Semantic Web 1 (2009) 1–5
IOS Press

Linked European Television Heritage

Editor(s): Pascal Hitzler, Kno.e.sis Center, Wright State University, Dayton, OH, USA; Krzysztof Janowicz, University of California, Santa Barbara, USA
Solicited review(s): Aidan Hogan, DERI, University of Galway, Ireland; Michael Hausenblas, DERI, University of Galway, Ireland; Emanuele Della Valle, Politecnico di Milano, Italy

Nikolaos Simou [a,*], Jean-Pierre Evain [b], Nasos Drosopoulos [a] and Vasillis Tzouvaras [a]
[a] *School of Electrical and Computer Engineering, National Technical University of Athens, Zographou 15780, Greece*
E-mail: {nsimou,ndroso,tzouvaras}@image.ntua.gr
[b] *Metadata and Workflow Processes Groups EBU, Switzerland*
E-mail: evain@ebu.ch

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 its creation was 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: Linked Open Data, Metadata Ingestion, TV on the Web

1. Introduction

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. Following the increasing support at the European level, as well as the emerging involvement of major IT companies, there has been 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[1] as a unique point of access to European Cultural Heritage, has been one of the major achievements of these efforts, while a recent trend in the area is the publication of cultural datasets as Linked Open Data[2].

Television content is regarded as a vital component of Europe's heritage, collective memory and identity - all our yesterdays - but it remains difficult to access. Even more than with the museum and library collections, the dealing with copyrights, encoding standards, costs for digitization and storage make the process of its aggregated and contextualized publishing on the Web extra challenging.

The EUscreen project [3] aims at the creation of a representative collection of television programs, secondary sources and articles permitting in this way access to students, scholars and the general public. However, providing access to large integrated digital collections of cultural heritage objects is a challenging task involving the resolution of various issues. Firstly, the

*Corresponding author. E-mail: nsimou@image.ntua.gr
[1] http://www.europeana.eu

[2] http://www.openimages.eu,http://semanticweb.cs.vu.nl/lod/am,http://www.europeana.eu/portal/thought-lab.html
[3] http://euscreen.eu/

aggregation of metadata together with a harmonizing process - since different content providers adopt different types of models - must be considered. 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 their consumption and re-use in a machine understandable manner.

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. 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 (see Fig. 1).

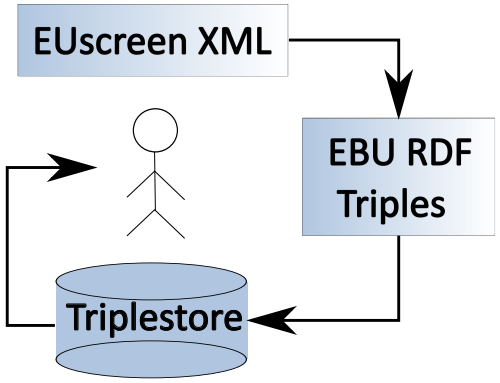

Fig. 1. The Overall Architecture

The content providers of EUscreen have been using various collections and content management systems that stored and exported different types of knowledge in a range of metadata models. In order to achieve semantic interoperability within the aggregation and with external repositories, a harvesting schema was implemented based on EBU Core[4] [3], which is an established standard in the area of audiovisual metadata. The MINT aggregation platform[5] was used for the ingestion and transformation of the metadata. MINT is a suite of web services that facilitate the mapping and transformation of providers' proprietary, legacy or standardized metadata to a reference representative model.

[4] http://tech.ebu.ch/lang/en/MetadataEbuCore http://www.ebu.ch/metadata/ontologies/ebucore/

[5] Metadata Interoperability Services - http://mint.image.ece.ntua.gr/

Following to the transformation of the content's metadata, was the establishment of a Linked Open Data publication procedure. In order to achieve this objective, the conversion of the harvested metadata to RDF - using an expressive data model - was required. In our case, the EBU Core ontology was the most appropriate to guide this semantic transformation. Finally, internal and external linking of the EUscreen content has been performed and the resulting repository was made accessible through a SPARQL endpoint.

