Date submitted: 03/06/2009

Video active – European television heritage online

Johan Oomen
Netherlands Institute for Sound and Vision
The Hague, Netherlands
joomen@beeldengeluid.nl

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

Marco Rendina
Istituto Luce,
Italy
mrendina@maas.ccr.it

Meeting: **193. Information Technology**

WORLD LIBRARY AND INFORMATION CONGRESS: 75TH IFLA GENERAL CONFERENCE AND COUNCIL
23-27 August 2009, Milan, Italy
http://www.ifla.org/annual-conference/ifla75/index.htm

Abstract:

Many audiovisual archives are in the process of digitising their material and are exploring the new possibilities this brings to publish their content online. This paper provides insight into the background and development of the award winning Video Active Portal (thousands of video items are accessible through www.videoactive.eu).

The Video Active project has used the latest advances in Semantic Web technologies in order to provide expressive representation of the metadata, mapping heterogeneous metadata schema in a common metadata schema based on Dublin Core, and advanced query services. As one of the main outcomes, the project successfully integrated their data to Europeana.

In this paper, the work done in Video Active is presented focusing on the system architecture and the Semantic Web technologies used.

Keywords: *Semantic Web; European Digital Library; Audiovisual Archives; Streaming media; Open Archives Initiative.*

1. Introduction: online access to audiovisual heritage

The greatest promise of the internet as a public knowledge repository is to create seamless access for anyone, anywhere, to all knowledge and cultural products ever produced by mankind. Mainly due to increased bandwidth availability, web sites offering online video material have managed to mature and in a short period have become extremely popular. Web sites like YouTube, MySpace, Revver and many others show how the idea of making and manipulating images (once mostly the preserve of professionals) has been embraced as a way of broadcasting who we are to anyone prepared to watch.

It is evident that the potential for releasing material from audiovisual archives online is enormous. However, from the many millions of hours in these archives online a few percent can be found online. Many of the existing online services are based on user generated content. And if professional content is offered (i.e. Joost, Miro, Blinkx) the focus is rather on recent material.

Audiovisual archives need to overcome several obstacles before they can set up meaningful online services. These include: managing intellectual property rights, technological issues concerning digitisation and metadata standardisation and issues related to the way the sources are presented to users. The latter is even more challenging if the aim is to present material from several countries in a structured way, in fact the starting point of the Video Active project.

The main challenge of Video Active is to remove the main barriers listed above in order to create multilingual access to Europe’s television heritage. Video Active achieves this by selecting a balanced collection of television archive content, which reflects the cultural and historical similarities and differences of television from across the European Union, and by complementing this archive content with well-defined contextual metadata. Video Active is invited member of EDLnet, the network was initiated in 2006 to built consensus to create the European Digital Library. Video active will be made available though the Europeana.eu portal.

This paper provides an insight into the background and development of the Video Active Portal (www.videoactive.eu) which offers access to television heritage material from 14 archives across Europe. The Video Active project, a content enrichment project under the eContentPlus programma, has used all the latest advances of the Semantic Web technologies in order to provide expressive representation of the metadata, mapping heterogeneous metadata schema in the common Video Active schema, sophisticated query services, interactive presentation modes (e.g. timeline). Using these technologies, Video Active is fully compliant with the Europeana's interoperability specifications.

The Video Active system comprises of various modules, all using web technologies. The whole workflow from annotating, uploading material, transcoding material, keyframe extraction, metadata storage and searching is managed by these components. Video Active provides multilingual annotation, search and retrieval of the digital assets using the ThesauriX technology. ThesauriX is a web-based multilingual thesauri tool based on the IPTC standard. The system also exploits Semantic Web technologies enabling automation, sophisticated query services (based on the SPARQL standard) and semantic interoperability with other heterogeneous digital archives. The metadata have been represented using the Resource Description Framework (RDF) and the Simple Knowledge Organizational System (SKOS) and are stored in the Sesame semantic metadata repository . The use of Semantic Web technologies enables light reasoning services (use of implicit knowledge through subsumption and equivalence relations), merging/aligning metadata from

heterogeneous sources and sophisticated query facilities using the SPARQL RDF query language. Additionally, relational databases have been used to speed up some processes where semantic processing is not required. Finally, the Video Active metadata are public and ready to be harvested using the OAI-MPH technology.

