AUTOMATIC METADATA ENRICHMENT IN NEWS PRODUCTION

E. Mannens, R. Troncy#, K. Braeckman+,*
D. Van Deursen, W. Van Lancker*, R. De Sutter+ and R. Van de Walle**

*Ghent University – IBBT, Multimedia Lab, Ledeberg-Ghent, Belgium
#CWI, Amsterdam, The Netherlands
+VRT-medialab, Brussels, Belgium

ABSTRACT

News production is characterized by complex and dynamic workflows in which it is important to produce and distribute news items as fast as possible. In this paper, we show how personalized distribution and consumption of news items can be enabled by automatically enriching news metadata with open linked datasets available on the Web of data, thus providing a more pleasant experience to fastidious consumers where news content is presented within a broader historical context. Further we present a faceted browser that provides a convenient way for exploring news items based on an ontology of NewsML-G2 and rich semantic metadata.

1. INTRODUCTION

The Internet – and information technology in general – is having a major impact on the broadcast industry. Physical carriers of audio-visual material are being replaced by files, and analogue networks by IP-based networks. Therefore, international news agencies can now distribute multimedia news items enriched with metadata to their customers as soon as they are produced. Consumer expectations have also raised far beyond the traditional radio and television medium. Cross-functional and configurable content [1], cleverly engineered to scale from a three inch screen up to high definition, will gracefully complement current uni-size news television and radio programmes. Eventually, mechanized and sequential news manufacturing methods will be more and more often complemented by scalable, configurable, and agile methods [2] capable of supporting exactly the window of interactivity desired by the end-user [3]. The latest multimedia news stories can be personalized for the consumers and presented within a broader historical context using enhanced semantics [4].

Furthermore, the metadata lifecycle in the audiovisual content creation news environments has undergone a significant development during the last decade. The digital evolution has changed the way the metadata is generated and how agents are involved in the metadata workflow [5]. Until recently, the annotations of the broadcasters' archives were mainly managed by the archivists but rarely generated and enriched automatically. Consequently, the metadata related with the production and distribution, and the content itself were not treated uniformly. Despite of the existence of standards for representing and querying metadata, most of the solutions in the broadcast industry are proprietary and customized to specific needs.

In the PISA project[1] we have investigated how, given a file-based media production and broadcasting system with a centralized repository of metadata, many common production, indexing and searching tasks could be improved and automated. In particular, we aim to automate the production of multiple versions of news bulletins for different consumption platforms since news broadcasters generally aggregate and produce more material than is required for broadcast or online distribution. We have shown in [6] how an up-to-date news bulletin can be dynamically created and personalized to match the consumer's categories preferences by merging different news sources and using the NewsML-G2 specification [7]. In this paper, we build on this work and exploit further the semantic capabilities of NewsML-G2. We show how to automatically enrich the metadata with knowledge available on the Web of data and we present a faceted browser for exploring news items within a broader historical context.

This paper is organized as follows. In Section 2, we present the NewsML-G2 standard and its conceptualization in an OWL ontology. In Section 3, we explain how news metadata can be automatically enriched with knowledge about people, organizations, places and events available in large linked datasets. In Section 4, we propose to use a faceted browser for presenting personalized news content to the end-user. Finally, we give our conclusions and outline future work in Section 5.

The research activities that have been described in this paper were funded by Ghent University, VRT-medialab, CWI, the Interdisciplinary Institute for Broadband Technology (IBBT) through the PISA-project (50% co-funded by industrial partners), 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.

