AVResearcher: Exploring Audiovisual Metadata

Bouke Huurnink
Nederlands Instituut voor Beeld en Geluid
Sumatralaan 45
Hilversum, The Netherlands

Amit Bronner
Nederlands Instituut voor Beeld en Geluid
Sumatralaan 45
Hilversum, The Netherlands

Marc Bron
University of Amsterdam
Science Park 904
Amsterdam
The Netherlands

Jasmijn van Gorp
Utrecht University
Muntstraat 2A
Utrecht
The Netherlands

Bart de Goede
Dispectu
Julianweg 61
Wijk aan Zee
The Netherlands

Justin Wees
Dispectu
Julianweg 61
Wijk aan Zee
The Netherlands

{bhuurnink, abronner}@beeldengeluid.nl, m.m.bron@uva.nl
j.vangorp@uu.nl, {bart, justin}@dispectu.com

ABSTRACT

In this demonstration we present AVResearcher, a prototype aimed at allowing media researchers to explore metadata associated with large numbers of audiovisual broadcasts. It allows them to compare and contrast the characteristics of search results for two topics, across time and in terms of content. Broadcasts can be searched and compared not only on the basis of traditional catalog descriptions, but also in terms of spoken content (subtitles), and social chatter (tweets associated with broadcasts). AVResearcher is a new and ongoing valorisation project at the Netherlands Institute for Sound and Vision, and as such is under active development. At DIR 2013 we will present the current version of the software.

1. INTRODUCTION

In this demonstration we present AVResearcher, a prototype aimed at allowing media researchers to explore the professional, content-based, and social metadata associated with a collection of hundreds of thousands of broadcasts. With the continuous online production and storage of audiovisual broadcasts, a challenge for media researchers has arisen. There is a large amount of archival metadata about broadcasts becoming available. In addition, metadata from additional sources is becoming available. For example, the Netherlands Institute for Sound and Vision has over 960,000 catalog entries, and has an archive of subtitles for a subset of television broadcasts going back to 1989. In addition, members of the public microblog about broadcasts on Twitter, in the Netherlands sometimes amounting to tens of thousand of tweets for an individual program. Our prototype addresses this challenge, allowing media researchers to examine the metadata characteristics of sets of broadcast results.

DIR 2013 Delft, The Netherlands

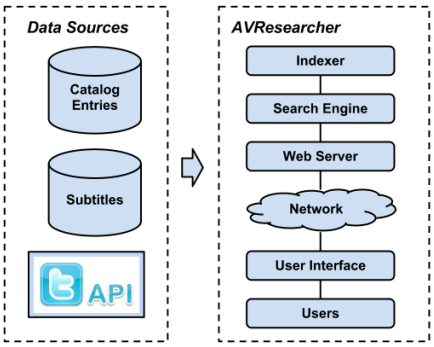

Figure 1: AVResearcher system overview.

AVResearcher is based on the Media Researchers Data Exploration Suite (MeRDES) [1], which was developed specifically to support media studies researchers to explore audiovisual catalog entries. In addition to the professional catalog entries supported by MeRDES, AVResearcher allows media researchers to explore social chatter in the form of tweets, and spoken content in the form of subtitles. In addition, the code of AVResearcher has been completely rewritten for improved speed and scalability. It is a new valorisation project at the Netherlands Institute for Sound and Vision, and as such is under active development. It is undergoing iterative development using the AGILE methodology: user feedback is used to determine the requirements and their prioritisation for each iteration. After the second iteration has been accepted the prototype will be made available to media professionals through an online portal of the archive. At DIR 2013 we will present the current version of the software, which is expected to be the second iteration.

