Television meets the Web: a Multimedia Hypervideo Experience

José Luis Redondo García, Raphaël Troncy

EURECOM, Sophia Antipolis, France,
{redondo, raphael.troncy}@eurecom.fr

Abstract. Nowadays, more and more information on the Web is becoming available in a structured and easily searchable way. Multimedia and particularly television content has to adopt those same Web principles in order to make the consumption of media items and Web browsing seamlessly interconnected, no matter if the content is coming live from a local broadcaster or from an online video streaming service. This research proposes the creation of a methodology for representing, searching and interlinking multimedia items with other existing resources. This model has to be easily extensible and Web compliant in order to make the generated annotations available to the outside world and reused in other similar applications.

Keywords: Hypervideo, Television, Multimedia Ontology, NER, Second Screen

1 Problem Statement

The amount of multimedia content available is huge: there are billions of pictures and videos spread in very different ecosystems that keep growing. Since each of those ecosystems has is own peculiarities, it is not easy to identify which media items are relevant in a particular context, and how they can be effectively consumed or re-used by users. Television content constitutes a subset of this multimedia world where the presence of isolated ecosystems becomes even more obvious. Most of the television providers are media silos ruled by their own idiosyncrasy: broadcasters such as the German RBB or the French TF1 groups are the owners of tons of television content that are not conveniently exposed and interlinked with the rest of the world.

Other open issue is the way those media items are described. In most of the cases, the content is considered like an unitary piece that does not need to be further fragmented. However, in many situations, a more fine-grained decomposition of the media resource is needed, in order to point to particular fragments where something of interest is happening. The broad variety of non-interoperable standards for representing those descriptions, such as TV-Anytime[1] or MPEG-7[2], has been largely recognized.

[1] http://tech.ebu.ch/tvanytime
[2] http://mpeg.chiariglione.org/standards/mpeg-7

Finally there is a need of complementing seed video programs with additional resources on the Web, such as domain related web sites, encyclopedic sources or even conversation happening on social platforms. Users' status updates and tweets enable people to share their activities, feelings and emotions opening a window to a world of fresh information that can successfully illustrates what is being broadcasted in the television screen.

2 Relevancy

Multimedia content is rapidly increasing in scale and ubiquity but it still remains largely poorly indexed and unconnected with other related media from other sources. The current state of the TV domain clearly reflects this fact: there are no clear approaches for going further than a simple PC-like browsing experience on a full screen. Users wish to have new functionalities such as getting access to information not explicitly present in the television content itself, like browsing from a local news show to an open government data portal about a particular location in order to understand voting patterns, or learning more about animals and plants shown in a nature documentary without leaving that show.

The information available in social platforms is growing and becoming more and more attached to television programs. Filtering this massive amount of data and trying to make sense out of it is an extremely challenging task due to the heterogeneity and dynamics of the information. Such an analysis is, however, very valuable for explaining or illustrating what is going on in a video using similar audiovisual content or encyclopedia knowledge, for completing missing information, or for providing other users' point of view about the same fact or topic.

3 Related Work

Several approaches for describing multimedia content in general and television content in particular have been proposed. One of the most relevant ones is the model proposed by the British broadcaster, the BBC Programmes Ontology[3]. This ontology includes concepts about various aspects of a television program such as series, brands, episodes, etc. In [1], the authors describe how the BBC is working to integrate data and linking documents across their domains by using Semantic Web technology. The latest version of schema.org includes also classes related to the TV domain, such as `TVSeries`, `TVSeason` and `TVEpisode` which belong to the SchemaDotOrgTV[4] proposal. This set of classes include some improvements over the BBC model, like a better relationship between episodes, series and seasons or the addition of clips and broadcast services.

Those attempts consistently consider the television program as a whole and do not consider its sub-parts or fragments. The possibility of working with pieces

[3] `http://purl.org/ontology/po/`
[4] `http://www.w3.org/wiki/SchemaDotOrgTV`

of different granularities is a crucial requirement for implementing true hypervideo systems. There are standards that use non-URI based mechanisms to identify parts of a media resource, such as MPEG-7 or the Synchronized Multimedia Integration Language (SMIL)[5]. In the group of URI-based approaches, temporalURI[6] was the first to define media fragments using the query parameter in a URI. The W3C Media Fragment Working Group has edited the Media Fragment URI 1.0 specification[7], which defines a hash URI syntax for identifying and referencing fragments of audiovisual content in the Web. Some systems, such as Synote [3], rely on these media fragments for representing information about subtitles and entities in videos. In [2], we have used media fragments and entities for classifying videos from Dailymotion and YouTube.