[1] `http://projects.ibbt.be/pisa`

2. SEMANTIC WEB BASED INFRASTRUCTURE FOR NEWS METADATA

Broadcasters receive news information from various sources. For example, the Flemish public service broadcaster in Belgium – VRT (*Vlaamse Radio- en Televisieomroep*) gathers its material from both its own news crews and from several international news agencies such as *Reuters* and *EBU Eurovision*. The rough-cut and mastered essence created by the news crews is then stored into the VRT Media Asset Management (MAM) system. Reporters use the AVID's iNews application[2] to enlarge the essence by adding descriptive information such as captions, anchor texts and other metadata, and to organize the rundown of a classical television news broadcast represented in the NewsML-G2 format [7] . Then the essence needs to be packed into an Material Exchange Format (MXF) instance before the MAM can process it. Afterwards, the essence is transcoded into a consumer format, such as H.264/AVC. As the use of NewsML is key within our framework, it is being elaborated on in the next sections.

2.1. NewsML-G2 Overview

The IPCT News Architecture framework (NAR[3]) is a generic model that defines four main objects (*newsItem*, *packageItem*, *conceptItem* and *knowledgeItem*) and the processing model associated with these structures. Specific languages such as NewsML-G2 or EventsML-G2 are built on top of this architecture. For example, the generic *newsItem*, a container for one particular news story or *dope sheet*, is specialized into media objects (textual stories, images or audio clips) in NewsML-G2.

Within a *newsItem*, the elements *catalog* and *catalogRef* embed the references to appropriate taxonomies; *rightsInfo* holds rights information such as who is accountable, who is the copyright holder and what are the usage terms; *itemMeta* is a container for specifying the management of the item (e.g. title, role in the workflow, provider). The core description of a news item is composed of administrative metadata (e.g. creation date, creator, contributor, intended audience) and descriptive metadata (e.g. language, genre, subject, slugline, headline, dateline, description) grouped in the *contentMeta* container. A news item can be decomposed into parts (e.g. shots, scenes, image regions and their respective descriptive data and time boundaries) within *partMeta* while *contentSet* wraps renditions of the asset. Finally, semantic inline markup is provided by the *inlineRef* container for referring to the definition of particular concepts (e.g. person, organization, company, geopolitical area, point of interest).

2.2. NewsML-G2 Ontology

NAR is a generic model for describing news items as well as their management, packaging, and the way they are exchanged. Interestingly, this model shares the principles underlying the Semantic Web: *i)* news items are distributed resources that need to be uniquely identified like the Semantic Web resources; *ii)* news items are described with shared and controlled vocabularies. NAR is however defined in XML Schema and has thus no formal representation of its intended semantics (e.g. a *NewsItem* can be a *TextNewsItem*, a *PhotoNewsItem* or a *VideoNewsItem*). Extension to other standards is cumbersome since it is hard to state the equivalence between two XML elements. We have proposed to model an OWL ontology of NewsML-G2[4] to address these shortcomings and we have discussed the design decisions regarding its modeling from existing XML Schemas [4].

Flattening the XML structure. XML Schema provides the means to have very rich structure but is rather limited when expressing the meaning of this structure as the language is (only) concerned with providing typing and structuring information for isolated chunks of data. Consequently, the NAR model defines intermediate structures and *containers* whose only goal are to group a number of properties without particular semantics. These structures should not be represented in the ontology, as they will generate blank nodes in the RDF graph at the instance level, complexifying its visualization in any Semantic Web browser. While modeling the ontology, we therefore advocate to flatten the XML structure keeping only the properties that will be instantiated.

Linking with Media Ontologies. Many other multimedia standards such as EXIF, Dublin Core, XMP, DIG35 or MPEG-7 are used in the media industry. These standards have generally been converted into OWL ontologies and can thus be integrated within our ontology infrastructure.Therefore, we have added OWL axioms stating the relationship between resources defined in different but strongly overlapping ontologies. For example, the NAR ontology contains the following axioms: *nar:subject owl:equivalentProperty dc:subject* and *nar:Person owl:equivalentClass foaf:Person*.

