Towards Metadata Interoperability

Peter Wittenburg
MPI for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen, Netherlands
Peter.Wittenburg@mpi.nl

Daan Broeder
MPI for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen, Netherlands
Daan.Broeder@mpi.nl

Paul Buitelaar
DFKI
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
paulb@dfki.de

Abstract

Within two European projects metadata interoperability is one of the central topics. While the INTERA project has as one of its goals to achieve an interoperability between two widely used metadata sets for the domain of language resources, the ECHO project created an integrated metadata domain of in total nine data providers from five different disciplines from the humanities. In both projects ad hoc techniques are used to achieve results. In the INTERA project, however, machine readable and ISO compliant concept definitions are created as a first step towards the Semantic Web. In the ECHO project a complex ontology was realized purely relying on XML. It is argued that concept definitions should be registered in open Data Category Repositories and that relations between them should be described as RDF assertions. Yet we are missing standards that would allow us to overcome the ad hoc solutions.

1 Introduction

Metadata is a key source of information towards realization of the Semantic Web that could be exploited in many different ways. Several projects are starting to focus on exploiting rich metadata in and between projects and disciplines. For instance, the ECHO (European Cultural Heritage Online)[1] project brings together metadata for resources from the History of Arts, History of Science, Linguistics, Ethnology and Philosophy. One aspect of the work in ECHO is to create a cross-disciplinary domain for resource discovery. In the INTERA (Integrated European Language Resource Area)[2] project one of the tasks is to establish a foundation for a more flexible definition and use of metadata for language resources.

We can distinguish two types of metadata. The first one concerns its use as "data about data". This definition of metadata includes for example text that describes images, sounds, videos and other texts. Such metadata can exist in different forms like complex annotations of media recordings as discussed for example by Bird (2001) and Brugman (2001). A second type of metadata consists of keywords describing objects that form the catalogues of the increasingly large digital collections, e.g., of linguistic data. This type of metadata was introduced by initiatives such as Dublin Core[3] for general type web-resources, OLAC[4] for general type linguistic resources and IMDI[5] for more elaborate linguistic resource descriptions that are useful not only for discovery but also for management purposes.

Although the first type of metadata is very important for the above mentioned use in content descriptions, in this paper we will focus on aspects that are related to the second, keyword type of metadata. It is obvious that this type of metadata

- contains amongst others important information about a resource that cannot be retrieved from its content;
- are especially relevant for the discovery and management of multimedia resources since speech and image recognition are still far away from being applicable in most cases;
- includes a reduced set of descriptive elements and requires classification such that content information in many cases is richer;
- offers a limited set of semantically well-defined data categories (ISO 12620) that can be related with other concepts.

[1] ECHO: http://www.mpi.nl/echo
[2] INTERA: http://www.elda.fr/rubrique22.html
[3] Dublin Core: http://dublincore.org
4 OLAC: http://www.language-archives.org
[5] IMDI: http://www.mpi.nl/IMDI

In this paper we will describe the problems that we encountered in the INTERA and the ECHO projects to come to interoperable metadata domains, the structural and semantic solutions that were chosen to solve the tasks and the solutions we are aiming at in the long run. In this context we will also refer to the intentions within ISO TC37/SC4[6].

2 Current tasks

The INTERA task

One focus of the work in the INTERA project is on the integration of metadata elements that are used in describing language resources for open data category repositories. Two metadata sets are being used currently for the discovery and management of language resources. The OLAC set is used for discovery purposes and aims to be used for all kinds of language resources. The set was derived from the Dublin Core set, i.e., on purpose it only includes a limited set of elements.

The IMDI set was designed bottom-up and is used for discovery and management purposes. It is a rich and structured set especially derived for annotated resources and lexica. The distributed IMDI domain was extended in the INTERA and ECHO projects to more than 27 participating European institutions sees itself as an OLAC data provider, i.e., the OLAC harvester can read all IMDI records that are offered via the Open Archives Initiative metadata harvesting protocol[7] (OAI MHP). A wrapper is used to map the IMDI elements to the OLAC elements, i.e., the mapping relations are hardwired into a server-based program.

