ABSTRACT
This paper describes the integration of two new services aimed at assisting into the retrieval of video content from a Multimedia Asset Manager (MAM). The first tool suggests tags after an initial textual query, and the second ranks the keyframe of retrieved assets according to their visual similarity. Both applications were implemented as web services that are accessed from a Rich Internet Application via REST calls.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—information filtering, query formulation; C.2.4 [Computer Systems Organization]: Computer-Communication Networks—Client/server, Distributed applications

General Terms
Management

Keywords
Tag suggestion, content-based retrieval, image databases, similarity retrieval, RIA, Web service, REST

1. INTRODUCTION
Video retrieval from archives is nowadays a challenging need for TV broadcasters. The growing amount of data daily ingested into their repositories has made no longer feasible relying only on the manual search and indexing of expert documentalists. This paper presents the integration of state of the art retrieval tools in the Multimedia Asset Manager (MAM) of the Catalan Broadcast Corporation (CCMA), a media group that produces four TV stations. The main goal of the new systems is the assist journalists in a fast retrieval of video assets to be edited and included in the video clips for the news programs.

The starting point of the presented work is Digition, the MAM previously developed by the CCMA whose search system is based only on matching query keywords and the textual metadata manually generated by the documentalists. Digimatge is a new search system that adds two new services to the ones previously offered by Digition: a tag suggestion for text-based queries and a ranking of retrieved keyframes based on image similarity.

The addition of these new plug-ins into a MAM in exploitation requires a modular architecture of multiple systems interconnected with a variety of network protocols. Web services coming from the Internet offer nowadays mature technologies designed to be deployed in distributed systems, offering user experiences on the cloud as rich as if they were being run at the client side. This application are referred as Rich Internet Applications (RIA) and their technologies are also valid in the framework of a corporate network. A second contribution of the presented work is the adoption of this paradigm to extend the features of the existing MAM.

The paper is structured as follows. Section 2 presents the new services added to the existing system and how to use them from the graphical user interface. Section 3 introduces the concept of Rich Internet Application and described the system architecture and communications protocols. Finally Section 4 draws the future research lines and provides the project conclusions.

2. DESCRIPTION OF THE NEW SERVICES
The video retrieval problem is normally solved by assessing the similarity between a set of query descriptors and the target descriptors associated to the assets stored in the repository. A similarity score is computed for each comparison and a ranked list of the assets is returned to the user as a response to the query. The nature of the descriptors can be diverse and, in the case of video data, they usually represent semantic, visual and audio information.

2.1 Text and Image-based Searches
Text and image descriptors are two types of features that have been largely used for image retrieval. State of the art solutions provide examples of each case and the combination of both.

Semantic information is normally expressed under the form of text descriptors and it is the most used modality in search engines. Users can directly formulate their queries by entering keywords or full sentences to the system, and these
generated query descriptors are compared to the text or semantic descriptors previously associated to the asset. The text-to-text video search tends to provide good results as humans can be pretty precise when using text to express semantics. Nevertheless, this approach requires generating textual metadata for the multimedia content. A first option is the manual annotation of the content, a very consuming task when manually performed by a human. As an alternative, descriptors can also be automatically parsed from the image filenames or contextual text, or be generated through signal processing like OCR-based solutions [5] or semantic classifiers [9].

A second family of retrieval systems are based only on visual descriptors, low-level features automatically generated by applying signal processing algorithms on the content. The automatic nature of the process allows processing a larger amount of content and the generated descriptors are as accurate and complete as the set of feature extractors. This solution, though, presents two drawbacks when compared to text annotation. Firstly, the correlation between perceptual descriptors and semantic concepts is not as close as in the textual case and, secondly, the user needs to become familiar with a new interface that formulates queries in terms of perceptual descriptors instead of text. There are mainly two options for the user to formulate the query, whether by directly providing a quantified value of the descriptor or by providing the system some sort of visual content from which the query descriptors can be extracted, such as an example or a sketch [1] [3] [4].