3. ENRICHING NEWS METADATA WITH THE LINKED DATA CLOUD

Once the NAR ontology has been modeled and linked to other media ontologies, the conversion of the metadata of individual news items into RDF according to this ontology is straightforward. However, we advocate a further step aiming at enriching semantically the news metadata following the linked data principle[5]. In our case, we apply linguistic processing on

[2] http://www.avid.com/products/iNews/index.asp

[3] http://www.iptc.org/NAR/

[4] http://newsml.cwi.nl/ontology/

[5] http://linkeddata.org/

the plain text contained into some XML elements of the metadata such as *title*, *caption*, *description* and the main body of the news stories.

The linguistic processing consists in extracting named entities such as persons, organizations, companies, brands, locations and events. We use the OpenCalais infrastructure[6] for extracting these named entities. For example, the processing of the headline "Barack Obama for president of the U.S.A." will result in three named entities: 'Barack Obama', 'president' and 'U.S.A.' together with their type (i.e. person, location, etc.). Once the named entities have been extracted, we map them to formalized knowledge on the web available in GeoNames[7] for the locations, or in DBPedia[8] for the persons, organizations and events. The string 'Barack Obama' is therefore mapped to its URI in DBPedia[9] that provides *i)* a unique identifier for the resource and *ii)* formalized knowledge about this person such as his biography, career and genealogy in multiple languages. Therefore, the use of the OpenCalais web service allows us to populate the knowledge base by providing a list of possible instances for all named entities discovered.

The main challenge in this semantic enrichment step is then to deal with the ambiguity. For example, the GeoNames web service tends to return primarily a US city when a single string is passed as an argument. Fortunately, news items contain always information about the city and the country yielding accurate recognition of the location mentioned in the story. We are currently implementing and evaluating more sophisticated disambiguation heuristics such as the IdentityRank algorithm [8] and a hybrid statistical approach [9] to minimize the disambiguation errors. When the accuracy is the primary concern, we envision a semi-automatic approach where suggestions will be proposed to the journalist during the annotation process. Furthermore we will integrate the extracted metadata from shot segmentation tools, scene detection tools, and face recognition tools of the PISA project to be able to also further enrich these extracted named entities.

4. PRESENTING PERSONALIZED NEWS CONTENT

In order to demonstrate the appropriateness of our proposed ontology infrastructure, we present a web-based faceted browser for presenting personalized news items. The web-based user interface uses Google's Web Toolkit and connects to a SPARQL endpoint where all RDF metadata of the news items is stored. The top panel in Fig. 1 shows examples of possible facets.

Similarly to [10], the facets are dynamically fetched and can be configured to match the users needs. They correspond to the datatype and object properties of the class hierarchy that has the root *AnyNewsItem*. Within each facet (*the predicate*), a list (*the object range*) is composed of instances (for object properties) or literal values (for datatype properties) corresponding to the news items available (*the subject domain*). Selecting facets and their values enable to build complex queries made of multiple filters that further constrain the search of particular news stories.

