EASING THE CREATION OF MAPPINGS BETWEEN METADATA FORMATS

Kurt Majcen*, Werner Bailer, Martin Höffernig, Werner Preininger, Silvia Russegger

DIGITAL – Institute for Information and Communication Technologies, JOANNEUM RESEARCH Forschungsgesellschaft mbH, Steyrergasse 17, 8010 Graz, Austria
(kurt.majcen, werner.bailer, martin.hoeffernig, werner.preininger, silvia.russegger)@joanneum.at

KEY WORDS: metadata, mapping, conversion, crosswalk, formats, archives, museums, configuration, tool, web service

ABSTRACT:

Being able to exchange metadata is the key to ensuring access to collections, establishing interoperability among collections, and between different types of cultural heritage institutions, such as across libraries, museums and audiovisual archives. Motivated by two use cases, one for audiovisual archives and one for museums and general archives, we present an approach for automating mapping between different metadata formats. The mapping approach uses an intermediate ontology and formalises the relations to each of the metadata formats supported. An intuitive web-based configuration user interface is provided in order to build and customise mappings. Based on the two use cases, we discuss two ways of applying the mapping approach: as a web service, which can be included in processes of an audiovisual archive's preservation system and integrating of the generated mapping instructions into collection management applications for museums and archives. The proposed approach reduces the effort for defining metadata conversions. It thus allows overcoming interoperability issues between cultural heritage institutions and facilitates content provision to portals like Europeana and Archives Portal Europe.

1. INTRODUCTION

Preserving cultural heritage does not only require ensuring the integrity of the objects to be preserved, but also includes making them accessible and usable, i.e. providing technologies for long-term access and use in changing contexts. Being able to exchange metadata is the key to ensuring access to collections, establishing interoperability among collections, and between different types of cultural heritage institutions, such as across libraries, museums and audiovisual archives. Metadata exchange is often hindered by the diversity of metadata formats and standards that exist in the different communities. Thus metadata interoperability needs to be established between the different parties involved.

The problem of metadata interoperability exists on two levels:

On a *syntactic level*, metadata can be accessed and processed in the same syntactic format, today typically some XML format. This does not imply that all metadata are already XML, only that they can be rendered as such (with services or wrappers). On a *semantic level*, metadata can (partially) be interpreted within the same semantic frame of reference. Meaning of metadata of one institution (often coded in in-house metadata vocabularies) needs to be linked with metadata from another institution. Thus, it requires alignment of archive vocabularies, which might be incomplete as vocabularies differ in scope and perspective.

Tools for metadata mapping are needed to overcome these interoperability issues. However, with n formats existing in a given environment, we need in the worst case $\mathcal{O}(n^2)$ mappings if we go for a simple approach considering only pair-wise mappings. Chaining mappings is also not a useful approach, as due to the incompleteness of mappings transitivity of relations cannot be ensured. We thus propose an approach that uses a high-level intermediate representation, together with mapping templates on data type level, from which the code for a mapping problem between a pair of standards can be derived. This would ideally allow us to solve the problem with $\mathcal{O}(2n)$ definitions.

The need for mapping between different metadata representations comes from diverse scenarios. They include the conversion of legacy technical metadata in preservation scenarios, access to legacy content descriptions, extracting metadata embedded in digital file headers and converting it to the data structures needed in a SIP/AIP (in OAIS terminology, see ISO 14721:2003) of a preservation system, ingest of metadata from non-/semi-professional content creators, outsourcing of annotation and access services, with possibly different data models between customer and service provider's infrastructure and content provision to Europeana (http://www.europeana.eu), Archives Portal Europe (http://www.archivesportaleurope.eu/) or similar portals.

In this paper, we analyse two specific use cases of metadata mapping: one specific to the domain of audiovisual archives, and the one targeting museums and general archives. We propose a mapping approach that starts from schemata of metadata standards or in-house metadata models, in contrast to approaches like e.g. the mapping tools of the MINT framework (Kollia I. (et al.), 2012) that start from individual metadata documents. We describe the configuration user interface for defining and customising mappings and discuss the application of the tools in the two use cases.