Crawling social platforms for enriching a media resource has been proposed in several work. Liu *et al.* combine semantic inferencing and visual analysis to automatically find media to illustrate events [4]. Visual, temporal and spatial similarity measures are used for attaching photo streams with events in [9]. We developed a generic media collector for retrieving media items shared on social platforms [6]. Extra insights about the facts and events illustrated in the media items collected are inferred by performing non-supervised clustering and labeling processes on the result set.

4 Research Questions

The first challenge is to find an appropriate ontology model for correctly representing the metadata about a particular media resource. We hypothesis that Media Fragment URI can be used to identify part of media content, but it is still a key topic of discussion how the relationship between media fragments and a media resource should be represented and what are the implications for the annotations.

The second challenge is how to enrich and improve the available metadata by retrieving extra information from the Web and media items published in social platforms. Some questions are: (i) what are the anchors or properties inside a media resource that can best describe a particular fragment, (ii) what is the best way to materialize links between different media fragments, and (iii) how to formalize the context surrounding an annotation including its provenance, the motivation that leads to the creation of the annotation, etc., in order to better support search and hyperlink operations.

Finally, we aim to investigate how this enriched television content with data from the web interconnected to its subparts should be consumed and displayed to the end-user. The presence of multiple links to other resources opens a window to a great variety of new television applications where a much minimalist interaction and interface design should be implemented.

[5] `http://www.w3.org/AudioVideo/`
[6] `http://annodex.net/TR/draft-pfeiffer-temporal-fragments-03.txt`
[7] `http://www.w3.org/TR/media-frags/`

5 Hypotheses

This research proposed the use of a semantic graph metadata representation for implementing innovative hypervideo systems. The resulting RDF data is flexible enough to include different types of content description in a structured way: it can be completed with information from external resources, it naturally supports links with other pieces of content, and its web nature enables to bring hypermedia experience to the TV field.

A video program can be decomposed into segments, either automatically or manually, in order to create a hierarchy (structure) of media fragments which can be further indexed and semantically described with resources from the Web of Data. In our hypothesis, those anchors are Named Entities [8] spotted by different extractors used on timed texts coming with a media resource. Those entities represent a bridge between the audiovisual content and related information in the Web of Data that is potentially relevant for the viewer. By filtering and ranking those entities, the important parts of the video can be identified, modifying and further adjusting an existing segmentation result obtained via a visual analysis. Finally, the set of relevant entities inside a particular Media Fragment becomes an appropriate way of characterizing it, which allows to infer how similar a part of the video is to other fragments from other multimedia resources. By making explicit relationships between analogous media fragments, the expected hypermedia experience can be effectively created.

6 Approach

This sections shows a first implementation of the proposed hypotheses made in the context of the LinkedTV project[8], which will be further refined during the subsequent phases of this doctoral research.

RDF conversion. In a first step, some legacy metadata or some results coming from automatic multimedia analysis processes (e.g. face detection, shot segmentation, scene segmentation, concept detection, speaker identification, automatic speech recognition, etc.) are converted into RDF and represented according to the LinkedTV Ontology[9]. We have developer a REST API service named *tv2rdf*[10] to perform this operation. The video content is structured into parts, with different degrees of granularity, identified using Media Fragments URIs. For better classifying those different levels of segmentation, the LinkedTV ontology includes classes such as `Chapter`, `Scene` or `Shot`.

Those instances of the *ma:MediaFragment* class are anchors where entities will be attached in the following serialization step. The media fragment generation introduces a very important level of abstraction that opens many possibilities when annotating certain parts of the analyzed videos and makes possible to associate to fragments other metadata with temporal references. The underlying

[8] `http://www.linkedtv.eu/`
[9] `http://data.linkedtv.eu/ontologies/core`
[10] `http://linkedtv.eurecom.fr/tv2rdf`

model also relies on well-known ontologies such as the *The Open Annotation Core Data Model*[11], the *Ontology for Media Resources*[12] and the *NERD ontology*. A schema of the ontology is depicted in Figure 1.

