An aggregation system for cultural heritage content

NasosDrosopoulos, VassilisTzouvaras, NikolaosSimou,
Anna Christaki, Arne Stabenau, Kostas Pardalis,
FotisXenikoudakis, EleniTsalapati and StefanosKollias

Department of Electrical and Computer Engineering,
National Technical University of Athens,
Zografou 15780, Greece
{ndroso,tzouvaras,nsimou,achristaki,stabenau,cpard,fxeni,etsalap}@image.ntua.gr

Abstract

Ongoing activities for thedigitisation, cataloguing and preservation of cultural heritage are taking place in Europeand the rest of the world. They involvea variety of content holdinginstitutions, following the established practices ofmuseums, libraries and archives. In parallel, aggregation and indexing initiatives, such as Europeana, illustrate the benefits and added value of metadata interoperability for repository owners and the end user. Respective modeling efforts are directed in facilitating the aggregation of diverse, and mostly proprietary, metadata records under well-defined, machine understandable, reference data models. The mapping and transformation procedure is rarely a straightforward task, varying according to the existing infrastructure and data, while requiring the continuous involvement of domain experts and content providers. In this paper we present an intuitive platform that offers a set of web services to managetheaggregation of metadata records, effectively handling the complexity of mapping cases and the need to maintain and evolve the alignment of content holding repositories. The architecture and interfaces of the system areoutlined, focusing oningestion management, the mapping editor and available functions, the resulting repository and respective publishing interfaces. The system is currently deployed for several European and national aggregation and digitisation projects, as well as for prototyping efforts regarding LIDO and the Europeana Data Model.

Introduction

Digital evolution of the Cultural Heritage field has accelerated rapidly in the past few years. Massive digitisation and annotation activities are in progress all over Europe and the world,following the early developments atthe European level and the Lund principles[1]. Furthermore, the strong involvement of companies like Google, and the positive reaction and increasing support of the European Union, have led to a variety of, rather converging, actions towards multimodal and multimedia cultural content generation from all possible sources (i.e. galleries, libraries, archives, museums, audiovisual archives etc.). The creation and evolution of Europeana[2], as a unique point of access to European Cultural Heritage, has been one of the major achievements ofthese efforts. At the moment, more than 19 million objects, expressing the European cultural richness, are accessible through the Europeana portal, and it is expected that this number will be doubled within the next five years.

Nevertheless, despite the creation and availability of numerous digital collections, only a small proportion of the available cultural heritage material has been processed to date. For this reason, there is a significant commitment to further digitisation at national and institutional levels across Europe [3]. An estimate of the vast amount of data (around 77 million books, 358 million photographs, 24 million hours of audiovisual material, 75 million of works of art, 10,5 billion pages of archives) still to be digitized and the related cost, that is about 100 billion euro, is provided in the recent European Report of the Comité des Sages[4]. Furthermore, a substantial amount of cultural heritage related, born-digital material isgenerated, such as data originatingfrom scientific research or digital analysis of cultural resources.

Due to the diversity of content types and the numerous metadata schemas used to annotate the content, interoperability plays an important role, having been identified and treated as a key issue during the last five years [5]. The main approach to interoperability of cultural content metadata has been the adoption of established standards in the specific museum, archive and library sectors (e.g. Dublin Core, CIDOC-CRM, LIDO, EAD, METS etc.) and their mapping to a common data model used - at the Europeana level: Europeana Semantic Elements (ESE)[6] andEuropeanaData Model (EDM)[7]–in order to provide unified access to the distributed repositories. Still, the above procedure is far from trivial, since the heterogeneity and individuality of the cultural content has led to metadata descriptions that differ a lot from a syntactic (based on technologies used for the representation) as well as a semantic (based on the meaning of the information provided) point of view.

This paper presentsthe MINT platform [8], which provides to users and content providers the ability to perform, in an effective way, the required mapping of their own metadata schemas to reference domainor aggregation models, like LIDO and EDM respectively. MINT follows a typical web-based architecture, offering an expanding set of services for metadata aggregation and remediation. It addressesthe ingestion of metadata from multiple sources, the mapping of the imported records to a well-defined machine-understandable reference model, the transformation and storage of the metadata in a repository, and the provision of services that consume, process and remediate these metadata. Although its design was also guided by expediency, the system has been developed using established tools and standards, embodying best practices in order to animate familiar content provider procedures in an intuitive and transparent way.

