Television Heritage and the Semantic Web: Video Active and EUscreen

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

Anna Christaki
National Technical University of Athens, Greece
achristaki@image.ntua.gr

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

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) and its successor EUscreen; initiatives offering access to television heritage material from archives across Europe. 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. The work of Video Active will be continued in the EUscreen Best Practice Network, to be launched in September 2009. In this three-year project, more fine-grained access to video objects will be provided for, using the EBU Core Set of Metadata, released by the EBU metadata working group at the end of 2008. In this report, the authors will firstly outline the work done in Video Active and will elaborate on the architecture of EUscreen that will expand the possibilities for connecting data across cultural heritage organizations in a meaningful way.

Keywords: Semantic Web; European Digital Library; audiovisual archives; streaming media; EBU Core; Open Archives Initiative; Linked Data.

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 and affordability of video cameras, web sites offering online video material have managed to mature and in a short period have become extremely popular. Web sites like YouTube, Vimeo, Blip.tv, Hulu and many others show how the idea of making and manipulating images (previously done exclusively done by professionals) has been embraced as a way of broadcasting who we are to anyone prepared to watch. The most popular site to date, YouTube, was launched in early 2005 and serves 75 billion videos this year. (Manjoo 2009). The number of user generated video uploads per day is expected to go from 500,000 in 2007 to 4,800,000 in 2011. (Ireland 2008) Recent studies indicate that 77 percent of U.S. internet users had viewed online videos in 2008, and that the average online video viewer watched 273 minutes of video. A staggering13 billion videos are served monthly in the U.S. alone. (Musil 2009)

Looking at these numbers, it's evident that the potential for releasing material from audiovisual archives online is enormous. To date, however, from the many millions of hours in these archives a small percentage can be found online. (Wright 2007) Although audiovisual content is now being digitised and some of it is available via the Internet, online access is mainly at the national institutional level, resulting in a wide range of conflicting and competing access routes. Many broadcasters in Europe have their own websites, with a growing collection of video items. Much of it is distributed across programme and subject-related 'pages'. There is little in the way of

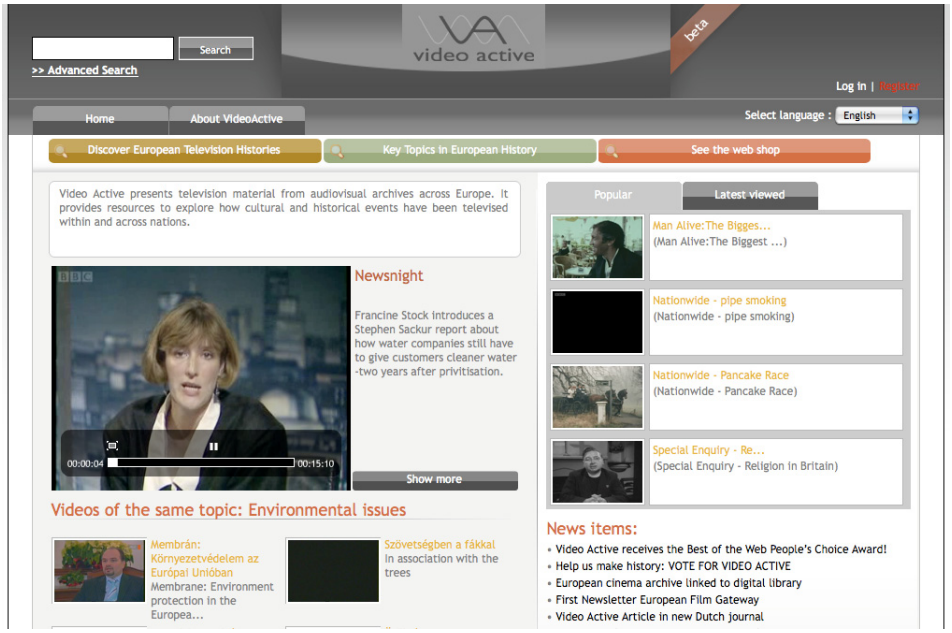
https://doi.org/10.23106/dcmi.952109604

cataloguing for the bulk of material online and also alignment with standards for integrated access (developed in the Digital Library domain) is in its infancy still. 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.

