Enabling Semantic Search in a News Production Environment

Pedro Debevere[1], Davy Van Deursen[1], Dieter Van Rijsselbergen[1], Erik Mannens[1], Mike Matton[2], Robbie De Sutter[2], and Rik Van de Walle[1]

[1] Ghent University - IBBT, Multimedia Lab, Belgium,
pedro.debevere@ugent.be, davy.vandeursen@ugent.be,
dieter.vanrijsselbergen@ugent.be, erik.mannens@ugent.be,
rik.vandewalle@ugent.be
[2] VRT-medialab, Belgium,
mike.matton@vrt.be, robbie.desutter@vrt.be

Abstract. News production is characterized by a complex and dynamic workflow, in which it is important to produce and broadcast reliable news as fast as possible. In this process, the efficient retrieval of previously broadcasted news items is important, both for gathering background information and for reuse of footage in new reports. This paper discusses how the quality of descriptive metadata of news items can be optimized, by collecting data generated during news production. Starting from a description of the news production process of the Flemish public service broadcaster in Belgium (VRT), information systems containing valuable metadata are identified. Subsequently, we present a data model that uniformly represents the available information generated during news production. This data model is then implemented using Semantic Web technologies. Further, we describe how other valuable data sets, present in the Semantic Web, are connected to the data model, enabling semantic search operations.

Keywords: News item retrieval, News production, Semantic Web

1 Introduction

Efficient media search applications can highly improve productivity in various domains. However, an important requirement for efficient media retrieval is the availability of high quality metadata documenting the archived media [5, 6, 8, 16]. This is also the case within a news production environment, where professional archive users spend considerable amounts of time searching in the media archive in order to find useful media for reuse in news broadcasts [14]. However, in a news production environment, where it is important to produce and distribute news as soon as possible to as many channels as possible, in an audiovisual quality as good as possible [9], metadata generation is often reduced to an absolute minimum [12]. Consequently, dedicated archivists are responsible for the generation of high quality metadata as a last step in the news production chain in order to facilitate efficient media retrieval.

In this paper, we investigate the news production process of the Flemish public service broadcaster in Belgium (i.e., Vlaamse Radio- en Televisieomroep (VRT[3])). In large news production enterprises, such as VRT, several databases and information systems (e.g. subtitling and rundown management systems) are used during the news production process, each containing valuable information about the news and related media being produced. These information sources are often not or barely coupled. In addition, every information source has its own information storage facility resulting in a non-uniform data representation. However, most information contained in these information systems can often serve as valuable contributors of metadata describing the archived audiovisual material.

Therefore, the 'Medialoep' project[4] investigates how news-related audio and audiovisual content can be found in a more effective, efficient, and easy way. One of the goals of this research project is to increase the amount of quality metadata by capturing and structuring the available information during the news production. This automatically retrieved metadata can then be used by archivists as a starting point for further enrichment, leading to faster availability and more accurate search results, which improves the overall efficiency and productivity of news program editors.

As a first step, we investigate VRT's news production process and identify information sources containing valuable metadata. Subsequently, a data model is developed which covers and represents the information present in the various identified information sources in a uniform way. The model is implemented using Semantic Web technologies, which enable a formal and machine-understandable representation of the metadata. Further, the identified information sources need to be mapped to our data model. Finally, in order to enrich our metadata even more, we elaborate on the potential of connecting information stored according to the data model with other (external) data sets, which are available in the Semantic Web [3].

2 News Production Process

The following subsections describe VRT's news production process in order to facilitate the identification of valuable data sources for later retrieval of audiovisual content. This process is also illustrated in Fig. 1, where important steps in the news production process are labeled and referred to in the following subsections.

2.1 Editorial Planning and Management

The main tool used for news production at VRT is Avid's iNEWS[5]. This tool is amongst others used by directors and editors to create and manage news rundowns. A rundown consists of a list of items that will be covered during a

