EDITE - ED 130

# Doctorat ParisTech

# T H È S E

**pour obtenir le grade de docteur délivré par**

# TELECOM ParisTech

## Spécialité « Computer Science and Multimedia »

*présentée et soutenue publiquement par*

## José Luis REDONDO GARCÍA

le 04 Mars 2016

# Semantically Capturing and Representing

# Contextualized News Stories on the Web

Directeur de thèse : **Dr. Raphaël TRONCY**

**Jury**
**Fabien GANDON**, Prof, Inria, France                                          Reviewer
**Peter MIKA**, Dr, Yahoo ! Research Labs, U.K.                                  Reviewer
**Lora AROYO**, Dr, VU University Amsterdam, The Netherlands                     Examiner
**Enrique ALFONSECA**, Dr, Google Research, Switzerland                         Examiner
**Bernard MERIALDO**, Prof, EURECOM, France                                      Examiner

**TELECOM ParisTech**
école de l'Institut Télécom - membre de ParisTech

# Acknowledgments

Working as a PhD student in EURECOM was a great experience that would not be achieved without the help and support of many people, who I would like to acknowledge here.

First and foremost, I would like to thank my supervisor Dr. Raphaël Troncy for giving me the chance to join this big family at EURECOM and offer me invaluable support throughout my studies. Together with the methodic and shrewd Giuseppe Rizzo, they have become my research guidance during those three years and a half, providing me a priceless feedback along the way. This work would not have been possible without their scientific knowledge, and their great support.

I would like to thank my committee members, the reviewers Dr. Fabien Gandon and Dr. Peter Mika, and furthermore the examiners Dr. Lora Aroyo, Dr. Enrique Alfonseca and Dr. Bernard Merialdo for their precious time in reviewing this manuscript, shared positive insight and guidance.

I owe my deepest gratitude to my parents, Dr. José Antonio Redondo González and María del Carmen, to my brother Jesús, my strong inspiration Hélène, and to my friends Juan Luis, and Ghislain for their infinite help and encouragement, devotion and love and for helping me in this long but passionating way. Last but not least, special thanks go to my buddies at EURECOM Ahmad, Giovanni, Robin, Rajeev, Vuk, Leela, Mariella, Julien, Chiara, Ester, Pasquale and many others for their vital support and unconditional friendship.

# Abstract

Multimedia content is one of the most widely used, intuitive ways of consuming information. During their daily lives, people watch TV, share videos on social networks, and use different Web streaming sources to learn things, know other people's opinions, keep informed about latest events happening around, or just for entertainment. There is no doubt multimedia content is making its way to the Web, with hundred thousands hours of video content being uploaded every day to platforms like Youtube[1]. But, can this video content and its metadata become available in a way that is easily accessible and addressable, not only as a whole but also at different levels of granularity? Is the information contained in those video items sufficiently contextualized so it can be effectively consumed by humans with reduced prior knowledge about what is being told in it? In this thesis we investigate how to make the most of available standards like Media Fragments URI[2] and widely used vocabularies to turn multimedia content into a first class citizen of the Web where media stories are fully contextualized and integrated with the rest of the knowledge in the Web.

As any citizens of the Web, those multimedia fragments need to also live together with the other resources already offered by the Web, like textual documents, images, charts and diagrams, etc. Having them cohabiting on equal terms can bring to the table a symbiotic effect: on one hand the power of the video to express complex ideas and effectively communicate a message to viewers can nicely illustrate and reinforce other information already available in the Web. On the other hand, videos can be further complemented with other already existing documents, in order to have a better contextualized consumption or bring a deeper knowledge about the expressed facts. To make this symbiosis effective, we need to annotate this multimedia content and build bridges from video to different resources on the Web. In this thesis, we advocate the use of semantic technologies as the tool to allow machines to automatically perform these tasks. We present approaches relying on different information retrieval and machine learning algorithms for filling the gap between the low-level visual features obtainable via traditional analysis techniques, and the meaningful high-level annotations ready to be consumed. Aligning those concepts to standard vocabularies on the domain make it possible for machines to interpret and reason over the knowledge contained in those documents, in order to offer innovative operations for browsing, enriching and hyperlinking media fragments, and ultimately, improve the way multimedia information is consumed.

Another factor that increases the complexity of the annotation process is dealing with the lack of context that a sole multimedia document can provide to properly

---

[1] http://www.onehourpersecond.com/
[2] http://www.w3.org/TR/media-frags/

understand the story being reported. International news items published or broad-casted on different channels are a good example of such phenomena. Being able to extract concepts that are occurring on those videos (because they appear during one scene, are written on some banner, mentioned in the audio, etc.) is already a difficult task, but unveiling other aspects that, even not being explicitly present in the content itself, are crucial to fully capture the backstory, are even more challenging.

To deal with this problem, we propose an innovative conceptual model called News Semantic Snapshot (NSS) that is designed to be able to capture the wide context of a news event. This structure intends to be flexible enough to give support to differ-ent applications targeting the complex process of displaying news stories, from very different angles, targeting particular needs, focusing on certain parts of the story and exploiting certain visualization paradigms. Following a process called Named Entity Expansion we query the Web to bring other viewpoints about what is happening around us, from the thousands of news articles, posts, and social media interactions where we can potentially find those missing story details. Once the additional infor-mation is collected, we perform a selection process in order to filter out the duplicated or unrelated information and keep only those annotations that are relevant for the news story. During the research work presented in this thesis, we have analyzed the multi-dimensionality of the entity relevance by considering different aspects like their frequency on related document, popularity on external sources, expert's opinions, informativeness, interestingness, etc. in order to promote the most adequate ones and make them part of the final NSS. In addition, we have proposed an innovative concentric model that better deals with those different relevancy dimensions to more effectively spot the important entities. It formalizes a duality found in the news anno-tations that distinguishes between the so-called *Core*, which contains representative entities that are spottable via frequency functions, and the relevant entities shaping up the *Crust*, which become important because of particular semantic relationships with the *Core* and can be incorporated into the model in the form of concentric layers. The different implemented algorithms have been evaluated against a ground truth of international news videos annotated with sets of named entities that have been found to be relevant to recreate their context.

Finally, we have analyzed existing storytelling prototypes to verify how the NSS can respond to their particular needs and even empower advanced prototypes in the future, probing also how this model can feed very different applications assisting the users before, during, and after the consumption of the news story.

# Contents

# List of Figures

# List of Tables

# List of Publications

## Journal

1. **José Luis Redondo García** and Adolfo Lozano-Tello: **OntoTV: an Ontology Based System for the Management of Information about Television Content**. International Journal of Semantic Computing, 6(01), 111-130, 2012.

2. **José Luis Redondo García**, Vicente Botón-Fernández and Adolfo Lozano-Tello: **Linked data methodologies for managing information about television content**. International Journal of Interactive Multimedia and Artificial Intelligence, 1(6), 2012.

## Conferences

1. **José Luis Redondo García**, Giuseppe Rizzo and Raphaël Troncy: **The Concentric Nature of News Semantic Snapshot**. 8th International Conference on Knowledge Capture (K-CAP 2015), October 7-10, 2015, Palisades, USA.
<div align="center"><span style="color:red">**Best Paper Award**</span></div>

2. Giuseppe Rizzo, Raphaël Troncy, Oscar Corcho, Anthony Jameson, Julien Plu, Juan Carlos Ballesteros Hermida, Ahmad Assaf, Catalin Barbu, Adrian Spirescu, Kai-Dominik Kuhn, Irene Celino, Rachit Agarwal, Cong Kinh Nguyen, Animesh Pathak, Christian Scanu, Massimo Valla, Timber Haaker, Emiliano Sergio Verga, Matteo Rossi, **José Luis Redondo García**: **3cixty@Expo Milano 2015 enabling visitors to explore a smart city**. 14th ISWC 2015, 14th International Semantic Web Conference (ISWC 2015), October 11-15, 2015, Bethlehem, USA.
<div align="center"><span style="color:red">**Winner of the Semantic Web Challenge**</span></div>

3. **José Luis Redondo García**, Giuseppe Rizzo and Raphaël Troncy: **Capturing News Stories Once, Retelling a Thousand Ways**. 8th International Conference on Knowledge Capture (K-CAP 2015), October 7-10, 2015, Palisades, USA.

4. **José Luis Redondo García**, Giuseppe Rizzo, Lilia Pérez Romero, Michiel Hildebrand and Raphaël Troncy: **Generating Semantic Snapshots of Newscasts using Entity Expansion**. 15th International Conference on Web Engineering (ICWE 2015), June 23-26, 2015, Rotterdam, The Netherlands.

5. Lilia Pérez Romero, Michiel Hildebrand, **José Luis Redondo García** and Lynda Hardman: **LinkedTV News: A Dual Mode Second Screen Companion for Web-Enriched News Broadcasts**. TVX 2014, ACM International Conference on Interactive Experiences for Television and Online Video, June 25-27, 2014, Newcastle, UK.

6. **José Luis Redondo García**, Michiel Hildebrand, Lilia Pérez Romero and Raphaël Troncy: **Augmenting TV newscasts via entity expansion**. ESWC 2014, 11th Extended Semantic Web Conference, Demos Track, vol. CCIS 476, May 25-29, 2014, Anissaras, Crete.

7. **José Luis Redondo García** and Raphaël Troncy: **Television meets the Web: a Multimedia Hypervideo Experience**. ISWC 2013, 12th International Semantic Web Conference, Doctoral Consortium Track, CEUR Proceedings, Volume 1045, October 22, 2013, Sydney, Australia.

8. Evlampios Apostolidis, Vasileios Mezaris, Mathilde Sahuguet, Benoit Huet, Barbora Cervenková, Daniel Stein, Stefan Eickeler, **José Luis Redondo García**, Raphaël Troncy and Lukas Pikora: **Automatic fine-grained Hyperlinking of Videos within a Closed Collection using Scene Segmentation**. ACMMM 2014, 22nd ACM International Conference on Multimedia, November 3-7, 2014, Orlando, Florida, USA.

9. Daniel Stein, Alp Öktem, Evlampios Apostolidis, Vasileios Mezaris, **José Luis Redondo García**, Raphaël Troncy, Mathilde Sahuguet and Benoit Huet: **From Raw Data to Semantically Enriched Hyperlinking: Recent Advances in the LinkedTV Analysis Workflow**. NEM Summit 2013, Networked & Electronic Media, 28-30 October 2013, Nantes, France.

10. **José Luis Redondo García**, Mariella Sabatino, Pasquale Lisena and Raphaël Troncy: **Finding and Sharing Hot Spots in Web Videos**. ISWC 2014, 13th International Semantic Web Conference, Demo Track, October 21-23, 2014, Riva del Garda, Italy.

11. Vuk Milicic, **José Luis Redondo García**, Giuseppe Rizzo and Raphaël Troncy: **Tracking and Analyzing the 2013 Italian Election**. ESWC 2013, 10th Extended Semantic Web Conference, Demos Track, May 26-30, 2013, Montpellier, France.

