



Video Active – Le patrimoine télévisuel européen en ligne

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

Traduction :

Rosa María Gómez de Regil
Ingénierie documentaire des contenus numériques
CNDP, France

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>

Résumé

Plusieurs archives audiovisuelles sont en train de numériser leurs matériels et ils explorent les nouvelles possibilités existantes afin de mettre en ligne ces contenus. Ce papier donne un aperçu du contexte et du développement du portail Video Active (milliers des documents vidéo accessibles sur le site www.videoactive.eu) qui a gagné un prix important.

Le projet Video Active a utilisé les dernières avancées des technologies du web sémantique afin de donner une représentation expressive des métadonnées, la mise en correspondance des schémas de métadonnées hétérogènes dans un schéma commun basé sur le Dublin Core et des systèmes d'interrogation avancés. Un des principaux débouchés du projet est l'intégration des données dans Europeana.

Dans ce papier, le travail fait sur Video Active est présenté en se focalisant sur l'architecture du système et sur les technologies du web sémantique utilisées.

Mots clés : *web sémantique ; bibliothèque numérique européenne ; archives audiovisuelles ; transmission multimédia en continu ; Initiative des Archives Ouvertes.*

1. Introduction : accès en ligne au patrimoine audiovisuel

La plus grande promesse d'Internet en tant qu'entrepôt de connaissances public est de créer un accès global pour tous et partout, à toute la connaissance et produits culturels créés par l'être humain. Les sites offrant des matériels vidéo en ligne sont aujourd'hui arrivés à maturité et dans une courte période sont devenus extrêmement populaires, principalement à cause de la disponibilité de la bande passante. Les sites comme YouTube, MySpace, Revver et autres montrent comment l'idée de création et de manipulation d'images (branche avant réservée aux professionnels) a été adoptée en tant que forme de diffusion générale accessible à tous ceux qui sont prêts à regarder.

Il est évident que le potentiel de rendre publics les archives audiovisuelles en ligne est énorme. Cependant, de tous les millions d'heures de ces archives en ligne, seulement un faible pourcentage peut être réellement trouvé. Plusieurs services en ligne existants se basent sur les contenus générés par les utilisateurs. Si l'offre de contenus professionnels existe (ex. Joost, Miro, Blinkx) la priorité est plutôt donnée aux matériels récents.

Les archives audiovisuelles ont besoin de surmonter plusieurs obstacles avant de mettre en place des services en ligne signifiants. Cela inclut : la gestion des droits de propriété intellectuelle, les aspects techniques en relation avec la numérisation et la normalisation des métadonnées et les aspects de présentation des sources aux utilisateurs. Ce dernier point représente un grand défi si l'objectif est de présenter du matériel de différents pays dans une forme structurée. Ceci est le point de départ du projet *Video Active*.

Le principal défi de *Video Active* est d'enlever les principales barrières listées ci-dessus afin de créer un accès multilingue au patrimoine télévisuel européen. *Video Active* atteint son objectif en sélectionnant une collection équilibrée des contenus des archives de la télévision, qui reflètent les similarités et différences culturelles et historiques de la télévision à travers l'Union Européenne ; les contenus des archives sont complétés avec des métadonnées contextuelles bien définies. *Video Active* est un membre invité de EDLnet, réseau initié en 2006 afin de construire un consensus pour créer la bibliothèque numérique européenne. *Video Active* sera disponible à travers le portail Europeana.eu.

Ce document donne un aperçu du contexte et du développement du portail *Video Active* (www.videoactive.eu) qui propose l'accès au patrimoine des matériels télévisuels de 14 archives en Europe. Le projet *Video Active*, un projet d'enrichissement de contenus du programme eContentplus, a utilisé les dernières avancées des technologies du web sémantique afin de donner une représentation expressive des métadonnées, la mise en correspondance des schémas de métadonnées hétérogènes dans un schéma commun propre à *Video Active*, des systèmes d'interrogation sophistiqués, des modes de présentation interactifs (ex. ligne du temps) des contenus. L'utilisation de ces technologies permet à *Video Active* d'être compatible avec les spécifications d'interopérabilité d'Europeana.