[3] http://www.vrt.be

[4] http://www.vrtmedialab.be/index.php/english/project/medialoep/

[5] http://www.avid.com/US/products/inews/

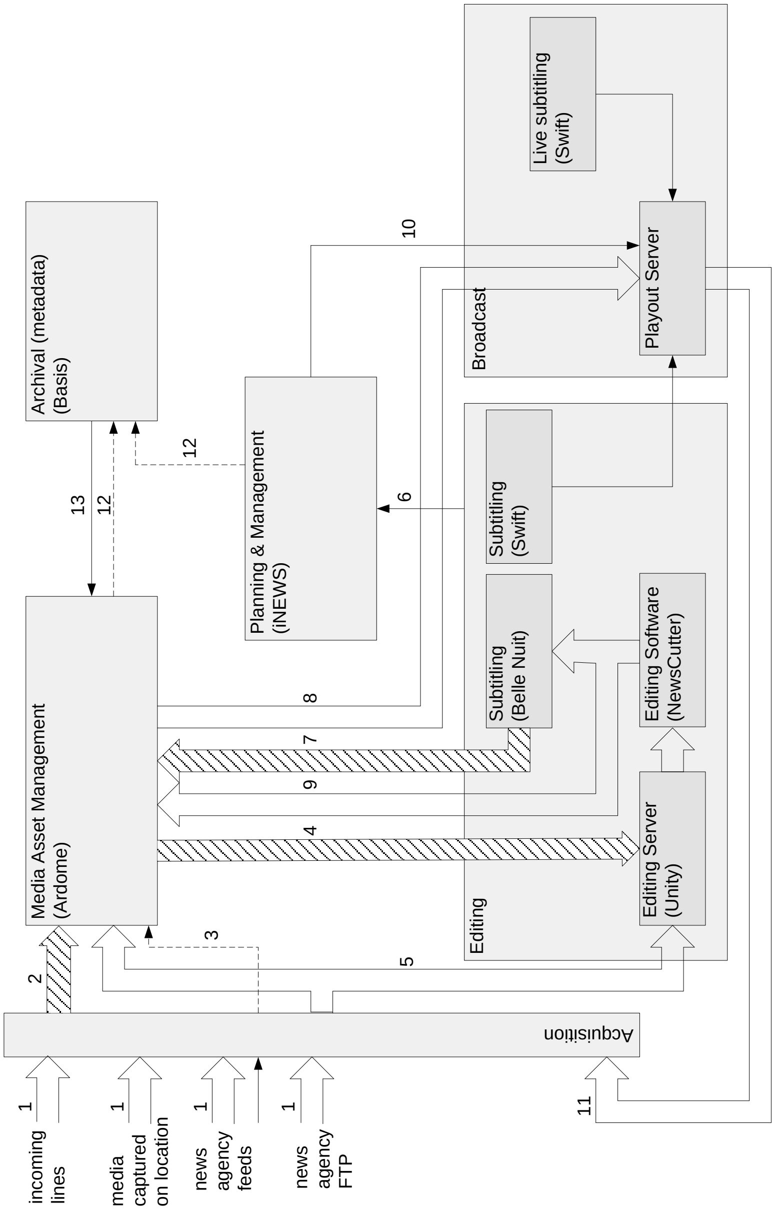


Fig. 1. News production process.

news broadcast. An item can be an introductory text displayed on the autocue for the news anchor, a report made by a journalist, a live interview, etc.

2.2 Acquisition

For most planned news items, relevant audiovisual content is needed. There are multiple options to obtain content related to a news item (label (1) in Fig. 1):

- audiovisual content captured by a news crew on location (typical for national news);
- content from incoming lines shot by other news providers (e.g., Reuters, EBU Eurovision and APTN);
- news feeds from news agencies, such as EBU Eurovision and Reuters, containing media files and additional metadata (typically represented according to the NewsML-G2 standard [10]);
- reuse of suitable audiovisual material obtained from the archive.

2.3 Ingest and Storage

Captured media is ingested and stored on servers managed by Ardome (Vizrt)[6]. Ardome is a Media Asset Management (MAM) system and is one of the core components in VRT's media production process. Consequently, it has many links with other information systems. Ardome contains all the produced media resources, both unfinished material as well as finished footage. In addition to media storage, Ardome provides other functionalities such as browsing, rough cut editing, and searching. In order to facilitate browsing and searching, low resolution versions of stored media are generated and accompanying metadata such as the title, audio track information and episode number can be inserted. However, during ingest, metadata insertion is kept to a minimum as this is currently a manual operation and consequently takes too much time. Therefore, metadata is often limited to a filename and a path indicating the location where the media is to be stored. Because captured content often needs further editing, content is also ingested in an editing server (Avid Unity ISIS[7]), which contains all media to be edited. As the editing server only hosts rough content that needs editing, the lifetime of this media is restricted to 72 hours.

