Linked Television Heritage

Vassilis Tzouvaras
National Technical University of Athens, Greece
tzouvaras@image.ntua.gr

Jean-Pierre Evain
Metadata and Workflow Processes Groups EBU, Switzerland
evain@ebu.ch

Nikolaos Simou
National Technical University of Athens, Greece
nsimou@image.ntua.gr

Athanasios Drosopoulos
National Technical University of Athens, Greece
ndroso@image.ntua.gr

ABSTRACT

The EUscreen project represents the European television archives and acts as a domain aggregator for Europeana, Europe's digital library. The main motivation for it is to provide unified access to a representative collection of television programs, secondary sources and articles, and in this way to allow students, scholars and the general public to study the history of television in its wider context. In this paper, we present the methodology followed for publishing the EUscreen dataset as Linked Open Data.

Keywords

Metadata, Interoperability, Linked open Data

1. INTRODUCTION

In this paper, we present the workflows and respective tools used for the ingestion and manipulation of Europe's Television Heritage content as well as the methodology adopted for its publication as Linked Open Data.

Massive digitization and aggregation activities all over Europe and the world have shaped the forefront of digital evolution in the Cultural Heritage domain during the past few years. Europe (i.e. galleries, libraries, archives, museums and audiovisual archives and major IT companies,) has developed a wide range of converging actions supporting the aggregation and capture of knowledge via multimodal and multimedia cultural content generation combined with Linked Open Data.

The creation and evolution of Europeana (www.europeana.eu) as a unique point of access to European Cultural Heritage, has been one of the major achievements of these efforts. At the moment, more than 20 million metadata records, representing 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. The Europeana portal currently provides access to various cultural objects and their digital representations, of which the majority is text or images; audio-visual collections are underrepresented. However, recent analysis of query logs from the Europeana portal indicated that users have a special interest for audiovisual content, as is the trend throughout the web. A lot of attention has therefore been paid to facilitate wider access to such potential users including professional. Enriching metadata and facilitate queries remains an important goal. Involving users is an necessary step before data from social network become part of Linked Open Data enrichment supporting richer and more user targeted searches.

Television content is regarded a vital component of Europe's heritage, collective memory and identity – all our yesterdays – but it remains difficult to access. Copyrights, the mulitplicity of audio and video formats, digitization costs and storage issues make the process of its aggregated and contextualized publishing on the Web more challenging than for museums and library collections. The Euscreen (www.euscreen.eu) project has created a representative collection of television programs, secondary sources and articles facilitating access to students, scholars and the general public.

Providing access to large integrated digital collections of cultural heritage objects is challenging. The aggregation of metadata from different content providers using different models requires to take some harmonisation actions. After that, the metadata must be made available to the public in a consistent way, not only offering a user friendly navigation and preview but also allowing its consumption and re-use by machines. Specifically the overall workflow consists of three main steps, the metadata ingestion, their transformation to a common reference schema, and finally their publication as Linked Open Data.

2. Building a consistent metadata framework

As it was identified by the EUscreen project's survey and reports, content providers use various collection and content management systems that store and export different types of knowledge in a range of varying metadata models. It was therefore required to propose a common format in which metadata would be submitted by content providers or transformed into by the metadata aggregator.

In order to achieve semantic interoperability within the aggregation and with external repositories, a harvesting schema was implemented based on EBUCore [4], which is an established standard in the area of audiovisual metadata. An extensive evaluation of standards in this area (such as MPEG7, DCMI, TV Anytime etc.) has been conducted [9] and led to the adoption of EBUCore, a schema that has been purposefully designed as an extension of the Dublin Core to describe audio and video resources for a wide range of broadcasting applications including archives, exchange and publication. The MINT aggregation platform (http://mint.image.ece.ntua.gr/) was used for the ingestion and transformation of the metadata. MINT is suite of a web services that facilitate the mapping and transformation of provider's proprietary, legacy or standardized metadata to a reference representative model, i.e. EBUCore in the case of EUscreen.