The rest of the paper is organized as follows. The first section shortly introduces the EUscreen content. The next two sections present the MINT platform that was used for the aggregation and transformation of the metadata, and the procedure followed for their publication as Linked Open Data, respectively. Finally, we conclude by presenting the significance of the dataset along with its known shortcomings and by discussing the future directions and the ongoing work.

2. The EUscreen Content

The EUscreen project aims to create access to over 30,000 items of programme content and information while its consortium comprises of 20 collection owners, technical enablers, legal experts, educational technologists and media historians of 20 countries.

Every programme is described with explanatory information including the title and the series title in both the original and the English language, the genre, the subject, the provider, the dates of production and broadcast and a summary in English. In addition, technical information such as the aspect ratio, the duration, the type of sound and the color types for every programme is provided.

In collaboration with leading television historians EUscreen has defined a content selection policy[6], divided into three strands:

- Historical Topics: 14 important topics in the history of Europe in the 20th Century (70% of content);
- Comparative Virtual Exhibitions: two specially devised topics that explore more specialized aspects of European history in a more comparative manner (10% of content - include documents, stills, articles);
- Content Provider Virtual Exhibitions: Each content provider selects content supported with other

digital materials and textual information on subjects or topics of their own choosing (20% of content).

EUscreen has written a set of guidelines regarding management of intellectual property rights. The copyright situation of each and every item has been investigated prior to uploading.

3. Metadata Aggregation and Transformation

This section introduces MINT [6], the system that has been used for the metadata aggregation and transformation. MINT is an open source, web based platform for the ingestion, mapping and transformation of metadata records. Interoperability is achieved through the use of well-defined metadata models - like the EUscreen harvesting schema that was used in this 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. Registered users can start by uploading their metadata records in XML or CSV serialization, using the HTTP, FTP and OAI-PMH protocols. 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. After that, users can transform their selected collections to the desired target schema.

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 aggregated metadata transformed to the EUscreen harvesting schema (XML to RDF), and we proceed to the way the produced resources are linked to external sources.

4.1. Semantic Representation of the EUscreen Content

For instantiating the EUscreen data as Linked Data resources, a machine readable representation in RDF was necessary. Hence, a decision that was made in accordance to the items described in the homogenized XML documents, was the selection of the EBU Core ontology[7] as the vocabulary used for the RDF representation. The EBU Core ontology is an RDF representation of the EBU Class Conceptual Data Model (CCDM). CCDM defines a structured set of audiovisual classes (e.g. groups of resources, media resources, parts, media objects but also locations, events, persons and organizations). The EBU Core ontology also defines the semantic relationships (objectProperties) between these classes as well as the properties (dataProperties) characterizing these classes. A lot of the knowledge gathered in the EBU CCDM and EBU Core RDF was used to develop the W3C Media Annotation ontology[7] (W3C MAWG). Reciprocally, EBU Core RDF is linked to W3C MAWG and has implemented the RDF modeling options chosen by it, in a subsequent version.

The next step, after the selection of the appropriate vocabularies for the RDF representation of the EUscreen content, was the creation of resources for the described programmes. In other words, the fulfillment of the first principle of Linked Data [1] that states the use of URIs for things. There are various guidelines for creating cool URIs for the semantic web [2,10] and the two basic characteristics they must have are uniqueness for every item, and consistency. 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.euscreen.eu`) and the assigned unique identifier of the item is part of the URI. The corresponding set of URIs for an example of an EUscreen item are shown below.