In order to save time, content from incoming lines is often simultaneously ingested in the MAM and the editing server, resulting in a dual ingest (label (5) in Fig. 1). If dual ingest is not possible, the media is first ingested in the MAM (label (2) in Fig. 1) and is then sent from the MAM to the editing server (label (4) in Fig. 1). When content is captured from news feeds, the additional metadata is also inserted in the MAM (label (3) in Fig. 1). Note that the latter is again a manual operation (which is also indicated by a dashed line in Fig. 1).

[6] `http://vizrt.com/products/article138.ece`

[7] `http://www.broadcastautomation.com/products/unityISIS/index.asp`

2.4 Editing

After ingest, editing can be performed. The journalist who has been given the task to cover a news item retrieves the corresponding media from the editing server and starts editing it using Avid NewsCutter[8]. Simultaneously, anchor text (appearing on the autocue during broadcast) is provided in iNEWS by the journalist which afterwards is reviewed by the news anchor. Textual information that must appear as graphics on screen during the broadcast of a news item (e.g., the name of the interviewed person) is also provided in iNEWS (label (6) in Fig. 1).

An important task during editing is the generation of subtitles. Two types of subtitles can be considered. The first type, referred to as open subtitles, are subtitles that are 'burned' into the picture and therefore always appear on screen. A typical example of the use of open subtitles is when a translation is needed for an interviewed person speaking a foreign language. The second type of subtitles, referred to as closed subtitles, are by default not displayed but can be retrieved when requested (e.g., via Teletext).

A journalist is responsible for creating open subtitles and inserting these in iNEWS. A copy of the edited media is rendered containing the open subtitles on the screen using the Belle Nuit tool[9]. The edited media with subtitles is transferred to the MAM (label (7) in Fig. 1) and afterwards sent to the playout server for broadcasting (label (8) in Fig. 1). Also a link is provided in iNEWS relating the edited media with the corresponding item in the rundown. A version without subtitles is also sent to the MAM for later archiving (label (9) in Fig. 1).

If there is time left, journalists also create the closed subtitles using Swift[10]. If this is not the case, live subtitling is performed during broadcast by the subtitling department.

2.5 Broadcast

During broadcast, the edited media (with open subtitles) is available on the playout server. iNEWS executes the news rundown (label (10) in Fig. 1) and sends the anchor text to the autocue. Textual information (e.g., the name of the interviewed person) is also displayed on screen when needed using the Character Generator (CG).

The broadcast is again captured on an incoming line (label (11) in Fig. 1) by the media management department in order to have a copy of the integral news broadcast (which includes the open subtitles and textual information displayed by the CG on the screen). This captured broadcast is then also marked for archival.

[8] `http://www.avid.com/US/products/NewsCutter-Software/index.asp`

[9] `http://www.belle-nuit.com/subtitler/index.html`

[10] `http://www.softelgroup.com/product_info_1.aspx?id=0:53799&id=0:53783`

2.6 Archival

Archival is the last step in the news production process. An archivist who has been given the task to archive a news broadcast retrieves the rundown of the news broadcast from iNEWS and then searches for the corresponding media fragments in the MAM (label (12) in Fig. 1). Note that this is a manual operation and therefore indicated as a dashed line in Fig. 1. If a related fragment is found, the archivist watches it and generates a record containing relevant metadata using Open Text's Basis. Basis is the main tool used to document archived media at VRT in order to facilitate media retrieval. It is also the main tool for media search, e.g., when searching for archived content for reuse. Typically, a user searches for a relevant media fragment in Basis and subsequently retrieves it from the MAM.

As the archival step is performed as the last one in the media production process, it typically takes a few days before a media fragment is documented in Basis. When a Basis record is generated, some metadata fields that are also present in the MAM are transferred from Basis to the MAM (label (13) in Fig. 1), overwriting previously entered metadata in the MAM (e.g., metadata taken from the NewsML-G2 documents).

3 Production Process Evaluation

As can be seen from the discussion of the news production process, data is generated and spread across the entire production chain. Unfortunately, during a media search operation, only information generated by an archivist at the end of this chain is currently used. Other, possibly valuable information, is not used when searching for relevant media. We identified the following information sources containing valuable additional information for media search:

- **Ardome:** Ardome is the central MAM system used by VRT, containing all produced media. During ingest, metadata such as title, episode number, descriptive information of the audio tracks, aspect ratio, and video format can be inserted. However, as already noted in the previous section, this manual metadata insertion is always limited due to time constraints. Once the media has been documented in Basis, metadata is copied from Basis to Ardome.
- **News Agency provided metadata:** Incoming media from news agencies, such as EBU Eurovision and Reuters, is accompanied by metadata documenting the media in the NewsML-G2 format. This metadata can be inserted into the MAM during ingest. Typical metadata contained in a NewsML-G2 document are titles and textual descriptions (in English) of the media content.
- **iNEWS:** iNEWS contains important information as it is VRT's main tool for managing news productions. It is used to create the entire rundown of a news broadcast. Every news item from the rundown can be provided with anchor text, open subtitles, and textual information that must appear on the screen during broadcast by the CG.