Recent advances in encoding standardisation and semantic web technologies and the widely adopted Dublin Core metadata element set have provided archives with the tools necessary to build meaningful services that provide unified access to archive videos online. A leading example in the audiovisual archive domain, the Video Active portal demonstrates that unified access to archives across different counties is possible. 14 of Europe's leading archives are publishing parts of their digitised assets on this platform, that is also linked to Europeana (www.europeana.eu); the European digital library. The project is supported by the eContentplus programme from the European Commission. The project started in 2006 and will run until September 2009.

2. Video Active

Video Active has created a pool of television archive content (10.000 videos) 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.)



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

https://doi.org/10.23106/dcmi.952109604

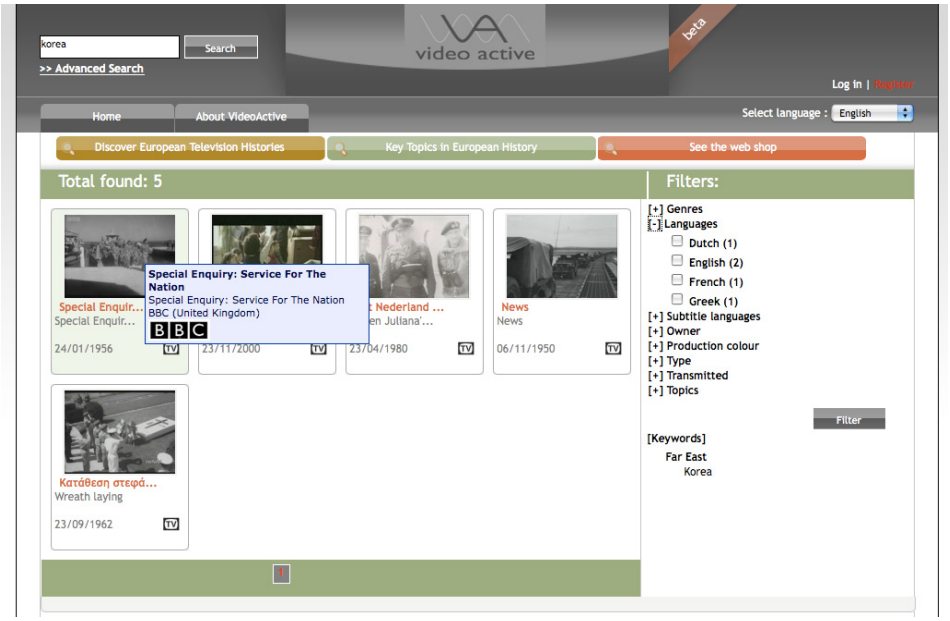


FIG. 1a and 1b. The Video Active homepage and results page

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 consists 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, sophisticated query services (i.e. using the Video Active SKOS thesaurus and ontology reasoning tools) 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 query services based on SPARQL RDF query language. Also, the use of Semantic Web technologies enabled us to transform the Video Active thesaurus in the Simple Knowledge Organisation System (SKOS) standard. In this way, the metadata and the thesaurus terms exist in RDF format in the Sesame repository enabling us to perform combined queries and retrieve implicit knowledge (i.e. not explicitly defined in the RDF store) (see § 2.3). Additionally, relational database and full-text search technologies have been used to store data where semantic information is not required and to improve the querying performance of the overall system. Finally, the Video Active metadata is public and ready to be harvested using the OAI Metadata Harvesting Protocol.