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

[6] `http://mint.image.ece.ntua.gr/`

[7] `http://www.w3.org/2008/WebVideo/Annotations/`

– http://lod.euscreen.eu/data/EUS_55F569268ACA42B186682960875F862B.rdf

At this point it must be noted that except for the URIs that are constructed for the unique things described in the dataset (i.e. the videos) additional URIs are made for information shared among the dataset. Such information is the actors, the countries, the subjects, the topics, the organizations and other in the EUscreen dataset. (For example a country can be the location of production of more than one video item.) Therefore new resources have been created for these elements' values without any identifier - and only by using their name - since those are already unique. Hence, in the case of the Netherlands the shared resource constructed is http://lod.euscreen.eu/resource/Netherlands (note that in Fig. 2 the Netherlands resource is both the coverage of the video as well as its location i.e. country of production)

After specifying the method for minting present and future URIs, we proceeded to identify the things described according to appropriate EBU Core classes and properties that would be used for their representation in RDF. More specifically, the type of item, as it is defined in the XML schema can be a document, a video, an audio or a still (i.e. image). Depending on this information the resource created for it can be an instance of the EBU Core classes Document, TVProgramme, RadioProgramme or Image respectively. The additional characteristics of the resources are represented in RDF by using EBU Core properties ranging from typed literals (e.g. original title is represented by ebu:title) to other internal resources (e.g. video publishers are instances of ebu:Organisation). Furthermore, in the case of string literals their language is also provided - allowing the consumers of the EUscreen dataset to perform queries for language specific mash-ups - while whenever required, typed literals are used. Figure 2 illustrates an excerpt of the graph that presents a programme's metadata transformed in RDF[8].

Finally, other recommendations that are very important for the publication of Linked Data and have been considered are the ownership of resources, the licensing and the 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, either one of "Rights Reserved - Free Access", "Rights Reserved - Paid Access" or "Restricted Access" applies to each item. The selection is done by the data provider 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 Model [5], together with the corresponding Europeana rights.

4.2. Linking of EUscreen Resources

As already mentioned, Linked Data is simply about using the Web to create typed links between data from different sources, therefore after the RDF representation of the EUscreen content, 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 among the internal resources originating from providers' data sources and the ones connecting to external repositories.

In the case of internal linking, specific elements of the harvesting schema that relate items have been used. As such, the value of the harvesting schema's element "isRelatedToItem" is an EUscreen item identifier. Respectively, in the RDF representation the EBU Core property "isRelatedTo" was used having as range the resource of the specific item. Furthermore, additional internal linking was implemented for the countries, the actors, the subjects, the topics and the organizations. As mentioned in the previous section, URIs are created for those used as the object of a triple. (e.g. the Netherlands resource can be the object of a triple having as predicate the EBU Core property "hasLocation" and as subject the TV programme resource see Fig. 2)

The resources implemented for the countries and languages are also externally linked, since information on them is served by many data sources. For the creation of external links DBpedia[9] has been used. The names of the local dataset countries and languages are compared, using SPARQL [9], to the names of the

[8] The complete graph can be found at http://lod.euscreen.eu/page/EUS_55F569268ACA42B186682960875F862B.svg while the mapping of all the harvesting schema's elements to the set of properties and classes used for the representation of a programme's metadata in RDF can be found at https://docs.google.com/spreadsheet/ccc?key=0Akruw5a0_oaLdEQyMl85NVQxZ2lmT00wcVU4ZVRJZ0E#gid=8.