Le système de *Video Active* est composé de plusieurs modules, tous utilisent des technologies web. Le workflow global dès l'annotation, le chargement des matériels, la conversion du code, l'extraction d'images clés, le stockage des métadonnées et la recherche sont gérées par ces composants. *Video Active* propose l'annotation multilingue des contenus, la recherche des contenus numériques utilisant la technologie ThesauriX. ThesauriX est un outil web de gestion des thesaurus multilingues basé sur la norme IPTC. Le système exploite également les technologies du web sémantique qui permettent l'automatisation, des services de requêtes sophistiqués (basés sur la norme SPARQL) et l'interopérabilité sémantique avec d'autres archives numériques hétérogènes. Les métadonnées ont été représentées avec le *Resource Description Framework* (RDF) et le *Simple Knowledge Organizational System* (SKOS), elles sont stockées dans l'entrepôt sémantique de métadonnées Sesame. L'utilisation des technologies sémantiques permet d'avoir des services de raisonnement légers (utilisation de la connaissance implicite à travers la subsomption et les relations d'équivalence), la fusion et alignement des métadonnées des sources hétérogènes et des fonctionnalités de recherche sophistiquées qui utilisent le langage de requêtes

SPARQL RDF. Des bases de données relationnelles ont été aussi utilisées afin d'accélérer quelques processus où les traitements sémantiques ne sont pas nécessaires. Finalement, les métadonnées de *Video Active* sont publiques et prêtes à être moissonnées avec la technologie OAI-PMH.

2. *Video Active*

Video Active a créé un fonds de contenus des archives de la télévision (10000 vidéos) et des données contextuelles (articles, instantanés, programmes) qui est représentatif des spécificités nationales et culturelles des différents pays européens à travers différents thèmes et événements historiques. Les archives qui ont contribué sont : BBC (UK), INA (FR), DR (D), DW (D), ORF (AT), NAVA (HU), Sound and Vision (NL) et beaucoup d'autres¹. Le portail répond à différents modes de recherche textuelle ainsi qu'à la consultation par facettes, par thématique et par la ligne du temps. (Voir la figure 1)