- **Swift:** Closed subtitles are generated using the Swift tool. In addition to the subtitle text, Swift contains subtitle layout, spoken language indication, and timestamps indicating the appearance and disappearance time of a subtitle.
- **Basis:** Basis is considered to be the main tool for annotating and describing media resources at VRT. Basis has a relational database structure and currently contains over 600 000 records documenting media fragments spanning a period of over 20 years. A Basis record defines metadata fields such as title, duration, keywords, textual description, journalist, and identification number of the corresponding media in the MAM. Some fields, for example the keywords and journalist fields, can only contain controlled values, defined by manually maintained thesauri. The keywords thesaurus is the largest and contains over 300 000 terms. In addition to defining terms, relationships between terms are defined indicating 'narrower than', 'related' and 'used for'-relationships. Unfortunately, keywords are not categorized in terms of types, such as persons or locations.

By unifying all the information generated during news production, more efficient search applications can be realized. Also, by capturing information generated during news production, metadata can be generated and already filled in when creating an archival record, allowing an archivist to focus on the further enrichment of this metadata.

Subtitle information can significantly improve the media search operation [5, 6]. For example, when a certain keyword appears in a subtitle, timestamp information can be used to start playback from the indicated timestamp. Closed subtitles, generated with the Swift tool, are provided with such timing information. Unfortunately, the appearance and disappearance timestamps for open subtitles are not available, as journalists use the Belle Nuit tool only to paste the subtitles onto the relevant frames and no related timing information is stored during this operation. However, this timing information can be obtained after media production through the use of speech analysis or optical character recognition (OCR) tools.

Timing information indicating the appearance time of CG text can also improve media search operations. For example, when a user searches for media related to a person, timing information related to the CG text containing the name of this person can be used to start playback at this point. However, the exact moment CG text is displayed is decided during live broadcast and this information is currently not stored. This timing information could be reconstructed through the application of feature extraction tools on the generated media.

It is clear that although much of the data generated during the news production process can be used for media search operations, not all this information is currently stored, and therefore introduces the need for post-production operations to reconstruct this information. In order to avoid the need for these costly post-production operations, production systems should be changed in order to store this information.

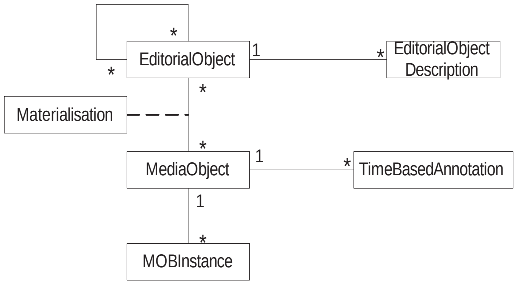


Fig. 2. Core concepts of the PISA media production model.

In addition to the fact that not all information is stored, the current news production process does also not guarantee the preservation of inserted metadata. For example, when a news item from a news agency is ingested, a manual insertion of metadata provided in the accompanying NewsML-G2 document is performed into the MAM. However, when an archival record is generated in Basis documenting the archived media, an update operation is performed which overwrites the original metadata inserted into the MAM. This introduces the loss of the metadata obtained from the NewsML-G2 document although this metadata might be important when searching for media and should therefore be preserved too.

4 Integrating Information Sources: Data Model

In order to represent the information present in the various identified information sources in a uniform way, we propose a data model specifically designed to preserve metadata along the news production process. The data model is based on the PISA data model [17] that was developed as part of the IBBT PISA research project[11]. Note that this model is also compatible with the P/Meta data model [7] defined by the European Broadcasting Union (EBU) and shares business objects with the BBC SMEF model [1]. The core concepts of the PISA data model are illustrated in Fig. 2.

The PISA data model centers around four elemental types of processes that contribute to media production, as described in [17]. The lifecycle of an entire product (e.g, a news broadcast) starts with the product planning process. During product planning, initial product requirements are specified. Because this information is typically delivered from an external Enterprise Resources Planning (ERP) system, the data model only represents an abstract notion of this process. The product planning process will, however, prepare a number of business objects for elaboration during the subsequent production processes.