The rest of this paper is organised as follows. After discussing the use cases in Section 2, we describe the proposed automatic mapping approach in Section 3, and then present in Section 4 a user interface for building and configuring mappings. The application of the proposed approach in the two use cases is discussed in Section 5 and Section 6 concludes the paper.

2. USE CASES

Memory institutions (such as archives, museums, libraries and so forth) are hosting collections including very different kinds

of objects and archival material. These materials are used within the context of an organisation (maybe with various departments) but are also transferred to other organisations, a variety of professionals or maybe to the interested public. Metadata – the information about these objects – differ also among the various producers and consumers.

2.1 Audiovisual archives

Both audiovisual archives acting as depository institutions for specific types of audiovisual content as well as those linked to media production organisations (e.g. public broadcasters) face the issue of interfacing with processes that have diverse requirements in terms of metadata, both in ingest and access. Digitisation has blurred the boundaries between traditional types of audiovisual media, opening new options for reusing and repurposing content. This huge amount of content can be generally accessed either via standardised and proprietary metadata formats, which are often incompatible between the parties involved. As a result, the content is often locked in within silos preventing an effective search across these sites and making it complicated to exchange rich metadata for audiovisual content.

While many of the multimedia metadata formats in use overlap in their functionality, they are at the same time dissimilar in many ways.

Coverage. Some formats aim to be domain independent while others focus on specific domains (e.g. film) or usage scenarios (e.g. broadcast metadata for consumers, such as EPG information).

Comprehensiveness. Some formats aim to provide comprehensive descriptions of multimedia content ranging from low-level features that can be extracted automatically to fine-grained semantic description of a scene, while other formats provide a simple list of general annotation properties, that only refer to the entire media item.

Complexity. Metadata formats also differ in the complexity of their description syntax. Some formats only support free text for specific properties (e.g. names of creators), while others support structured content and/or references to controlled vocabularies.

Due to the differences between the formats, mappings can only be partial in many cases, e.g. when properties do not exist in one of the formats involved in mapping. If the mapping target is a format with a strict definition that does not allow extensions, information can be lost during the mapping steps.

For allowing the exchange of data between different data models some transformation of structure of the description – the metadata mapping, we focus on – and also translation between vocabularies (a potential enhancement for the future) is needed. The mapping process itself is more intended to be an underlying activity (except content provision for public portals) of the IT-systems in use for workflows in archives which usually need not become explicitly visible to the user.

2.2 Museums and general archives

Nowadays archive and collection management systems based on information technologies are widely used and have proven to provide valuable support for the management of objects in the cultural heritage domain. Archiving of cultural data is still an important issue for museums and archives. But this is only one side of the medal.

An important fact is that the presentation of the data becomes more and more important after archiving and putting a lot of effort into the scientific preparation of the data. After several years of data entry into various databases, it is now possible and important for cultural institutions to keep track, develop and make real use of these data repositories.

As museums and archives begin to transform their data management applications into smaller and more manageable application modules, it seems likely that data management will become more and more relevant. This is already considered in the *imdas/archivis pro* software package (www.imdas.at and www.archivis.at) that was developed at JOANNEUM RESEARCH. The programme can be customised to individual user needs and can be adapted to different types of objects and collections. It supports a combination of visual representations (text, images, symbols, multimedia data, and maps) and intelligent collection management. This concept of customisation enables a flexible software solution for museums and archives and offers multiple ways of accessing, analysing and presenting the data.