[9] http://dbpedia.org

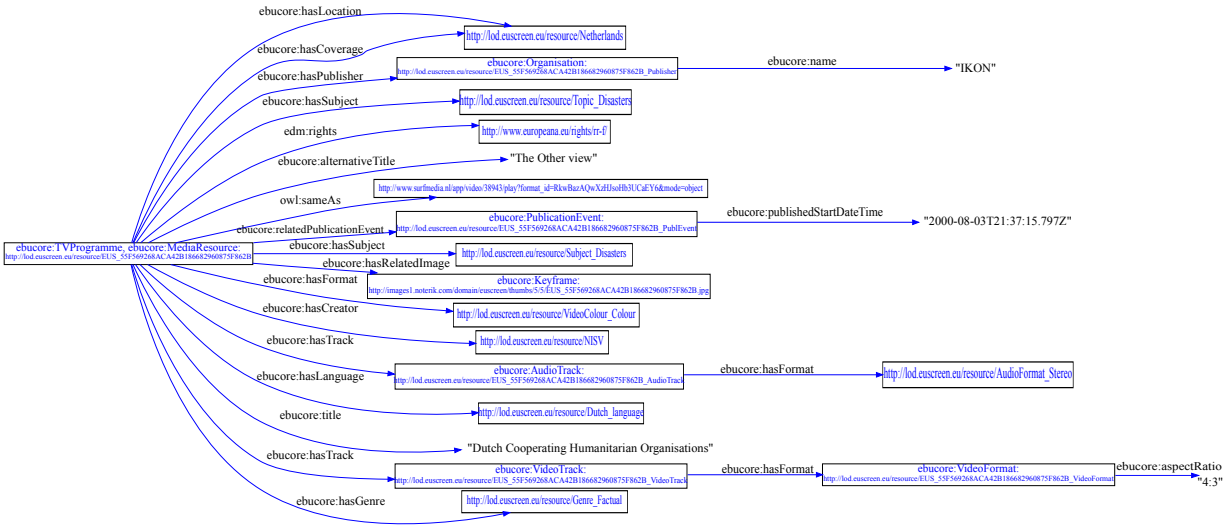

Fig. 2. A graph excerpt of a programme representation in RDF

countries and languages' resources served by DBpedia. After the establishment of a link to DBpedia, additional linked data resources are discovered by retrieving the links of each link. In that way the EUscreen repository is linked to more datasets of interest other than DBpedia, like Freebase[10], Eurostat[11] and NYTimes[12]. We have preferred manual linking from a semi automatic approach like Silk [11] or Limes [8] because in our dataset we had a small number of distinct countries and languages. This fact allowed for the application of manual linking and also gave us the ability to examine the validity of the links established to the external data sources. Therefore, the correctness of all the links created to external data sources is guaranteed. In addition to these links, new external links are extracted from the video summaries by using DBpedia spotlight [13]. 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 are extracted, providing very useful additional information about the video and therefore improving its searchability. Finally, external links were made -whenever possible- to the provider's portals that also serve the programmes.

[10] http://www.freebase.com/
[11] http://eurostat.linked-statistics.org/
[12] http://data.nytimes.com/
[13] http://dbpedia.org/spotlight

Table 1
EUscreen dataset statistics

Dataset Resources	
Programme Resources	41,622
Person Resources	18,995
Countries Resources	586
Languages Resources	46
Subject Resources	1,397
Topic Resources	17
Total Resources	511,816
Linking Statistics	
External Links to Countries Resources	5,940
External Links to Language Resources	618
Programmes from which person resources extracted from summaries	905
Person resources extracted from summaries using Spotilight	1,081
Total External Links	15,036

4.3. 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 [1] we have deployed the EUscreen linked open data pilot available at `http://lod.euscreen.eu`. This pilot was first deployed on the 29th of September 2011 and since then it has been

visited by more than 824[14] unique visitors around the world. Both machine (RDF) and human (HTML) understandable information (the HTML representation of the items is given through the EUscreen portal[15]) are served. More specifically, the aggregated and transformed metadata by MINT are converted to RDF and published weekly as Linked Open Data. Table 1 illustrates the EUscreen dataset statistics after the publication made on the 13th of January 2013. Finally, the data are uploaded to fuseki[16] - a purpose built database- in order to provide SPARQL 1.1[4] access for making their consumption easier. In that way the data can be consumed through the SPARQL endpoint [17].

5. Conclusion

In this paper, we presented 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. Audiovisual content is very popular in web users and the main advantage of EUscreen content is that it is very well annotated and of great significance. By its publication as Linked Data it can be easily consumed, making in that way, the implementation of various applications that use it much simpler. Such applications may serve educational or historical purposes since the majority of the content covers historical topics. Moreover, the EUscreen content is enriched by its linking to external data sources such as the DBpedia, Eurostat, Freebase and NY Times allowing in that way for more expressive search and retrieval. For example a consumer can query for videos about actors that have played the James Bond character by the following query

```
PREFIX ebu:<http://www.ebu.ch/metadata
/ontologies/ebucore/ebucore#>
PREFIX db:<http://dbpedia.org/resource/>
PREFIX db-on:<http://dbpedia.org/
ontology/>

SELECT ?video ?actor
WHERE
{
 SERVICE <http://dbpedia.org/sparql> {
  db:James_Bond db-on:portrayer ?actor.
 }
 ?video ebu:isMentionedinSummary ?actor.
}
```