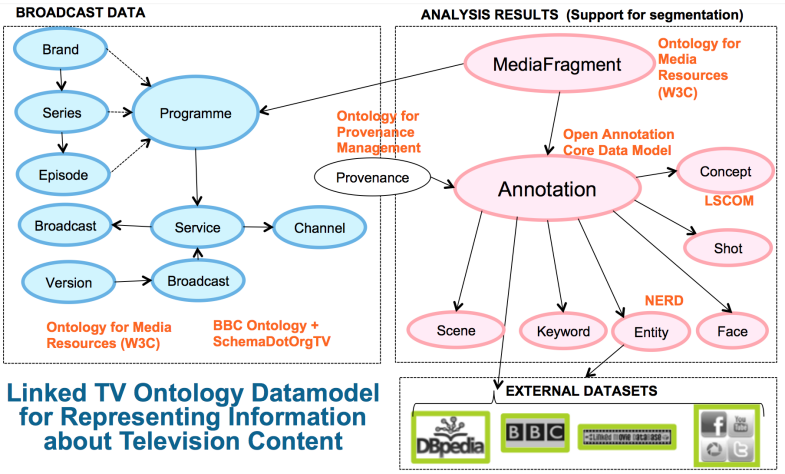

Fig. 1. LinkedTV Ontology.

The listing below corresponds to the description (Turtle serialization) of a media fragment from the *Tussen Kunst en Kitsch* show. The temporal references are encoded using the NinSuna Ontology[13]. The fact that one media fragment belongs to a larger media resource is made via the property `ma:isFragmentOf`.

```
<http://data.linkedtv.eu/media/e2899e7f-67c1-4a08-9146-5a205f6de457#t
    =1492.64,1504.88>
      a  nsa:TemporalFragment , ma:MediaFragment ;
      nsa:temporalStart "1492.64"^^xsd:float ;
      nsa:temporalEnd "1504.88"^^xsd:float ;
      nsa:temporalUnit "npt" ;
      ma:isFragmentOf <http://data.linkedtv.eu/media/e2899e7f-67c1-4a08
          -9146-5a205f6de457> .
```

Broadcasters generally provide basic legacy metadata related to the TV content such as EPG information (title, description, tags, channel, category, duration, language) and subtitles. Those items are also included in the RDF graph during the serialization process.

Name Entity Extraction. Once the RDF graph is built, some nodes are further interlinked with the Linked Open Data Cloud. Named entity extractors are used over the transcripts of the TV content (either the subtitles of the television program or the automatic speech recognition (ASR) results). The tv2rdf REST service launches this task by relying on a *NERD Client*, part of the

[11] `http://www.openannotation.org/spec/core`
[12] `http://www.w3.org/ns/ma-ont`
[13] `http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna`

NERD[14] framework. A multilingual entity extraction is performed over the video transcript and the output result is a collection of entities that are temporally related to a video. The entities are classified using the core NERD Ontology[15] and attached to the right Media Fragment according to their temporal appearance.

Both Dublin Core[16] and LinkedTV properties are used in order to specify the entity label, confidence and relevance scores, the name of the extractor used, the entity type and the disambiguation URI for this entity (for example, a resource from DBPedia). This entity extraction process can be extended to be also applied over other textual resources such as users comments or notes from the publisher.

Enrichment. In a third step, the named entities extracted in the previous step are used to trigger the enrichment process that consists in retrieving additional multimedia content that can illustrate what is shown or discussed in a seed television program. The logic for accessing the external datasets where this information can be collected is implemented inside the LinkedTV REST service MediaCollector[17]. MediaCollector gets as input the label of entities spotted by NERD and provides as result a list of media resources (photos and videos) grouped by source [7]. Those sources include mainly social platforms such as Twitter or YouTube but also domain-related web sites provided as white list of content that should be mined. When serializing the information into RDF, every item returned by the MediaCollector is represented as a new `ma:MediaResource` instance according to the Ontology for Media Resources. The entity used as input in the media discovery process is linked to the retrieved items through an `oa:Annotation` instance from the Open Annotation Ontology.

Search and Hyperlinking. Entities and related media items can then be used to describe and illustrate a particular media fragment of the television program. Given some input parameters from the viewer and analyzing the entities that are relevant for a video fragment, it is possible to filter out the ones that can be potentially interesting for the user. When different media fragments share similar named entities, they can be explicitly interrelated through the creation of hyperlinks that allow the user to navigate from one multimedia content to the other.

7 Reflections

We aim to publish the RDF metadata following the linked data principles. Hence, the resulting RDF graph can not only be stored and queried to enable data visualization such as a second screen application [5], but it can also be re-used by other RDF consuming applications. Once the metadata about a particular content has been gathered, serialized into RDF, and interlinked with other resources in the Web, it is better suited to be used in the subsequent consumption phases such as an editorial review or a data visualization. The creation of a hierarchy

[14] `http://nerd.eurecom.fr/`
[15] `http://nerd.eurecom.fr/ontology/nerd-v0.5.n3`
[16] `http://dublincore.org/documents/2012/06/14/dces`
[17] `http://linkedtv.eurecom.fr/api/mediacollector/`

of media fragments with different levels of granularity provides a flexible model for allowing different interpretations of the data depending on the user or the particular context.

