Publishing Europe's Television Heritage on the Web

Johan Oomen[1], Vassilis Tzouvaras[2],

[1] Nederlands Instituut voor Beeld en Geluid, Sumatralaan 45, Hilversum, the Netherlands
joomen@beeldengeluid.nl
[2] National Technical University of Athens, Iroon Polytexneiou 9, 15780 Zografou, Greece
tzouvaras@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 man 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. The main goals of EUscreen are to (i) develop a state-of-the-art workflow for content ingestion, (ii) define content selection and IPR management methodology, and (iii) provide a front-end that accommodates requirements of several user groups.

Keywords: Information integration, TV on the Web, Metadata Interoperability, Linked Open Data, Visualization, Europeana

1 Introduction

Providing access to large integrated digital collections of cultural heritage objects is a challenging task. Multiple initiatives exist in different domains. For example, Europeana manages a state-of-the art technical infrastructure to manage the ingestion and management of data from a wide variety of content providers. It aims to give access to all of Europe's digitised cultural heritage by 2025. Europeana focuses on two main tasks (i) to act as a central index, aggregating and harmonising metadata following a common data model [1], and (ii) to provide persistent links to content hosted by trusted sources. The portal currently provides access to 15 million objects, primarily books and photographs; audiovisual collections are underrepresented. However, recent analysis of query logs from the Europeana portal indicated users have a special interest for this type of content. Television content is regarded 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 makes the process of its aggregated and contextualized publishing on the Web extra challenging.

In this paper, we will focus in outlining the ingestion workflow; the projects' main technical achievement. In Section 2, we outline the motivation of our work. In Section 3, we elaborate on different components that make up the ingestion workflow.

2 Motivation

The main motivation for our work is to overcome the current barriers and provide a 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 The multidisciplinary nature of the EUscreen project is mirrored in the composition of the socio-technical nature of the consortium; comprising of 20 collection owners, technical enablers, legal experts, educational technologists and media historians of 20 countries. EUscreen represents all major European television archives and acts as one of the key domain aggregators providing content to Europeana.
Several public reports on our work can be downloaded from the project blog. This paper reports on the results of the work performed over the past one and a half years, leading up to the launch of the first version of the portal. Notably, we analyse the design decisions from a Web Science perspective; zooming in on the interplay between user requirements, technical possibilities and societal issues, including intellectual property rights. We will show how EUscreen contributes to a so-called 'Cultural Commonwealth' [2] that emerges by bringing content from memory institutions and the knowledge of its heterogeneous constituency together.
Conceptually, EUscreen is built on the notion that knowledge is created through conversation [3]. Hence, ample attention is given to investigating how to stimulate and capture knowledge of its users. Combining organizational, expert and amateur contributions is a very timely topic in the heritage domain, requiring investigation of the technical, organizational and legal specificities.

The goals of the project are to (i) develop a state-of-the-art workflow for content ingestion, (ii) define content selection and IPR management methodology (35.000 items will be made available), and (iii) design and implement a front-end that accommodates requirements from several user groups. To reach these goals, close cooperation between the different stakeholders in the consortium is essential. For example, the selection policy needs to take in to account the available content, wishes from media historians and the copyright situation. The workflow will need to study the existing metadata structures, should support aggregation by Europeana and provide support for multilingual access.

2.1 Define content selection methodology

In collaboration with leading television historians EUscreen has defined a content selection policy [4], divided into three strands:
1. Historical Topics: 14 important topics in history of Europe in the 20th Century (70% of content);
2. Comparative Virtual Exhibitions: two specially devised topics that explore more specialised aspects of European history in a more comparative manner (10% of content – include documents, stills, articles);
3. 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 guideless regarding management of intellectual property rights. The copyright situation of each and every item is investigated prior to uploading.

2.2 The Front-end

Representatives of the four primary user groups, e.g. secondary education, academic research, the general public and the cultural heritage domain were consulted in order to define user requirements and design front-end functionality. The main challenge for the portal's front-end is to include advanced features for specific use cases without overwhelming the users with a complex interfaces. The Helsinki University of Arts and Design adapted a component-based conceptual model that accommodates this requirement (Figure 1.)

