MediaLoep: Optimizing Search in a Broadcaster Archive

Debevere P., Van Deursen D., Mannens E., Van de Walle R., Braeckman K., De Sutter R.

In: Proceedings of the 12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011) (Apr 2011)

MEDIALOEP: OPTIMIZING SEARCH IN A BROADCASTER ARCHIVE

P. Debevere, D. Van Deursen,
E. Mannens, R. Van de Walle

Ghent University - IBBT
ELIS Dept. - Multimedia Lab
Gaston Crommenlaan 8 bus 201
B-9050 Ledeberg-Ghent, Belgium

K. Braeckman, R. De Sutter

VRT-medialab
Gaston Crommenlaan 10 bus 101
B-9050 Ledeberg-Ghent, Belgium

ABSTRACT

In this demonstration, we introduce 'MediaLoep', a media search platform developed to enable efficient media retrieval from a broadcaster archive. We illustrate how media search can be optimized by capturing valuable data generated during the media production process. The MediaLoep platform is based on Semantic Web technologies, allowing us to connect to the Linked Open Data cloud. This facilitates the introduction of search functionalities such as advanced faceted search and semantic query suggestion.

1. INTRODUCTION

Efficient search applications can greatly improve productivity in a broadcast environment. For example, in a news production environment (which is characterized by very strict deadlines), it is important to find relevant archived content very fast as these are used by editors for reuse in new news broadcasts and by journalists to collect background information related to a certain topic. However, as the amount of archived media items grows significantly, this becomes an increasingly challenging task.

The media archive of the Flemish public service broadcaster in Belgium (i.e., Vlaamse Radio- en Televisieomroep (VRT)) contains thousands of hours of archived audiovisual material. Media search at VRT is currently enabled through manual annotation of the archived content. For each archived media item, a metadata record is generated containing typical fields such as a title, description, keywords, rights information, etc. It is clear that the quality of the metadata has a major impact on search efficiency. The 'MediaLoep' project[1] tries to optimize media retrieval by capturing and structuring valuable data generated during the media production process. In addition, information available from the Linked Open Data (LOD) cloud is used to further improve media retrieval.

[1] http://www.vrtmedialab.be/en/projects/medialoep/

2. INTEGRATING MEDIA PRODUCTION DATA

The generation of a metadata record is the last step of the media production process. However, during media production, a lot of data is generated that can be considered valuable metadata during media search. For example, the news production process contains many steps where valuable data is generated. An important source of information in news production is the rundown, which defines the items covered in a news broadcast. For each item in the rundown, anchor text is prepared during editing (which is displayed on the autocue or teleprompter during broadcast) and Character Generator (CG) text is generated. During broadcast, the exact time each topic was aired is stored, which during search can be used to jump to the exact location in the integral broadcast. In drama productions, the script contains valuable information such as the dialogues and cast information. In addition, the Electronic Program Guide (EPG) contains a short description of the episode.

Both for news and drama productions, often subtitles are available[2] which can significantly improve media search. For example, when a keyword (entered in a search box) appears in a subtitle, timing information can be used to start playback from the indicated timestamp.

The current MediaLoep platform aggregates the manually provided metadata with the extracted media production information described above. A data model was developed in order to represent the information present in the various information sources in a uniform way [1]. Fig. 1 illustrates available information for an archived news item that appears as a search result.

3. LINKED OPEN DATA

Media search can be further improved by using information available in external data sets. By using Semantic Web technologies, information available in the LOD cloud can be inte-

[2] VRT currently generates subtitles for about 85 % of the broadcasted items.