2. Video Active

Video Active has created a pool of television archive content (10.000 video's) and contextual data (articles, stills, program guides), representing national and cultural specificities of different European countries over a range of themes and historical events. Contributing archives include: BBC (UK), INA (FR), DR (D), DW (D), ORF (AT), NAVA (HU), Sound and Vision (NL) and many others.1 The portal supports various textual search modes as well as faceted, thematic and timeline-based browsing. (see Figure 1.)

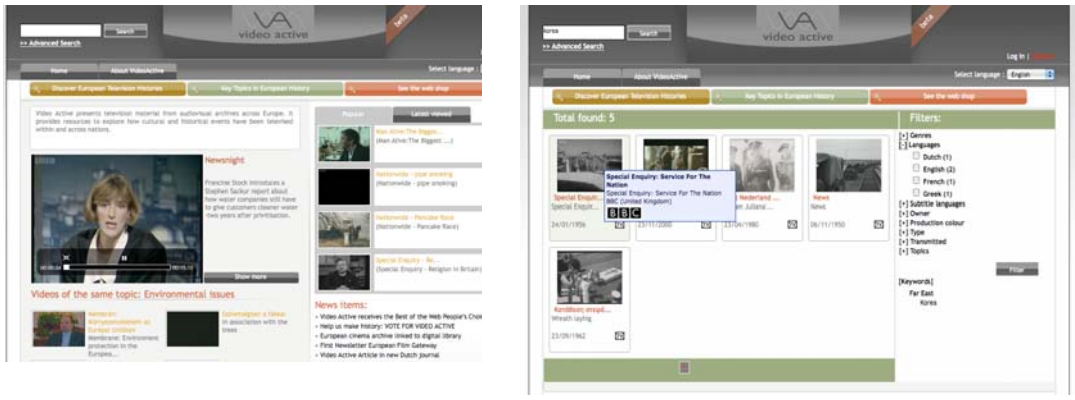

FIG 1. The Video Active homepage and results page

2.1 Defining requirements for Video Active

The demand for access to audiovisual content online has been growing, in a number of distinct sectors: education the general public and the heritage sector. For instance, digitization of archive content transforms cultural heritage into flexible 'learning objects' that can easily be integrated into today's teaching and learning strategies. These user groups have specific expectations and profiles, and the Video Active project had to understand and encompass these to ensure user satisfaction and revisits. Surveys, interviews and desk research have been executed in the initial stages of the project. The resulting insights in user requirements became fundamental to define the technical specifications and hence the technical architecture. Usability tests have been executed on the two consecutive releases of the portal. The excellence of the portal was acknowledged during the Museums and the Web conference 2009, where Video Active won the Best of the Web award.

2.2 The high level architecture:

The Video Active system comprises of various web modules. The whole workflow from annotating, uploading material, transcoding material, keyframe extraction, metadata storage and searching is managed by these components. Figure 2 shows the architecture of the Video Active portal. The architecture exploits semantic web technologies enabling automation, intelligent query services and semantic interoperability with other heterogeneous digital archives. In particular, a semantic layer has been added through the representation of its metadata in Resource Description Framework (RDF). The expressive power of RDF enables light reasoning services, merging/aligning metadata from heterogeneous sources and sophisticated query facility based on SPARQL RDF query language. Additionally Relational database and full-text search technologies have been used to store

1 For a complete list: http://videoactive.wordpress.com/the-consortium/

data where semantic information is not required and to improve the quering performance of the overall system. Finally, the Video Active metadata is public and ready to be harvested using the OAI Metadata Harvesting Protocol.