12. Vuk Milicic, Giuseppe Rizzo, **José Luis Redondo García** and Raphaël Troncy: **Grab your Favorite Video Fragment: Interact with a Kinect and Discover Enriched Hypervideo**. EUROITV2013, 11th European Interactive TV Conference, June 24-26, 2013, Como, Italy.

13. Vuk Milicic, Giuseppe Rizzo, **José Luis Redondo García**, Raphaël Troncy, and Thomas Steiner: **Live Topic Generation from Event Streams**. WWW 2013, 22nd International World Wide Web Conference, Demos Track, May 13-17, 2013, Rio de Janeiro, Brazil.

## Workshops

1. Giuseppe Rizzo, Thomas Steiner, Raphaël Troncy, Ruben Verborgh and **José Luis Redondo García** : **What fresh media are you looking ? for Extracting media items from multiple social networks**. SAM 2012, ACM International Workshop on Socially Aware Multimedia, In conjunction with ACM Multimedia 2012, 29 October 2012, Nara, Japan.

2. Yunjia Li, Giuseppe Rizzo, **José Luis Redondo García** and Raphaël Troncy: **Enriching Media Fragments with Named Entities for video Classification**. WWW 2013, 1st Worldwide Web Workshop on Linked Media (LiME'13), May 13, 2013, Rio de Janeiro, Brazil.

3. Raphaël Troncy, Vuk Milicic, Giuseppe Rizzo, and **José Luis Redondo García**: **MediaFinder: Collect, enrich and visualize media memes shared by the crowd**. WWW 2013, 2nd International Workshop on Real-Time Analysis and Mining of Social Streams (RAMSS'13), May 14, 2013, Rio de Janeiro, Brazil.

4. Mathilde Sahuguet, Benoit Huet, Barbora Cervenková, Evlampios Apostolidis, Vasileios Mezaris, Daniel Stein, Stefan Eickeler, **José Luis Redondo García**, Raphaël Troncy and Lukas Pikora: **LinkedTV at MediaEval 2013 search and hyperlinking task**. MEDIAEVAL 2013, Multimedia Benchmark Workshop, October 18-19, 2013, Barcelona, Spain.

5. **José Luis Redondo García**, De Vocht Laurens, Raphal Troncy, Erik Mannens and Rik Van de Walle: **Describing and contextualizing events in TV news show**. WWW 2014, 2nd International Workshop on Social News on the Web (SNOW 2014), April 7, 2014, Seoul, South Korea.

## Technical Reports

1. **José Luis Redondo García**, Raphaël Troncy and Miroslav Vacura: **Linking Hypervideo to Web Content**. Television Linked To The Web (LinkedTV), June, 2012, FP7-ICT-2011-7.

2. **José Luis Redondo García**, Raphaël Troncy, Dorothea Tsatsou and Dimitrios Panagiotou: **Annotation and Retrieval Module of Media Fragments**. Television Linked To The Web (LinkedTV), October, 2013, FP7-ICT-2011-7.

3. Jan Bouchner, Lukáš Pikora, Barbora Červenková, Tomáš Kliegr, Jaroslav Kuchař, Ivo Lašek, Mathilde Sahuguet, Benoit Huet, **José Luis Redondo García**, Raphaël Troncy, Jan Thomsen, Ali Sarioglu, Lyndon Nixon, Evlampios Apostolidis, Vasileios Mezaris, Daniel Stein and Stefan Eickeler: **Specification of the Linked Media Layer**. Television Linked To The Web (LinkedTV), October, 2013, FP7-ICT-2011-7.

4. Tomáš Kliegr, Jan Bouchner, Barbora Červenková, Milan Dojchinovski, Jaroslav Kuchař, **José Luis Redondo García**, Raphaël Troncy, Jan Thomsen, Dorothea Tsatsou, Georgios Lazaridis, Pantelis Ieronimakis and Vasileios Mezaris: **LinkedTV Framework for Generating Video Enrichments with Annotations**. Television Linked To The Web (LinkedTV), October, 2014, FP7-ICT-2011-7.

5. Tomáš Kliegr, Jan Bouchner, Barbora Červenková, Milan Dojchinovski, Jaroslav Kuchař, Ivo Lašek, Milan Šimůnek, Ondřej Zamazal, Raphaël Troncy, **José Luis Redondo García**, Giuseppe Rizzo, Benoit Huet, Maria Eskevich, Bahjat Safadi, Mathilde Sahuguet, Hoang An Le, Quoc Minh Bui, Jan Thomsen, Adrian M.P. Brasoveanu, Lyndon J.B. Nixon and Lilia Perez Romero: **Final Linked Media Layer and Evaluation**. Television Linked To The Web (LinkedTV), April, 2015, FP7-ICT-2011-7.

# Glossary

Here are the main acronyms used in this document. The meaning of an acronym is usually indicated once, when it first appears in the text.

| | |
|---|---|
| API | Application Programming Interface |
| CSV | Comma Separated Values |
| DI | Data Integration |
| FOAF | Friend of a friend |
| HTML | Hyper Text Markup Language |
| HTTP | Hypertext Transfer Protocol |
| IR | Information Retrieval |
| JSON | JavaScript Object Notation |
| KB | Knowledge Base |
| LD | Linked Data |
| LDA | Latent Dirichlet Allocation |
| LOD | Linked Open Data |
| ML | Machine Learning |
| NE | Named Entity |
| NER | Named Entity recognition |
| NERD | Named Entity Recognition and Disambiguation |
| NLP | Natural Language Processing |
| OD | Open Data |
| OWL | Web Ontology Language |
| RDF | Resource Description Framework |
| RDFS | Resource Description Framework Schema |
| REST | Representational State Transfer |
| SKOS | Simple Knowledge Organization System |
| SPARQL | Protocol and RDF Query Language |
| URI | Universal Resource Identifier |
| URL | Universal Resource Locator |
| W3C | World Wide Web Consortium |
| XML | Extensible Markup Language |

# Introduction

As the saying goes, a picture is worth a thousand words. Even the most simple image can express many divergent facts and evoke significantly different feelings and opinions in humans who interpret them. Let's think now what would happen if we instead consider just a single shot and move to an audiovisual world, where two new dimensions come into play: the time, since the perceived visual information evolves as is consumed; and the audio, which complements and extends what is being depicted. Then, how many words is a video worth?

Some people have already formulated such question and bravely given an approximate number as an answer. In a report for Forrester Research[1] entitled *How Video Will Take Over the World*, James McQuivey claimed that a video is worth 1.8 million words. The maths behind the article are not rocket science: being a picture equals to 1,000 words, and considering video shoots 30 frames per second, every second of video is worth 30,000 words. Multiply 30k by 60 seconds gives this final astonishing number.

Of course this soundbite is purely anecdotal and exhaustively quantifying the number of words a video can express is just unfeasible and pointless. But this analogy is perfect for understanding the complexity of the challenges tackled in this thesis: to capture the essence of what is being told in an audiovisual content is an extremely audacious task even for humans, imagine how difficult it can be for machines. And it is not only about finding the right terms to describe it, it is also about discovering how they relate to each other in order to build up the story behind. Not to mention that sometimes the story being told is so vast that we need to focus on certain portions of the video that matters to us the most. Then, the main concepts summarizing the fragment can be very different than the ones that stood out when looking at the entire video, revealing the importance of the granularity of the information unit being consumed.

News items are a kind of content particularly difficult to interpret. Specially when it comes to an international scope, where many places and agents are involved, the story behind the news may require additional knowledge not explicitly represented in the video to be properly understood and reconstructed. News videos offer indeed a partial view of the happened facts, so viewers in general and machines in particular

---

[1] https://www.forrester.com/How+Video+Will+Take+Over+The+World/fulltext/-/E-RES44199

will miss certain aspects which are crucial to make right sense out of them. To come back to our original metaphor: there will be concepts that are not available inside the million words that a multimedia document is expressing, but they are still important for consumers to understand the story. Therefore we would need to bring them from other knowledge silos where they can be found and fetched. This is what journalists and experts in the domain manually achieve through a tedious and thorough process of consulting other sources. Ordinary people cannot afford such amount of efforts to interpret a story, but they still require a proper understanding of the news that reach us on a daily basic.

We cannot turn our back to those challenges, no matter how daring it may be. More than ever, video is everywhere. According to experts' predictions[2] a 74% of all internet traffic in 2017 will be video. Last year, the amount of video from people and brands in Facebook's News feed increased by a factor of 3.6[3], highlighting the presence of audiovisual content posted in social networks. Main platforms like Facebook, aware of the increasing demand of such kind of material, start offering new features that enhance the way of interacting and consuming videos[4]. Another strong indication is how Youtube, the by far biggest video provider in terms of content and active viewers (currently it has more than 1 billion users), is getting new 300 hours[5] of video every minute. Videos that you could potentially be interested in, but you have high chances to miss in the ocean of multimedia content, which is fast engulfing the Web.

We need to provide machines with mechanisms that allow to autonomously process this huge amount of content that is available on the Web, for improving the way users discover the documents they want to consume. In this thesis we apply semantic techniques to replicate way humans reason over the knowledge when consuming and interpreting the available knowledge. The meaning of the concepts being displayed on a video can help to decide which are the most important ones, abstract from low level facts to more general terms and the other way around, and discover semantic relations between entities to find related information that can generate full/partial summarizations of the story being told so it is easier to be digested. In this manuscript we present the achievements obtained after more than 3 years experimenting with such techniques over the multimedia domain, opening the window to a new landscape of overwhelming possibilities for the future that will bring the video consumption experience to a new level.

---

[2]http://syndacast.com/video-marketing-statistics-trends-2015
[3]http://media.fb.com/2015/01/07/what-the-shift-to-video-means-for-creators/
[4]https://www.facebook.com/facebookmedia/best-practices/facebook-video
[5]https://www.youtube.com/yt/press/en-GB/statistics.html

## 1.1 The Landscape of Semantic Multimedia

Information in the Web is becoming semantically annotated using different techniques in order bring new ways of programmatically exploiting the available knowledge and further reason over it. Some initiatives showcasing this phenomena are the Google Knowledge Graph powered in part by Freebase[6] or DBpedia[7], which have become widely used vocabularies acting as a crystallization point for structured data in many different domains: e-commerce, enterprise management, governmental affairs, etc.

There are already tools that automatically transform documents on the Web (textual, in most cases), and annotate them following some of the aforementioned vocabularies. Some of them like[8] are able to spot particular words in a textual documents and link them to concepts in knowledge bases like DBpedia. Those concepts, also frequently known as entities, are already citizens of the Web where other annotated documents are pointing to. And they are contextualized in a bigger cloud of entities and interlinked with other resources, laying the foundation for a new generation of algorithms that reason over the semantic annotations to potentially replicate the way the humans would perform the same operations. There is still a lot of work to do in refining those concept recognition and disambiguation techniques, but the first pieces of the puzzle are already available to be used in order to build intelligent systems working on top of the obtained annotations.

However the scenario is much less mature when it comes to multimedia documents. To apply similar techniques to such documents would require the use of advanced audio transcription and visual analysis methods that introduce a higher level of complexity to the workflow. Because of this, in most cases the available video annotations are manually generated, they correspond to very general aspects of the video, or they are just materialized in some raw text accompanying the audiovisual content.