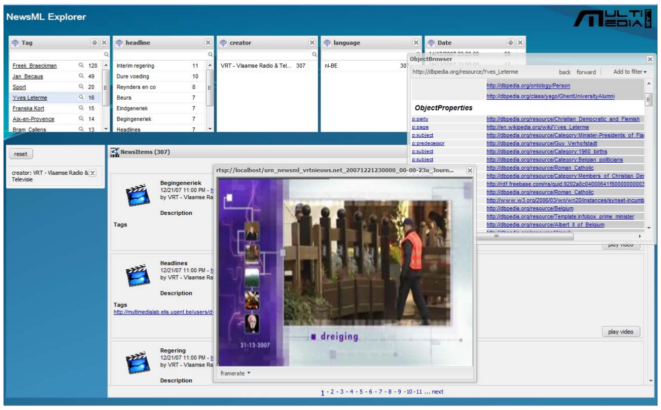
NewsML Explorer
Tag
Freek_Braeckman 120
Jan_Becaus 49
Sport 20
Yves Leterme 16
Franska Kort 15
Aix-en-Provence 14
Bram_Callens 13
headline
Interim regering 11
Dure voeding 10
Reynders en co 8
Beurs 7
Eindgeneriek 7
Begingeneriek 7
Headlines 7
creator
VRT - Vlaamse Radio & Tel... 307
language
nl-BE 307
Date
ObjectBrowser
http://dbpedia.org/resource/Yves_Leterme
back
forward
Add to filter
http://dbpedia.org/ontology/Person
http://dbpedia.org/class/yago/GhentUniversityAlumni
ObjectProperties
p:party
p:page
p:subject
p:predecessor
p:subject
p:subject
http://dbpedia.org/resource/Christian_Democratic_and_Flemish
http://en.wikipedia.org/wiki/Yves_Leterme
http://dbpedia.org/resource/Category:Minister-Presidents_of_Fla
http://dbpedia.org/resource/Guy_Verhofstadt
http://dbpedia.org/resource/Category:1960_births
http://dbpedia.org/resource/Category:Belgian_politicians
http://dbpedia.org/resource/Roman_Catholic
http://dbpedia.org/resource/Category:Members_of_Christian_Der
http://rdf.freebase.com/ns/guid.9202a8c04000641f80000000003
http://dbpedia.org/resource/Roman_Catholic
http://www.w3.org/2006/03/wn/wn20/instances/synset-incumb
http://dbpedia.org/resource/Belgium
http://dbpedia.org/resource/Template:infobox_prime_minister
http://dbpedia.org/resource/Albert_II_of_Belgium
reset
creator: VRT - Vlaamse Radio & Televisie
NewsItems (307)
Begingeneriek
12/21/07 11:00 PM
by VRT - Vlaamse Ra
Description
Tags
Headlines
12/21/07 11:00 PM
by VRT - Vlaamse Ra
Description
Tags
http://multimedialab.elis.ugent.be/users/d
Regering
12/21/07 11:00 PM
by VRT - Vlaamse Ra
Description
rtsp://localhost/urn_newsml_vrtnieuws.net_20071221230000_00-00-23u_Journ...
21-12-2007
dreiging
framerate
play video
play video
1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 10 - 11 ... next

Fig. 1. Faceted browser for presenting personalized news

The bottom panel in Fig. 1 depicts the list of news items that match the filters currently selected with the facets. Some relevant item object properties and datatype properties are shown together with the extra records that were captured during the enrichment phase (see Section 3) and visualized as a tag cloud. This bottom panel is built using a Fresnel lens [11] which describes the formatting of the properties to be displayed.

Each facet is finally searchable either in plain text or for more complex objects (i.e. another web resource with a type). Similarly to [10], the faceted browser can pop up and display the *local view* for any objects. The RDF triples are then retrieved from the store and formatted according to an appropriate stylesheet. Any property can therefore be added as a filter on its own, thus re-iterating the cycle of selecting other resources within the pool of existing news items.

In [12], the authors propose an approach that goes beyond Fresnel where one does not only specify how RDF resources should be visualized, but also what should be the behavior of each specific widget in the interface (e.g. graph search algorithm, facet queries, auto-completion, etc.). As a result, the whole user interface design is reduced to a mapping between the RDF data and a generic but extensible interface model. We are investigating how to embed this model within Ninsuna, our generic format-independent multimedia content adaptation platform [13].

[6] http://www.opencalais.com/

[7] http://www.geonames.org

[8] http://dbpedia.org

[9] http://dbpedia.org/resource/Barack_Obama

5. CONCLUSION

In this paper, we have presented an OWL ontology of the NewsML-G2 standard and we have proposed to automatically enrich news metadata using large linked datasets available on the Web of data. News stories become more interlinked and we have shown how a faceted browser interface can be used by an end-user to compose an enriched personalized news bulletin. The knowledge added during the enrichment step allows to view and understand news stories within a broader historical context thus enhancing the overall user experience.