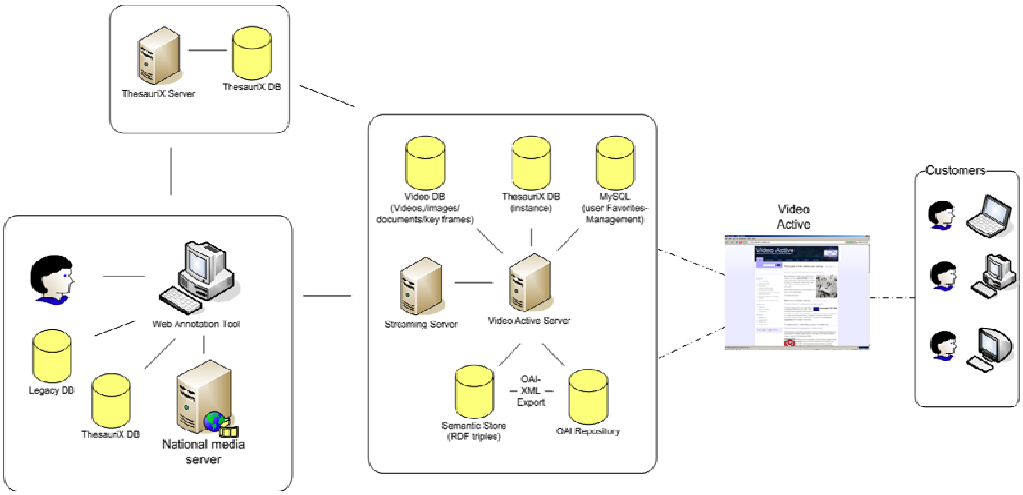

FIG. 2 Video Active: High Level Architecture

2.3 Storing and querying data the semantic way

The Video Active metadata schema is based on the Dublin Core set of metadata schema with additional elements (i.e. Genre, English Title) essential in capturing the specific properties of the resources. (Venetis 2007) The video metadata are generated automatically and are represented in a schema that is based on MPEG-7. In order to enable semantic services, the metadata is transformed in RDF triples and stored in a semantic metadata repository.

The annotation process is either manual or semi-automatic. In the semi-automatic process, the archives export their metadata (the ones that have mappings to the Dublin Core elements) using a common XML schema. The elements that cannot be mapped to the Video Active schema (or are missing from the legacy databases) are inserted manually using the Web Annotation Tool. This tool allows entering and managing the metadata associated with the media and also handles the preparation of the actual content. It contains the Transcoding Factory module that transcodes the original format of the source material to Flash and Windows Media streaming formats, creates low and medium bit rates for the streaming service and performs keyframe extraction for thumbnail creation. The Web Annotation Tool produces an XML file that contains metadata, based on Dublin Core, as well as content encoding and key frame extraction information. The XML is then transformed into RDF triples and stored in the Sesame semantic repository. Sesame is an open source Java framework for storing, querying and reasoning with RDF (Broekstra 2002). It allows storing RDF triples in several storage systems (e.g. Sesame local repository, MySQL database). The use of an ontology language, such as RDF that has formal semantics enables rich representation and reasoning services that facilitates sophisticated query, automation of processes and semantic interoperability. Search and retrieval in VideoActive is performed using a combination of structured RDF queries in SeRQL (optimization of SPARQL query language for Sesame) and full text search queries using the high-performance, full- text search engine library Lucene.

All metadata stored in Sesame are exposed to external systems/archives with the help of an OAI-PMH compliant repository. Europeana, bringing together hundreds of collections of resources throughout Europe, has already indexed the data from the Video Active repository. (see Figure 3.)