Nevertheless advanced ways of consuming multimedia documents are more and more demanded. In the news domain for example, there are already different prototypes which try to offer stories in an innovative way. News stories happening on the Web can be shared online via some platforms [9] [10] [11] that assist the users in the process of integrating different media elements and text into a single visualization so it can be easily consumed and shared with other people. Other examples are some television projects led by the British Broadcasting Corporation, like HomeFront [12] that allows to explore stories happened in Britain during the First World War via different axes: timeline, characters, and interactive diagrams mixed with video docu-

---

[6]http://freebase.com
[7]http://wiki.dbpedia.org/
[8]http://nerd.eurecom.fr/
[9]https://storify.com/
[10]http://seen.co/
[11]http://eventifier.com/
[12]http://homefront.ch.bbc.co.uk/

ments. Others like Highlights [13] empower sport footages by identifying what are the important milestones within a game so users do not miss a thing. Those prototypes would clearly benefit of the existence of such kind of annotations. Unfortunately at the moment, most of them are based in the output of manual annotators or crowd-sourcing initiatives. To make those solutions scalable to the amount of new content being daily published we need unsupervised approaches that can get close or match the human driven ones.

Sometimes semantic annotations are not directly used to be displayed to the potential consumers, but to serve as a basis for launching other intelligent operations over the original content. For example, we could be interested in creating links between media documents that share certain commonalities, or combine content based annotations with user preferences to generate recommendations according to what the viewer is watching or has decided to watch before. Hence, we can build algorithms that autonomously exploit those annotations by filtering and selecting the most important ones given the context and the final objective, and are ruled by different selection criteria that those aiming to just display those annotations to the users.

Finally, the media on the Web is evolving from being searched, consumed and shared as a whole to be available at different levels of temporal and spatial granularity thanks to standards like the Media Fragments URI [14]. This way, we can address a particular fragment inside a longer video or even highlight a particular spatial region on the screen that we want to make emphasis into. Aforementioned media operations like hyperlinking, recommendation, or story browsing can be based on smaller yet more reusable and context-suited media pieces that can improve the way media on the Web is consumed.

## 1.2   Research Challenges

After having introduced the current state of the presence of multimedia documents on the Web and the possibilities that semantic technologies bring to the table when applied over video content in general and news items in particular, in this section we list the key challenges that have been tackled in this thesis. Some of them have been already addressed in previous research works or are in process of being refined. Together they pursue the objective of effectively improving the Web multimedia annotation and consuming experience.

---

[13]http://www.bbc.co.uk/rd/projects/highlights
[14]http://www.w3.org/TR/2012/REC-media-frags-20120925/

### 1.2.1 Common Methodology for Media Representation

The Web is an open ecosystem, where documents can be published and consumed in very different manners. And we want it to keep it this way, ensuring that remains as a heterogeneous environment where even antagonist approaches can coexist. But this freedom does not exempt us from wanting to promote widely adopted vocabularies and standards that make possible to seamlessly consume audiovisual information on the Web. Following good principles in Linked Data publishing we can allow different agents to deal with the vast amount of data available, which makes it easier to spot what is relevant in a timely manner.

- **Vocabularies and Standards for Representing Multimedia Documents.** We need mechanisms that allow us to address audiovisual content not only at the level of the entire content but also specific parts inside it. At the same time we need to categorize those fragments according to different application requirements: fragments can correspond to shots, scenes, chapters, introductory parts, openings, etc. At the end of the day they are all fine-grained pieces inside the video, but they have different characteristics that need to be annotated with the correct semantic.

- **Vocabularies for Representing Information inside the Multimedia Documents.** Inside every fragment there can be different agents and actions being depicted or mentioned, about any imaginable topic. We need to annotate highly heterogeneous information, from general concepts the fragment is talking about, to particular actors playing a role in the action or the location where the facts are taking place. Those important concepts and entities can belong to any imaginable domain and materialized using different vocabularies, but they can be classified according to their video provenance: things mentioned in the audio, entities appearing in the video, concepts that are relevant to the whole fragment, visual clues, etc.

- **Vocabularies for Materializing Connections.** The different annotations describing a video item are not isolated. Instead, they are associated each other via some relationships: at the level of fragments inside the same video when some video chunks complement others, at the level of the annotations inside those fragments like two entities involved in the same facts occurred in the past, or connections between annotations inside the video and external resources on the Web like media posts talking about a particular entity depicted in the video. Here again it is crucial to reuse already existing vocabularies to formalize the natural relations that are established between entities in all kind of domains, and emphasizing predicates materializing the storytelling aspect to be able to uncover how the different multimedia fragments play a role in the underlying

story.

## 1.2.2 Automatically Generating Semantic Media Annotations

All the annotations introduced in previous subsections could be generated by humans and experts in the domain, but at a high cost in time and resources. Thoroughly consuming a video item, properly digesting the information extracted from it, bringing together the different pieces of the story being told, and browsing the vast amount of knowledge available in the Web to find related information and possible relationships between the content, make this annotation process an extremely difficult task. We need autonomous information processing algorithms able to produce similar annotations in a shorter amount of time and less human intervention, so we can deal with a much bigger variety of multimedia content to be processed and interlinked with other Web sources in a timely manner.

In more detail, we need techniques to fragment the video in different pieces that make sense from the user point of view, recognize what is being said in the audio and depicted in the video, and find links between the spotted annotations and external content available on the Web. The complexity of such tasks is high in all the steps of the workflow and the final results are error-prone given the subjectivity of the annotation process. To get proper data to be presented to the consumers a manual curation would be still required. But the objective is to reduce as much as possible the time needed to generate such annotations while increasing the amount of additional sources and potential concepts considered to annotate and link to. Interpolating to the news domain, journalist would no longer need to read lot of additional documents, which in a large part are irrelevant. Instead, they can concentrate on a compressed and summarized snapshot where most of the irrelevant information has been filtered out, and have time to dive deeper into what is really important.

## 1.2.3 Exploiting Semantic Media Annotations

Properly identifying concepts and entities of a video is an important task, but the annotation process should not end there. It is still missing how to detect which of those descriptions are worth to be showing to the viewer or can trigger other operations like recommendation and interlinking.

- **Filtering and Ranking Annotations.** During the entire duration of the video, a huge amount of concepts can be depicted in the video or mentioned in the audio, even lasting just some seconds. Tools annotating those videos can detect them even if they just briefly appear, so they become part of the multimedia content representation. However sometimes they are just not interesting or redundant, so we need a process of filtering the unimportant items and keeping only those that we can potentially leverage on. In addition the candidate

annotations can be ranked according to different considerations depending on the final objective of the application that is going to consume them: bringing knowledge to the user (informativeness), try to bring their attention (interestingness), offer other points of view about the explained facts (opinioness), etc.

- **Extending Content with more Content.** Most prominent annotations summarizing the content or fragments inside the content can be used to find multimedia documents that talk about similar matters. This way users can jump from one video to other based on certain similarities found in them. At the same time, certain parts of the video can talk about concepts that users would like to know more about. Interlinking operations could help to find additional content where those particular annotations are deeply depicted so that initial story is easier to understand.

- **Recreate the Context of News Stories** A very important step in exploiting the semantic annotations of a news media document is to be able to filter and rank annotations in order to recreate the context of the story being told. This story can be conceived in many ways: like a summary of the facts that assist viewers in getting a quick understanding about what the news story is about, or like more elaborated conceptualization diagrams where users can browse and explore the details of the different facts being told in the video. The complexity of building up such story is huge because it requires a deep understanding about the agents involved, the actions performed by them, and how those actions evolve in time.

### 1.2.4   Present Media Content to the Users.

Once the media annotations are available and we have leveraged on them for enriching, interlinking, and summarizing the content, we need to show the resulting information to the user in a way that it becomes profitable for him without being too intrusive or distracting. It is extremely important to understand what the user needs and how we can provide solutions to him/her, in order to fully unveil the potential of the generated semantic annotations.

## 1.3   Thesis Contributions

Considering the challenges in semantic media annotation and news description described in previous section, this thesis has tackled different research questions bringing the following contributions:

### 1.3.1 Contributions on Media Representation

- Design of the LinkedTV ontology. In this ontology, multimedia content in general and television programs in particular can be annotated not only at the upper level of the entire program but also with different degrees of granularity thanks to the use of the Media Fragments URI 1.0 specification. The instances of the MediaFragment class are the anchors where the other information is attached: legacy metadata from the providers, results obtained by automatic analysis of the video file, and even more important, links it to other resources in the Web where extra information about the content can be found.

- Development of the service TV2RDF for generating annotations according to the LinkedTV Ontology. This service takes as input already-existing non-Web-compatible formats in the multimedia television domain, like subtitles in SRT [15] format, metadata in TVAnytime[16], and different analysis results in EXMAR-aLDA[17] files and converts them to instances of the Ontology for Media Resources[18] where different annotations are attached via the Open Annotation Data Model[19].

- Model for Linking Media Annotations with Resources on the Web. We have extended the LinkedTV model for serializing different relations between annotations in the video content and resources on the Web that complement or further describe them. We interlink video fragments in the main video with items in different Social Networks via the TVEnricher service [20].

### 1.3.2 Contributions on Semantic Annotation Exploitation

- Enriching and Hyperlinking. Complementing linear videos by offering continuative and related information via, e.g., audio streams, web pages, as well as other videos, is typically hampered by its demand for massive editorial work. In this thesis we introduce our approach to leverage on video annotations for identifying similar fragments available not only in the same original collection where the video belongs to, but also different platforms on the Web.

- Multimodal Approaches. The primary techniques for semantically annotating media content are based in the processing of textual metadata. However some visual analysis algorithms are already producing good results regarding object detected in the video itself. In this thesis we showcase how the combination of

---

[15] https://en.wikipedia.org/wiki/SubRip
[16] https://tech.ebu.ch/tvascope
[17] http://www.exmaralda.org/en/tool/exmaralda/
[18] http://www.w3.org/TR/mediaont-10/
[19] http://www.w3.org/ns/oa
[20] http://linkedtv.eurecom.fr/tvenricher/api/

those text based techniques with the results of visual analysis can improve the way we interlink fragments of video that are potentially similar.

- Video Classification. The semantic annotations detected inside the videos can be used to infer the category of a video. Taking as input the temporal distribution of named entities detected in a video and their type, we have developed methods that outperform the state the art approaches for classifying videos, specially when it comes to Web platforms where the classification process has to be done in a short time.

- Media Fragment Summarization and Promotion. We use annotations as the insights to identify those fragments that are more relevant to the the main topics being discussed. We propose a set of automatically annotated media fragments called Hot Spots, which intend to highlight the main ideas of the video and make easier for the user to decide which fragment can be interesting for him to watch or share.

- Named Entity Expansion. Existing Named Entity Extraction techniques spot entities in the video transcripts. However, and specially when it comes to videos about international news happening in the world, their results are too much limited to what is being said in a program so they become insufficient for describing the entire context of the story being told. Therefore the Named Entity Expansion algorithm uses a subset of entities from Named Entity Expansion as seeds for retrieving and analyzing additional documents from the Web where the same event has been also described.