We are in the process of integrating various feature analysis & extraction tools from the PISA project to get more relevant metadata from the audio-visual news items in order to get an even bigger set of named entities that can be used for further enrichment. The added-value brought by the editorial work of the professionals themselves when producing news should not be underestimated. Getting feeds and selecting individual news items automatically using rich semantic metadata is a first step but it does not build a complete story with a flow. For example, we are investigating how story telling approaches can be adapted to this aim.

6. REFERENCES

[1] D. Van Deursen, F. De Keukelaere, L. Nachtegaele, J. Feyaerts, and R. Van de Walle, "A Scalable Presentation Format for Multichannel Publishing Based on MPEG-21 Digital Items," in *International Workshop on Multimedia Content Representation, Classification and Security (MRCS'06)*, Istanbul, Turkey, 2006, pp. 650–657.

[2] E. Mannens, M. Verwaest, and R. Van de Walle, "Production and Multi-Channel Distribution of News," *Multimedia Systems Journal, Special Issue on Canonical Processes of Media Production*, vol. 14, no. 6, pp. 359–368, 2008.

[3] F. Pereira and I. Burnett, "Universal multimedia experiences for tomorrow," *IEEE Signal Processing Magazine*, vol. 20, no. 2, pp. 63–73, 2003.

[4] R. Troncy, "Bringing the IPTC News Architecture into the Semantic Web," in *7th International Semantic Web Conference (ISWC'08)*, Karlsruhe, Germany, 2008, pp. 483–498.

[5] J. A. G. Avilés, B. Leon, K. Sanders, and J. Harrison, "Journalists at digital television newsrooms in Britain and Spain: workflow and multi-skilling in a competitive environment," *Journalism Studies*, vol. 5, no. 1, pp. 87–100, 2005.

[6] R. De Sutter, E. Mannens, D. Van Rijsselbergen, R. Van de Walle, and M. Verwaest, "Automatic News Production," in *International Broadcasting Conference (IBC'08)*, Amsterdam, The Netherlands, 2008, pp. 158–165.

[7] International Press Telecommunications Council (IPTC), "NewsML-G2 Specification - version 2.2," 2008, `http://www.iptc.com/std/NewsML-G2/NewsML-G2_2.2.zip`.

[8] N. Fernández, J. M. Blázquez, L. Sánchez, and A. Bernardi, "IdentityRank: Named Entity Disambiguation in the Context of the NEWS Project," in *4th European Semantic Web Conference (ESWC'07)*, Innsbruck, Austria, 2007, pp. 640–657.

[9] H. T. Nguyen and T. H. Cao, "Named Entity Disambiguation: A Hybrid Statistical and Rule-Based Incremental Approach," in *3rd Asian Semantic Web Conference (ASWC'08)*, Bangkok, Thailand, 2008, pp. 420–433.

[10] M. Hildebrand, J. van Ossenbruggen, and L. Hardman, "/facet: A Browser for Heterogeneous Semantic Web Repositories," in *5th International Semantic Web Conference (ISWC'06)*, Athens, USA, 2006, pp. 272–285.

[11] E. Pietriga, C. Bizer, D. Karger, and R. Lee, "Fresnel: A Browser-Independent Presentation Vocabulary for RDF ," in *5th International Semantic Web Conference (ISWC'06)*, Athens, USA, 2006, pp. 158–171.

[12] M. Hildebrand and J. van Ossenbruggen, "Configuring Semantic Web Interfaces by Data Mapping," in *IUI Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW'09)*, Sanibel Island, USA, 2009.

[13] D. Van Deursen, W. Van lancker, T. Paridaens, W. De Neve, E. Mannens, and R. Van de Walle, "NinSuna: a Format-independent Multimedia Content Adaptation Platform based on SemanticWeb Technologies," in *International Symposium on Multimedia (ISM'08)*, Berkeley, USA, 2008, pp. 491–492.