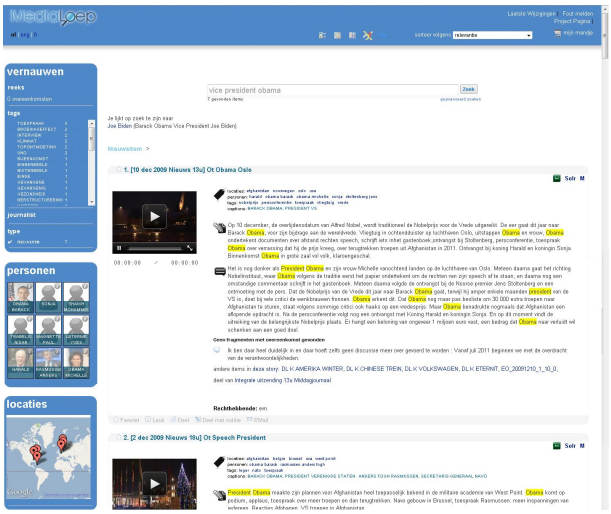

Fig. 1. Overview of the MediaLoep search application.

grated in the search application. As already mentioned in Section 1, each archived media item is annotated with keywords by an archivist. These keywords are mapped to the corresponding resources in DBpedia[3], an important linking hub in the LOD cloud. This allows the introduction of general facets such as 'person' and 'location'. Locations are mapped to the corresponding resource in GeoNames[4], which contains useful information such as geo-coordinates and type indication (e.g. country, forest, etc.). Using the geo-coordinates, search results can be visualized on a map. Keywords that appear to be persons are enriched with additional type information defined in DBpedia and YAGO[5] in order to enable advanced faceted search (e.g., facets for politicians, writers, actors, etc.). For every person that is mapped with DBpedia, additional information such as a short abstract, a depiction of the person, the URL of their homepage, etc. can be obtained if present.

DBpedia contains a vast amount of formalized knowledge which is used in the MediaLoep platform to enable semantic query suggestions. When a user for example enters the query 'producer nevermind', the MediaLoep search application suggests 'Butch Vig' as a relevant suggestion. In addition, a formalized version of WordNet[6] is used to enable synonym suggestion. In order to enable synonym suggestion in Dutch, the Cornetto[7] lexical data set is also used, which is linked with WordNet.

[3]http://dbpedia.org
[4]http://geonames.org
[5]http://www.mpi-inf.mpg.de/yago-naga/yago/
[6]http://semanticweb.cs.vu.nl/lod/wn30/
[7]http://www2.let.vu.nl/oz/cltl/cornetto/

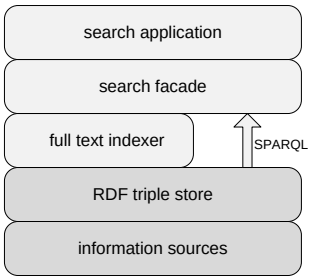

Fig. 2. Overview of the MediaLoep architecture.

4. ARCHITECTURE

The current architecture of the MediaLoep platform is illustrated in Fig. 2. The data model was formalized into a OWL ontology and all information from the information sources described in Section 2 was converted to RDF and stored in a RDF triple store. This triple store is then queried using SPARQL. In order to allow fast text search operations, Solr is used to index textual data [2]. The search application is built on top of a search facade which executes queries against Solr and the RDF triple store.

5. DEMONSTRATION

We demonstrate the MediaLoep platform, a web-based search application illustrated in Fig. 1. The current demonstration allows the retrieval of all video items and photographs produced at VRT during 2010, resulting in a collection of about 25000 video items and over 20000 photos.

6. ACKNOWLEDGEMENTS

The research activities as described in this paper were funded by Ghent University, VRT-medialab, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO-Flanders), and the European Union.

7. REFERENCES

[1] P. Debevere, D. Van Deursen, D. Van Rijsselbergen, E. Mannens, M. Matton, R. De Sutter, and R. Van de Walle, "Enabling Semantic Search in a News Production Environment," in *Proceedings of the 5th International Conference on Semantic and Digital Media Technologies*, Saarbrücken, Germany, December 2010.

[2] K. Braeckman, "Unlocking a broadcaster video archive using Solr," *Apache Lucene Eurocon*, 2010.