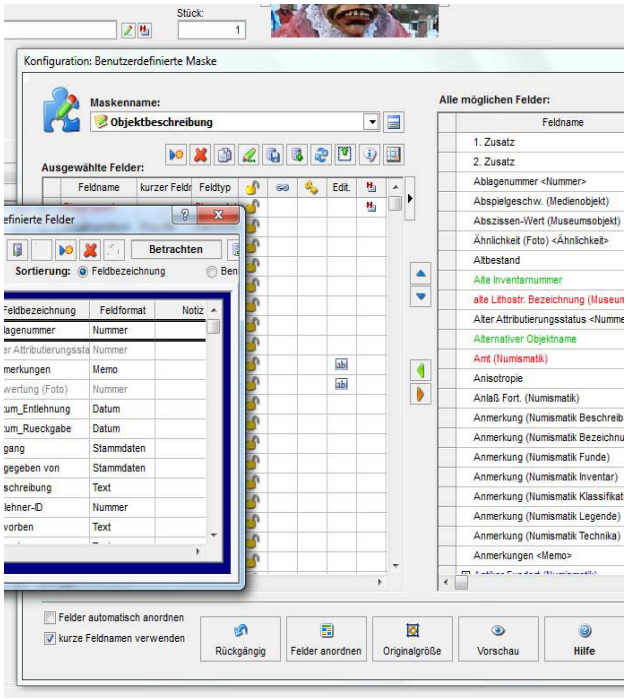

Figure 1: Customisation editor in *imdas/archivis pro*

The customisation of an information system in the domain of cultural heritage leads to individualisation also regarding metadata and metadata format. Therefore processes and systems for metadata mapping are important and necessary if data should be presented and made available via public portals like Europeana.

In order to enable different organisations with customised *imdas/archivis pro* applications to exchange data with these kind of portals it is necessary to do the mapping definition and further on the export of the data in house – without additional implementation of individual software pieces. The aim must be to have a flexible configuration tool that allows specifying the mapping between individualised (often relational) database formats and common public portals.

3. MAPPING

Our mapping approach uses a high-level intermediate representation of generic metadata elements serving as a hub for mappings between metadata formats. Therefore metadata elements from a specific metadata format are formalised in terms of this intermediate representation. Then, mapping relations between format-specific and generic concepts are described. Combining these two sets of mapping relations, mapping relations between a pair of metadata formats can be inferred. These inferred mapping relations are the basis to create mapping instructions in order to map a metadata document from one format to another. Since these mapping relations are based on a conceptual level only, additional information about data types together with context information is required. After linking this information with an appropriate set of mapping templates provided by a library, mapping instructions expressed as XSL (Kay M. (ed.), 2007) templates are created. Finally, executing these XSL templates enables a metadata mapping between a pair of metadata formats. The overall workflow of this approach is visualised in Figure 2. A detailed description of the approach can be found in (Höffernig M. (et al.) 2010).

Since our mapping approach features an intermediate representation of generic metadata elements serving as a hub for mapping between formats, hand-crafted one-to-one mappings between each pair of metadata formats are avoided and the mappings can be created automatically. Therefore mapping relations are easier to maintain as well as adding a new metadata format is done without side effects to existing definitions.

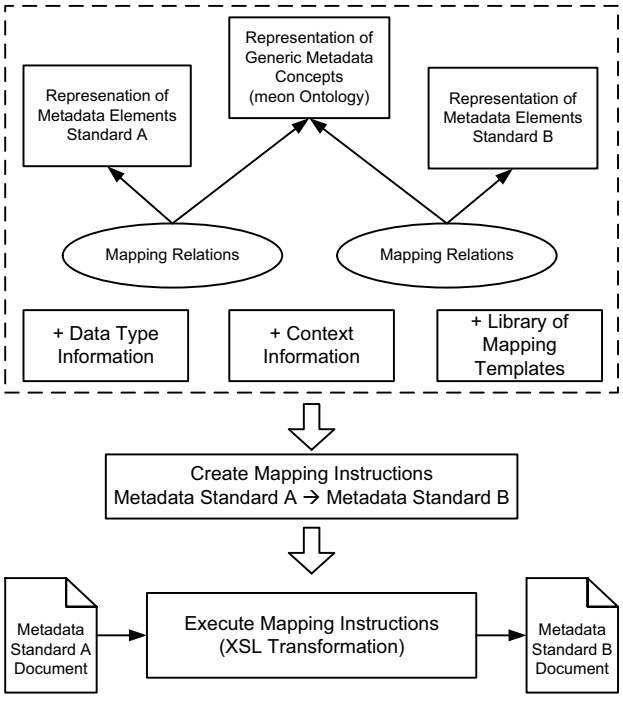

Figure 2: Metadata mapping approach

The core of this approach is the *meon* ontology (Höffernig M. (et al.) 2009) which describes generic metadata elements and the relations between them. *meon* was originally developed to model metadata elements used throughout the audiovisual media production workflow in a format independent way in order to support content exchange and its automation. The *meon* ontology has been extended to express mapping relations between metadata formats. In addition to the ontology of generic metadata concepts, specific ontologies are created for each format taken into account. Then it is possible to infer how concepts from different metadata formats are related by observing the relations among generic concepts and to the format-specific concepts.

