC37/SC4 is XML based. XML has all features that are necessary to allow us to add attributes and to point to subclasses within such a registry.

Also for representing the relations we have chosen XML although RDF is the most obvious choice to represent assertions that can be exploited by inference engines. Yet the inferencing possibilities are very limited so we chose to rely on a straightforward DORA engine. It is only a little effort to generate RDF from the XML assertions. It is very easy to modify the relations that were defined by the MPI team. This is important since people will very much disagree with the mapping choices made by others. We see many practical ontologies emerging that will focus on the exploitation of special relations. The strict split between concept definitions and relations will facilitate this kind of usage.

Results

Yet it is too early to come up with final results. Very interesting is the exploitation of geographic overlap and the integrated domain gives hits from various domains. In many cases languages can be linked to geographical areas. The same is true for the time periods although the temporal coverage is very much diverging. Typical cultural heritage objects date back to early time periods while linguistic recordings for example only come from the last century. Very interesting is the exploitation of overlap in the content description. There are many examples of partial relations between the thesauri and the other content descriptions. However, since content description is the most time-consuming part during the encoding phase only few metadata repositories include rich content information. Further tests will have to be carried out and we will report about the results.

Conclusions

Within the ECHO project an attempt was made to create a cross-disciplinary metadata domain. Many technical obstacles had to be overcome to create this integrated domain indicating that many repositories still are not prepared to be linked together. A complex ontology was created mostly by manual inspection and its components are represented in a well-structured form that will facilitate re-usage by others. As can be expected, the overlap between the disciplines is very important for fruitful exploitation. However, metadata is still created too often only with the own objectives in mind and this limits the enormous potential. The DORA approach in ECHO is seen as one of the first steps in the direction of cross-disciplinary exploitation of metadata repositories from which we can learn a lot on the way to the Semantic Web.

References

Wittenburg, P. (2004a). Note on ECHO's Digital Open Resource Area. WP2-TR013-2003 – Version 5. http://www.mpi.nl/echo

Wittenburg, P. (2004b) Note on an ECHO Ontology. WP2-TR014-2004 – Version 1. http://www.mpi.nl/echo

6 ISO TC37/SC4: http://www.tc37sc4.org
7 RDF: http://www.w3.org/RDF
8 OWL: http://www.w3.org/2001/sw/WebOnt