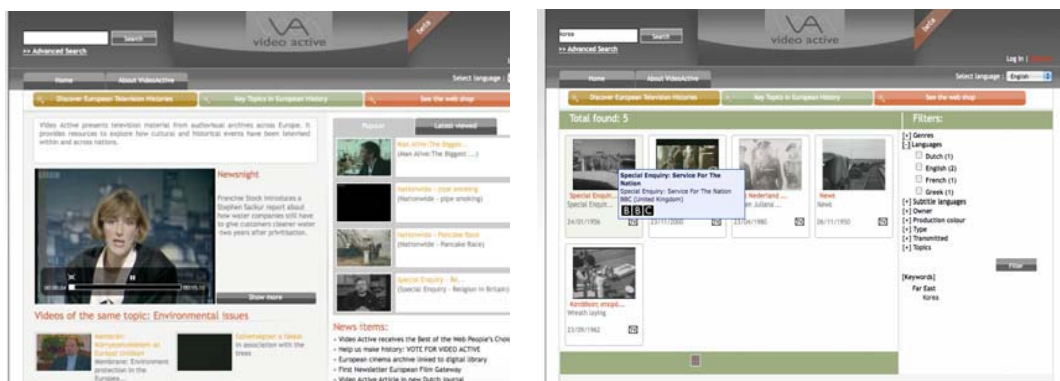


FIG 1. Page d'accueil et des résultats de *Video Active*

2.1 Définition des besoins pour *Video Active*

La demande d'accès aux contenus audiovisuels en ligne ne cesse pas de croître dans les secteurs tels que l'éducation du grand public et le patrimoine. Par exemple, la numérisation du contenu des archives transforme le patrimoine culturel dans des objets pédagogiques flexibles qui peuvent être facilement intégrés dans l'enseignement d'aujourd'hui et dans les stratégies d'apprentissage. Ces groupes d'utilisateurs ont des attentes et des profils différents, le projet *Video Active* a dû comprendre et inclure ceci afin d'assurer la satisfaction des utilisateurs et le trafic du site. Des enquêtes, des interviews et des recherches documentaires ont été réalisées lors des premières étapes du projet. Les résultats de l'étude de besoins ont été fondamentaux pour définir les spécifications techniques et par conséquent l'architecture technique. Des tests d'usabilité ont été menés dans les deux versions suivantes du portail. L'excellence du portail a été reconnue pendant la conférence *Museums and the Web* en 2009, où *Video Active* a gagné le prix « Best of the Web ».

2.2 Architecture de haut niveau

Le système de *Video Active* est composé de plusieurs modules web. Le workflow global du système dès l'annotation, le chargement des matériels, la conversion du code, l'extraction d'images clés, le stockage des métadonnées et la recherche sont gérés par ces modules. La figure 2 montre l'architecture du portail *Video Active*. L'architecture exploite les technologies du web sémantique qui permettent l'automatisation, des services de requêtes intelligents et l'interopérabilité sémantique avec d'autres archives numériques hétérogènes. Une couche sémantique a été ajoutée avec la représentation des métadonnées avec le *Resource Description Framework* (RDF). Le pouvoir d'expression de RDF permet d'avoir des services de raisonnement légers, la fusion et alignement des métadonnées des sources hétérogènes et des fonctionnalités de recherche sophistiquées basés sur le langage de requêtes SPARQL RDF. Des bases de

¹ Voir la liste complète sur : <http://videoactive.wordpress.com/the-consortium/>

données relationnelles ont été utilisées également afin de stocker les données où l'information sémantique n'est pas requise et afin d'optimiser les performances du système. Finalement, les métadonnées de *Video Active* sont publiques et prêtes à être moissonnées en utilisant le protocole OAI-PMH.

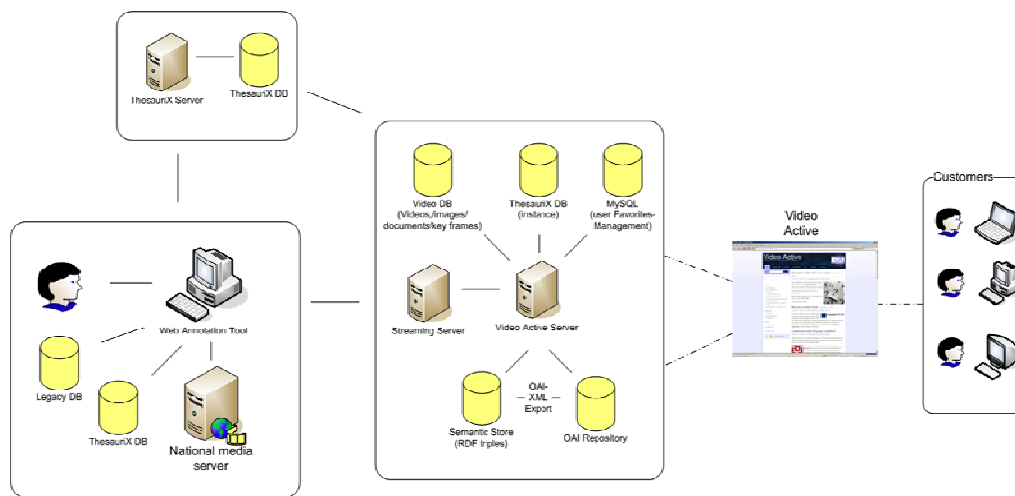


FIG. 2 *Video Active*: Architecture de haut niveau

2.3 Stockage et recherche des données selon le web sémantique

Le schéma de métadonnées de *Video Active* est basé sur le schéma Dublin Core auquel ont été ajoutés des éléments complémentaires (ex. Genre, Titre en anglais) nécessaires à l'enregistrement des propriétés spécifiques des ressources. (Venetis 2007) Les métadonnées vidéo sont générées automatiquement et sont représentées dans un schéma basé sur le MPEG-7. Afin d'activer les services sémantiques, les métadonnées sont transformées en triplets RDF et stockées dans un entrepôt de métadonnées sémantiques.