[11] http://projects.ibbt.be/pisa

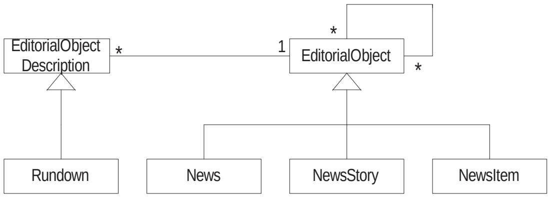


Fig. 3. The *EditorialObject* class, with subclasses *News*, *NewsStory*, and *NewsItem*, and associated *Rundown* description class.

The first of these processes is the product engineering process, in which the content of the product is determined by the editorial staff based on the initial product specifications. After the product engineering process, the product is defined as a composition of logically and editorially constituent parts, whereby each logical unit of creative or editorial work is represented by an editorial object. In the following manufacturing engineering process, information related to the manufacturing of a product is specified and represented in manufacturing objects. Finally, during the manufacturing process, the manufacturing objects are realized in the form of audiovisual material.

The following subsections describe how the news production process is represented using this generic media production model. We introduce a number of extensions specific for news production and indicate how the corresponding information sources are mapped onto the model.

4.1 Product engineering

An editorial object, as defined in [17], represents a unit of editorial work defined during the product engineering phase. For the news production process, we implemented three specialization subtypes of editorial object: *News*, *NewsStory*, and *NewsItem*. A news story represents a topic that is to be covered during a news broadcast. A story can consist of several news items. A news item is the atomic editorial object for news production and can, for example, correspond to an interview or a news anchor reading an introductory text from the autocue. Information contained in editorial objects is for example the story title, broadcast date, relevant keywords, etc.

Fig. 3 depicts the UML class diagram of the introduced editorial objects. The *EditorialObject* class can be reflexively associated through a many-to-many relationship. As a result, a story can be part of a news broadcast. However, a story does not need to be related to a news editorial object, as some stories are developed without eventually being part of a news broadcast. Note that application logic is responsible for prohibiting the occurrence of invalid relations, such as the fact that a news broadcast cannot be part of a news item.

The *Rundown* class is used to specify the editorial content of a news broadcast, which includes the list of news items to be broadcast, and the order in

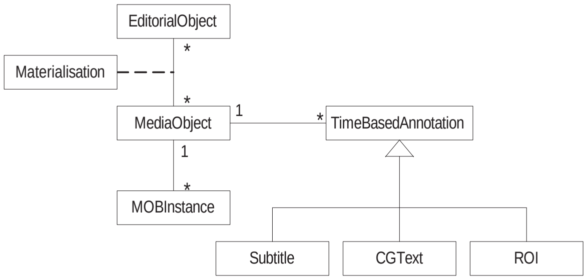


Fig. 4. The *MediaObject* class and it's relation with the *EditorialObject* class.

which these items are scheduled. Hence, a Rundown is attached to the News editorial object as a *EditorialObjectDescription* subclass. Similarly, anchor texts and other metadata that define the editorial content of the more granular news item are also incorporated into an EditorialObjectDescription. The data model supports the representation of iterative versions of EditorialObjectDescriptions to model rundowns and news items that change over time. However, in VRT's current news production process, only the final versions are effectively stored.

4.2 Manufacturing engineering

Considering the strict deadlines and well-established format of news production, there is little time and need for extended preparations and premeditation concerning the cinematography of news items. As such, the manufacturing engineering layer of the data model is of limited use since the translation of editorial object semantics to a media object is straightforward and does not require e.g., the accurate definition of camera positions by means of storyboarding or previsualization, as is often the case with more elaborate media production processes such as drama production.

4.3 Manufacturing

Due to the lack of manufacturing engineering in current day news production, objects from the product engineering layer can be related directly to objects in the lowest manufacturing layer. In the manufacturing phase, editorial objects are materialised into (audiovisual) *Media Objects*. In news production, different versions of a media object can be generated and stored. For example, a High Definition (HD) version and a downscaled version can be realized, containing the same audiovisual content. Every version is then represented by an instance of the *MediaObjectInstance* class and related to the corresponding Media Object, as illustrated in Fig. 4.

The *Materialisation* class associates an *EditorialObject* with a *MediaObject*, and contains information specifying how the editorial object was manufactured into the media object in question. An example of information that belongs to a materialization object is the name of the camera operator. The *TimeBasedAnnotation* class provides media objects with time-based annotations. For example, a media object can be annotated with a subtitle or a CG text together with its associated appearance and disappearance time.