https://doi.org/10.23106/dcmi.952109604

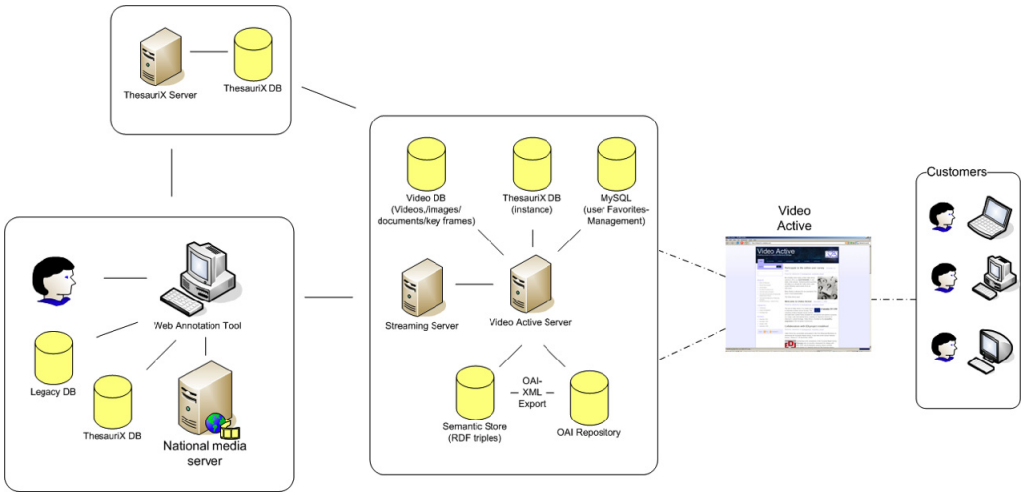


FIG. 2. Video Active: High Level Architecture

2.3. Storing and Querying Data the Semantic Way

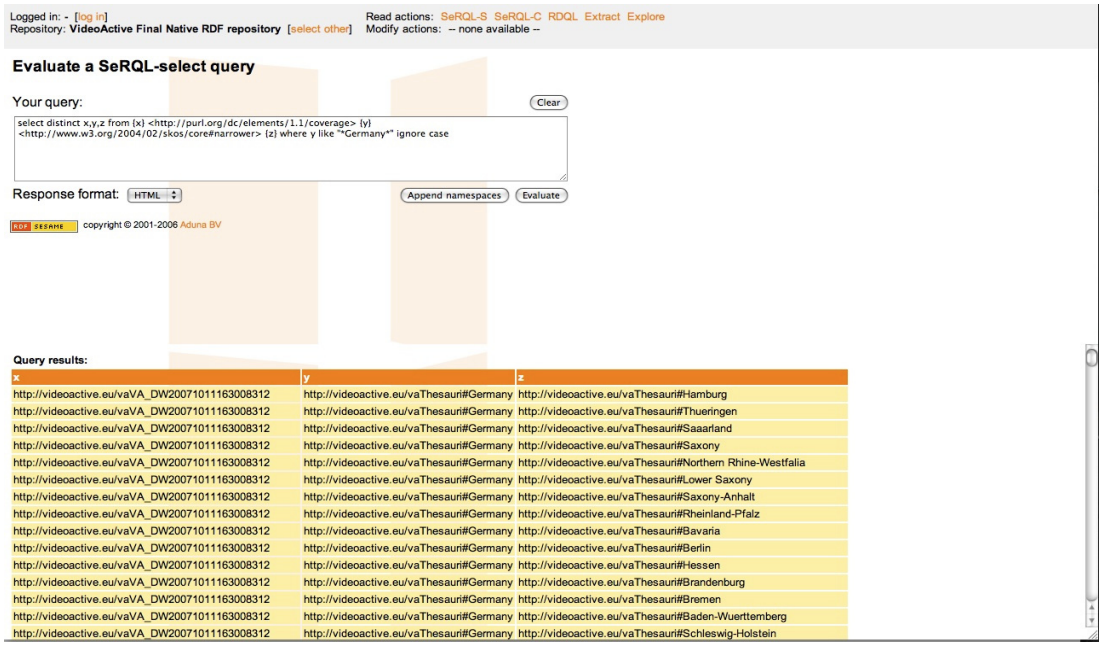
The content providers are using various approaches regarding managing their metadata, so Video Active needed to accommodate heterogeneous source metadata. It was decided to use the Dublin Core set of metadata schema as basis for the Video Active metadata schema. Additional elements, essential to capture the specific properties of the collection (i.e. genre, English title) are added to the Dublin Core element set. (Venetis et al. 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 have created mappings from their in-house catalogues to the Dublin Core elements. The legacy metadata is exported using a common XML schema. Elements that cannot be mapped to the Video Active schema (or are missing from the source databases) are inserted manually using the Web Annotation Tool. This tool also allows entering and managing the metadata associated with the media and also handles the preparation of the actual content. For example, archives are requested to attach keywords from the multilingual subject thesaurus and the geographical reference list in use. The Web Annotation Tool also 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 et al. 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 Video Active 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.