The system has been customized and deployed for several horizontal, thematicor regional aggregators,whose diversity has guided the support for various domain metadata models and approaches, mapping cases, and consuming services such as OAI-PMH deployment for harvesting by Europeana or Lucene indexing for portal services.It is important to notice that the system is currently used in the framework of several European aggregation and digitisation projects, such as ATHENA, EUscreen, CARARE, Judaica, ECLAP, DCA and Linked Heritage,having ingested more than 4 million objects to Europeana until now [9].

Metadata Aggregation

The key concept behind the aggregation part of the system has been that, although ’low-barrier’ standardssuch as Dublin Core were used in the first stages of Europeana (ESE data model) to reduce the respective effort and cost, a richer and better-defined model could reinforce the domain’s conceptualization of metadata records, at least for the mainly descriptive subset of their cataloguing elements. Moreover, since the technological evolution of consuming services for cultural heritage is greater than that of most individual organizations, a richer schema would at least allow harvesting and registering of all annotation data regardless of the current technological state of the repositories or their intended (re-) use.

The developed system has been deployed for several standard or specialized models such as LIDO, Dublin Core, ESE, CARARE's MIDAS-based schema, EUscreen'sEBUCore-based approach etc. It facilitates the ingestion of semi-structured data and offers the ability to establish crosswalks to the reference schema in order to take advantage of a well-defined, machine understandable model. The underlying data serialisation is in XML, while the user's mapping actions are registered as XSL transformations. The common model functions as an anchor, to which various data providers can be attached and become, at least partly, interoperable. Some of the key functionalities are:

- Organization and user level access rights and role assignment.
- Collection and record management (XML serialisation).
- Direct import and validation according to registered schemas (XSD).
- OAI-PMH based harvesting and publishing.
- Visual mapping editor for the XSLT language.
- Transformation and previewing (XML and HTML).

- Repository deployment and remediation interfaces.

The metadata ingestion workflow, as illustrated in**Error! Reference source not found.**,consists of four main procedures. First is the *Harvesting/Delivery* procedure, which refers to the collection of metadata from content providers through common data delivery protocols, such as OAI-PMH, HTTP and FTP. Following is the *Schema Mapping*procedure, during which the harvested metadata are mapped to the common reference model. A graphical user interface assists content providers in mapping their metadata structures and instances to a rich, well defined schema (e.g. LIDO), using an underlying machine-understandable mapping language. Furthermore, it provides useful statistics about the provider's metadata while also supportingthe share and reuse of metadata crosswalks and the establishment of template transformations. The third step is the *Transformation* procedure, which also aims at the transformation of the content provider's list of terms to the vocabularies and terminologies introduced by the reference model. The last step is the *Revision/Annotation*procedure that enables the addition and correction of annotations, the editing of single or group of items in order to assign metadata not available in the original context and, further transformations and quality control checks according to the aggregation guidelines and scope(e.g. for URLs).

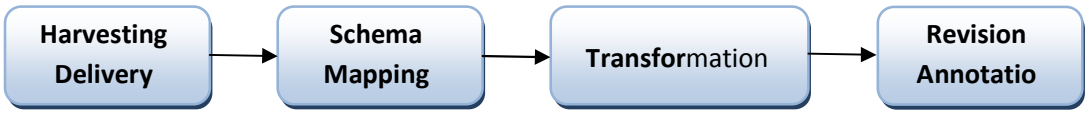

Figure 1: Ingestion Workflow

Mapping Editor

Metadata mapping is the crucial step of the ingestion procedure. It formalizes the notion of a metadata crosswalk, hiding the technical details and permitting semantic equivalences to emerge as the centrepiece. It involves a user-friendlygraphical environment (**Figure** shows an example mapping opened in the editor)where interoperability is achieved by guiding users in the creationof mappings between input and target elements. User imports are not required to include the respective schemadeclaration,while the records can be uploaded as XML or CSV files. User's mapping actions are expressed through XSLT stylesheets, i.e. a well-formed XML document conforming to the namespaces in XML recommendation. XSLT stylesheets are stored and can be applied to any user data, exported and published as a well-defined, machine understandable crosswalk and, shared with other users to act as template for their mapping needs.

The structure that corresponds to a user's specific import is visualized in the mapping interface as an interactive tree that appears on the left hand side of the editor. The tree represents the snapshot of the XML schema that is used as input for the mapping process. The user is able to navigate and access element statistics for the specific import while the set of elements that have to be mapped can be limited tothose that are actually populated.The aim is to accelerate the actual work, especially for the non-expert user, and to help overcome expected inconsistencies between schema declaration and actual usage.