Although the visual-based similarity provides reasonably good results from a semantic point of view, most system exploiting these technologies also rely on textual descriptors, providing hybrid search solutions that combine both. Fast indexing and retrieval algorithms are more mature in the text mode than in the visual, so many solutions use text queries to retrieve a first set of results and these are later refined by a second search based on visual descriptors. This third family of retrieval systems are the natural evolution of the two previous, introducing multimodality to the search experience. Figure 1 shows an example of combined text and visual query, in which the semantic class of the object dog is expressed through text, but the color of its hair is represented through a palette choice. Commercial examples like Google Similar Images, Pixolu or Xcavator have applied this strategy.

2.2 Proposed solution
The presented system also combines text and visual descriptors in the framework of a broadcaster MAM. Textual annotations are generated by a team of expert documentalists according to some guidelines and recommendations. The manual annotations can refer to two different scales of a video asset: a global or local scale. Global annotations describe the whole shot as, for example, "Soccer game: F.C.Barcelona 3 - Athletic Bilbao 1", while local annotation are limited to the segment between to two time codes, as it would be "Player X scores a goal". The textual descriptors are complemented with new visual descriptors automatically computed from the video keyframes. Using keyframes instead of full videos reduces the amount of data generated during indexing and that needs to be processed at search time. In most cases, a few keyframes of a video sequence are enough to effectively represent its semantic content. The system computes four MPEG-7 visual descriptors from each keyframe: Color Layout, Color Structure, Edge Histogram and Texture Homogeneous.

The present work focuses on a subset of perceptual descriptors, those computed on still images. Although the final goal of the system is the retrieval of videos, the evaluation of the similarity between queries and target content is performed at the keyframe level. This approach is taken to reduce the amount of data generated during indexing and that needs to be processed at search time. In most cases, a few keyframes of a video sequence are enough to effectively represent its semantic content.

The search process begins with the input of a textual query in a box located at the upper left side of the graphic interface. While typing the query, an autocompletion combobox appears suggesting terms from a thesaurus maintained by the documentalists. This tool speeds up the typing process as well as decreases the probability of spelling errors. The words in the thesaurus are preferably used by documentalists, but during asset metadata may contain many more words are used. For this reason, the user can also enter keywords that are not included in the thesaurus.

Once the textual query is introduced, the search is executed and results are displayed in the graphic interface shown in Figure 2. The upper part of the screen shows a table with the numeric ID of the retrieved video assets as well as their title. The lower half contains the ten first keyframes of each retrieved video object. A double-click on any row in the results table repaints both upper and lower parts of the interface with more data about the selected video object: the results table is replaced by a full textual metadata associated to the asset, while at the lower part a new tab is created to show all keyframes from the selected video object.

A new tag suggestion service [2] is also executed at query time to retrieve similar terms to the entered keyword. This service is based on a statistic analysis of terms in the textual
metadata and shows its results in a word cloud. By doing so, the system is proposing new query terms aimed at assisting the user when its search idea is vague or fuzzy.

In addition, a new visual search service is available by double clicking on any thumbnail. When doing so, an image similarity search is assessed between the selected keyframe and the rest of the thumbnails of the retrieved assets. The process can be understood as a reranking of thumbnails, but in this case, all thumbnails of the retrieved assets are considered, instead of the first five thumbnails that are shown after the initial text query. The results are displayed in a new tab, where any click on a thumbnail will show the asset summary on the upper part of the interface.

3. IMPLEMENTATION

3.1 Web Services and Rich Internet Applications

Video content in the broadcasting industry is normally stored in Multimedia Asset Managers (MAMs), a distributed system to access a content through queries on an indexed database or catalog navigation. The addition of new services to an existing system is an operation that must face several technical restrictions due to the existing installations but, at the same time, offers the opportunity to introduce state of the art technologies. In the presented work, the distributed architecture of the Internet inspired the chosen approach to access the new services through a Rich Internet Application (RIA).

RIAs [8] are a conceptual framework for developing applications based in two basic trends of computing during the last years. In one hand, applications run on local computers search for information sources on distributed remote servers; on the other hand, many applications that used to run at the client-side are now accessed on a remote server through a web browser. Moving local resources and computations to remote servers is a new paradigm for the user to prepare all required data before connecting to the server. This approach is possible thanks to the graphic elements being persistent after the first download.