Le processus d'annotation est manuel ou semi-automatique. Dans le cas du processus semi-automatique les archives exportent leurs métadonnées (celles qui ont des correspondances avec les éléments du Dublin Core) en utilisant un schéma XML commun. Les éléments qui ne trouvent pas de correspondance dans le schéma *Video Active* (ou qui manquent dans les bases de données patrimoniales) sont insérés manuellement avec l'outil d'annotation web. Cet outil permet d'entrer et gérer les métadonnées associées au média et traite aussi la préparation du contenu en cours. L'outil contient la *Transcoding Factory*, module qui convertit le code de la vidéo source au format original vers les formats de diffusion en flux Flash et Windows Media, crée des débits binaires bas et moyens pour le service de diffusion en flux et extrait les images clés pour la création des imajettes. L'outil d'annotation web produit un fichier XML qui contient les métadonnées, basés sur le Dublin Core, les contenus encodés et les informations de l'extraction des images clé. Le XML est alors transformé dans des triplets RDF et stocké dans l'entrepôt sémantique Sesame. Sesame est un cadre conceptuel disponible en code source libre et écrit en Java qui sert à stocker, interroger et raisonner avec RDF (Broekstra 2002). Il permet de stocker les triplets RDF dans d'autres systèmes de stockage (ex. entrepôt local Sesame, base de données MySQL). L'utilisation d'un langage ontologique, comme RDF, qui dispose d'une sémantique formelle permet d'enrichir la représentation et les services de raisonnement qui fournissent les fonctionnalités de recherche sophistiquées, l'automatisation des processus et l'interopérabilité sémantique. La recherche en *Video Active* est opérée avec une combinaison de requêtes RDF structurées en SeRQL (optimisation du langage SPARQL pour Sesame) et la recherche en plein texte utilise la librairie de moteur de recherche plein texte à haute performance Lucene. Toutes les métadonnées stockées dans Sesame sont exposées aux systèmes / archives externes avec l'aide

d'un entrepôt OAI-PMH. Europeana, agrégateur de centaines de collections de ressources en Europe, a déjà indexé les données de l'entrepôt *Video Active*. (Voir figure 3).