On the right hand side, buttons correspond to high-level elements of the target schema and are used to access their corresponding sub-elements. These are visualized on the middle part of the screen as a tree structure of embedded boxes, representing the internal structure of the complex element. The user is able to interact with this structure by clicking to collapse and expand every embedded box that represents an element, along with all relevant information (attributes, annotations) defined in the XML schema document. To perform an actual (one to one) mapping between the input and the target schema, a user has to simply drag a source element from the left and drop it on the respective target in the middle.

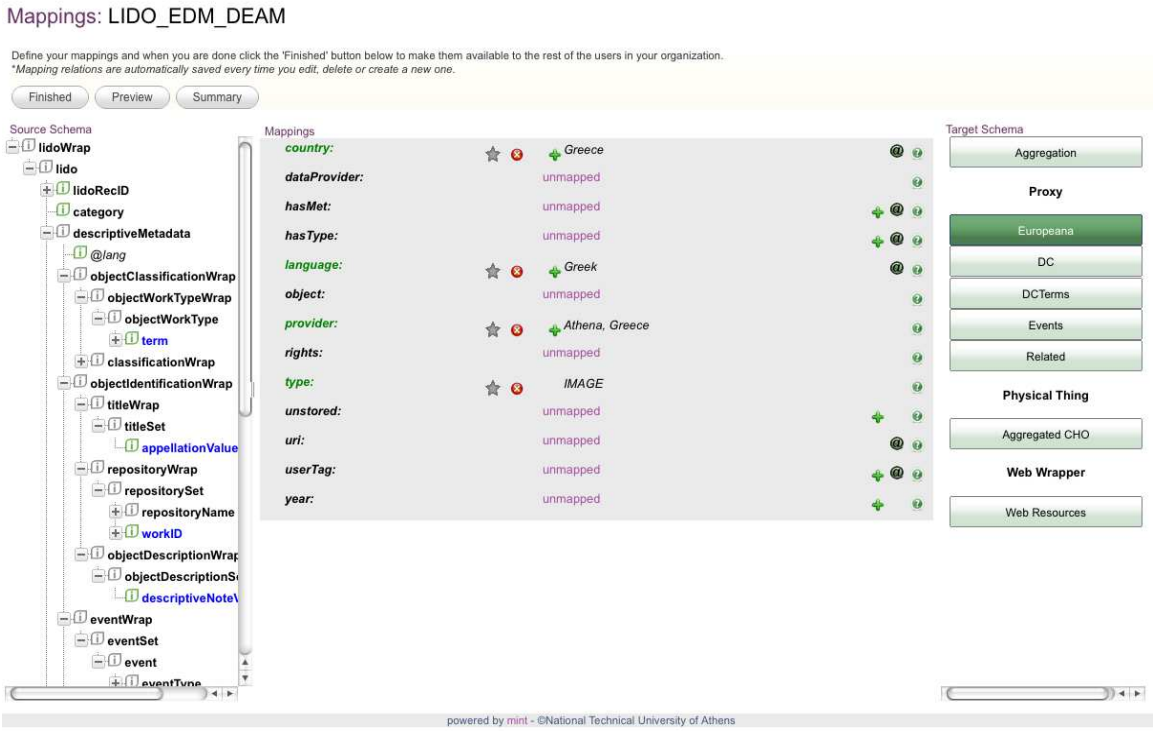

Figure 2: Screenshot of the mapping editor

The user interface of the mapping editor is schema aware regarding the target data model and enables or restricts certain operations accordingly, based on constraints for elements in the target XSD. For example, when an element can be repeated then an appropriate button appears to indicate and implement its duplication. Several advanced mapping features of the language are accessible to the user through actions on the interface, including:

- String manipulation functions for input elements.
- m-1 mappings with the option between concatenation and element repetition.
- Structural element mappings.
- Constant or controlled value assignment.
- Conditional mappings (with a complex condition editor).
- Value mappings editor (for input and target element value lists).

Knowledge Management using MINT

The World Wide Webis currently evolving from a global information space of interconnected documents to one where both documents and, most importantly, data are linked. In this framework, effort is directed towards aggregating cultural content from different providers through unifying models thatallow for semantic interoperability. Moreover, semantic linking of content descriptions with rich terminological knowledge published on the webaims to provide users with the ability to pose expressive queries in terms of this knowledge.

Following the ongoing efforts to investigate the usage of the semantic layer as a means to improve user experience, we are facing the need to provide a more detailed semantic description of cultural content. This information, accessible through its metadata, would be of little use if users were not in position to pose their queries in terms of a rich integrated ontological knowledge. Currently this is performed through a data storage schema, which highly limits the aim of the query. Semantic query answering refers to finding answers to queries posed by users, based not only on string matching over data that are stored in databases, but also on the implicit meaning that can be found by reasoning based on detailed domain terminological knowledge. In this way, content metadata can be terminologically described, semantically connected and used in conjunction with other, useful, possibly complementary content and information, independently published on the web.