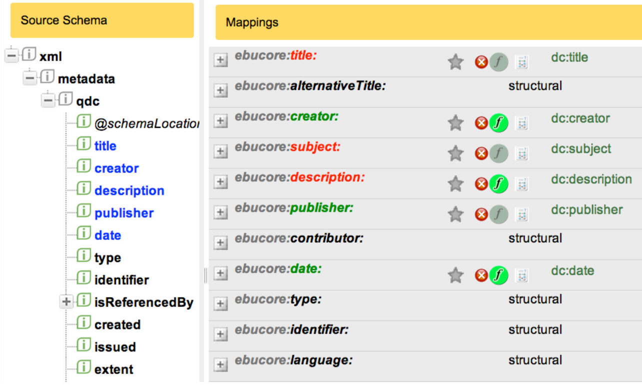
Following the metadata format harmonisation, a Linked Open Data publication procedure has been established. This required the conversion of the harvested metadata to RDF using an expressive data model. The RDF representation of EBUcore (http://tech.ebu.ch/lang/en/MetadataEbuCore) was used. Finally, internal and external linking to the EUscreen content has been performed and the resulting repository was made accessible through a SPARQL query endpoint. Once available, SPARQL endpoints will need to be masked by a GUI (graphical User Interface) hiding the complexity of SPARQL queries. The GUI will provide a framework for search specific to a domain of application or user profile (e.g. public access point vs. professionals or academics)

3. Metadata Aggregation and Transformation

Metadata aggregation has performed using the Metadata Interoperability (MINT) toolset. MINT is an open source, web based platform for the ingestion, mapping and transformation of metadata records. Interoperability is achieved aligning provider's records through the use of well defined metadata models – EBUCore in the EUscreen case.

More specifically, the platform offers a user and organization management system that allows the deployment and operation of different aggregation schemes with corresponding user roles and access rights. Users can start by uploading their metadata records in XML or CSV serialization, using the HTTP, FTP and OAI-PMH protocols. Users can also directly upload and validate records in a range of supported metadata standards (XSD). XML records are stored and indexed for statistics, previews, access from the mapping tool and subsequent services. Handling of metadata records includes indexing, retrieval, update and transformation of XML files and records. XML processors are used for validation and transformation tasks as well as for the visualization of XML and XSLT.

The most important step is the implementation of crosswalks for the providers' metadata, for which MINT introduces a visual mapping editor for the XSL language. Mapping is performed through drag-and-drop and input operations which are translated to the corresponding code. The editor visualizes (Figure 1) the input and target XSDs, providing access and navigation of the structure and data of the input schema, and the structure, documentation and restrictions of the target one. Mappings can be applied to ingested records, edited, downloaded and shared as templates.



Source Schema

- xml
  - metadata
    - qdc
      - @schemaLocatio
      - title
      - creator
      - description
      - publisher
      - date
      - type
      - identifier
      - isReferencedBy
      - created
      - issued
      - extent

Mappings

ebucore:title:	dc:title
ebucore:alternativeTitle:	structural
ebucore:creator:	dc:creator
ebucore:subject:	dc:subject
ebucore:description:	dc:description
ebucore:publisher:	dc:publisher
ebucore:contributor:	structural
ebucore:date:	dc:date
ebucore:type:	structural
ebucore:identifier:	structural
ebucore:language:	structural

Figure 1. Screenshot of the MINT mapping editor

After that, users can transform their selected collections using complete and validated mappings in order to publish them in available target schemas for the required aggregation and remediation steps. Preview interfaces present the steps of the aggregation such as the current input xml record, the XSLT code of mappings, the transformed record in the target schema, subsequent transformations from the target schema to other models of interest (e.g. Europeana's metadata schema), and available html renderings of each xml record.

Finally, the last step corresponds to 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).

4. EUscreen Linked Open Data Pilot

In this section we present the steps followed for the publication of the EUscreen content as Linked Open Data. We start by illustrating the production of the RDF instances from the metadata aggregated and transformed in compliance with to the EBUCore based schemas (XML to RDF via XSLT). Then semantic kwnowledge constructed from metadata records is linked to external open data sources.