Recently, a new version of the IMDI metadata set (version 3.0.3) was provided. In parallel, also the new version of the OLAC metadata set (August 2003) was worked out. Both metadata sets are described by human readable definition documents available in the web. New mapping rules have to be constructed which for short-term needs will again be hard-wired into a server-based program.

But this is not seen as being sufficient to serve future needs. New ways have to be developed for making the mapping more transparent and to prepare the metadata domain for Semantic Web applications. Therefore, as a first step, the IMDI metadata concepts are entered into the open data category registry that is currently emerging within ISO TC37/SC4.

The ECHO task

In the ECHO project one of the tasks is to create a metadata domain that covers five disciplines and several institutions within each discipline. In total we were confronted with nine different metadata sets.

The table below gives an overview of the metadata types that we were confronted with. One of the sets is DC compliant, two produce descriptions that are close to DC, two provide true OAI compliance including the delivery of DC records. Most of the data is extracted from relational databases, encoding other types of data as well. In many cases the elements used were not well defined, possibly leading to differences in usage by the metadata creators.

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

History of arts (HoA), History of Science (HoS), Linguistics (L), Ethnology (E), Phylosophy (P)

6 ISO TC37/SC4: http://www.tc37sc4.org
7 OAI MHP: http:// www.ukoln.ac.uk/cd-focus/presentations/ cldprac/sld020.htm

Also the way in which the content of resources is described differs substantially. In 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".

A variety of description options is used for the indication of geographic regions. In the RMV case a geographic thesaurus is used. Others use descriptors such as "country" and "region". In some instances language names have to be used to indicate a geographical overlap.

When creating an interoperable metadata domain one has to cope with problems at each layer: character encoding, data harvesting, syntactical aspects and semantic integration. Only the last point is of relevance in the context of this paper.

To enable semantic integration an ontology was built that covers

- nine metadata repositories;
- a file where all metadata concepts relevant for the integrated domain ECHO domain are listed including their description in a number of major languages (the setup is similar to the one used within ISO TC37/SC4);
- a file that includes all mappings between these concepts where each individual set presents a view that is mapped to all others;
- two geographic thesauri containing different types of geographic information with cross-links between them;
- two category thesauri describing the content of the resources;
- two mapping files containing one-directional cross-links between the two thesauri;
- a file that contains all content type of descriptions that occur in the metadata records and which do not use one of the big thesauri with mappings to these two.

As we are currently using the existing files simply as exchange formats they have been represented in XML (rather than RDF for instance). To implement fast search, specially optimized internal representations are chosen and combined with fast indexes. The representations are such that all occurring references are expanded in preparation time and not during execution time. A special engine was programmed that can operate on these extended representations.

To illustrate this we use an example with geographic thesaurus information. A search for "Country=Italy" should result in hits for all objects that have to do with "Italy" either as the creation site or as the site where the scene takes place. The metadata records are now extended such that for all locations that are within "Italy" the nodes appearing higher up in the thesaurus hierarchy are added. This assures that a record containing for example "Rome" will also be indicated as a hit when "Italy" was entered in the query.

Exploiting all repositories during run-time by intelligent crawlers would require fast parallel algorithms. Only parallelism would yield the execution speed needed to satisfy the users.

Relation types

We have discovered different types of relations between the concepts used in the INTERA and ECHO projects.

In the INTERA project we can indicate internal relations within the structured IMDI metadata set, i.e., structure conveys semantic relations. An example can be given by the many attributes of a participant. A certain participant has a "name" as an identifier and various attributes such as "age", "role" and "education". Between the IMDI and OLAC concepts there are three types of relations: (1) For some concepts one can speak of equality and it was agreed that the controlled vocabularies will be unified where possible. (2) There are also hierarchical relations such as "subClass" and "superClass" between some of the concepts. (3) There is a type of relation where we can speak about a semantic overlap that we cannot specify in more detail. Finally, there are concepts such as "age" or "education" of a participant that do not map at all.