Different entities spotted within the same media fragment of a video resource can be considered simultaneously for obtaining new insights about what is happening in that part of the media resource. For example, let's imagine that an entity corresponding to a particular painting has been spotted inside a media fragment, while another entity known to be an artist in DBpedia is also spotted nearby. It would then be possible to infer that this person is the author of the painting with some confidence level, or at least, that this person is somehow related with it. Similar deductions can be done by relying on other annotations in the model such as keywords and LSCOM concepts (i.e. visual concepts as popularized by the TRECVID benchmarking campaign).

8 Evaluation plan

The cooperation with the AVRO (via Sound and Vision) and RBB broadcasters as partners of the LinkedTV project opens many possibilities. Apart of having more material to be processed, their viewers and editors can potentially test new features and provide feedback about the quality of the annotations and the usefulness of the enrichment process described in this paper.

The evaluation of the entire approach will be based on three complementary dimensions. First, the accuracy of the named entity extraction process. Are entities correctly spotted and disambiguated? This will be evaluated by applying standard metrics from the NLP domain. Second, the adequacy of the media fragment temporal boundaries and their relevance for the spotted entities. What is the right temporal window to consider around particular entities? Are they meaningful according to the story being told in the video? We aim to compare those results with with ground truth annotations manually generated by users. Finally, we will measure the precision and recall of the search and hyperlinking operations over the generated Media Fragments: for a particular search term given by a user, are the media fragments retrieved relevant? Are the links between media fragments interesting from a user point of view? The evaluation of this part is probably the most subjective and complex one but we plan to rely on standard datasets and in particular on the MediaEval Search and Hyperlinking Task[18] that has exactly this goal.

We have already performed some very preliminary evaluations. We computed some basic statistics about the number of named entities per NERD type and the number of media fragments in a 55 minutes episode of the show *Tussen Kunst en Kitsch* from the Dutch broadcaster Avro (Tables 1 and 2).

[18] `http://www.multimediaeval.org/mediaeval2013/`

Table 1. Number of entities per type

NERD type	Entities
Person	37
Location	46
Product	3
Organization	30

Table 2. Number of MediaFragment's

Serialized Item	MediaFragment
Shots&Concepts	448
Subtitles	801
Bounding Boxes	4260
Spatial Objects	5

Acknowledgments

This work was partially supported by the European Union's 7th Framework Programme via the project LinkedTV (GA 287911).

References

1. G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media Meets Semantic Web — How the BBC Uses DBpedia and Linked Data to Make Connections. In *6th European Semantic Web Conference (ESWC'09)*, pages 723–737, Heraklion, Crete, Greece, 2009.
2. Y. Li, G. Rizzo, J. L. R. García, R. Troncy, M. Wald, and G. Wills. Enriching media fragments with named entities for video classification. In *1st Worldwide Web Workshop on Linked Media (LiME'13)*, pages 469–476, Rio de Janeiro, Brazil, 2013.
3. Y. Li, M. Wald, T. Omitola, N. Shadbolt, and G. Wills. Synote: Weaving Media Fragments and Linked Data. In *5th Workshop on Linked Data on the Web (LDOW'12)*, Lyon, France, 2012.
4. X. Liu, R. Troncy, and B. Huet. Finding Media Illustrating Events. In *1st ACM International Conference on Multimedia Retrieval (ICMR'11)*, Trento, Italy, 2011.
5. V. Milicic, J. L. R. García, G. Rizzo, and R. Troncy. Grab your Favorite Video Fragment: Interact with a Kinect and Discover Enriched Hypervideo. In *11nd European Interactive TV Conference (EuroITV'13), Demo Track*, Como, Italy, 2013.
6. V. Milicic, G. Rizzo, J. L. R. García, R. Troncy, and T. Steiner. Live Topic Generation from Event Streams. In *22nd International World Wide Web Conference (WWW'13), Demo Track*, pages 285–288, Rio de Janeiro, Brazil, 2013.
7. G. Rizzo, T. Steiner, R. Troncy, R. Verborgh, J. L. R. García, and R. V. de Walle. What Fresh Media Are You Looking For? Retrieving Media Items from Multiple Social Networks. In *1st International Workshop on Socially-Aware Multimedia (SAM'12)*, Nara, Japan, 2012.
8. G. Rizzo and R. Troncy. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, Avignon, France, 2012.
9. J. Yang, J. Luo, J. Yu, and T. S. Huang. Photo stream alignment for collaborative photo collection and sharing in social media. In *3rd ACM International Workshop on Social Media (WSM'11)*, pages 41–46, Scottsdale, Arizona, USA, 2011.