In Figure 3, we illustrate an example of a query that utilizes RDF and SeRQL languages. In this example we show also the benefit of using a SKOS thesaurus and how this can be combined

https://doi.org/10.23106/dcmi.952109604

with the metadata to retrieve richer and better results. The example is illustrated in the interface of the Video Active Sesame repository.



x	y	z
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Hamburg
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Thueringen
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Saaarland
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Saxony
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Northern Rhine-Westfalia
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Lower Saxony
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Saxony-Anhalt
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Rheinland-Pfalz
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Bavaria
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Berlin
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Hessen
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Brandenburg
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Bremen
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Baden-Wuerttemberg
http://videoactive.eu/vaVA_DW20071011163008312	http://videoactive.eu/vaThesauri#Germany	http://videoactive.eu/vaThesauri#Schleswig-Holstein

FIG. 3. An example query in the Sesame repository

In this example query we simulate the user query "bring me items that are dealing with Germany". This query is formulated in SeRQL "bring me items that has "geographical coverage" the term "Germany" and also bring me items of the narrow terms of "Germany". The field "geographical coverage" gets values form the Video Active thesaurus. In this way, instead of getting only the items that have been explicitly defined that have geographical coverage the term "germany", we also retrieve the items from all the narrow terms of Germany (i.e. Hamburg, Thuerlinger, Saxony etc.) using the ontology reasoning services provided the OLWIM Sesame plugin. This is feasible using the SKOS:narrow and SKOS:broader transitive relations to associate the terms in the Video Active thesaurus. Additionally, we can perform conjunctive queries such as "bring me item form Germany that contain the word "Soccer" in the tile field.

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 4.)