4.1 Semantic Representation of the EUscreen Content

EUscreen Linked Data resources have been created as machine readable representation in RDF transforming EBUCore XML metadata into EBUCore RDF (see below the description of the EBUCore ontology [3])

4.2 EBUCore ontology

The EBUCore ontology is an RDF representation of the EBUCore object model, which forms part of the EBU Class Conceptual Data Model (CCDM) also used as class model of W3C MAWG's Media Annotation ontology (http://www.w3.org/TR/mediaont-10/). CCDM and EBUCore define a minimum structured set of audiovisual classes (inc. groups of resources, media resources, parts, media objects but also locations, events, persons and organizations). The EBUCore and CCDM ontologies also define the semantic relationships (objectProperties) between these classes as well as properties (dataProperties) characterizing these classes. A lot of the knowledge gathered in the EBU CCDM and EBUCore RDF was used to develop the W3C Media Annotation ontology (W3C MAWG). Reciprocally, EBUCore RDF has implemented in a subsequent version the RDF modeling options chosen by W3C MAWG.

The EBUCore ontology (expressed in RDF) is not a conversion from EBUCore XML to RDF. It is a representation of all EBUCore XML mapped to the corresponding part of the EBU CCDM model.

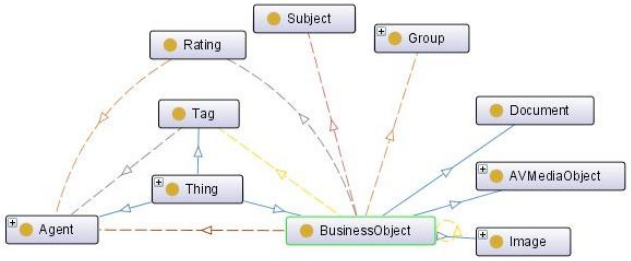


Figure 2. Snapshot of the EBUCore ontology

As shown in figure 2, the general concept of BusinessObject corresponds to the content being described and made available for consultation i.e. a document (e.g. PDF), an image, and audio and/or video file. All these BusinessObjects can be associated through a variety of relations and can also be grouped.

The ontology offers several Linked Open Data connections to the social web via user tagging and rating.

EBU also proposes additional concepts such as for example 'genre', 'role' and 'target audience' (target groups and parental guidance) as Linked Open Data in the form of SKOS Classification Schemes. Figure 3 shows a snapshot of the display of the EBU SKOS genre list In Protégé (Stanford University).

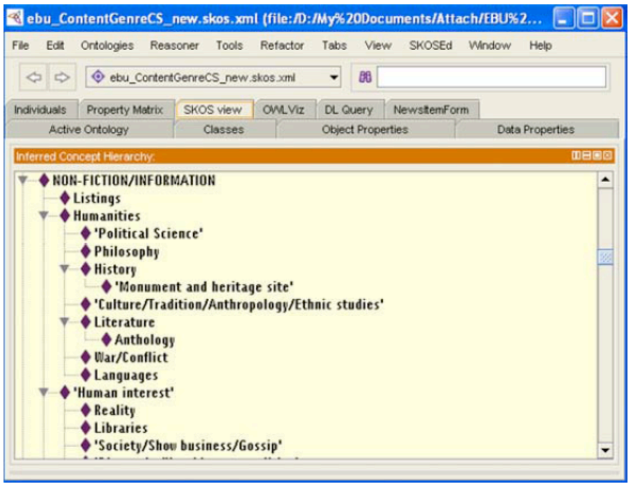


Figure 3. Schreenshot of the EBU RDF/SKOS Genre list on protégé

4.3 Implementing Linked Open Data

The RDF representation of the EUscreen metadata and linkage to content, was followed by the creation of additional resources and to fulfill the first principle of Linked Data [2], the use of URIS for things. There are various guidelines for creating cool URIs for the semantic web [1,8] and the two basic characteristics they must have are to be unique for every item, and consistent. According to these guidelines every entity represented in our data set leads to the minting of at least three URIs:

- a URI for the real-world object itself
- a URI for a related information resource that describes the real-world object and has an HTML representation (dereferencable)
- a URI for a related information resource that describes the real-world object and has an RDF/XML representation

To ensure the uniqueness of the URIs, web resources are served under a domain administered by the project (lod.euscreeen.eu) and the assigned unique identifier of the item is part of the URI. The corresponding set of URIs for an example EUscreen item are shown below.