### 1.3.3 Contributions on News Annotation Generation

- Applying News Entity Expansion to International News. We have performed some preliminary studies showing how the news video items can benefit from the execution of a process able to bring into the table other annotations that are not explicitly presented in the video but are valuable to understand the backstory.

- News Semantic Snapshot. We propose a new data representation called News Semantic Snapshot (NSS), in order to explicitly capture in a single model the knowledge about the context of a News video item, bringing the necessary information for the audience to understand what is being described on the news.

- Ground Truth of News Semantic Snapshots. We have produced a set of Gold Standard annotations that reproduce the ideal NSS for 23 different videos following an exhaustive methodology that analyzed the main elements building

up the context of the story told in the video. This methodology and the corresponding ground truth annotations per video are available at [21].

- Explore the Multidimensionality of Entity Relevancy in News. The original NSS generation algorithm looked only at the relevancy of the entities in terms of their frequency in the related documents that are retrieved during the process. We have probed that relevancy is a much broader concept that needs to be tackled from different perspectives and dimensions like entity distribution along the documents, popularity, opinions of the experts, etc.

- Probe the Concentric Nature of the Semantic Snapshot. The entities that compose the Semantic Snapshot should not be simply considered as isolated units. Instead, those entities can be disposed following a more complex representation strategy that transforms the bag of entities into a concentric sphere, where the entities in the *Core* drive the main aspects of the story during the whole duration of the show and are highly repeated along all the documents, and others (the so called *Crust*) get relevant because very particular and one-time relevant relationships between those entities and the ones in the *Core*.

### 1.3.4   Contributions on Advanced News Consumption

- Study on News Consumption Phases. Despite the great variety of alternative ways applications present news to the users (they follow a different user interaction philosophy, target a different audience, or illustrate the main facts from a different angle), we propose a classification that includes all them and highlights how the viewers' information consumption needs vary along time in terms of specificity and diversity, and how the NSS can effectively respond to those needs.

- Development of different Prototypes. In order to apply the techniques and fundamentals derived from the contributions of this thesis, we have worked on the implementation of different applications and prototypes, like second screen applications showcasing active and passive experiences for television users consuming semantic annotations, and also advanced prototypes experimenting with summarization of news events and and timeline representations to see how news stories evolve.

## 1.4   Thesis Outline

This thesis is divided in two main parts. In Part I, we focus on presenting the different aspects that together compose the world of semantic multimedia on the

---

[21]https://github.com/jluisred/NewsConceptExpansion/wiki/Golden-Standard-Creation

Web in order to effectively improve the way video content is consumed. We highlight the different contributions we have done in every research challenge identified in Section 1.2 for bringing to the multimedia scene new features like media fragment based Web experience, intelligent annotation and classification of fragments according to resources in the Web of Data, and advanced semantic enriching and hyperlinking with other videos and documents by exploiting textual and visual features. The contributions of this part have been published in [159, 180, 6, 114, 115, 148, 116, 193, 99, 142, 147, 146]

- Chapter 2: **An Ontology Model for Multimedia on the Web**, we specify the requirements for building an ontology model that gives support to the publication of media content and its metadata in a Web of Data compliant way. Examples of instances serialized according to this model will be studied, together with use cases on how to access this information.

- Chapter 3: **Generating Video Annotations**, conducts a comprehensive summary of the different semantic approaches that have been used and developed for tackling the complexity of annotating multimedia content, considering textual based, pure visual and multimodal techniques.

- Chapter 4: **Exploiting Annotated Media Fragments**, we present the various approaches for exploiting and bringing value to the multimedia content, like media enriching, key-fragment detection, video classification, or hyperlinking. Those operations leverage on the semantic annotations produced according to techniques in previous Chapter 3.

In part II, we focus on the challenge of automatically generating high quality annotations for describing international news by considering not only the annotations that can be derived from the initial documents, but also other relevant entities that further describe the context of the news story. The contributions of this part have been published in [140, 143, 144, 145, 58, 57]

- Chapter 5: **The Semantic Snapshot of a News Item**, we introduce the concept of the News Semantic Snapshot (NSS) of a news item, as a semantic representation composed by named entities that allows to reconstruct the context of a news story so we can fully better make sense out of it. In this chapter we introduce a first approach for recreating the NSS of a news video. We also describe the methodology followed for creating a gold standard of international news' NSS that will be used for evaluating the different NSS generation techniques.

- Chapter 6: **The Multidimensionality of the News Entity Relevance**, we present a better formalized and tuned approach for automatically generating

the News Semantic Snapshot of a news item. Given the multidimensionality of the relevance scores of the entities composing the ideal NSS, we have different functions working over different domains in order to properly recreate the news context in the most accurate way possible.

- Chapter 7: **The Concentric Nature of the News Semantic Snapshot**, emphasizes again on the multidimensionality of the NSS, but brings to the table an innovative way of dealing with the different entity relevancy aspects by organizing the candidate entities into two concentric layers: the *Core*, composed of the most representative entities that are well-connected each other and are spottable via frequency measures, and the *Crust*, which includes not necessary frequent entities that are attached to the Core via particular semantic relationships.

- Chapter 8: **The NSS in the in the News Consumption Paradigm**, we study the numerous ways of consuming a news story. Every existing tool or application for displaying information about a news item follows a different philosophy, targets a different audience, and presents the main facts from a different angle. Despite this variety of alternatives we propose a model for classifying those news consumption approaches, analyzing how the NSS can give support to their particular needs.

# Part I

# Towards a Semantic Multimedia Web

# Overview of Part I

In Part I, we will lay the foundations for a Semantic Web driven multimedia consumption, grounded in three pillars: (1) designing a model for representing media fragments and their associated descriptors, (2) automatically populating this model with adequate annotations, and (3) leveraging on the generated knowledge and other information on the Web to provide advanced features over the multimedia content.

In Chapter 2, we conduct a survey on different multimedia representation formats available in the industry and research domains like TV-Anytime, MPEG-7 or EBUCore, passing through some ontology Web models that are called to support the transition from multimedia content stored in individual silos to a seamless integration of multimedia content in the Web. After identifying specific requirements for a knowledge model that could make effective such transition, we describe the LinkedTV ontology as a means to provide multimedia content at the level of fragment and annotated with very broad visual and textual descriptions that can turn video into a first citizen of the Web.

In Chapter 3, Generating Video Annotations we introduce different techniques for semantically annotating the textual dimension of the video content, obtained from subtitles, automatic transcripts and results from text recognition over video. In addition, we cover some visual-based algorithms for automatically recognizing text (ASR, OCR) over images, and other video analysis able to split video in temporal fragments, detect static and moving objects, or identifying faces of people at the time they show up on the screen.

In Chapter 4, Exploiting Annotated Media Fragments, we describe in deep how to exploit multimedia documents annotated at the level of fragment in order to enrich particular parts of the video, identify the chunks that are worth to be watched, classify them according to different categories an automatic way, hyperlink them with other relevant content, and ultimately provide the viewers with a better contextualized, easy to browse and consume multimedia content.

# An Ontology Model for Multimedia on the Web

## 2.1 Introduction

Textual based documents are present on the Web since the beginning of its existence: encyclopedic articles, blogs, newspapers, tutorials, etc. populate most of the knowledge available online. Those documents are interlinked so we can browse them, and they are constantly being more and better syntactically and semantically annotated in order to improve the way they are searched and exploited by both humans and machines.

Unfortunately, multimedia documents are still being published in a much more primitive way. Video and audio files are becoming available online as indivisible units or embedded inside textual documents so they are not easily addressable and therefore difficult to annotate and interlink with other resources. The objective of this Chapter is to propose different semantic techniques in order to make those media documents first class citizens of the Web. This way we can reference not only the whole content but also particular fragments inside them, which can therefore be described and interlinked with other resources on the Web, opening room to new advanced ways of consuming information the Web that were not imaginable before.

In the following sections we lay the foundations for a multimedia based Web experience that give the video content to the same privileges than textual documents already have, bringing to the table the best of those two worlds for improving information sharing: the precision and accuracy of text for explicitly and exhaustively describing things, and the intuitive power of illustrating ideas in the form of multimedia documents. In first Section 2.2, we review the state of the art in multimedia annotation in traditional scenarios, like broadcast and multimedia industry. In Section 2.3 we establish the requirements of a knowledge model able to bring multimedia to the Web at different levels of granularities, and supporting different visual and semantic techniques. Finally in Section 2.4 we present the LinkedTV ontology, available at http://data.linkedtv.eu/ontologies/core/, which intends to fulfill the aforementioned requirements and illustrate some examples where this model is used for representing television content.

## 2.2 State of the Art in Multimedia Metadata Models

A large number of multimedia metadata standards exist, coming from different application areas, focusing on different processes, supporting different types of metadata and providing description on different levels of granularity and abstraction.

In [202] the authors discuss requirements for semantic content description and review the capabilities of standards coming from the W3C and MPEG communities. In the second part of their article [123], they focus on the formal semantic definition of these standards which determines the expressiveness for semantic content description and enables mapping between descriptions. The report in [38] surveys multimedia ontologies and related standards and defines requirements for a multimedia ontology of which many are also relevant for multimedia metadata standards. A comprehensive overview on multimedia metadata standards and formats has been prepared by the W3C Multimedia Semantics XG [73]. A more recent review in [138] focuses on the standards from MPEG but also discusses interoperability issues between standards in the context of general multimedia metadata application scenarios.

This section provides a comprehensive overview of the most commonly used multimedia metadata standards. We start by introducing several metadata formats developed in the broadcast industry in the Section 2.2.1. The low-level features that can be detected in videos are interpreted into higher level semantic metadata information that can be used to describe the program content. We review the standards developed by the multimedia analysis community in the Section 2.2.2. But the seed video content can be also enriched with additional multimedia content hosted and shared on the Web. We review the current formats used to describe the Web multimedia content in the Section 2.2.3. The news and photo industry has a long tradition of developing metadata standards for photos and videos. We review those formats in the Section 2.2.4. Since two years, the W3C has created a working group to standardize an ontology that will bridge the gap between the numerous standards used on the Web to develop multimedia content. We describe in details the resulting Ontology for Media Resources, a W3C recommendation in the Section 2.2.5, that will be heavily used in to build the proposed data model. Events are also commonly used to describe multimedia resources: event models can be used to describe the fact of broadcasting a program on a channel at a particular moment in time, or they can be used to describe the content itself at a very fine grained level. We review the event models proposed by the semantic Web community and beyond in the Section 2.2.6. We also present the ongoing effort of the Web community under a W3C community group to propose a generic annotation model with several extensions in the Section 2.2.7. Finally, we conclude this survey by describing several commonly used vocabularies such as FOAF or the Provenance Ontology in the Section 2.2.8.