https://doi.org/10.23106/dcmi.952109604

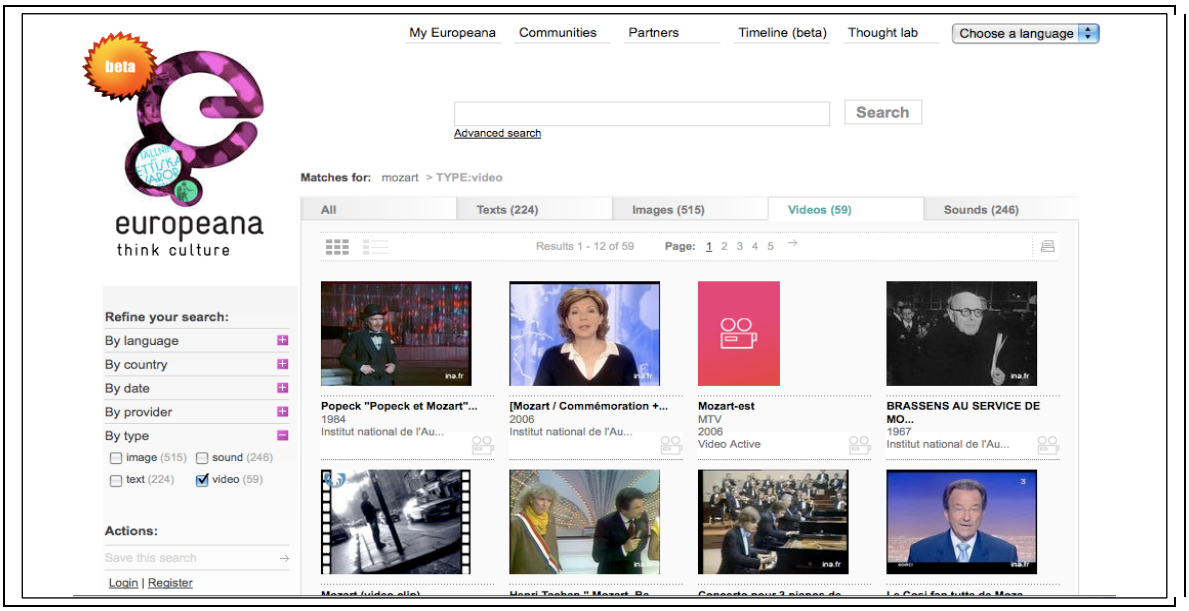


FIG 4. Video Active metadata indexed by Europeana

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, this number will grow to 15 within EUscreen. 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 the (before mentioned) 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 et al. 2007)

3. EUscreen: Introducing Linked Data and EBU Core

The EUscreen Best Practice Network brings together 28 institutions (including 20 archives) from across Europe. To a large extent, EUscreen follows the same workflow as Video Active. Much of the software components will be reused. One of the differences is the fact that EUscreen adopts the EBU Core set of metadata that has been identified as being the minimum information needed to describe radio and television content. (EBU 2008) An XML representation is provided in case this metadata would be implemented in archive exchange projects using the Open Archive

https://doi.org/10.23106/dcmi.952109604

Initiative's Protocol for Metadata Harvesting. Also, the integration with Europeana will change. Since Europeana and EUscreen will have their metadata represented in OWL/RDF and stored in semantic stores, a SPARQL end-point can be created to enable remote access. This way, Europeana will have the ability to directly perform SPARQL queries to the EUscreen repository (and vice versa). The novel features of the EUscreen system compared to other similar systems notably include 1) the combination of Web 2.0 with Semantic Web technologies, 2) the different viewing services, 3) the SPARQL end-point way of making available the metadata in Europeana and generally in the web, 4) the metadata export system for exploiting EUscreen metadata in e-learning applications, and 5) the metadata enrichment system. The high-level architecture of the EUscreen system components is illustrated in Figure 5.