2. AVRESEARCHER SYSTEM

An overview of the AVResearcher system is given in Fig-

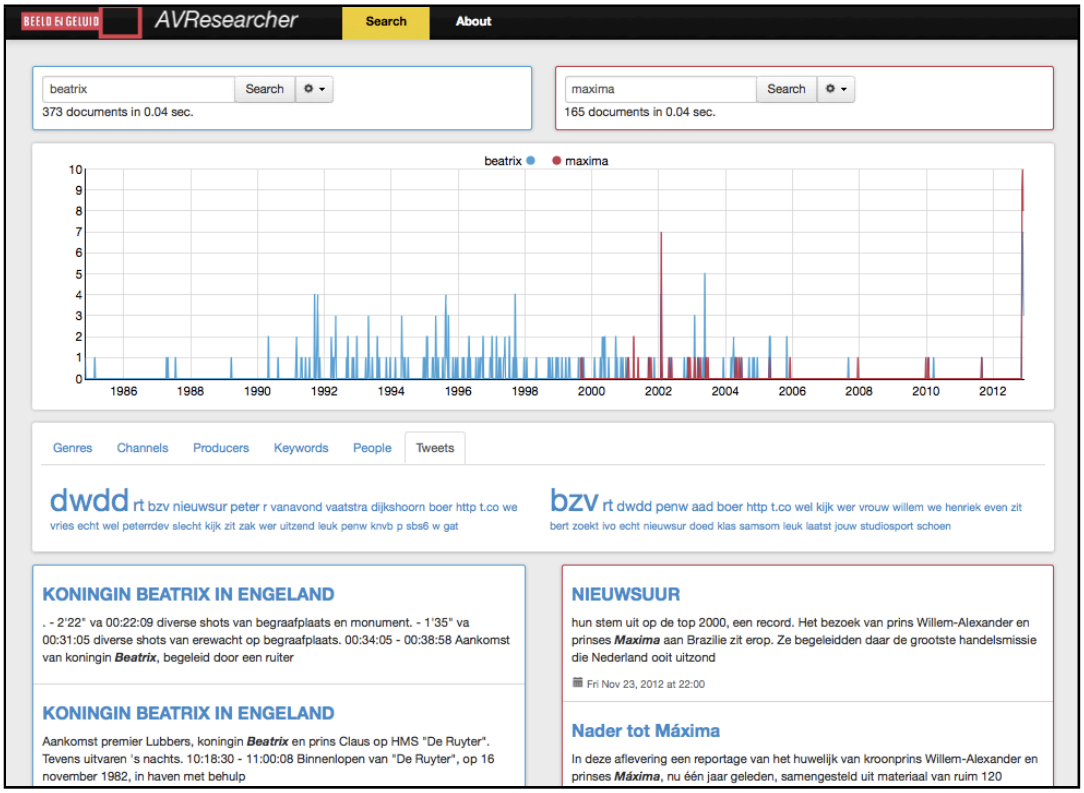

Figure 2: Search results screen of AVResearcher

ure 1. Here we briefly summarize the system in terms of the underlying data set, architecture, and visualization.

Data Set Catalog descriptions of the broadcasts are obtained from the archive of the Netherlands Institute for Sound and Vision: at the time of writing the collection consists of just over 960,000 broadcasts. Subtitles are obtained through an agreement with the Netherlands public broadcasters from November 2012 onwards. In the future we also plan to incorporate a legacy database of subtitles dating back to December 1989. Tweets about programs also date from November 2012 onwards. They are obtained using the Twitter Streaming API: we monitor a collection of official hashtags for 25 Dutch television shows, obtained from the website `http://hekjeplekje.nl`. If a tweet occurs during a television broadcast, it is associated with that broadcast.

Architecture The AVResearcher architecture is illustrated in Figure 1. Data for television broadcasts is collected from three different sources: catalog entries maintained by the archive, subtitles obtained from the Netherlands Public Broadcasting system, and tweets connected to broadcasts obtained through the Twitter API. The data is stored and indexed for use by an open source search system.[1] The user interface is made available on the webserver. Users can interact with the interface over a secure network connection

Visualization The AVResearcher interface, shown in Figure 2, allows users to issue two search queries and compare the results side-by side. For each query the user can view:

- The number of broadcasts containing the query terms on a timeline. The hits for each query are visualised on the same timeline, and given a different color. This allows researchers to see how two given topics (represented by queries) have evolved over time.
- Term clouds of term frequently occurring within search, divided into facets from the catalog entries (genres, channels, producers, keywords, and people), as well as words frequently occurring in subtitles and tweets.
- The list of search results used to generate the timeline and term cloud. When users click search result they can see more details for that particular broadcast.

[1]We use the ElasticSearch search engine, which scales to our needs.

3. CONCLUSION

AVResearcher is a prototype that addresses the problem of exploring different kinds of broadcast metadata on a large scale. It allows media studies researchers to explore and compare metadata for two different topics in a collection of hundreds of thousands of broadcasts. It includes subtitles and tweets, as well as professional catalog data, and in this way allows media studies researchers to explore spoken content and social chatter about broadcasts. The system is under active development, and will be used to perform user studies aimed at improving archival access. At DIR 2013 we will present the current version of the prototype.

4. REFERENCES

[1] M. Bron, J. van Gorp, F. Nack, M. de Rijke, and S. de Leeuw. A subjunctive exploratory search interface to support media studies researchers. In *SIGIR '12: 35th international ACM SIGIR conference on Research and development in information retrieval,*, pages 425–434, Portland, Oregon, 2012. ACM, ACM.