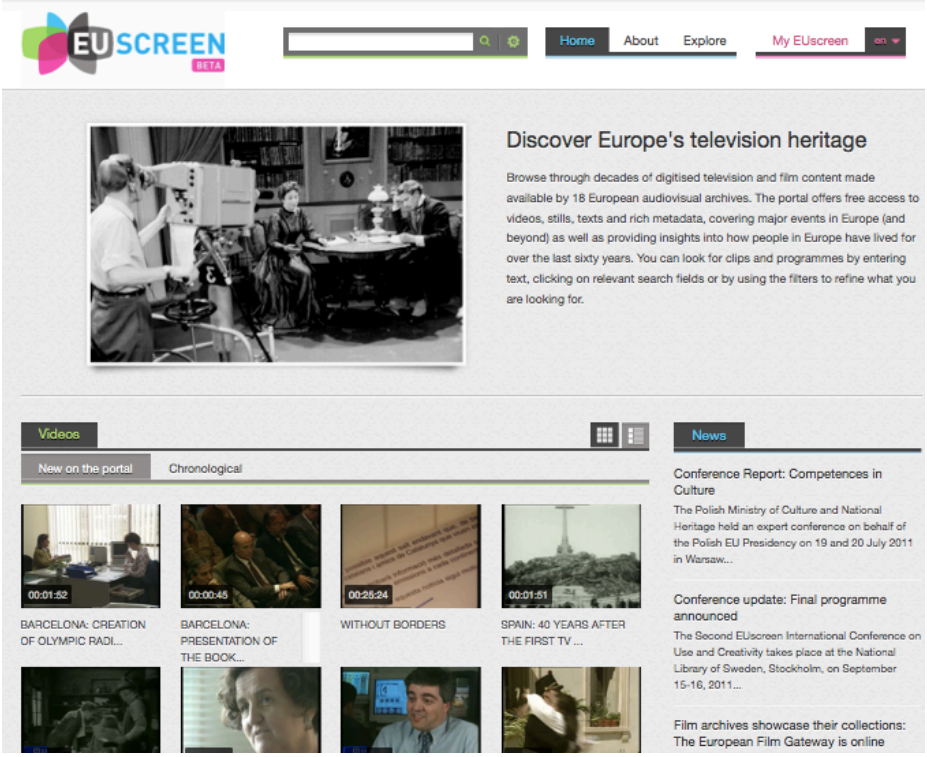

Fig. 1. EUscreen Homepage.

Implementation of the front-end services is not done in the traditional way using serverside programming language like php, java or asp. EUscreen implemented a

'server-less' front-end APIs where a javascript/flash proxy system handles the communication with the back-end services. The result will be a front-end system that can be 'installed' on any plain html web server without any need for server-side technologies. This means it can be hosted and moved to any location or multiple locations. It also means partners can use these APIs to integrate parts of the functionality in their own intranet and internet systems using simple 'embed' ideas. This method is gaining more ground, for example companies like Google who provides these types of APIs for services like Google Maps.

3 Metadata Ingestion and Video Playout

The technical standards enabling interoperability form an important dimension of the technical achievements. In order to achieve semantic interoperability, a common automatic interpretation of the meaning of the exchanged information is needed, i.e. the ability to automatically process the information in a machine-understandable manner. The first step of achieving a certain level of common understanding is a representation language that exchanges the formal semantics of the information. Then, systems that understand these semantics can process the information and provide web services like searching, retrieval.

Many different metadata schemas or in a broader sense, sets of elements of information about resources, are being used in this domain, across a variety of technical environments and scientific disciplines. EUscreen has developed an ingestion mechanism providing a user friendly environment that allows for the extraction and presentation of all relevant and statistical information concerning input metadata together with an intuitive mapping service that uses the EUscreen Metadata schema, and provides all the functionality and documentation required for the providers to define their crosswalks. The workflow (Figure 2) consists of four phases, each responsible for specific services to ensure the quality of the ingestion process:

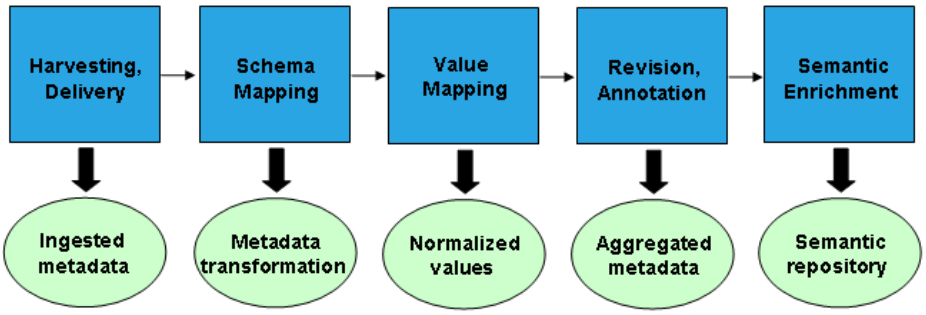

Figure 2. Metadata Ingestion Workflow

The Workflow consists of five steps. The first is *harvesting/delivery*, which refers to collection of metadata from content providers through common data delivery protocols, such as OAI-PMH, HTTP and FTP. The service is implemented as a web

service, where authentication is required to perform a series of tasks that correspond to work flow steps. The harvesting service is an application written in the Java and hosted on a web server by the Tomcat servlet engine. Data is imported into a PostgreSQL database in xml format. Once uploaded, the xml structure is parsed and represented in a relational database table.

Second is the *Schema Mapping* that aligns harvested metadata to the common reference model. A graphical user interface assists content providers in mapping their metadata structures and instances to the EUscreen metadata model, using an underlying machine-understandable mapping language. It supports sharing and reuse of metadata crosswalks and establishment of template transformations.

The next step is *Value Mapping*, focusing on the alignment and transformation of a content provider's list of terms to the authority file or external source introduced by the reference model. It provides normalisation of dates, geographical locations or coordinates, country and language information or name writing conventions.

Revision/Annotation, being the fourth step, enables the addition of annotations, editing of a single or group of items in order to assign metadata not available in the original context and, further transformations and quality control checks (e.g. for URLs) according to the aggregation guidelines and scope.

Finally, the *Semantic Enrichment* step focuses on the transformation of data to a semantic data model, the extraction and identification of resources and the subsequent deployment of an RDF semantic repository.

3.1 EBUcore, Solr and Multilinguality

In order to achieve semantic interoperability with external web applications, EUscreen metadata are exported in EBUcore [5], which is an established standard in the area of audiovisual metadata. An extensive evaluation of alternative standards in this area (MPEG7, DCMI, TV Anytime) has been conducted [6] before choosing the EBUcore. EBUCore has been purposefully designed as a minimum list of attributes to describe audio and video resources for a wide range of broadcasting applications including for archives, exchange and publication.

It is also a Metadata schema with well-defined syntax and semantics for easier implementation. It is based on the Dublin Core to maximise interoperability with the community of Dublin Core users. EBUCore expands the list of elements originally defined in EBU Tech 3293-2001 for radio archives, also based on Dublin Core. The metadata is stored in RDF format to improve the search functionality and enable the alignment with external resources.

In EUscreen portal, retrieval is performed using the Solr framework. Solr is an open source enterprise search platform from the Apache Lucene project. Its major features include powerful full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document handling. Providing distributed search and index replication, Solr is highly scalable. Solr uses the Lucene Java search library at its core for full-text indexing and search, and has REST-like HTTP/XML and JSON APIs that make it easy to use from virtually any programming language. Solr's powerful external configuration allows it to be tailored to EUscreen retrieval

application without Java coding, and it has an extensive plugin architecture for more advanced customization.

Finally, EUscreen has created a SKOS multilingual thesaurus (15 languages) based on the subject terms of IPTC standard and the geographical places of GeoNames. The baseline of the thesaurus is the *Descriptive NewsCodes vocabulary* from The International Press Telecommunications Council [7]. Translations are made with a software solution for the creation and administration of multilingual thesaurus cslled Thesaurix, as licensed by Joanneum Research. The thesaurus supports multilingual retrieval services and links to open data resources that could be used for enrichment and to contextualise the collection.

3.2 Video Playout

EUscreen requires content providers to provide MPEG 4 part 10 (normally known as H.264). EUscreen advises to encode in a bit rate between 500 and 1000 kb/sec, as this resembles SD quality video. Since the client playback method will be a Flash player with h.264 streaming, EUscreen demands that providers have streaming servers that are capable stream videos to a Flash client. In practice this means using one of the available Flash streaming servers.

This will leave room for the content providers themselves to add HTML5 or Silverlight server programs to create an 100% coverage of the possible technologies.