For the mappings in ECHO we have identified four useful types of relations: (1) "isEqualTo" defines semantic equivalence, (2) "isSubclassOf"

defines a hyponymy relation, (3) "isSuperclassOf" defines the inverse and (4) "mapsTo" is used to express a semantic overlap. In most cases, the "mapsTo" relation type was used – a one-directional relation indicating semantic overlap that should be exploitable. It is not clear yet in how far it makes sense to define the fuzzy "mapsTo" relation in terms of the standard types provided by RDF(S)[8] and/or OWL[9]. All concepts that do not map to others or that are too special (for example "size of an image") were excluded in the ontology definition process.

Examples from ECHO

Using the described ECHO interoperability framework a number of experiments were carried out for evaluation purposes. A few examples will be discussed here.

Example 1

Simple Search "dogon"
1 match was found: NECEP: 1
Complex Search "dogon"
View NECEP - society name: 1 in NECEP
View IMSS - language: 1 in NECEP
View DC - language: 1 in NECEP
View Language - language: 1 in NECEP
Complex Search "mali"
View Language - country: 1 in NECEP

This example demonstrates the effect of the mapping between the metadata sets and of the geographical thesaurus. The language element is mapped to the society name element in NECEP although this is semantically not correct. Entering "mali" in the country specification yields a hit since "mali" is seen as a superclass to "dogon". Here a relation type such as "has_language" would be semantically more appropriate.

Example 2

Simple Search "inuit"
2 matches are found: Language: 1, NECEP: 1
Complex Search "inuit"
*View Language - *: 0 in Language (could not be found in the Language domain)*
View Language – language: 1 in NECEP
Complex Search "greenland"
View Language – language: 1 in NECEP

The results are similar compared to example 1. It indicates that the element including "inuit" in the language domain is not an element that is used for mapping. It was used as avalue of an optional element by one specific researcher. This example shows that simple search covering all metadata elements can lead to improved results.

Example 3

Simple Search "agriculture"
75 matches are found: Language: 73, Fotothek: 2
Complex Search "agriculture"
View Fotothek - iconography: 2 in Fotothek
View RMV – content: 2 in Fotothek
View IMDI – content: 2 in Fotothek

These results are misleading and demonstrate the weakness of simple search. The 73 hits for language result from matching with the recording place ("southern agriculture kindergarten") and the affiliation of an actor ("ministry of agriculture"). These results obviously do not refer to documents the user was serching for. In the case of Fotothek the hits make sense since it is about "harvesting". The mapping in complex leads to the expected results, the misleading hits from the language domain are not found.

Example 4

Simple Search "clothing"
22 matches: Language: 8, RMV: 8, Fotothek: 6
Complex Search "clothing"
View RMV – content: 8 in RMV, 6 in Fotothek
View Fotothek – iconography: 8 in RMV, 6 in Fotothek
View Language – content: 8 in RMV, 6 in Fotothek

Again the rich annotations that are used in various free-text fields in the language domain lead to wrong hits. They are about chats at the bakery shop and the clothes people are wearing – so it's not about clothing as an object which may be intended by the person specifying the search. The results for complex search from different domains shows the correctness of the mappings.

Example 5

Simple Search "horses"
7 matches: Fotothek: 2, Language: 2, IMSS: 3
Complex Search "horses"
View Fotothek – object title: 3 in IMSS
View Fotothek – iconography: 2 in Fotothek
View Lineamenta – title: 3 in IMSS
View Lineamenta – keywords: 2 in Fotothek
View IMSS – title: 3 in IMSS
View IMSS –subject: 2 in Fotothek
View Language – title: 3 in IMSS
View Language – content: 2 in Fotothek

This example clearly indicates the strength of simple search and the weakness of complex search. The pattern used by complex search can

[8] RDF: http://www.w3.org/RDF
[9] OWL: http://www.w3.org/2001/sw/WebOnt

be compared with a narrow path in the complex semantic space. If selecting the title element the hits of IMSS are found, if the content element is chosen the Fotothek hits are found. Both, however, are leading to useful hits where "horses" are central concepts in the resources. The reason for the indicated results are partly caused by very sparsely encoded metadata. In the case of IMSS the term "horses" is only mentioned in the title, the content element is yet not used. In the language case thesaurus information is used to infer from the string found in the title element ("spatial layout task, farm scenarios") to "horses".