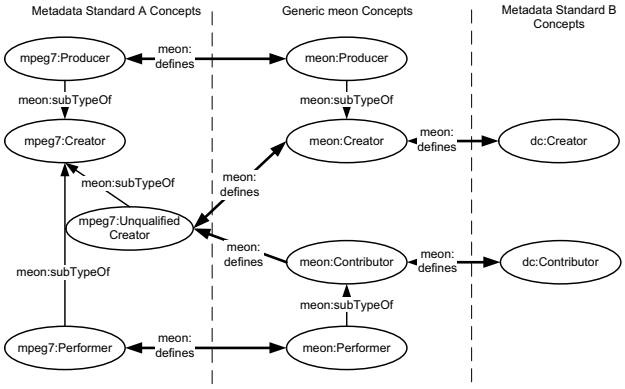

Figure 3: Example of the *meon* ontology for describing metadata elements and their relations

The *meon* ontology, expressed in OWL-DL (Coburn E. (ed.), 2010. LIDO - Lightweight Information Describing Objects. Version 1.0. http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf (accessed 14 June 2012)

Dean M. (ed.), 2004), defines properties to describe definition and equivalence relations (property `meon:defines`) as well as subtype relations (property `meon:contains`). These properties can be applied on instances of class `meon:Concept` with its subtypes `meon:AtomicConcept` and `meon:CompoundConcept`.

In order to express mapping relations between metadata formats, the *meon* ontology has been extended. In addition to the ontology of generic metadata concepts, schema specific ones are created for each format following the *meon* pattern. Figure 3 shows a schematic example for the mapping of metadata elements from a metadata format A to another metadata format B via generic *meon* concepts, in this case `meon:Creator`, `meon:Producer`, `meon:Contributor`, and `meon:Performer`. It also models their relations, i.e. `meon:Producer` is a subtype of `meon:Creator`, `meon:Performer` is a subtype of `meon:Contributor`. In the same manner the format-specific concepts are defined and mapping relations via `meon:defines` properties are established (e.g. expressing equivalence between `mpeg7:Producer` and `meon:Producer`).

In order to retrieve mapping instructions between formats it is necessary to model the definition relations in more detail. Therefore additional data type information as well as context information is attached to the *meon* ontology. Then it is possible to select appropriate mapping templates to generate mapping instructions expressed as XSL templates which are applied to a given input document.

4. CONFIGURATION USER INTERFACE

As described in our mapping approach, any information which is necessary to determine mapping relations between a pair of metadata formats has been formalised with using the *meon* ontology. Additional information such as data type and context information as well as referencing mapping templates has been formalised by extending the *meon* ontology. Furthermore a logical reasoner is employed to infer new knowledge needed during the mapping process.

In order to hide the complexity of describing all this data in OWL style, we have developed a web-based GUI for managing all the required data for creation of mapping instructions. A screenshot of this configuration tool is depicted in Figure 4. This user interface enables editing and inspecting mapping relations between metadata formats and *meon* concepts as well data type and context information management needed to create mapping instructions.

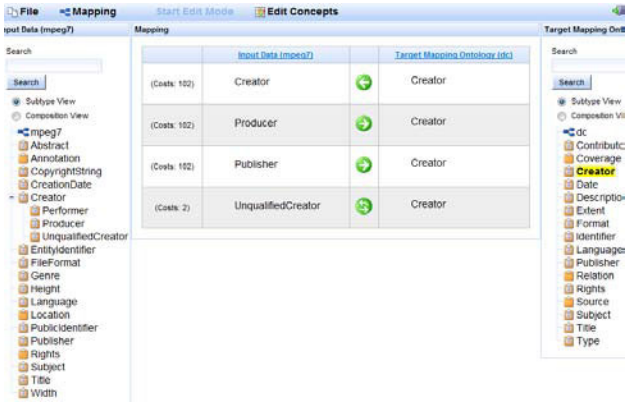

Figure 4: Screenshot of configuration user interface

In a typical use case the user selects an existing formalisation of a metadata format to work with or start to create a new one from scratch. On one side of the screen, concepts of the format-specific metadata representation are displayed using a tree structure, respectively concepts coming from the *meon* ontology are displayed on the other side (cf. Figure 4). In the middle of the screen current mapping relations are displayed. In this view new concepts can be added and existing ones can be modified. Furthermore mapping relations can be created by drag and drop or modified. In case a second format-specific representation has been selected, mapping relations between format-specific ones are inferred via the *meon* concepts and finally can be inspected as well.