EUscreen supports four scenarios:

1. Content provider transcodes and files are hosted by service provider Noterik
2. Content provider transcodes and the content provider hosts
3. Noterik transcodes and Noterik hosts
4. Noterik transcodes, and the content provider hosts

3.3 The Mapping Tool

Metadata mapping is a crucial step of the ingestion procedure. It formalizes the notion of `crosswalk' by hiding technical details and permitting semantic equivalences to emerge as the centrepiece. It involves a graphical, web-based environment where interoperability is achieved by letting users create mappings between input and target elements. User metadata imports are not required to include the adopted XML schema. Moreover, the set of elements that have to be mapped are only those that are populated. As a consequence, the actual work for the user is easier, while avoiding expected inconsistencies between schema declaration and actual usage.

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 (Figure 3). The tree represents the snapshot of the XML schema that the user is using as input for the mapping process. The user is able to navigate and access element statistics for the specific import.

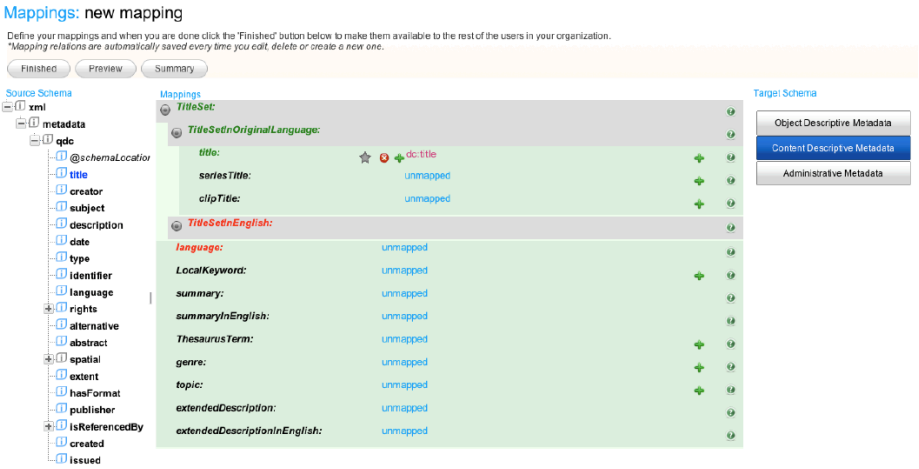

Figure 3. Mapping Interface

The interface provides the user with groups of high-level elements that constitute separate semantic entities of the target schema. These are presented on the right hand side as buttons, which are then used to access the set of corresponding sub-elements. This set is 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 mapping between the input and the target schema, a user has to simply drag a source element and drop it on the respective target in the middle.

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. 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, can be exported and published as a well-defined, machine understandable crosswalks and shared with other users to act as template for their mapping needs. Features of the language that are accessible to the user through actions on the interface include:

- string manipulation functions for input elements;
- 1-n mappings;
- 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).

4 Future Work

The first version of the portal has been launched in August 2011. It is followed by a period of extensive evaluations with end-users. Also, the selection policy will be reviewed. Outcomes of this process will form the basis of the development of the second release, scheduled for early 2012. The major enhancements will be related to the front-end. For instance, EUscreen will support the on-line creation of on so-called virtual exhibitions, consisting of media objects of various archives.

Acknowledgments. EUscreen is co-funded by the European Commission within the eContentplus Programme.

References

1. Isaac, Antoine.: Europeana Data Model Primer. Europeana v1.0 Technical report, http://group.europeana.eu, 2010
2. Scott, B.: Gordon Park's conversation theory: a domain independent constructivist model of human knowing. In: Foundations of Science 6(4):343-360, 2001
3. Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyber infrastructure for the Humanities and Social Sciences.American Council of Learned Societies, 2006
4. Kaye, Linda. D3.1 Content Selection and Metadata Handbook, Euscreen BPN project, euscreen.eu, 2011
5. Evain, Jean Pierre.: European Broadcasting Union Core Metadata, http://tech.ebu.ch/publications, 2009
6. Schreiber, Guus.: D2.2.2 Metadata Models, Interoperability Gaps, and Extensions to Preservation Metadata Standards, PrestoPrime FP7 project, http://www.prestoprime.org/project/public.en.html, 2010
7. Descriptive NewsCodes of International Press and Telecommunication Council. http://www.iptc.org/cms/site/index.html?channel=CH0103, 2010