The content served by EUscreen can prove to be valuable data source for linking to many organizations. The last few years the wed TV has gained great popularity with many channels broadcasting part of or their whole programme to the web along with metadata. The very next step for these channels is the publication of their content as Linked Open Data and the linking to other data sources of relevant content such as EUscreen. We must note at this point that a very important contribution of the EUscreen LOD is the creation of 18.995 resources for persons individuals referring to the contributors of the programme. Different kinds of organizations that could link to EUscreen LOD are cultural institutions like Europeana that has currently published part of its content as LOD.

At the moment one of the main shortcomings of the dataset is its size that can be considered small (41.622 items) compared to other cultural datasets like Europeana ($\sim$ 22 million items). However, having in mind that the type of content served by EUscreen is European rare television programmes dating back to the early days of television, we can say that its size is significant. Furthermore due to EUscreen's success a follow up project has been accepted by the European Commission, hence the dataset size will be at least doubled within the next few years.

Current work includes the further improvement of the MINT services by extending it to offer more advanced normalization and refine functionalities, permitting in that way better results in resource discovery from literals. In addition we intend to further link the EUscreen dataset to cultural data sources like Europeana and also EBU Core ontology to other media specific vocabularies.

[14] info from Google Analytics

[15] http://euscreen.eu/

[16] http://jena.apache.org/documentation/serving_data/index.html

[17] http://lod.euscreen.eu/sparql/

References

[1] T. Berners-Lee. *Linked Data*. W3C Design Issues 27 July 2006, http://www.w3.org/DesignIssues/LinkedData.html.

[2] T. Berners-Lee. *Cool URIs don't change*. W3C Style Guide for online hypertext, http://www.w3.org/Provider/Style/URI.html, 1998.

[3] J.-P. Evain. *EBU – Core Metadata Set*. European Broadcasting Union, http://tech.ebu.ch/docs/tech/tech3293v1_4.pdf, 2009.

[4] S. Harris and A. Seaborne. *SPARQL 1.1 Query Language*. W3C Proposed Recommendation 08 November 2012, http://www.w3.org/TR/sparql11-query/.

[5] A. Isaac and R. Clayphan. *Europeana Data Primer*. Europeana, http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5, 2010.

[6] L. Kaye. *Content Selection and metadata handbook*. EUscreen Deliverable, http://blog.euscreen.eu/wp-content/uploads/2010/10/Content-Selection-and-Metadata-Handbook_public.pdf, 2011.

[7] W. Lee, W. Bailer, T. Bürger, P.-A. Champin, J.-P. Evain, V. Malaisé, T. Michel, F. Sasaki, J. Söderberg, F. Stegmaier, and J. Strassner. *Ontology for Media Resources 1.0*. W3C Recommendation 09 February 2012, http://www.w3.org/TR/2012/REC-mediaont-10-20120209/.

[8] A.-C. N. Ngomo and S. Auer. Limes – a time-efficient approach for large-scale link discovery on the web of data. In T. Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, pp 2312-2317*. IJCAI/AAAI, 2011.

[9] E. Prud'hommeaux and A. Seaborne. *SPARQL Query Language for RDF*. W3C Recommendation 15 January 2008, http://www.w3.org/TR/rdf-sparql-query/.

[10] L. Sauermann and R. Cyganiak. *Cool URIs for the Semantic Web*. W3C Interest Group Note 31 March 2008. http://www.w3.org/TR/2008/NOTE-cooluris-20080331/.

[11] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Silk – A Link Discovery Framework for the Web of Data. In *Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, April 20, 2009, CEUR Workshop Proceedings, ISSN 1613-0073, online CEUR-WS.org/Vol-538/ldow2009_paper13.pdf.*, 2009.