Another view in the configuration tool provides the possibility to edit data type and context related information such as XPath references (Berglund A. (ed.), 2010) and attach these data to format-specific concepts. Editing the library of mapping templates is also addressed by the configuration tool. After all required mapping relations as well as additional data type and context information have been provided by the user, the configuration tool creates the resulting XSL document, which can be integrated in our applications.

5. APPLICATIONS & SERVICES

The mapping tools are available in an online and an offline manner depending on the current application where it is going to be used. This also depends on other matters as organisational structure, number of departments/persons involved and on the technical possibilities of the application site.

5.1 Web services for audiovisual archive systems

Archive systems of audiovisual archives allow ingest of media and their metadata, i.e. importing Submission Information Packages, (SIP) in OAIS terminology (ISO 14721:2003). After that often media are updated (e.g. for preservation purposes) resulting in updates of the corresponding metadata (i.e. change of Archival Information Packages, AIP). For consumption media files are extracted and packed with the necessary metadata into a Dissemination Information Package (DIP).

The first and last steps are in many cases automated processes. Media files are transferred from one place to another. Their locations along with the metadata are stored into the archive database. Before this can take place the transformation of metadata (i.e. the mapping) is needed. Thus a web service is provided which can be called by the overall ingest process before import into the metadata.

The use of this service may take some time depending on the size of the metadata description (which may become rather large e.g. in the case of MPEG-7 based descriptions). Further the service may be consumed by large parallel batch jobs. Therefore the service was implemented as a non-blocking interface. Calls are basically performed for triggering a mapping job and subsequent polls detect the status of pending jobs. The mapping service is implemented as a RESTful HTTP service (Fielding R., 2000) which can be used in a rather flexible and suitable way with most programming languages and system.

After a first step to find out which format identifiers exist the general sequence of calls for using the metadata mapping web service as shown in Figure 5 has to be performed: (2) define the new project-specific environment containing the necessary conversion settings; (3) upload the XML document available in the input format for mapping; (4.1) start the conversion process; (4.1.1) repeatedly check the status of the conversion process; keeping track of triggered conversions and result handling are part of the application which uses the service; the outcomes of these checks are foreseen as not finished/failed/success; (4.1.1.1) fetch the resulting XML which is available in the output format.

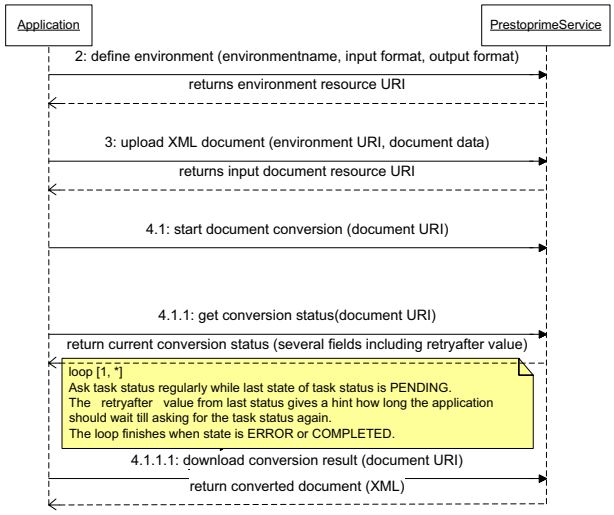

Figure 5: Sequence diagram of one conversion

Beside the aforementioned service calls the web service interface further includes functionalities to delete documents from the server and also to stop previously initiated jobs.

For purpose of visualisation and providing others with a test system a web front end (Figure 6) to the mapping service was developed. It includes some exemplary mappings and is available at prestoprime.joanneum.at.