4.4 Data source mapping

The information from the selected data sources needs to be mapped to our proposed data model. Note that we keep the sources of the different kinds of information, since the provenance of this information is crucial for future reuse[12].

- **iNEWS:** As already mentioned, iNEWS is used as a management tool for news production. Therefore, iNEWS is used for defining instances of the *News*, *NewsStory* and *NewsItem* editorial objects. The order of the news items is defined through an instance of the *Rundown* object. Anchor text belongs to the *NewsItem* editorial object as this is generated during the product engineering process. Open subtitles and CG text are implemented as instances of the corresponding subclasses of the *TimeBasedAnnotion* class, because this information has associated timing information.
- **Ardome:** Items stored in Ardome are represented through instances of the *MediaObjectInstance* class. Metadata present in Ardome belongs to the corresponding editorial objects that are related to these instances.
- **News Agency Provided Metadata:** Information taken from the NewsML-G2 documents such as title and textual descriptions is added to the corresponding instance of the *NewsItem* editorial object.
- **Swift:** Closed subtitles are represented as instances of the *Subtitle* class.
- **Basis:** Descriptive information present in a Basis record such as title, description, and keywords is associated with the corresponding editorial object. Information related to the realisation such as the camera operator, director, and recording date is provided as part of the *Materialisation* class. Technical information such as the video coding format is provided to the corresponding instance of the *MediaObjectInstance* class. Information that is common for all versions of a materialisation is represented as part of the *MediaObject* class. An example of information present in the *MediaObject* class is rights information.

5 Enabling Semantic Search

Unifying all generated information through the use of a common data model improves the potential of a search application significantly. However, by connecting this information with external data sets, even more intelligent search operations are enabled. The following sections describe the used architecture and give an overview of how external data sets are connected. The subsequent sections then illustrate some added functionality that is obtained by following this approach.

[12] `http://www.w3.org/2005/Incubator/prov/wiki/Presentations_on_State_of_the_Art`

5.1 Architecture

The data model was formalized into an OWL ontology and all relevant information from the selected data sources discussed in Section 3 was converted into a representation according to this formalized data model and stored in a RDF triple store. The resulting RDF data present in the triple store can then be queried using the SPARQL query language. By formalizing the data model and the corresponding information from the information sources, an unambiguous and machine processable representation of the available information is obtained. This in turn enables the use of reasoners, which can derive new facts based on the provided information.

The use of Semantic Web technologies also allows us to connect with other data sources using the Linked Open Data principles [2]. This enables us to use information from external data sets in order to make more intelligent search applications. Therefore, the search application makes use of a query facade which in turn uses data from several data sets according to the Linked Open data principles. The resulting architecture is illustrated in Fig. 5.

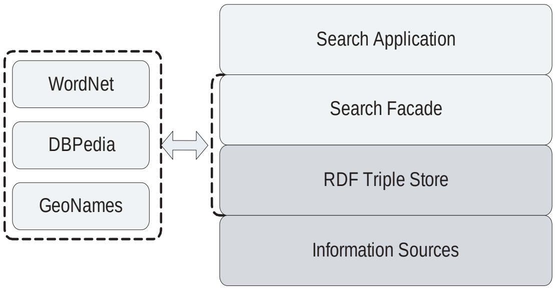


Fig. 5. Overview of the used architecture to enable semantic search.

5.2 Connecting with the Linked Open Data cloud

In order to be able to use information formalized in external data sets, concepts defined in our data model must be linked with the corresponding concepts defined in the external data sets. A valuable data set which is considered an important linking hub in the Linked Open Data cloud is DBPedia [4]. DBPedia is already used by other broadcasters [11] and proves to be a valuable data source. We use DBPedia to link concepts representing persons and other concepts, except for locations. For locations, the GeoNames[13] data set is used, providing a formalized representation of geographic locations.

[13] `http://www.geonames.org`

As a starting point, we use the Basis keywords as interlinking mechanism with the other data sets. As already mentioned, every Basis record contains a number of relevant keywords taken from the keywords thesaurus. The keywords can directly be linked to concepts defined in other data sets. For other fields containing textual data such as the Basis description field and Swift subtitles, Named Entitiy Recognition (NER) [15] must first be applied. NER is part of future work that will be performed in order to extend the set of relevant entities.