Summary

Only the first three relations (equality, hyponomy, hyperonomy) can be used in a strictly logical way. The fourth relation type is of a fuzzy nature but occurs most frequently. To prevent a semantic cycle during searching, the specially tailored inference engine is restricted to one inference step over this fuzzy relation and exploits all relations only in one direction[10]. It is evident that the existing ontology does not describe a complete logical system.

In case of the INTERA project we will continue to rely on a wrapper that will map IMDI to OLAC records to allow OAI style of harvesting. In the ECHO project we created optimized indexes such that searching can be executed fast, i.e., the knowledge components in XML are simply used as interchange formats allowing for the easy identification of all structural components and for their validation.

3 Foundation for Metadata-Interoperability

In the previous sections we described the current state of the practical work in two projects to achieve semantic interoperability. The way chosen has a number of disadvantages in the long-run:

- In the ECHO project there are no concept definitions that adhere to open and emerging standards such as ISO 11179 and ISO 12620, and which are available in validated machine-readable registries.
- The current definitions do not contain hierarchical relations, which could be part of the concept definitions if agreed upon by the community.
- A contribution from other experts, for example to improve the definitions and to add other language specific aspects, is largely excluded.
- The representation of the semantic relations between concepts is partly encapsulated in a program preventing any flexibility. In the ECHO case they are structurally described with the help of XML tags, however, it would be much better to provide them in a way that inference engines relying on RDF(S) and OWL could operate on them.

From the practical work we learned that often the semantic scope of the metadata elements is not specified as precisely as seems possible and also necessary. This will allow for a spectrum of usage that will have effects not only on human interpretation, but especially on the way of mapping relations to chose. It is obvious from this experience that users will not always agree on the interpretation of the definitions and on the types of mappings applied. At this moment we cannot make final statements in how far hierarchical relations will be effected by this that would constitute an implicit thesaurus as is expected within ISO TC37/SC4.

Open Data Category Repositories

Based on the experience so far it can be recommended to include into open repositories only concepts that have been used for a while and therefore have shown their semantic stability within a certain community. For the area of language resources ISO TC37/SC4 is on the way to create such a repository, which is compliant with widely recognized standards such as ISO 11179 and ISO 12620. Therefore, it makes sense to register all elements used within IMDI and OLAC as data categories in this repository.

This will open up several new possibilities for projects and initiatives: (1) IMDI and OLAC can create schemas that define their sets by referring to machine-readable definitions. For instance, an equality relationship can be directly indicated by referring to the same data category registry (DCR) entry. Search engines could make use of this information. (2) It is our experience that pro-

[10] It should be noted, however, that advanced inference systems can handle semantic cycles of this nature.

jects often like to tailor their own metadata sets due to their specific needs. In this case an open registry would simply allow to create a new schema and to re-use existing definitions as much as possible[11]. By referring to DCR entries again a direct form of interoperability is achieved.

We assume that we will have widely recognized DCRs as currently defined within ISO TC37/SC4. They should contain the concepts that are based on a wide agreement within communities. However, due to the slow acceptance processes within standardization bodies and the different needs that result for example from different languages there could be a need for researchers to set up their own temporary DCRs. We therefore foresee a large number of data category repositories.

For the ECHO project the usage of an open DCR is not yet an option. To be of use for the community there has to be a wide acceptance. The domain of "cultural heritage" addressed within ECHO covers too many different disciplines and the concepts are semantically mostly too different. Disciplines such as history of arts, history of science and ethnology have to start their discipline oriented discussion process to define useful concepts and to start building widely recognized registries. What seems necessary is to start creating files with concept definitions that can be easily integrated later into open registries and that are compliant to emerging standards.