5.2 Applications for museums and general archives

The products *imdas pro* and *archivis pro* are available as native Windows applications but also as web based applications. For the following we refer only to the native application but similar approaches apply to the web applications. Both applications allow the administration of museum objects and archival material. The metadata for such objects and references to accompanying media files are stored in a relational database.

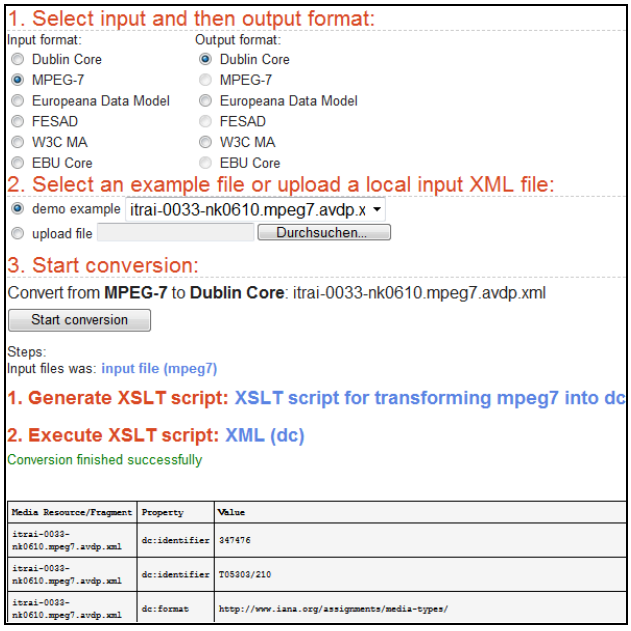

Figure 6: Web front end to the mapping service

The database schemata of the applications allow a very flexible approach with regard to metadata fields (their representation, cardinality and also to their labelling).

One of the wishes from customers is to export metadata to other systems, web portals like Europeana, Archives Portal Europe or other portals which accept metadata according to e.g. the LIDO (Coburn E. (ed.), 2010) or EAD (Encoded Archival Description, 2002) descriptions. Due to the flexibility of the data model nearly each installation has undergone some customisation. These customised software versions thus require a specific mapping to the potential export formats.

As described in the previous sections a mapping is mainly described through a number of XSL style sheets which are used to process an input document in a given format with (several) XSLT to create a document according to the desired output formats (e.g. LIDO, EAD, Europeana EDM (Europeana Data Model, 2012)). The XSL style sheets can be created manually which is a cumbersome and time consuming task. Furthermore it can be a hard or even impossible challenge to get the style sheets correct. Therefore the configuration tools allow creating

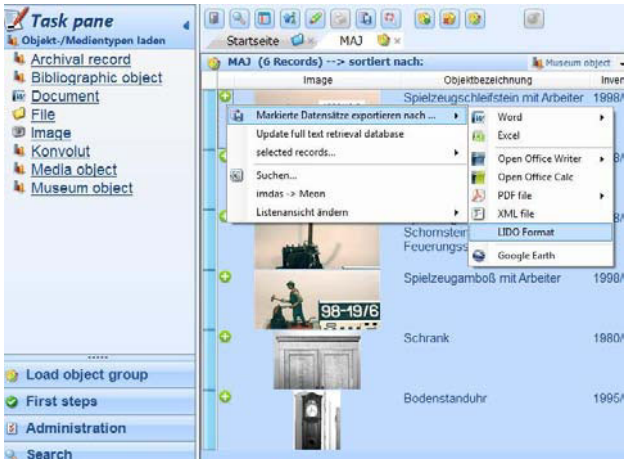

Figure 7: Exporting selected records according to the LIDO format within *imdas pro*

the necessary style sheets as an output of a user oriented graphically assisted definition process.

The *imdas pro* and *archivis pro* applications have implemented an XSL transformation engine which can use such style sheets. The style sheets themselves become available in the applications by importing into the product's database. Achieving the style sheets can be done according to one of three ways: (a) standard mapping directly available in the product database and download of some other pre-defined mappings from a server at JOANNEUM RESEARCH; (b) using the configuration tool on a server at JOANNEUM RESEARCH to create new mappings; (c) installation of the configuration tool within an institution's network and use of this tool to define mappings. Defining configurations as in (b) and (c) can be based on already existing ones or can start from scratch.