- http://lod.euscreen.eu/resource/EUS_55F569268ACA42B186682960875F862B
- http://www.euscreen.eu/play.html?id=EUS_55F569268ACA42B186682960875F862B
- http://lod.euscreen.eu/data/EUS_55F569268ACA42B186682960875F862B

At this point we must note that except the URIs that were constructed for the unique things described in the dataset (i.e. the videos) additional URIs were made for information shared within the dataset. Such information corresponds to e.g. the actors, the countries and the organisations in the EUscreen dataset. For example, a country can be the location of production of more than one video item. Therefore new reusable resources have been created for these elements using their unique names to carft unique identifiers (URIs. For example, in the case of the Netherlands the shared resource constructed is http://lod.euscreen.eu/resource/Netherlands

After specifying the method for minting present and future URIs, we proceeded to identify the things described according to appropriate EBUCore classes and properties that would be used for their representation in RDF. More specifically, the type of video, as is defined in the XML schema. can be either a part of a programme or the whole programme. Depending on this information the resource created for the video can either be an instance of the EBUCore class Part (i.e. one of several media fragments -audio, video, data- that composes an audiovisual media resource; in other ontologies fragment is often referred to e.g. as a 'part' or 'segment') or a MediaResource itself. The additional characteristics of the video resources are represented in RDF by using EBUCore properties having as range either typed literals (e.g. original title was represented by ebucore:originalTile) or in some cases other internal resources (e.g. for each video provider a new resource is made that is an instance of "ebucore:Agent"). Furthermore, in the case of strings the language in which they are epxressed is also provided. In this way the possible consumers of the EUscreen dataset can perform queries to generate language specific mash-ups. The complete set of properties and classes used for the mapping of all the harvesting schema's elements can be found at https://docs.google.com/spreadsheet/ccc?key=0Akruw5a0_oaLdEQyMl85NVQxZ2lmT00wcVU4ZVRJZ0E&hl=en_US#gid=3

Finally, another recommendation that is very important and has to be considered during Linked Data publication is ownership of resources, licencing and provenance of information. Therefore, for every RDF representation of an item provenance metadata is published including the publication date and the creator. In that way consumers can track the origin of particular data fragments. Regarding the rights that apply to the dataset, there are three kinds “Rights Reserved – Free Access”, “Rights Reserved – Paid Access” and “Restricted Access”. The data provider selects among them the one that applies to his/her dataset during the metadata mapping process. The rights are represented in the RDFized version of the metadata by using the “dc:rights” property, having one of the above values as filler, and also by using the property edm:rights, taken from the Europeana Data Model1, together with the corresponding Europeana rights.

4.4 Linking of EUscreen Resources

As already mentioned, Linked Data is simply about using the Web to create typed links between data from different sources, including the social web. Therefore after the RDF representation of the EUscreen content and related metadata, links to other resources had to be established. There are two distinct linking cases of interest for the scope of a cultural heritage aggregation repository like EUScreen, those between the internal resources originating from providers' data sources and ones connecting to external repositories. External RDF links are crucial for the Web of Data as they are the glue that connects data islands into a global, interconnected data space [5].

As mentioned earlier in this publication, users want to consume more AV content. At the same time it is important to convince content providers to submit more material. It is expected that this objective can be achieved by taking benefit of the association of Linked Open Data with social network tagging and recommendations.

For the case of internal linking, specific elements of the harvesting schema that relate items were used. As such, the value of the harvesting schema's element isRelatedToItem is an EUscreen item identifier. Respectively, in the RDF representation the EBUCore property isRelatedTo was used having as range the resource of the specific item. Furthermore, additional internal linking was implemented for the countries, the actors and the organizations. As mentioned in the previous section, URIs were created for them that are used as the object of a triple. For example, the Netherlands resource can be the object of a triple having as predicate the EBUCore property "createdIn" and as subject the video resource.

The resources implemented for the countries were also externally linked, since information about countries is served by many data sources. For the creation of external links DBpedia (http://dbpedia.org/About) has been used. The names of the local dataset countries were compared using SPARQL [7] to names of the countries resources served by DBpedia. After the establishment of a link to DBpedia additional linked data resources are discovered by retrieving the links of the link. In that way the EUscreen repository is linked to more datasets of interest other than DBpedia, like Freebase, Eurostat and NYTimes.

In addition to these links, new external links were extracted from the video summaries by using DBpedia spotlight, a tool that can extract resources from free text (http://dbpedia.org/spotlight). In the summary description of a video quite often names of persons are mentioned that either participate in the video or the video involves them in a way. By using spotlight, resources for such cases were extracted, providing very useful additional information about the video and therefore improving its searchability.