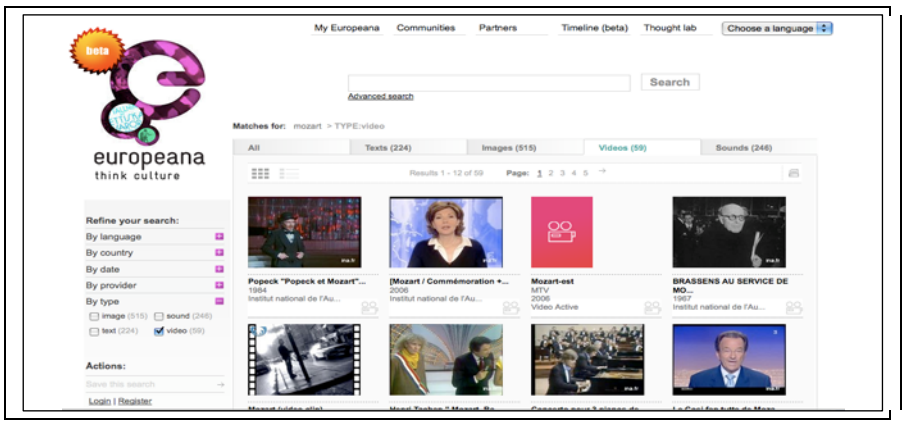

FIG 3. Video Active metadata indexed by Europeana

In order to exchange both the structure and the semantics of the metadata in a machine understandable way, distributed OWL/RDF query mechanisms will be employed in a future release.

2.4. Multilingual access: using SKOS

Eleven languages are supported in Video Active. The Video Active portal offers language support in four distinct ways. Firstly, Video Active has localized interfaces for each of the languages covered. Secondly, key metadata elements (i.e. DC Title, DC Description) are translated in English and thus provide the platform with a monolingual baseline. Thirdly, Video Active is using multilingual controlled vocabularies for the metadata elements Keywords, Genre and Location. The thesaurus from the International Press and Telecom Council is used as baseline for the keyword vocabulary. This 1500-term thesaurus has been translated by the Video Active project in 11 languages. For genres, the ESCORT 2007 EBU System of Classification of Radio and Television Programmes is used and for geographical names, the ISO 3166 English Country Names and Code Elements. Handling the translation of these terms and the export of these terms to machine readable XML is done in a specialized application called ThesauriX. (Janisch 2008) In order to achieve semantic interoperability the thesaurus taxonomy has been transformed into a semantic web language using the Simple Knowledge Organisation System (SKOS) standard. SKOS is a recommendation of the World Wide Web Consortium for the representation of thesaurus taxonomies. The SKOS standard is built on top of the RDF language and can be used to facilitate semantic retrieval of metadata and thesaurus alignment. Finally, a timeline view provides a visual overview over the milestones in the development of television in Europe using the SIMILE framework. (Alonso 2007)

3. Conclusion

In this paper the Video Active system was presented. Video Active has exploited all the recent advances of Semantic Web technologies to provide sophisticated web service using machine-understandable metadata. Semantic Web technologies, such as, RDF, SKOS, OWL and SPARQL have be used for the representation, query, presentation and exchange of the Video Active Metadata. Finally, Video Active metadata has already been included in Europeana using the OAI-PMH technology.

References

- Alonso, Omar, Gertz, Michael and Baeza-Yates, Ricardo (2007): Search results using timeline visualizations. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2007. p. 908
- Broekstra, J., Kampman, A., Harmelen, F. (2002) Sesame: A Generic Architecture for Storing

and Querying RDF and RDF Schema. In: 1st International Semantic Web Conference, Sardinia, Italy

- EBU – TECH 3293-2008 (2008). Core Metadata Set for Archives (EBUCore) Specification v.1.0. Retrieved April 20, 2009, from: http://tech.ebu.ch/lang/en/MetadataSpecifications
- Ireland, G. (2007) Transcoding Internet and Mobile Video: Solutions for the Long Tail, IDC, London
- Gerhard Janisch. (2008) Analyse von Rich Internet Application Frameworks am Beispiel einer Thesaurusverwaltung, Joanneum Research, Graz
- Manjoo, Farhad (2009) Do You Think Bandwidth Grows on Trees? Retrieved April 20, 2009, from: http://slate.com/id/2216162
- Musil, Steven (2009) Online video viewing jumps 34 percent. Retrieved April 20, 2009, from: http://www.cnet.com/profile/stevenmusil/?tag=mncol;txt
- Tim Berners-Lee (2006) Linked Data. http://www.w3.org/DesignIssues/LinkedData.html
- Venetis, Tassos, Anna Christaki and Vassilis Tzouvaras (2007) Video Active: Domain and Upper Harmonizing Ontology. Retrieved April 20, 2009, from http://videoactive.wordpress.com/press/
- Wright, Richard (2007) Annual Report on Preservation Issues for European Audiovisual Collections. Retrieved April 20, 2009, from: http://www.prestospace.org/project/deliverables/D22-8.pdf