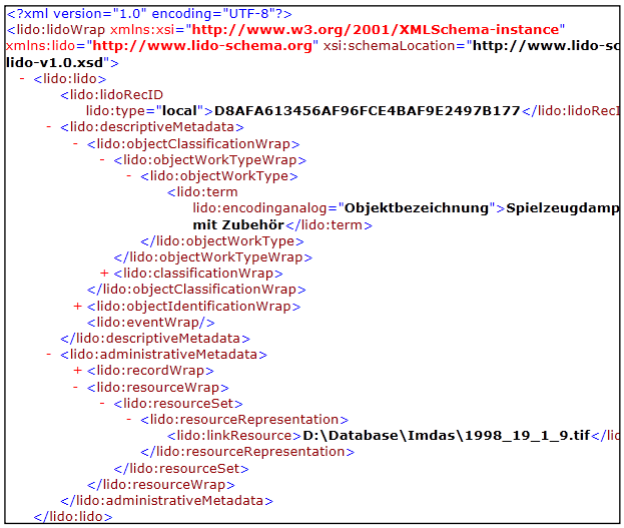

Figure 8: Record exported in LIDO format

After storing the style sheets in the database the new mapping can be chosen in the application (as shown in Figure 7). The specific transformations will be applied on the selected data sets in order to produce the desired output format. In a single license environment records are created and available on that one particular computer. In a client/server installation of

imdas pro style sheets are created, imported into the central database and available to clients from that moment.

An example output is shown in Figure 8. It was created through the selected records from Figure 7 and a basic mapping definition of the available source elements (e.g. "Objektbezeichnung"; attached media like "Image" are stored as references to the places where the files are stored) to the mandatory elements in LIDO.

6. CONCLUSION

In this paper we have presented an approach for automating mapping between different metadata formats, in order to overcome interoperability issues between cultural heritage institutions and facilitate content provision to portals like Europeana and Archives Portal Europe. The mapping approach uses an intermediate ontology and formalises the relations to each of the metadata formats supported. An intuitive web-based configuration user interface is provided in order to build and customise mappings. We have presented two applications of the proposed mapping approach: as a web service, which can be included in processes of an audiovisual archive's preservation system, and the integration of the generated mapping instructions into collection management applications for museums and archives. The proposed approach reduces the effort for defining metadata conversions and thus facilitates access to the memory institutions' collections.

7. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231161 (project PrestoPRIME, http://www.prestoprime.eu).

8. REFERENCES

Berglund A. (ed.), 2010. XML Path Language (XPath) 2.0 (Second Edition). W3C Recommendation, http://www.w3.org/TR/xpath20/ (accessed 14 June 2012)

Coburn E. (ed.), 2010. LIDO - Lightweight Information Describing Objects. Version 1.0. http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf (accessed 14 June 2012)

Dean M. (ed.), 2004. OWL Web Ontology Language: Reference. W3C Recommendation. http://www.w3.org/TR/owl-ref/ (accessed 13 June 2012)

Europeana Data Model, 2012. Definition of the Europeana Data Model elements, version 5.2.3. http://pro.europeana.eu/edm-documentation (accessed 14 June 2012)

Encoded Archival Description, 2002. http://www.loc.gov/ead/ (accessed 14 June 2012)

Fielding R., 2000. Architectural Styles and the Design of Network-based Software Architectures. http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm (accessed 14 June 2012)

Höffernig M. (et al.) 2009. Formal Metadata Semantics for Interoperability in the Audiovisual Media Production Process. In: Workshop on Semantic Multimedia Database Technologies (SeMuDaTe), Graz, Austria.

Höffernig M. (et al.) 2010. Mapping Audiovisual Metadata Formats Using Formal Semantics. In Proceedings of 5th International Conference on Semantic and Digital Media Technology, Saarbrücken, DE.

ISO 14721:2003. CCSDS, Reference Model for an Open Archival Information System (OAIS).

Kay M. (ed.), 2007. XSL Transformations (XSLT) Version 2.0. W3C Recommendation. http://www.w3.org/TR/xslt20/ (accessed 13 June 2012)

Kollia I. (et al.), 2012. A systemic approach for effective semantic access to cultural content. Semantic Web, 3(1), pp. 65-83.