### 2.2.1  Metadata Models from the Broadcasting Industry

The broadcast industry has developed several metadata formats for representing TV programs, their broadcast information or targeted audience and their content in order to generate Electronic Program Guides. In this section, we do a brief review of those different standards. Most of them are loosing ground against more standard and Web based formats, and other are even extinguished. First, we describe the XML-based formats such as DVB, BMF developed by the German broadcaster ARD and TV Anytime. Second, we present other models that are largely inspired by the Semantic Web technologies such as EBU (and its application in EU Screen and Europeana) or the BBC Programmes ontology 2.2.1.5.

#### 2.2.1.1  DVB metadata model

The Digital Video Broadcasting Project (DVB[1]) is an industry-led consortium of around 250 broadcasters, manufacturers, network operators, software developers, regulatory bodies and others in over 35 countries committed to designing open technical standards for the global delivery of digital television and data services. The DVB metadata model considers different classification schemes represented using the MPEG-7 standard, and various XML Schemas such as the DVB Classification Scheme schema: http://www.dvb.org/metadata/schema/dvbCSschema.xsd or the Content Item Information which uses mostly MPEG7 and TV Anytime content types: http://www.dvb.org/metadata/schema/ContentItemInformation.xsd, between others.

DVB is using the MPEG standard (Moving Picture Expert Group) to compress, en-code and transmit audio / video and data streams. Several variable bit rate data streams are multiplexed together to a fixed data stream. This makes it possible to transfer video and audio channels simultaneously over the same frequency channel, together with various services. These data services provide additional programme information to enable a complete EPG (Electronic Program Guide) for present and following schedule events. Digital television and services are broadcasted in various platforms and technologies with specified transmission and encoding standards. Each platform is specified by a standard by the European Telecommunications Standards Institute (ETSI)[2].

The DVB transport stream includes metadata called Service Information (DVB-SI). This metadata delivers information about transport stream as well as a description for service / network provider and programme data to generate an EPG and further programme information. In our context the EIT (Event Information Table) and the SDT (Service Description Table) are the ones containing interesting metadata to be brought to the Web. The content descriptor from the EIT table defines

---

[1]https://www.dvb.org/
[2]European Telecommunications Standards Institute, http://www.etsi.org
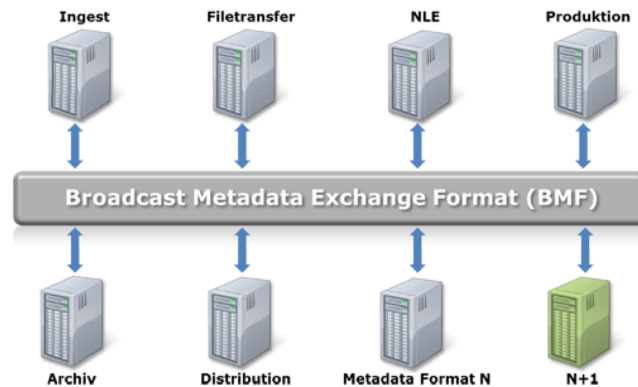
Figure 2.1: Broadcast Metadata Exchange Format (BMF) 2.0, courtesy of http://www.irt.de/en/activities/production/bmf.html

a classification schema for a programme event. It provides various genre categories using a two-level hierarchy with various different genres and subgenres.

#### 2.2.1.2 ARD BMF

IRT (Institut für Rundfunktechnik / Broadcast Technology Institute) is the primary research institute cooperating with public-broadcasting organisations in Germany, Austria and Switzerland. The Institute focuses on solutions which enhance the quality of radio, television and new media for the benefit of users and is committed to preserving broadcasting in Germany and abroad. IRT associates are the following public broadcasters: ARD, ZDF, DRadio, ORF and SRG/SSR.

The Broadcast Metadata Exchange Format Version 2.0 (BMF 2.0) has been developed by IRT in close cooperation with German public broadcasters with focus on the harmonization of metadata and the standardized exchange thereof. The standard particularly reflects the requirements of public broadcasters (figure 2.1). BMF contains metadata vocabulary for TV, radio and online content and defines a standardized format for computer-based metadata exchange. It facilitates the reuse of metadata implementations and increases the interoperability between both computer-based systems and different use case scenarios. BMF enables to describe TV, radio and online content as well as production, planning, distribution and archiving of the content. Metadata in BMF are represented in XML documents while the structure for the XML metadata is formalized in an XML Schema. The latest version of the format is the version BMF 2.0 Beta[3].

---

[3]http://bmf.irt.de/en
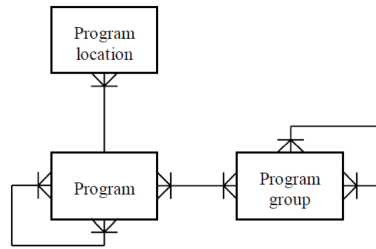
Figure 2.2: TV-Anytime Content Description Model

### 2.2.1.3   TV Anytime

The TV-Anytime Forum is a global association of organizations founded in 1999 in USA focusing on developing specifications for audio-visual high volume digital storage in consumer platforms (local AV data storage). These specifications for interoperable and integrated systems should serve content creators/providers, service providers, manufacturers and consumers. The forum created a working group[4] for developing a metadata specification, so-called TV-Anytime[5] and composed of:

- Attractors/descriptors used e.g. in Electronic Program Guides (EPG), or in Web pages to describe content (information that the consumer – human or intelligent agent – can use to navigate and select content available from a variety of internal and external sources).

- User preferences, representing user consumption habits, and defining other information (e.g. demographics models) for targeting a specific audience.

- Describing segmented content. Segmentation Metadata is used to edit content for partial recording and non-linear viewing. In this case, metadata is used to navigate within a piece of segmented content.

- Metadata fragmentation, indexing, encoding and encapsulation (transport-agnostic).

TV Anytime employs the MPEG-7 Description Definition Language (DDL) based on XML to be able to describe metadata structure and also the XML encoding of metadata. TV-Anytime also uses several MPEG-7 datatypes and MPEG-7 Classification Schemes. The TV-Anytime Content Description model is depicted on Figure 2.2[6] and its documentation provides the following definitions:

- Entity definitions:

---

[4]http://www.tv-anytime.org/workinggroups/wg-md.html
[5]http://www.tv-anytime.org
[6]Image taken from ftp://tva:tva@ftp.bbc.co.uk/pub/Specifications/COR3_SP003v13.zip, document SP003v13 PartA.doc
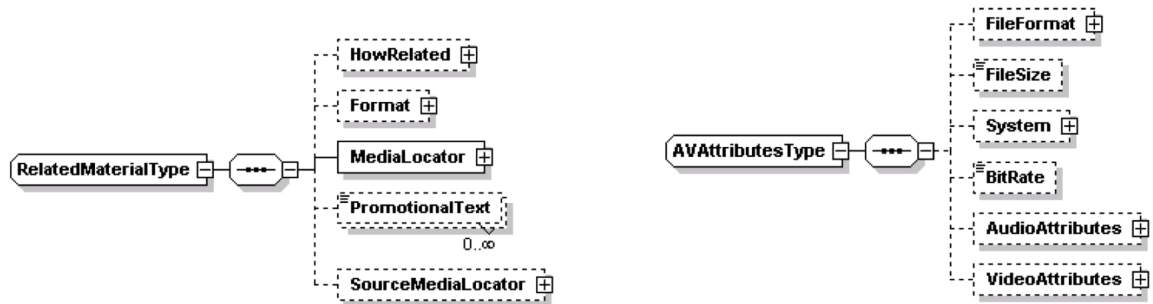
Figure 2.3: TV-Anytime Related Material and AV Attributes Type

- *Program* - the programme represents an editorially coherent piece of content.

- *Program group* - the programme group simply represents a grouping of programmes. A number of different types of group have been identified, such as series, show, aggregate (magazine) programme, and programme concept. Programme groups can also contain other programme groups.

- *Program location* - A programme location contains information about one instance (or publication event) of a programme. Multiple programme locations from the same service provider can be grouped to form a schedule.

- Relationship definitions: Program to Program location (zero to many), Program to Program Group (many to many), etc

As an example, we reproduce the XML schemas of some types structures in Figure 2.3.

NoTube[7] is a European research project that aims to show how Semantic Web technologies can be used to connect TV content and Web content using Linked Open Data. NoTube uses TV-Anytime as the persistent internal metadata format. The project participants argue that it was the only standardized format widely used in CE devices such as STBs, PVRs at that time. TV metadata interoperability[8] has also been well studied by the project.

#### 2.2.1.4   EBU Metadata Model

The EBU (European Broadcasting Union) is the collective organization of Europe's 75 national broadcasters claiming to be the largest association of national broadcasters in the world. EBU's technology arm is called EBU Technical. EBU represents an influential network in the media world[9]. The EBU projects on metadata are part

---

[7] http://notube.tv/
[8] http://notube.tv/tv-metadata-interoperability/
[9] http://tech.ebu.ch/aboutus
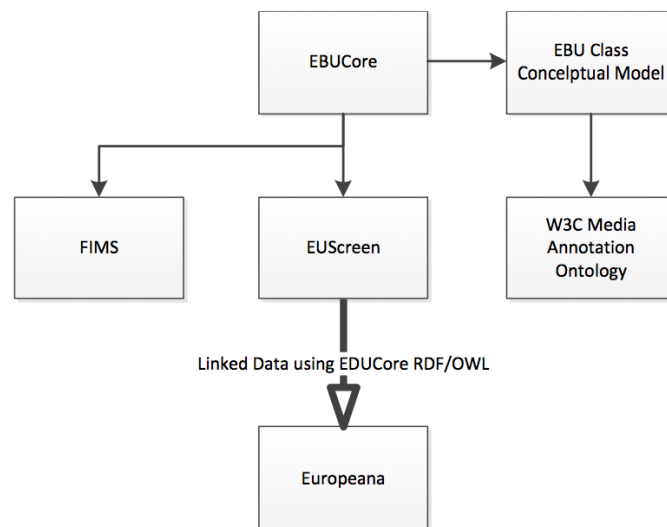
Figure 2.4: EBUCore and its relationships to other metadata models

of the Media Information Management (MIM) Strategic Programme. MIM benefits from the expertise of the EBU Expert Community on Metadata (EC-M), participation to which is open to all metadata experts, or users and implementers keen to learn and contribute[10].

**EBUCore.** The EBUCore (EBU Tech 3293) is the main result of this effort to date and the flagship of EBU's metadata specifications. It can be combined with the Class Conceptual Data Model of simple business objects to provide the appropriate framework for descriptive and technical metadata for use in Service Oriented Architectures. It can also be used in audiovisual ontologies for semantic Web and Linked Data environment. EBUCore has high adoption rate around the world. It is also referenced by the UK DPP (Digital Production Partnership). All EBU metadata specifications[11] are coherent with the EBU Class Conceptual Data Model (CCDM).

EBUCore has been the metadata schema of reference in the project EUScreen which delivers linked data to Europeana using EBUCore's RDF/OWL representation. EBUCore has been also published as AES60 by the Audio Engineering Society (AES)[12]. The W3C Media Annotation Ontology is based on EBU's Class Conceptual Data Model and is fully compatible with EBUCore which mapping has been defined and published as part of the W3C specification (Figure 2.4).