Open Relation repositories

Concept definitions in DCRs are one important aspect in defining metadata ontologies. Another aspect are repositories that store relations between these concepts. From our experience in the two projects mentioned, it seems required to separate these two types of information in order to achieve a high degree of independence and flexibility. However, other experiences as that of the GOLD initiative (Farrar, 2005) indicate that opinions on this vary largely.

Theoretically, it is possible to include all information that defines a concept into the DCR. The concept "country" that is used within IMDI is typically a sub-part of a "continent". However, the proper definition of the concept "country" in the context of language resources is not dependent on the availability of this hierarchical relation. But this again may be completely different for abstract linguistic concepts such as "transitive verb" where we know that the class relation "transitive-verb isSubClassOf verb" is part of the definition.

In general, we argue that whenever it is not strictly necessary for the proper definition of a concept, relation aspects should be kept outside of DCRs as much as possible, since they often form a constraint with only little agreement.

For the represention of relations in a machine-readable format, RDF(S) seems to be the most suitable choice. In RDF, all relations are represented as tertiary assertions as indicated in Figure 1. Actually, each of these RDF assertions defines a relation between two resources, since the value can be an arbitrary web-resource as well.

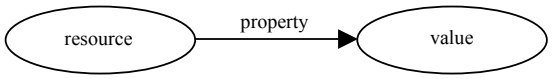

Figure 1 shows a basic RDF assertion specifying that a (web) resource identified by a URI has properties that may have values.

Obviously, this simple mechanism allows us to create complex repositories of semantic relations. Since all objects of such an assertion can be web-resources we can for example point to concepts defined in a DCR and relate them with each other.

From the two mentioned projects we can give two typical examples. From the INTERA project we notice that according to our interpretation the concept "IMDI:Participant.Role=Collector" is a sub-class of "OLAC:Creator" (Figure 2).

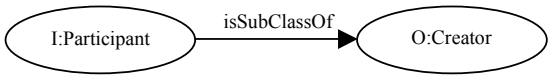

Figure 2 shows a typical relation that can be found in the INTERA project.

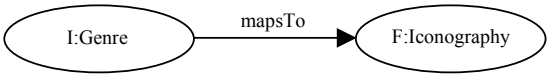

Figure 3 shows a typical relation that can be found in the ECHO project.

[11] IMDI already provides a step towards this kind of flexibility by allowing projects to define profiles or individuals to define new key-value pairs.

In the ECHO project we can identify a semantic overlap between "IMDI:Genre" and "Fotothek:Iconography" (Figure 3).

We can imagine that RDF will be used by some projects, initiatives and institutions to establish widely recognized and used repositories with mapping relations.

We also assume that many persons, projects and institutions will create their own mappings to tune their operations like searching according to their specific needs, i.e., a large variety of "practical ontologies" will emerge. These practical ontologies may re-use most of the semantics found in a repository, or they overwrite a certain number of relations or they introduce new relations that are not yet defined elsewhere.

In contrast to the ISO data category repository that is based on the experiences of the work about ISO 11179 and ISO 12620, there is no work yet of how to represent relations for the domain of language resources. For INTERA this creates the need of using ad hoc solutions. ISO TC37/SC4 should urgently take up this issue.

4 Registries and Engines

Given the discussion above, we can expect the Semantic Web era to produce a large number of data category definitions stored in different DCRs and mapping relations between these stored in other repositories. Amongst these components there will be some that deserve a larger interest by the language resource community, since they are maintained by recognized experts, but there will also be many others created within projects and institutions or even by individuals to satisfy only ad-hoc purposes. Therefore, we need an infrastructure for registering these components for making them visible and searchable.

Current inference engines such as provided by Jena[12] assume that there is one database of meaningful RDF triples. This would allow us to integrate all our mapping relations from the INTERA or ECHO ontologies (such as "Country isSubClassOf Continent" and "Place isSubClassOf Country"), that is currently part of an XML-based thesaurus. To arrive at an RDF-based database instead, we would need to harvest metadata from the XML-based thesaurus, i.e., we would first have to write a wrapper that converts XML structure information into RDF assertions.