We formalized the keywords thesaurus in SKOS [13]. Consequently, every concept defined in the thesaurus corresponds to a URI. Also, the relationships defined between different concepts are also present in the SKOS representation. The relations defined between concepts in the thesaurus are important in the disambiguation of concepts during mapping. For example, the concept 'apple' defined in the keywords thesaurus has the related concepts 'iMac' and 'taligent'. These related concepts are then used to select the correct corresponding concept defined in DBPedia (`http://dbpedia.org/resource/Apple\_Inc.` instead of `http://dbpedia.org/resource/Apple`).

As already mentioned, locations are mapped to the corresponding concept defined in GeoNames. Again, we make extensive use of the relations defined in the thesaurus. For example, the concept 'Parijs' (Eng. Paris), has a used-for-relationship with the concept 'Paris'. Also, 'Parijs' is defined as a narrower term of 'Frankrijk' (Eng. France). With this added information, it is possible to select the corresponding concept from GeoNames.

However, for many concepts little or even no relationships at all are defined in the thesaurus. In this case, selecting the corresponding concept from an external data set can be very difficult as there is no additional context present that can be used for disambiguation. For this reason, statistics are currently collected from the Basis records, in order to get a set of keywords that most frequently co-occur with another keyword. These keywords will then be used as additional context information in order to select the correct concept from the external data set. Note also that because of the fact that the majority of keywords are defined in Dutch, often an additional translation is needed for successful mapping with concepts from external data sets, e.g. DBPedia.

By formalizing the keywords thesaurus and connecting it's concepts with external data sets, new functionalities are obtained. Some of these are illustrated in the following sections.

5.3 Suggesting alternative queries

The result set of a query depends on the user input. When a user enters a general keyword, the result set can be too large. In order to limit this result set, other keywords having a narrower-than relationship with the entered keyword can be displayed as a suggestion. On the other hand, when a query results in an empty result set, an alternative keyword can be suggested which has results.

5.4 Lexical information

As the thesaurus is maintained by hand, it is difficult to include all possible information related to an introduced keyword. Consequently, related information such as abbreviations and synonyms are often not included in the thesaurus. However, this makes it more difficult for a user to find results, as they have to know which words are included in the thesaurus. In order to provide a better search experience, a connection was made with a formalized version of the WordNet lexical database[14]. Note that this data set is also linked with Cornetto[15], a lexical semantic database for the Dutch language.

When a user enters a keyword, additional information can then be retrieved and included in the query in order to optimize the search operation. For example, when a user enters 'populatie' (Eng. population) the result set of this query would be small as this concept is not included in the thesaurus. However, it can be found in Cornetto that the concept 'populatie' has similar meaning to 'bevolking' (Eng. inhabitants), which is included in the thesaurus.

5.5 User query evaluation

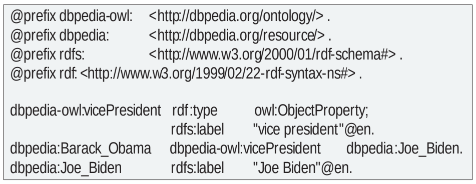
When a user enters keywords, reasoning can be performed in order to try to find out what a user really searches for. For example, when a user enters the keywords 'vice president Barack Obama', it can be derived that the user possibly searches media related to Joe Biden but maybe he does not recall his name. Although Joe Biden is present as a concept in the thesaurus, no relations are present indicating that Joe Biden is the vice president for Barack Obama.

In order to be able to show results having Joe Biden as a keyword, we use information available in external data sets as follows. When a query with multiple keywords is performed, we try to find a subject (or object) for which a property and a related object (or subject) exists corresponding with the entered keywords as follows.

From the entered keywords we retrieve the words Barack Obama and search for a corresponding concept in the formalized thesaurus having this as a label. This concept is present in the thesaurus and is already linked with the corresponding concept in DBPedia, thus allowing the use of that information available from DBPedia. In the following step, we try to map the other entered keywords with a property occurring in a triple having the concept of Barack Obama either as its subject or object. Fig. 6 illustrates that the property dbpedia-owl:vicePresident has as label vice president, allowing the identification of the needed property. Then we search for triples with the selected property and where Barack Obama appears either as subject or object. Finally, we obtain the concept dbpedia:Joe_Biden, which is also linked with the corresponding concept from the keywords thesaurus. This enables the inclusion of media provided with the keyword Joe Biden, as the user ultimately wanted.