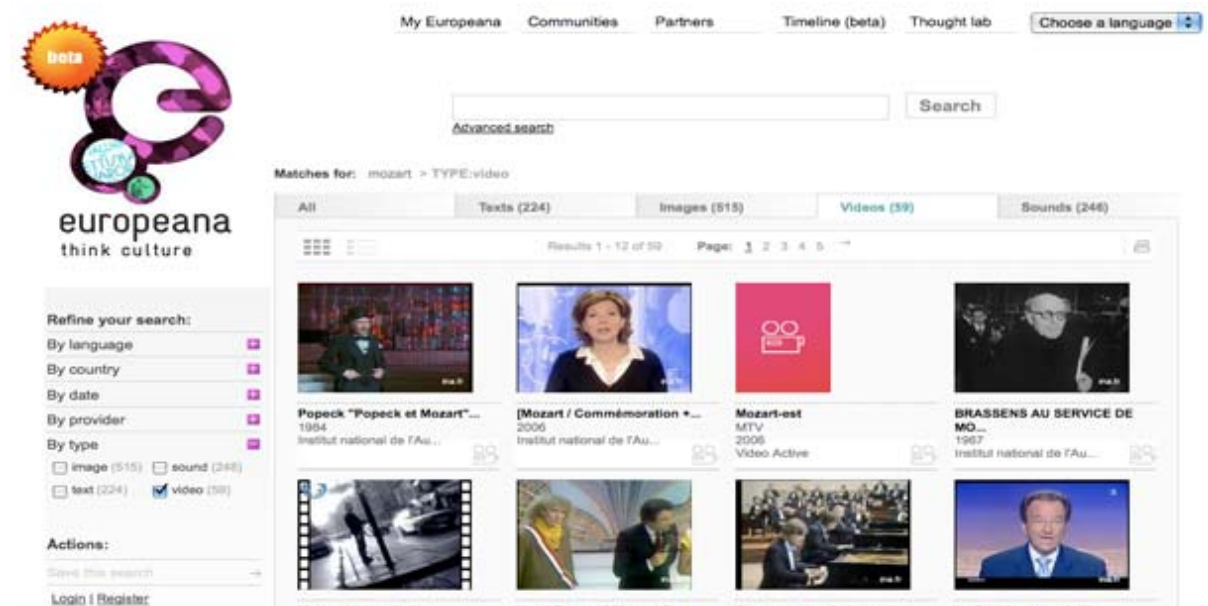


FIG 3. Métadonnées de *Video Active* indexées par Europeana

Afin de pouvoir échanger la structure et la sémantique des métadonnées de façon à ce que la machine l'interprète, des mécanismes de requête distribués OWL/RDF seront utilisés dans une prochaine version.

2.4 Accès multilingue : utilisation de SKOS

Onze langues sont prises en compte par *Video Active*. Le portail *Video Active* propose quatre façons de traiter les langues. En premier lieu, *Video Active* a localisé les interfaces de chacune des langues disponibles. Deuxièmement, les éléments de métadonnées clés (ex. DC Titre, DC Description) sont traduits en anglais afin de disposer d'une base monolingue dans la base de données. Troisièmement, *Video Active* utilise des vocabulaires contrôlés multilingues pour les éléments de métadonnées Mots clés, Genre, Localisation. Le thésaurus du conseil international de la presse et des télécoms est utilisé comme base pour le vocabulaire des mots clés. Ce thésaurus de 1500 termes a été traduit par le projet *Video Active* en 11 langues. Pour les genres, le système de classification *ESCORT 2007 EBU System of Classification of Radio and Television Programmes* est utilisé et pour les noms géographiques, la norme ISO 3166 noms de pays en anglais et éléments de codage est appliqué. La manipulation de la traduction de ces termes et l'export dans un format XML (interprétable par la machine) est fait dans l'application spécialisée ThesauriX. (Janisch 2008) Afin d'atteindre l'interopérabilité sémantique, la taxonomie du thésaurus a été transformée dans un langage du web sémantique en utilisant la norme SKOS (Simple Knowledge Organisation System). SKOS est une recommandation du W3C (World Wide Web Consortium) qui sert à représenter des taxonomies de thésaurus. La norme SKOS est construite à partir du langage RDF et peut être utilisée pour faciliter la découverte sémantique des métadonnées et l'alignement de thésaurus. Finalement, une ligne du temps est fournie afin de présenter visuellement une vue d'ensemble des événements clés dans le développement de la télévision en Europe avec le cadre conceptuel SIMILE. (Alonso 2007)

3. Conclusion

Dans cet article a été présenté le système de *Video Active* qui a exploité toutes les dernières avancées des technologies du web sémantique afin de proposer des services web sophistiqués qui utilisent des métadonnées destinées au système. Les technologies du web sémantique comme

RDF, SKOS, OWL et SPARQL ont été utilisées pour la représentation, la recherche, la présentation et l'échange des métadonnées de *Video Active*. Finalement, les métadonnées de *Video Active* sont déjà incluses dans Europeana par le biais de la technologie OAI-PMH.

Références

- 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>