4.5 Deployment of the linked open data pilot

So far we have described the main issues regarding the transformation of the harvested and homogenized XML items to RDF and their internal and external linking. However, for fulfilling the 4 main Linked Data principles [2] we have deployed the EUscreen linked open data pilot available at http://lod.euscreen.eu. This pilot was first deployed on the 29th of Semptember 2011 and since then it has been visited by more than 1000 unique visitors around the world (info from google analytics).

Both the machine (RDF) and the human (HTML) understandable information (a detailed description of the HTML representation of the items, that is given through the EUscreen portal - http://euscreen.eu/) are in operation. More specifically, the aggregated and transformed metadata by MINT are converted to RDF and published as Linked Open Data weekly. At the moment the pilot holds 22.190 programme resources while the total amount of resources is 114.142. Among the total resources, 13.158 are made for persons individuals referring to the contributor of the programme while 582 are made for countries - linked to 1439 externals- and 22 for languages –linked to 63 externals. In addition by using spotlight, 1490 person resources are extracted to which links are made from 1133 programmes' English summaries.

Finally, the data are uploaded to 4store (http://4store.org/) - a purpose built database - in order to provide SPARQL access to them making their consumption easier. In that way the data can be consumed through the SPARQL endpoint (http://oreo.image.ece.ntua.gr:10999/sparql/) and also by using the web interface of the 4store repository (http://oreo.image.ece.ntua.gr:10999/test/).

5. Conclusions

The Euscreen project is an operational portal for accessing broadcaster and national audiovisual library archives. It has now passed the level of proof of concept and more content providers are invited to join. RDF and Linked Open Data have been chosen has the preferred way to exhcnage metadata with Europana and provide access to a vast bank of content.

In this context the use of Linked Open Data is a natural step forward using available resources from e.g. DBPedia but also from social networks.

From a broadcaster point of view, the perspective is highly attractive but must be measured against the cost of operation. Solutions must be developed to automatise the association of resources using LOD mechanisms.

Other requirements such as editorial quality of the metadata, relevance of social network recommendations and tagging, as well as time persistence will require the utmost attention.

6. References

[1] T. B. Lee, "Cool URIs don't change," 1998. [Online]. Available: http://www.w3.org/Provider/Style/URI.html

[2] T. B. Lee, "Linked Data - Design Issues", 2006. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.htm

[3] T. Buerge, J-P. Evain, and P-A. Champin, "W3C Media Annotation Working Group RDF ontology" (ma-ont), 2011, [Online]. Available: http://www.w3.org/ns/ma-ont.rdf (see also http://www.w3.org/TR/2011/PR-mediaont-10-20111129/)

[4] J-P. Evain, "EBU Core Metadata Set EBU", 2009. [Onine]. Available: http://tech.ebu.ch/docs/tech/tech3293v1_3.pdf

[5] T. Heath, and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space" 2011 (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool DOI: 10.2200/S00334ED1V01Y201102WBE001 ISBN: 9781608454303 (paperback) ISBN: 9781608454310 (ebook).

[6] L. Kaye, "Content Selection and Metadata Handbook" , 2011. [Online]. Available: http://blog.euscreen.eu/wp-content/uploads/2010/10/Content-Selection-and-Metadata-Handbook_public.pdf

[7] E. Prud'hommeaux and A. Seaborne "SPARQL Query Language for RDF - W3C Recommendation", 2008, [Online] Available: http://www.w3.org/TR/rdf-sparql-query/

[8] L. Sauermann and R. Cyganiak, R. "Cool uris for the semantic web - w3c interest group note", 2008. [Online] Available: http://www.w3.org/TR/cooluris/

[9] G. Schreiber, "Metadata Models, Interoperability Gaps, and Extensions to Preservation Metadata Standards" 2010.

[10] B. Scott, "Gordon Park's conversation theory: a domain independent constructivist model of human knowing". In: Foundations of Science 6(4):343-360. National Centre for Biotechnology Information, 2001. [Online] Available: http://www.ncbi.nlm.nih.gov

[11] M. Welshons, "Our Cultural Commonwealth" The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. Connections, December 15, 2006. [Online] Available: http://cnx.org/content/col10391/1.2/