One of the main points that have guided the presented system's development is the apparent need for preservation and alignment of as much of the original data richness as possible. The aggregation is only the first effort on the part of providers and aggregators towards the efficient mediation and reuse of their knowledge bases. The support for semantic data models, such as the EDMor the EBUCore [10] ontologies, enables the repository for deployment and, most importantly, information reuse through knowledge modelling and data interoperability research activities. The aim is to allow for further resource linking between different collections, reconciliation across the repository and with external authorities and, enrichment of the information resources.It should be mentioned that it is only due to the achieved metadata aggregation, validated by the content providers or experts themselves, that semantic enrichment and semantic answering to the queries of the experts and users is possible.

The transformation of the data from content providers to RDF through the use of the mapping tool (e.g. see http://mint-projects.image.ntua.gr/europeanafor the EDM ontology)results in a set of RDF triples that correspond to an attribute-value set for each information resource. Since the targetmodel is a general ontology referring to metadata descriptions of each object, the use of thematic ontologies for different domains is necessary in order to add semantically processable information to each object. This process includes two steps. First, the thematic ontologies are adopted or created in collaboration with field experts. These include individuals that represent the information resources and concepts, which correspondto sets of objects and roles defining relationships between resources. After that the data values of the RDF instances are transformed to individuals of the thematic ontologies and these individuals are then grouped together to form concepts as imposed by the thematic ontologies. The transformation of the data values to individuals is performed, from a technical point of view, by mapping the data values to URIs. Following this transformation, the data are stored in a semantic repository from where they can be retrieved through queries.

An important issue that concerns the cultural community is the effective management of the aggregated interoperable metadata, with the publication of metadata as linked data being one of the most expected outcomes.Linked Data have gained great attention recently, aimingto make data accessible not only to human but also to software agents, building in that way a semantic layer and also making the consumption of information transparent and straightforward. Cultural heritage metadata constitute an ideal candidate for their publication as linked data, being capable of populating many different applications such as tourist guides or educational platforms. Furthermore, the fact that in most cases cultural metadata are produced by human annotation makes them trustworthy information, also guaranteeing analytical descriptions of the annotated content. Current developments of the MINT platform aim to facilitate the production and publication of interconnected information resources following the linked data principles, through the use of intuitive user interfaces for content providers and domain experts.

Conclusion

During the past few years,ongoing activities for digitization, cataloguing and preservation of cultural heritage have been taking place in Europe and the rest of the world, involving all types of cultural institutions, i.e., galleries, libraries, museums, archives, and all types of content.These activities have resulted to various content types and metadata schemata used to annotate the cultural content. Due to this diversity, interoperability has been identified as a very important requirement for cultural content metadata and variouspracticeshave been proposed. In this paper, the web-based platform MINT, that facilitates the ingestion of cultural metadata and their aggregation under adopted datamodels, is presented. Therespective metadata ingestion workflow is outlined, followed by a more detailed

illustration of the schema and value mapping functionalities. Finally, the objectives, first steps and current developments of MINTtowards the publication and reuseof cultural metadata in the web of dataare discussed.

References

1. The Lund Principles & the Lund Action Plan,http://cordis.europa.eu/ist/digicult/lund-principles.htm

2. Europeana,http://www.europeana.eu

3. NUMERIC - Developing a statistical framework for measuring the progress made in the digitisation of cultural materials and content (January 2010),http://cordis.europa.eu/fp7/ict/telearn-digicult/numeric-study_en.pdf

4. The New Renaissance Report of the European ReflectionGroup on Digital Libraries (Comite des Sages), January 10,2011.

5. SIEDL:FirstWorkshoponSemanticInteroperabilityinthe European Digital Library, 5th European Semantic Web Conference, Tenerife, Spain, June 2, 2008.

6. ESE - Europeana Semantic Elements, http://www.version1.europeana.eu/c/document_library/get_file?uuid=104614b7-1ef3-4313-9578-59da844e732f&groupId=10602

7. EDM - Europeana Data Model,http://www.version1.europeana.eu/technicaldocuments

8. Metadata Interoperability Services, http://mint.image.ece.ntua.gr/

9. Projects using mint, http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/Projects

10.Semantic Web @ EBU, http://tech.ebu.ch/semanticweb_ebu