EU Screen and EUscreenXL [13] are european projects in the fields of audiovisual archives, research and software technology. They have been focused on the promotion

---

[10]http://tech.ebu.ch/metadata
[11]http://tech.ebu.ch/MetadataSpecifications
[12]http://www.aes.org/
[13]http://www.euscreen.eu/

of exploration of Europe's rich and diverse cultural history by the use of television content. The **EUscreen metadata schema** is based on **EBUcore** schema that are backward compatible with the **Video Active** schema and fully mappable to the Europeana[14] Data Model (EDM 5.2)[15]. It includes 39 elements of which 18 are mandatory. Programme classification in EUscreen consists of seven main headings including mainly News, drama/Fiction, entertainment, factual programs, advertisements, interstitials and trailers and sport.

Over the last few years the European Broadcasting Union (EBU) and its members have developed several metadata specifications to facilitate the search and exchange of content:

- EBU Tech 3293 - EBUCore: http://tech.ebu.ch/docs/tech/tech3293v1_3.pdf

- EBU Tech 3295 - P-META: http://tech.ebu.ch/docs/tech/tech3295v2_2.pdf

- EBU Tech 3331 - Exchange: http://tech.ebu.ch/docs/tech/tech3331v1_1.pdf

- EBU Tech 3332 - Music: http://tech.ebu.ch/docs/tech/tech3332v1_1.pdf

- EBU Tech 3336 - Classification Schemes: http://tech.ebu.ch/docs/tech/tech3336v1_1.pdf

- EBU Tech 3340 - egtaMETA: http://tech.ebu.ch/docs/tech/tech3340.pdf

- EBU Tech 3349 - Acquisition Metadata: http://tech.ebu.ch/docs/tech/tech3349.pdf

- EBU Tech xxxxx - CCDM: http://tech.ebu.ch/Jahia/site/tech/classmodel

- EBU Eurovision - News Exchange: http://tech.ebu.ch/webdav/site/tech/shared/metadata/NMS_NewsML-G2_eng.pdf

### 2.2.1.5   BBC Programmes

The British Broadcasting Corporation (BBC) is the largest broadcaster in the world. One of the main resource they use to describe programs is the Programmes ontology[16]. This ontology provides the concepts of brands, series (seasons), episodes, broadcast events, broadcast services, etc. and is represented in OWL/RDF. The design of this ontology document is based on the Music Ontology[17] and the FOAF

---

[14]The Europeana Foundation aims at enhancing collaboration between museums, archives, audiovisual collections. It is developing a cross-domain portal providing access to Europe's cultural and scientific heritage. It also facilitates required formal agreement across museums, archives, audiovisual archives and libraries. http://pro.europeana.eu

[15]http://blog.euscreen.eu/wp-content/uploads/2010/11/D1.3.1-Annual-public-report-FINAL.pdf

[16]http://purl.org/ontology/po/

[17]http://www.musicontology.com/
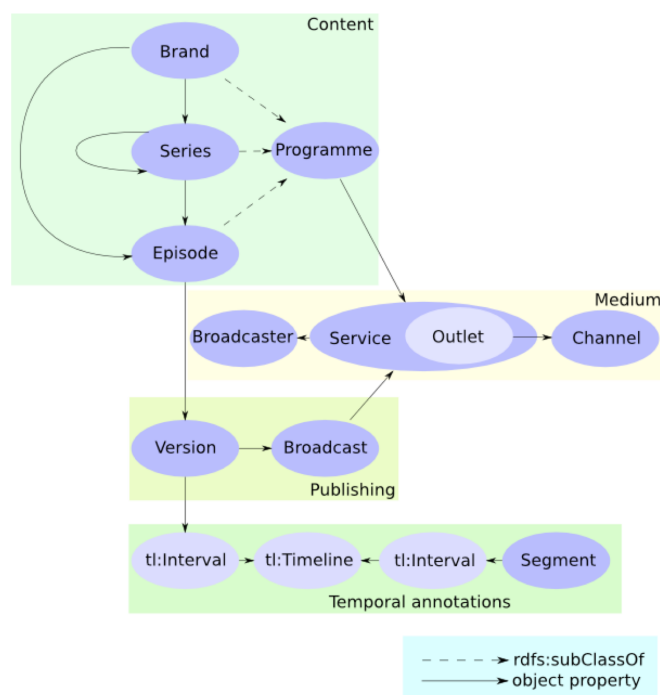
Figure 2.5: BBC Programme ontology model

Vocabulary[18]. The programs model is depicted on Figure 2.5[19] and in based on the
PIPS database schema used previously at the BBC. It describes content in terms of:
Brands, Series, Episodes and Programs.

Publishing is then described in terms of Versions of Episodes and Broadcasts.
Versions are temporarily annotated. Publishing of content is related to Medium, that
is described in terms of: Broadcaster, Service-outlet and Channel. This conceptual
scheme describes how brands, series, episodes, particular versions of episodes and
broadcasts interact with each other. The BBC Programmes ontology also re-uses
other ontologies such as FOAF to express a relationship between a programme to
one of its actors (a person who plays the role of a character). The exhaustive list of
classes available in the ontology is:

AudioDescribedVersion — Brand — Broadcast — Broadcaster — Category —
Channel — Clip — DAB — DVB — Episode — FM — FirstBroadcast — Format
— Genre — IPStream — LW — LocalRadio — MusicSegment — NationalRadio —
OriginalVersion — Outlet — Person — Place — Programme — ProgrammeItem —
Radio — RegionalRadio — RepeatBroadcast — Season — Segment — Series —
Service — ShortenedVersion — SignedVersion — SpeechSegment — Subject —
Subtitle — TV — Version — Web

---

[18]http://xmlns.com/foaf/spec/
[19]Image taken from http://www.bbc.co.uk/ontologies/programmes/2009-09-07.shtml

The exhaustive list of properties available in the ontology is:

actor — anchor — aspect_ratio — author — broadcast_of — broadcast_on — broadcaster — category — channel — clip — commentator — credit — director — duration — episode — executive_producer — format — frequency — genre — location — long_synopsis — masterbrand — medium_synopsis — microsite — news_reader — outlet — parent_series — parent_service — participant — performer — person — place — position — producer — schedule_date — season_broadcast — series — service — short_synopsis — sound_format — subject — subtitle_language — synopsis — tag — text — time — track — version

### 2.2.2 Metadata Models from the Multimedia Analysis Community

#### 2.2.2.1 MPEG-7

MPEG-7, formally named *Multimedia Content Description Interface* [121], is an ISO/IEC standard developed by the Moving Picture Experts Group (MPEG) for the structural and semantic description of multimedia content. MPEG-7 standardizes *tools* or ways to define multimedia *Descriptors* (Ds), *Description Schemes* (DSs) and the relationships between them. The descriptors correspond either to the data features themselves, generally low-level features such as visual (e.g. texture, camera motion) and audio (e.g. spectrum, harmonicity), or semantic objects (e.g. places, actors, events, objects). The description schemes are used for grouping the descriptors into more abstract description entities. These tools as well as their relationships are represented using the *Description Definition Language* (DDL), the core part of MPEG-7. After a requirement specification phase, the W3C XML Schema recommendation[20] has been adopted as the most appropriate syntax for the MPEG-7 DDL.

The flexibility of MPEG-7 is therefore based on allowing descriptions to be associated with arbitrary multimedia segments, at any level of granularity, using different levels of abstraction. The downside of the breadth targeted by MPEG-7 is its complexity and its ambiguity. Hence, MPEG-7 XML Schemas define 1182 elements, 417 attributes and 377 complex types which make the standard difficult to manage. Moreover, the use of XML Schema implies that a great part of the semantics remains implicit. For example, very different syntactic variations may be used in multimedia descriptions with the same intended semantics, while remaining valid MPEG-7 descriptions. Given that the standard does not provide a formal semantics for these descriptions, this syntax variability causes serious interoperability issues for multimedia processing and exchange [124, 131, 192]. The profiles introduced by MPEG-7 and their possible formalization [191] concern, by definition, only a subset of the whole standard.

---

[20]http://www.w3.org/XML/Schema

For alleviating the lack of formal semantics in MPEG-7, four multimedia ontologies represented in OWL and covering the whole standard were proposed [7, 59, 80, 198]. The Table 2.1 summarizes the main characteristic of these four ontologies.

| | Hunter | DS-MIRF | Rhizomik | COMM |
|---|---|---|---|---|
| Foundations | ABC | none | none | DOLCE |
| Complexity | OWL-Full[a] | OWL-DL[b] | OWL-DL[c] | OWL-DL[d] |
| Coverage | MDS+Visual | MDS+CS | All | MDS+Visual |
| Reference | [80] | [198] | [59] | [7] |
| Applications | Digital Libraries, e-Research | Digital Libraries, e-Learning | Digital Rights Management, e-Business | Multimedia Analysis and Annotations |

Table 2.1: Summary of the different MPEG-7 based Multimedia Ontologies.

[a]http://metadata.net/mpeg7/
[b]http://www.music.tuc.gr/ontologies/MPEG703.zip
[c]http://rhizomik.net/ontologies/mpeg7ontos
[d]http://multimedia.semanticweb.org/COMM/

#### 2.2.2.2   Hunter's MPEG-7 Ontology

In 2001, Hunter proposed an initial manual translation of MPEG-7 into RDFS (and then into DAML+OIL) and provided a rationale for its use within the Semantic Web [80]. This multimedia ontology was translated into OWL, extended and harmonized using the ABC upper ontology [91] for applications in the digital libraries [81, 82] and eResearch fields [83].

The current version is an OWL Full ontology containing classes defining the media types (Audio, AudioVisual, Image, Multimedia, Video) and the decompositions from the MPEG-7 Multimedia Description Schemes (MDS) part [121]. The descriptors for recording information about the production and creation, usage, structure and the media features are also defined. While the ontology has most often been applied in conjunction with the ABC upper model, it is independent of that ontology and can also be harmonized with other upper ontologies such as SUMO [136] or DOLCE [56].

#### 2.2.2.3   DS-MIRF Ontology

In 2004, Tsinaraki *et al.* have proposed the DS-MIRF ontology that fully captures in OWL DL the semantics of the MPEG-7 MDS and the Classification Schemes. The ontology has been integrated with OWL domain ontologies for soccer and Formula

1 [199] in order to demonstrate how domain knowledge can be systematically integrated in the general-purpose constructs of MPEG-7. This ontological infrastructure has been utilized in several applications, including audiovisual digital libraries and e-learning.

This ontology has been conceptualized manually, according to the methodology outlined in [198]. The XML Schema simple datatypes defined in MPEG-7 are stored in a separate XML Schema to be imported in the DS-MIRF ontology. The generalization of this approach has led to the development of a transformation model for capturing the semantics of any XML Schema in an OWL DL ontology [197].