[14] `http://semanticweb.cs.vu.nl/lod/wn30/`

[15] `http://www2.let.vu.nl/oz/cltl/cornetto/`



```
@prefix dbpedia-owl:    <http://dbpedia.org/ontology/> .
@prefix dbpedia:        <http://dbpedia.org/resource/> .
@prefix rdfs:           <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

dbpedia-owl:vicePresident   rdf:type       owl:ObjectProperty;
                            rdfs:label     "vice president"@en.
dbpedia:Barack_Obama     dbpedia-owl:vicePresident     dbpedia:Joe_Biden.
dbpedia:Joe_Biden           rdfs:label     "Joe Biden"@en.
```

Fig. 6. Selected triples from DBPedia used to optimize the result set related to the example query.

6 Conclusions and Future Work

In this paper, we proposed an architecture to optimize media search in a news production environment. Therefore, we evaluated VRT's news production process and identified data sources that contain information which can be used during media search. A data model was developed in order to uniformly represent the information contained in the data sources. The data model was implemented as an OWL ontology, enabling an unambiguous and machine processable representation of the information. This information is stored into a RDF triple store and queried by search applications using the SPARQL query language. Using Semantic Web technologies, a connection could be made with other valuable data sets that follow the Linked Data principles, e.g. DBPedia and GeoNames. Through the connection of all these information sources, advanced search functionalities are enabled and semantic search is obtained.

Future work includes an evaluation of the performance and efficiency of the developed search applications. Also, other valuable data sets from the Semantic Web will be evaluated. Relevant data sets will then be connected with the data model in order to further improve the media search. Currently, we are also researching how time-related concepts can be effectively represented and how new, inferred information can be automatically inserted and stored according to the data model. Named Entity Recognition (NER) tools will be evaluated in order to extract more relevant keywords from textual data such as subtitles. We are also implementing speech recognition software in order to extract timing information (e.g. when a certain keyword is spoken). Other feature extraction tools such as face recognition software will also be evaluated.

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.

References

1. British Broadcasting Corporation (BBC). Standard Media Exchange Framework (SMEF) Data Model, v1.10. 2005. Available at http://www.bbc.co.uk/guidelines/smef/.
2. Tim Berners-Lee. Design Issues: Linked Data. 2006. http://www.w3.org/DesignIssues/LinkedData.html,.
3. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 2009.
4. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, September 2009.
5. Martin G. Brown, Jonathan T. Foote, Gareth J. F. Jones, Karen Sparck Jones, and Steve J. Young. Automatic content-based retrieval of broadcast news. In *Proceedings of ACM Multimedia 95*, pages 35–43, San Fransisco, CA, USA, November 1995.
6. Franciska M. G. de Jong, Thijs Westerveld, and Arjen P. de Vries. Multimedia search without visual analysis: the value of linguistic and contextual information. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):365–371, March 2007.
7. EBU. P/Meta Metadata Exchange Scheme v1.1. Technical Report 3295. June 2005. Available at http://www.ebu.ch/en/technical/metadata/specifications/.
8. Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. Transcribing broadcast news for audio and video indexing. *Communications of the ACM*, 43(2):64–70, February 2000.
9. Ian Hargreaves and James Thomas. New news, old news. *ITC and BSC research publication*, October 2002.
10. International Press Telecommunications Council. NewsML-G2 v.2.2. November 2008. Available at http://www.iptc.org.
11. Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee. Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections. In *ESWC*, volume 5554 of *Lecture Notes in Computer Science*, pages 723–737. Springer, 2009.
12. Erik Mannens, Maarten Verwaest, and Rik Van de Walle. Production and multichannel distribution of news. *Multimedia Systems*, (14):359–368, July 2008.
13. Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. SKOS core: Simple Knowledge Organisation for the Web. In *Proceedings of the 2005 international conference on Dublin Core and metadata applications*, pages 1–9, Madrid, Spanien, 2005. Dublin Core Metadata Initiative.
14. Klaus Netter and Franciska de Jong. Olive: Speech based video retrieval. In *Proceedings of CBMI'99*, pages 75–80, Toulouse, France, October 1999.
15. Hien T. Nguyen and Tru H. Cao. Named entity disambiguation: A hybrid statistical and rule-based incremental approach. In *3rd Asian Semantic Web Conference (ASWC08)*, volume 5367, pages 420–433. Springer, 2008.
16. John R. Smith and Peter Schirling. Metadata Standards Roundup. *IEEE Multimedia*, 13(2):84–88, 2006.
17. Dieter Van Rijsselbergen, Maarten Verwaest, Barbara Van De Keer, and Rik Van de Walle. Introducing the Data Model for a Centralized Drama Production System. In *Proceedings of the IEEE Intl. Conference on Multimedia & Expo 2007*, pages 615–618, July 2007.