MediaHaven: Multimedia Asset Management with integrated NER and categorisation

Bruno Van Den Bossche
Zeticon
Ghent, Belgium
bruno.vandenbossche@zeticon.com

Brecht Vermeulen
INTEC - IBCN
Ghent University,
Ghent, Belgium
brecht.vermeulen@intec.ugent.be

Johannes Deleu
INTEC - IBCN
Ghent University,
Ghent, Belgium
johannes.deleu@intec.ugent.be

Thomas Demeester
INTEC - IBCN
Ghent University,
Ghent, Belgium
thomas.demeester@intec.ugent.be

Piet Demeester
INTEC - IBCN
Ghent University,
Ghent, Belgium
piet.demeester@intec.ugent.be

ABSTRACT

In order to allow for flexible search and asset management on the textual metadata of multimedia archives, the extraction of information and especially named entities is an essential step. Practically, they are of great help for applications like facetted search, input assistance, search suggestions, linking assets, etc. This paper describes MediaHaven, a Media Asset Management (MAM) system, commercialized by Zeticon, a spin-off of Ghent University-IBBT. MediaHaven incorporates an advanced NER and categorisation system to improve the user experience.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

1. INTRODUCTION

Zeticon's[1] MediaHaven was used a.o. as a Media Asset Management system in the BOM-Vlaanderen and the Vlaanderen in Beeld (VLIB) project. Both projects were funded by the IWT (agency for Innovation by Science and Technology), and focused on the archivation of multimedia content of the Flemish broadcasters. The total multimedia collection that was composed in the framework of VLIB, contains about ten thousand hours of video material, including audio data and textual metadata. Navitation through the archive is done with a search engine based on these metadata.

Many Flemish television broadcasters, but also a number of other players in the media sector, provided video material for the VLIB archive. These data, and especially the metadata, were rather inhomogeneous among the resulting archive, and in order to design a balanced search engine that would allow retrieving content from those several sources, the metadata had to be transformed to a more uniform format. A number of sources provided full text descriptions of the considered video material, as well as keywords, other sources did not, e.g., provide keywords, or keywords from a different thesaurus.

It became clear that we needed an automated system to extract information from the full text metadata. A promising option to this end was Named Entity Recognition (NER) appeared an attractive means for that goal, automatically extracting names of people, locations, and organizations. After a number of experiments with existing NER tools, we started to design our own system, since the existing tools did not at all perform as well as they did on news article collections (for which they were originally trained), due to the different characteristics of our archive material (lack of capitalization, very domain-specific terminology, condensely written...). The technical details of the system that we built, are described in a technical paper on this conference.

In this demo paper we will focus on the use of tools like NER and categorisation to make retrieval and annotation of content by users in a Media Asset Management system easier.

2. MEDIAHAVEN

MediaHaven was developed by Zeticon as a state of the art highly versatile multimedia asset management solution. All kinds of files –multimedia, graphic or office files– can be stored and managed together with metadata describing those files. MediaHaven is the ideal solution for fast and handy asset management of file-based multimedia content. MediaHaven was built from the start with scalability, robustness and future extensibility as its main drivers. The MediaHaven components are shown in Figure 1.

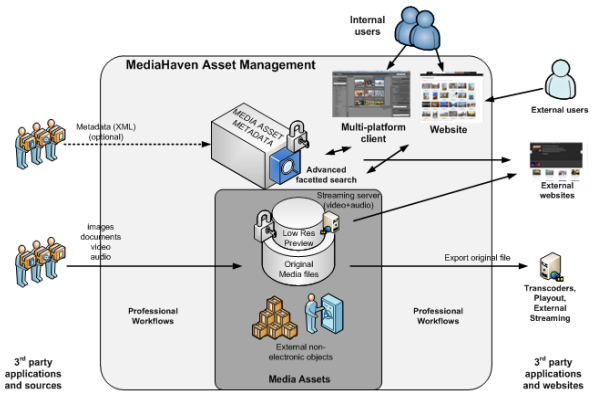

Figure 1: MediaHaven architecture and components

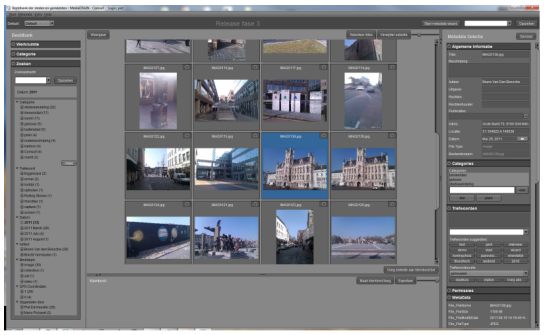

Figure 2: MediaHaven annotation module with facetted search at the left and metadata input at the right

3. NER AND CATEGORIES AS TOOLS

One of MediaHaven's strengths are the advanced (facetted) search and retrieval functions and annotation help (Figure 2). When a user adds keywords for a particular object, we show suggestions that continuously change according to the used keywords. For this we use NER extraction of already used metadata and cluster them.

In the website view of the prototype we made for Vlaanderen in Beeld, we have a tagcloud system where NER is used to suggest related terms and use colours for indicating persons, locations, events, ... (Figure 3).

Besides these NER based improvements, recently we added also categorisation to the MediaHaven product. By trying to put categories to each document or asset, we will in the near future improve the retrieval engine even more by making it possible for the users to dive in the archives in a category/thesaurus based way.

4. DEMO

The demo will show the NER and facetted search as separate demos and MediaHaven as a complete system in the Vlaanderen in Beeld prototype setup.

5. CONCLUSIONS

With this demo paper, we shortly presented the use of NER and categorisation tools in a commercial Media Asset Management product. Especially for the use in broadcasters' archives, where annotation was done on beforehand by professional archivists, the NER had to be adapted, because of the 'poor' text quality. The demo will highlight these specific improvements and strengths.

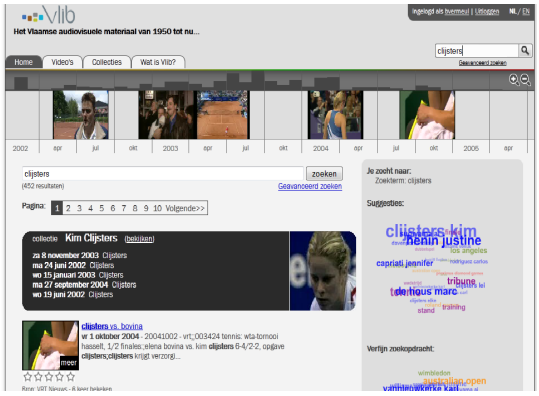

Figure 3: MediaHaven website prototype for Vlaanderen in Beeld project with NER tagclouds

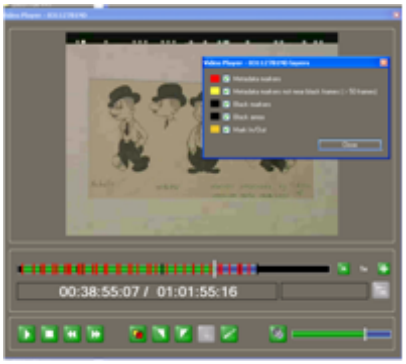

Figure 4: MediaHaven's videoplayer with annotation based on timecodes in the video

6. REFERENCES

[1] http://zeticon.com/.