### 2.2.2.4 Rhizomik Ontology

In 2005, Garcia and Celma presented the Rhizomik approach that consists in mapping XML Schema constructs to OWL constructs following a generic XML Schema to OWL together with an XML to RDF conversion [59]. Applied to the MPEG-7 schemas, the resulting ontology covers the whole standard as well as the Classification Schemes and TV Anytime. The Rhizomik ontology was originally expressed in OWL Full, since 23 properties must be modeled using an `rdf:Property` because they have both a data type and object type range, i.e. the corresponding elements are both defined as containers of complex types and simple types. An OWL DL version of the ontology has been produced, solving this problem by creating two different properties `owl:DatatypeProperty` and `owl:ObjectProperty`) for each of them. This change is also incorporated into the XML2RDF step in order to map the affected input XML elements to the appropriate OWL property (object or datatype) depending on the kind of content of the input XML element.

The main contribution of this approach is that it benefits from the great amount of metadata that has been already produced by the XML community. Moreover, it is implemented in the ReDeFer project[21], which allows to automatically map input XML Schemas to OWL ontologies and, XML data based on them to RDF metadata following the resulting ontologies. This approach has been used with other large XML Schemas in the Digital Rights Management domain, such as MPEG-21 and ODRL [61], or in the E-Business domain [60].

### 2.2.2.5 COMM Ontology

In 2007, Arndt et al. have proposed COMM, the Core Ontology of MultiMedia for annotation. Based on early work [190, 84], COMM has been designed manually by re-engineering completely MPEG-7 according to the intended semantics of the written standard. The foundational ontology DOLCE serves as the basis of COMM. More precisely, the Description and Situation (D&S) and Ontology of Information Objects

---

[21]http://rhizomik.net/redefer

(OIO) patterns are extended into various multimedia patterns that formalize the MPEG-7 concepts. The use of a upper-level ontology provides a domain independent vocabulary that explicitly includes formal definitions of foundational categories, such as processes or physical objects, and eases the linkage of domain-specific ontologies because of the definition of top level concepts.

COMM covers the most important part of MPEG-7 that is commonly used for describing the structure and the content of multimedia documents.

- *Decomposition.* COMM provides the equivalence of MPEG-7 decomposition to segments. MPEG-7 provides set of descriptors for spatial, temporal, spatiotemporal and media source decompositions of multimedia content into segments. A segment in MPEG-7 can refer to a region of an image, a piece of text, a temporal scene of a video or even to a moving object tracked during a period of time.

- *Annotation.* COMM provides equivalent of MPEG-7 descriptors used to annotate a segment. These descriptors can be low-level visual features, audio features or more abstract concepts. They allow the annotation of the content of multimedia documents or the media asset itself.

The OWL DL version of the core module is just an approximation of the intended semantics of COMM since the use of OWL 1.1 (e.g. qualified cardinality restrictions for number restrictions of MPEG-7 low-level descriptors) and even more expressive logic formalisms are required for capturing its complete semantics [22]. Due to the enormous amount of DL axioms that are present in DOLCE, this ontology is too complex to be used in less formal and real-life scenarios, including the Web datasets we want to target.

#### 2.2.2.6    Comparing Multimedia Ontologies

**Integration with Domain Semantics.** The link between a multimedia ontology and any domain ontologies is crucial. Hunter's MPEG-7 and COMM ontologies both use an upper ontology approach to relate with other ontologies (ABC and DOLCE). Hunter's ontology uses either semantic relations from MPEG-7, such as *depicts*, or defines external properties that use an MPEG-7 class, such as *mpeg7:Multimedia*, as the domain or range. In COMM, the link with existing vocabularies is made within a specific pattern: the *Semantic Annotation Pattern*, refining the DOLCE Ontology of Information Object (OIO) pattern. Consequently, any domain specific ontology goes under the *dolce:Particular* or *owl:Thing* class.

The DS-MIRF ontology integrates domain knowledge by sub-classing one of the MPEG-7 *SemanticBaseType*: places, events, agents, etc. Furthermore, it fully cap-

---

[22]The reification schema of DOLCE D&S is even not completely expressible in OWL 1.1
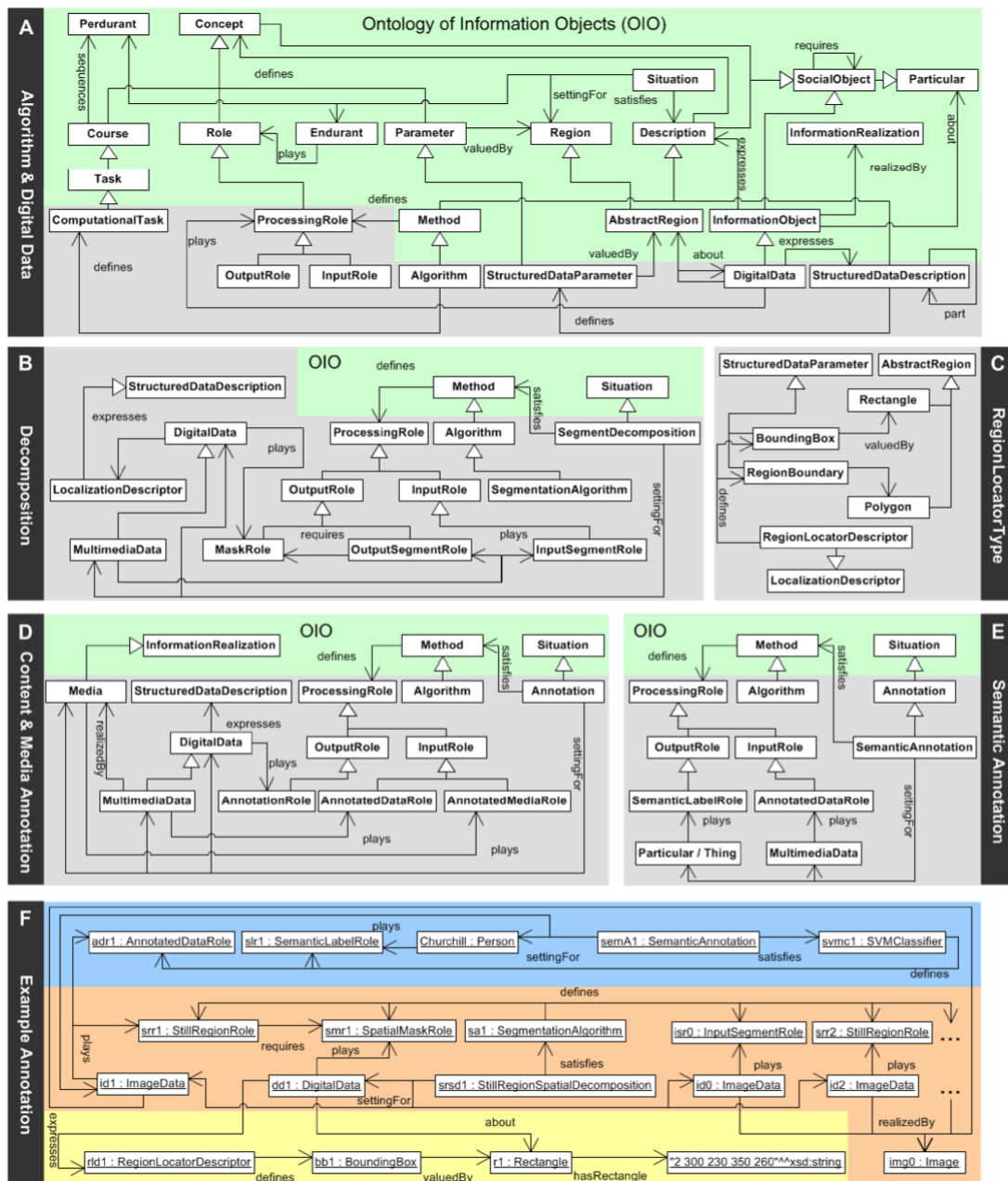
Figure 2.6: COMM: Core Ontology for Multimedia

tures the semantics of the various MPEG-7 relationships represented as instances of the *RelationType*. According to the standard, the value of these properties must come from some particular classification schemes: *RelationBaseCS*, *TemporalRelationCS*, *SpatialRelationCS*, *GraphRelationCS* and *SemanticRelationCS*. A typed relationship ontology extending DS-MIRF has been defined for capturing all these relationships.

**Coverage of a Multimedia Ontology.**    The four multimedia ontologies discussed here cover partially or totally MPEG-7 (see Table 2.1). They also extend sometimes the standard. For example, Hunter's MPEG-7 ontology has been extended for the description of scientific mixed-media data. Common terms used in signal processing and image analysis for describing detailed low-level features such as eccentricity, major axis length, lightest color, etc. are lacking in the MPEG-7 visual descriptors. These extra visual feature descriptors have been introduced as sub-properties of the visual descriptor and color properties, using the namespace mpeg7x to keep these extensions independent of the core MPEG-7 descriptors [78].

The modeling approach of COMM confirms that the ontology offers even more possibilities for multimedia annotation than MPEG-7 since it is interoperable with existing Web ontologies. The explicit representation of algorithms in the multimedia patterns describes the multimedia analysis steps (e.g. manual annotation, output of an analysis algorithm), something that is not possible in MPEG-7. The need for providing this kind of annotation is demonstrated in the use cases of the W3C Multimedia Semantics Incubator Group [23].

**Modeling Decisions and Scalability.**    An important modeling decision for each of the four ontologies is how much they are tied to the MPEG-7 XML Schema. These decisions impact upon the ability of the ontology to support descriptions generated automatically and directly from MPEG-7 XML output and on the complexity of the resulting RDF. Therefore the modeling choices also affect the scalability of the systems using these ontologies and their ability to handle large media data sets and cope with reasoning over very large quantities of triples.

Both the DS-MIRF and the Rhizomik ontologies are based on a systematic one-to-one mapping from the MPEG-7 descriptors to equivalent OWL entities. For the DS-MIRF ontology, the mapping has been carried out manually while for the Rhizomik ontology, it has been automated using an XSL transformation and it is complemented with an XML to RDF mapping. This has been a key motivator for the Rhizomik ontology and the ReDeFer tool where the objective is to provide an intermediate step before going to a more complete multimedia ontology, such as COMM.

The advantage of the one-to-one mapping is that the transformation of the RDF descriptions back to MPEG-7 descriptions may be automated later on. In addition,

---

[23]http://www.w3.org/2005/Incubator/mmsem/XGR-interoperability/

this approach enables the exploitation of legacy data and allows existing tools that output MPEG-7 descriptions to be integrated into a semantic framework. The main drawback of this approach is that it does not guarantee that the intended semantics of MPEG-7 is fully captured and formalized. On the contrary, the syntactic interoperability and conceptual ambiguity problems such as the various ways of expressing a semantic annotation remain.

The COMM ontology avoids doing a one-to-one mapping for solving these ambiguities that come from the XML Schemas, while an MPEG-7-to-COMM converter is still available for re-using legacy metadata. A direct translation from an MPEG-7 XML description using Hunter's ontology is possible. However, in practice, the multimedia semantics captured by the ontology have instead been used to link with domain semantics. Therefore rather than translating MPEG-7 XML descriptions into RDF, this ontology has been used to define semantic statements about a media object and to relate these statements to the domain semantics.