Further, we would like to harvest RDF triples from different sites, since we need to integrate already existing knowledge. Two problems can be foreseen here: (1) How do we know where to find useful RDF triple instances? We need mechanisms to register the existence of sites with that type of information and to semi-formally describe the content. (2) When we harvest triples from such a site we may include knowledge – metadata ontologies defined in RDF(S) - that is conflicting with what is already available. How can we deal with this and how can we be selective?

Currently, there are no answers to these questions. But they have to be addressed soon. Also here ISO TC37/SC4 could play an important role, since it is about infrastructure aspects that have to be worked out for the language resource community.

5 XML vs RDF

We explained why XML was chosen in representing the knowledge involved in the projects mentioned. Mainly short-term arguments guided us to take this decision. This may not be the correct decision in the long-term. Nevertheless, also ISO TC37/SC4 has chosen to represent data category definitions as XML structures including hierarchical references needed to properly define a concept.

The underlying data models of XML and RDF are very different. XML is based on a tree model, i.e., it has a strong bias towards hierarchies. All expressive power is gained from structural relations, which to a certain extent allow for the representation of semantic relations.

In contrast to this, RDF is based on a loose collection of relations. It is therefore very simple to combine relations from different RDF repositories into larger collections. Although implicit hierarchies will be difficult to recover.

Semantically, RDF Schema offers the user the option to define the value range of any user-defined relation (property) used in an RDF file with user-defined classes, while XML only offers basic data types. OWL has even more expressive

12 Jena: http://jena.sourceforge.net

power. A good overview is given by Gil and Ratnaker, 2001.

Summarizing, we would like to emphasize the following two points that need to be taken into account by any follow-up projects of INTERA and ECHO. Such a project should:

- represent all concept definitions of a resource metadata set in an ISO DCR compliant way and turn them over to RDF-based repositories that may emerge within the disciplines in the coming years;
- represent relations as much as possible in external RDF(S)-based metadata ontologies using all needed expressional power of RDF(S) and OWL so that users can easily add their own relations or reformulate existing ones.

6 Conclusion

The work on metadata interoperability in the two projects mentioned clearly indicate that this type of work is in its beginning phase. Ad hoc methods are used to achieve high speed and to guarantee efficient exchange of knowledge components, but they form obstacles on the way towards a flexible and open Semantic Web type of infrastructures. The examples indicate that the chosen mapping strategies lead to the expected results in many cases. They also indicate some of the problems that are associated with using specific elements for searching. Amongst others these are caused by sparsely filled in metadata descriptions, unawareness about the underlying element semantics, insufficient mappings between metadata elements and thesaurus concepts.

The usage of ISO 11179 and ISO 12620 compliant open Data Category Registries for machine readable definitions of metadata concepts within INTERA is a first step in the right direction. However, other disciplines than linguistics lack such a widely agreed registry type. For building up and combining repositories of RDF-based relations between registered concepts there is yet no infrastructure. Even in the linguistics domain yet there is no suggestion for standards. ISO TC37/SC4 should take up this issue, since Data Category repositories with concept definitions and relation repositories are mutually dependent on each other to form exploitable knowledge bases. Due to the many contributions from projects, institutions and even individuals that will disagree with proposed definitions and relations we will need an efficient infrastructure for discovering and combining useful knowledge components.

References

S. Bird and M. Liberman. 2001. A formal framework for linguistic annotation. http://www.ldc.upenn.edu/Papers/CIS9901_1999/revised_13Aug99.pdf

H. Brugman and P. Wittenburg. 2001. The application of annotation models for the construction of databases and tools. http://www.ldc.upenn.edu/annotation/database/papers/Brugman_Wittenburg/20.2.brugman.pdf

S. Farrar and D.T. Langendoen. 2003. Markup and the GOLD Ontology. http://saussure.linguistlist.org/cfdocs/emeld/workshop/2003/paper-terry.html

Y. Gil and V. Ratnaker. A Comparison of (Semantic) Markup Languages. In Proceedings of AAAI 2001. http://trellis.semanticweb.org.expect/web/semanticweb.comparison.html