RIAs at the client side can be implemented in multiple programming languages and development environments, also called frameworks. The first RIAs were implemented in Javascript and AJAX, and take advantage of the native adoption of Javascript support from most web browsers. However, this open approach results in certain cases in multiple implementations that make this option less stable than other closed solutions as Adobe Flex or Microsoft Silverlight. These two proprietary technologies use, respectively, Action Script and .NET as programming languages. The third version of Flex framework uses the AS3 and MXML languages and is supported by most web browsers through the Adobe Flash Plugin; while Microsoft Silverlight is based on .NET language and its middleware is shipped from Microsoft Windows 7. Java is another important language on the web and its compiled classes can be easily downloaded and safely executed on the client over the Java Web Start framework. By definition, RIAs must communicate with services running on remote machines, and these servers must respond to requests more complex than just sending static web pages. For this reason, it is also necessary to use at the server side languages that can keep the state, such as Java or .NET, or interpreted languages, like PHP or ASP.

The data exchange between web services and RIAs is achieved over a certain communication protocol that must be capable of transporting amounts of data significantly higher than in traditional websites. A family of protocols has evolved over the HTTP layer, as in the cases of SOAP [6] and REST [7]. SOAP uses XML files in every communication layer and it is an official standard by the W3C, while REST is a simplification of that it uses lighter messages. The overhead introduced by the XML-based messages in SOAP is removed in REST, and its implementation simplified by using the same GET/PUT native functions included in HTTP. A common practice in REST communications is to answer the REST requests with JSON, a lightweight data-interchange format based on JavaScript. A second family of communication protocols for RIAs are not based on HTTP. For example, the open sourced AMF (Action Message Format). This protocol allows RIAs implemented in Flex to use remote interfaces implemented in Java, AS3 and other several server languages.

3.2 System Architecture

The proposed architecture combines several technologies that exploit the advantages presented in the previous paragraphs. The main challenge in the design is to integrate heterogeneous different information sources in a modular approach that could be expanded in the future if new services are to be added in the future. Figure 3 shows and overview of the complete system architecture. A main server (1) connects to the different resources located inside the corporate intranet or outside through the public Internet.
Apart from connectivity, the main server (1) is the responsible of two other tasks. Firstly, it processes the text-based queries and retrieve the IDs of the resulting video assets. Secondly, it generates the standard definition keyframes from the time codes defined by the keyframe extractor. During the first tests on the system, it was detected that the text-based queries on the Informix database (2) could not be answered as quickly as required through classic Select queries to the SQL tables describing the 454,626 test assets. As a solution, a new platform capable of efficiently index and search text information was introduced (4) based on Solr, an open source enterprise search platform from the Apache Lucene project. The communication with this text-based search engine is established from the main server over REST. The second task controlled from the main server is the generation of keyframes from the time codes stored in the Informix database. These time codes are generated at ingest time by the keyframe detector. The extraction of keyframes requires accessing the time code database as well as the content repository to retrieve actual video. Keyframes are generated with ffmpeg, a free software tool for video coding, and stored in host (3).

The two remaining elements in the architecture are those providing new features to the system: the tag suggester (5) and image similarity (6) web services developed by Yahoo! and the Technical University of Catalonia (UPC), respectively. Both of them are accessed over REST, but the text suggester returns an XML data structure and the image ranker directly provides a JSON object for easier integration.

Finally, the Flex RIA is downloaded from the main server and can be executed on a web-browser with a Flash plug-in or as a stand-alone application over the Adobe Air middleware. The communication between client and server is over the AMP protocol, using the BlazeDS implementation. This option serializes Java and Flex objects at the two sides of the link.

4. CONCLUSIONS

This paper has presented the implementation of two additional video retrieval services in an existing MAM. Their integration in the existing system required the definition of a system in which data and applications are distributed in heterogeneous servers running different technologies. The access to the new services was implemented as a RIA, allowing simple access to all users through a conventional web browser. The additional services running in external hosts were successfully connected to the corporate systems, providing guidelines for the future deployment of similar extensions. After these first successful steps, the system will continue evolving to introduce new methods for results clustering and navigation, efficient indexing for visual descriptors and connectivity to external news sources.

5. ACKNOWLEDGMENTS

This work was partially founded by the Catalan Broadcasting Corporation (CCMA) and Mediapro through the Spanish project CENIT-2007-1012 i3media.

6. REFERENCES