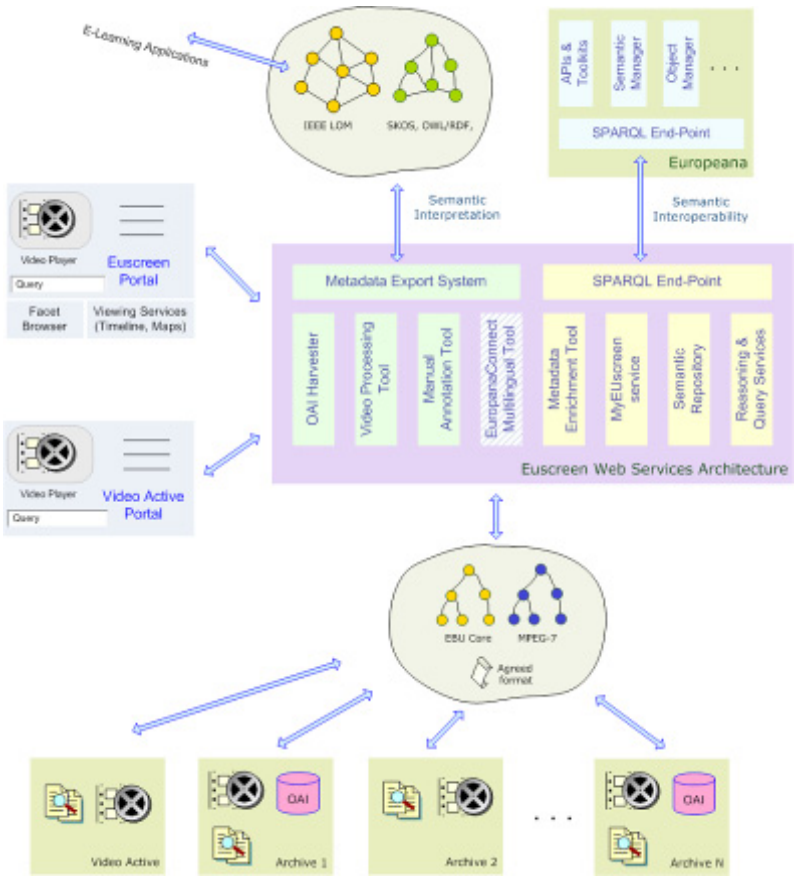


FIG 5. High-level architecture of the EUscreen system components

EUscreen will deliver a number of additional web services. EUscreen will focus on implementing advanced reasoning and query services on top of the Semantic Repository (see § 2.3) to enable metadata post-processing such as consistency checking and extraction of implicit knowledge. Automatic enrichment will be investigated, enabling metadata enrichment with complementary data, information, and knowledge from external web resources (e.g. Getty, UNESCO thesauri, Wikipedia) available as so-called Linked Data, using URIs and RDF (Berners-Lee 2006). For example, we can link the Creator field to DBPedia (uses RDF to represent information, and is part of the so-called LinkedData cloud) and create more links. A query on "Creator" Jonathan Ross will also result in links to related titles of programmes such as "Radcliffe and Maconie Show", "Sounds of the 60s, "Friday Night with Jonathan Ross" and so on. These titles can serve as basis for new queries to be sent to the retrieval system.

Also, using the metadata enrichment process, geographical coordinates will be assigned to all geographical places in order to visualize the results using the Google Maps API. Social

https://doi.org/10.23106/dcmi.952109604

networking functionalities will be developed, exploiting the recent advancements in Web 2.0 technologies in order to employ a framework where users can create their own galleries and participate in groups of interests, and where users can comment, annotate, tag, recommend or link up content items for personal or community use. Finally, the EUscreen metadata will be exported in other metadata schemas and in various language formats aiming at making it available in e-learning, leisure and research application scenarios.

4. Providing Content

EUscreen will take a pragmatic approach to the issue of rights and will draw on experience of the Video Active project to select, clear and deliver digitized content that is not hindered by restrictive IPR legislation, rules, precedents or contracts. This will mean that a critical mass of content can be delivered in a timely, efficient and cost-effective manner. At the same time, due to a lack of harmonization of legislation across the European Union, some countries will insist that material cleared for copyright restrictions under their national law must also be 'published' (streamed in the case of audiovisual materials on the internet) within its own borders. To achieve this EUscreen will have a flexible technical architecture, to allow material to be physically located (have its streaming server) in any of the partner locations – as well as supporting streaming from the central website server. For non-video material provided by EUscreen, notably the metadata, which will become available in RDF format, IPR regulations are less stringent. EUscreen will invest effort into understanding the complex rights issues related to online delivery including new emerging user cultures and media practices from a user point of view and will develop strategies, recommendations and guidelines of best practices for solving these issues.

At the same time, EUscreen will follow the recommendations put forward by the Europeana Connect project, that will provide a core set of interoperable licenses that cover rights information for objects in Europeana. This so-called Europeana Licensing Framework is based on ccREL rights expression language in RDF. (Abelson 2008)

5. Conclusion

Simply digitizing and uploading archive holdings does not release the full potential of audiovisual content. The added value of archives lies in their ability to place material in a context meaningful to different user groups and by enriching the metadata to allow interactive exploration. For a pan-European service, the infrastructure should meet very specific requirements, dealing with semantic and multilingual interoperability. Current developments (within the scope of W3C, DCMI and Europeana) provide the sector with the tools and examples necessary. As more archives join the EUscreen network, a vital part of our heritage will become available online for everybody to study and enjoy.

References

Abelson, Hal, Ben Adida, Mike Linksvayer, Nathan Yergler (2008) ccREL: The Creative Commons Rights Expression Language. Version 1.0. Retrieved July 12, 2009, from http:// wiki.creativecommons.org/images/d/d6/Ccrel-1.0.pdf

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. (2008) 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

https://doi.org/10.23106/dcmi.952109604

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

https://doi.org/10.23106/dcmi.952109604