### 2.2.3   Metadata Models from the Web Community

#### 2.2.3.1   hMedia

hMedia[24] is a format aiming to be simple and open for publishing metadata about Images, Video and Audio. It can be embedded in XML and HTML, XHTML, Atom or RSS formats. It is closely related to hCard [25] and uses its facilities for describing information about people, companies, organizations. The basic properties defined in this format are:

- fn: The name of a media.

- contributor: Using text or hCard.

- photo: Using the HTML IMG element (optional).

- player: Using any appropriate html element such as OBJECT (optional).

- enclosure: A URL using the rel-design-pattern.

#### 2.2.3.2   schema.org

Schema.org provides a collection of schemas freely available for marking up data. Some of them are usable for multimedia data annotation.

- schemas used for describing the encoding metadata of audiovisual data (MediaObject): AudioObject, ImageObject, VideoObject.

---

[24]http://microformats.org/wiki/hmedia
[25]http://microformats.org/wiki/hcard

- schemas designed for metadata specific to: Book, Movie, TVSeries, Recipe, CreativeWork, Painting, Photograph, Sculpture etc.

The example below used the Movie schema:

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Pirates of the Carribean: On Stranger Tides (2011)</h1>
  <span itemprop="description">Jack Sparrow and Barbossa embark on a quest to
  find the elusive fountain of youth, only to discover that Blackbeard and
   his daughter are after it too.</span>
  Director:
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">
    <span itemprop="name">Rob Marshall</span>
  </div>
  Writers:
  <div itemprop="author" itemscope itemtype="http://schema.org/Person">
    <span itemprop="name">Ted Elliott</span>
  </div>
  <div itemprop="author" itemscope itemtype="http://schema.org/Person">
    <span itemprop="name">Terry Rossio</span>
  </div>
</div>
```

### 2.2.3.3   MediaRSS

MediaRSS[26] is a RSS module developed to supplement the enclosure capabilities of RSS 2.0. Currently RSS enclosures are used to integrate audio files and images to RSS. Media RSS is an extension that allows handling other media types and enables media creator to provide additional metadata with the media.

MediaRSS defines its namespace to be http://search.yahoo.com/mrss/. The primary elements of MediaRSS are:

- **<media:group>** sub-element of <item>. Enables grouping of <media:content>.

- **<media:content** is a sub-element of either <item>or <media:group>. It defines the following properties:

    - url - specifies the direct URL to the media object.
    - fileSize - specifies the number of bytes of the media object.
    - type - the standard MIME type.
    - medium - specifies the type of media object (image — audio — video — document — executable).
    - isDefault - specifies if this is the default media object that should be used for the ¡media:group¿.
    - expression - specifies if the media object is a sample or the full version of the object, and if it is a continuous stream (sample — full — nonstop).

---

[26]http://www.rssboard.org/media-rss

- – bitrate - specifies the kb/sec rate of media.

- – framerate - specifies frames/sec rate of the media object.

- – samplingrate - specifies samples/sec (kHz) rate of media object.

- – channels - specifies number of audio channels in the media object.

- – duration - specifies playing time of the media object plays.

- – height - specifies the height of the media object.

- – width - specifies the width of the media object.

- – lang - specifies the primary language of the media object. Language codes are derived from RFC 3066, similarly to the xml:lang attribute detailed in the XML 1.0 Specification (Third Edition).

Some optional elements of MediaRSS are media:rating, media:title, media:description, media:keywords, media:thumbnails, media:category, media:hash, media:player, media:credit, etc.

### 2.2.3.4 YouTube metadata

YouTube[27] is an online video streaming service provided by Google. It provides documentation (YouTube Data API) for programmers who are writing client applications that interact with YouTube media content. It lists the different types of feeds that a user can retrieve and provides diagrams that explain how to navigate between them. It also defines the parameters used in YouTube Data API requests as well as the JSON tags returned in an API response. The YouTube API supports the following JSON schemas for the different API requests supported: Captions, ChannelBanners, Channels, ChannelSections, Comments, CommentThreads, GuideCategories, Languages, Regions, PlaylistItems, Playlists, Search, Subscriptions, Thumbnails, VideoCategories, Videos and Watermarks.

## 2.2.4 Metadata Models from News and Photo Industry

### 2.2.4.1 IPTC

The IPTC (International Press Telecomunications Council) is a consortium of more than 60 news agencies, news publishers and news industry vendors from all continents except South America and Oceania. It develops and maintains technical standards for improved news exchange that are used by the most of major news organizations in the world.

---

[27]https://developers.google.com/youtube/v3/

The latest specifications are part of the so-called G2 family of standards that are based on XML but created with the Semantic Web technologies idea[28]. The family of formats consists of:

- NewsML-G2 - standard to exchange news of any kind and media-type (XML).

- EventsML-G2 - standard for conveying event information in a news industry environment (XML).

- SportsML-G2 - standard for sharing sports data (XML).

Older News Exchange Formats are:

- NewsML 1 - IPTC's first standard to exchange multimedia news and packages of them (XML).

- NITF - format to define the content and structure of news articles (XML).

- IIM - first multimedia format of the IPTC (binary data).

- IPTC7901 - first news exchange format of the IPTC which is still widely used for simple text-only transmissions.

**IPTC Photo Metadata Standard** The IPTC Photo metadata standards are described in the CEPIC-IPTC Image Metadata Handbook[29]. IPTC provides also a free Adobe CS Metadata toolkit. IPTC issued the *Embedded Metadata Manifesto (2011)* document proposing guiding principles for embedding metadata in image formats[30].

**rNews** IPTC works recently on the rNews[31] format to embed metadata in online news. rNews is considered by IPTC to be at production level, i.e. the version 1.0 was approved in October 2011. rNews uses RDFa to embed semantic markup into HTML documents.

### 2.2.4.2   Metadata Working Group's Guidelines for Handling Image Meta- data

The Metadata Working Group[32] (MWG) is a consortium of companies in the digital media industry, focused on preservation and seamless interoperability of digital image metadata and interoperability and availability to all applications, devices, and services. Technical specifications published by MWG describe ways to effectively

---

[28]https://iptc.org/standards/
[29]http://www.iptc.org/site/Photo_Metadata/
[30]http://www.iptc.org/site/Photo_Metadata/Embedded_Metadata_Manifesto_(2011)
[31]http://dev.iptc.org/rNews
[32]http://www.metadataworkinggroup.com/specs
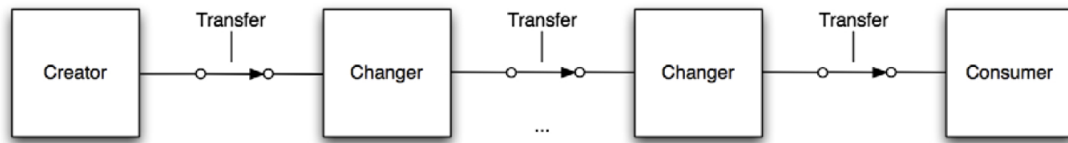
Figure 2.7: MWG Guidelines Actor state diagram

store metadata into digital media files. These specifications are freely available to software developers, manufacturers and service providers, ensuring that their use of metadata is consistent. It also allows consumers to better describe, organize and find their media. MWG specifications often rely on existing standards and the current document is the Guidelines For Handling Image Metadata version 2.0 [33].

MWG guidelines introduce the notion of different actors that play specific roles in metadata processing. There are essentially three types of actors: Creator, Changer and Consumer. The Guidelines For Handling Image Metadata specification also analyzes existing image metadata formats and their respective relation. They end up with result depicted on Figure 2.7.

### 2.2.5  W3C Ontology for Media Resources

The Ontology for Media Resources[34] is a core vocabulary of descriptive properties for media resources. Its main aim is to bridge the different descriptions of media resources and provide a coherent set of media metadata properties along with their mappings to existing metadata standards and formats. The Ontology for Media Resources provides also implementation compatible with Semantic Web paradigm in RDF/OWL form. It is a W3C recommendation since February 2012, produced the Media Annotations Working Group[35].

The Ontology for Media Resources provides mapping tables for metadata from many other standards such as CableLabs 1.1, DIG35, Dublin Core, EBUcore, EXIF 2.2, ID3, IPTC, LOM 2.1, Media RSS, MPEG-7, OGG, QuickTime, DMS-1, TTML, TV-Anytime, TXFeed, XMP, YouTube, 3GP, Flash (FLV, F4V), MP4, WebM. The following subsections provide an overview of the core properties, their descriptions and their relevance for the LinkedTV project.

#### 2.2.5.1  Identification Properties

As we plan to bring multimedia to the Web ecosystem, our approach will use URI identifiers to uniquely identify media resources. Properties such as title, language

---

[33]http://www.metadataworkinggroup.com/pdf/mwg_guidance.pdf
[34]http://www.w3.org/TR/mediaont-10/
[35]http://www.w3.org/2008/WebVideo/Annotations/

and locator will also be used. More details about those attributes are displayed on
Table 2.2.

Table 2.2: Ontology for Media Resources - Identification properties

| Name | Description |
|---|---|
| identifier | A URI identifying a media resource, which can be either an abstract concept (e.g., Hamlet) or a specific object (e.g., an MPEG-4 encoding of the English version of "Hamlet"). When only legacy identifiers are available, a URI must be minted, for example using the tag: scheme RFC 4151. |
| title | A tuple that specifies the title or name given to the resource. The type can be used to optionally define the category of the title. |
| language | The language used in the resource. We recommend to use a controlled vocabulary such as BCP 47. An BCP 47 language identifier can also identify sign languages e.g. using ISO 639-3 subtags like bfi (British sign language). |
| locator | The logical address at which the resource can be accessed (e.g. a URL, or a DVB URI). |

http://www.ietf.org/rfc/rfc4151.txt
http://www.rfc-editor.org/rfc/bcp/bcp47.txt

### 2.2.5.2   Creation Properties

The Ontology for Media Resources contains properties for describing the creator
of the media resource. Those properties together with more advanced provenance
information are needed in today's ecosystem for keeping track of the data quality
and trustiness.

### 2.2.5.3   Technical Properties

The Ontology for Media Resources also contains properties for describing the techni-
cal information of the media resource. This helps different devices dealing with the
content to choose the best configuration parameters for an optimal displaying.

### 2.2.5.4   Content Description Properties

The Ontology for Media Resources contains simple (free text) properties for describ-
ing the content of a media resource. For general properties, such as the description,
they can be used. In order to link to more structured pieces of information in dif-
ferent domains, we will study how to use the Open Annotation Model for attaching
media fragments with instances describing them via more elaborated relations.

Table 2.3: Ontology for Media Resources - Creation properties

| Name | Description |
|---|---|
| contributor | A tuple identifying the agent, using either a URI (recommended best practice) or plain text. The role can be used to optionally define the nature of the contribution (e.g., actor, cameraman, director, singer, author, artist, or other role types). An example of such a tuple is: `imdb:nm0000318`, director. |
| creator | A tuple identifying the author of the resource, using either a URI (recommended best practice) or plain text. The role can be used to optionally define the category of author (e.g., playwright or author). The role is defined as plain text. An example of such a tuple is: `dbpedia:Shakespeare`, playwright. |
| date | A tuple defining the date and time that the resource was created. The type can be used to optionally define the category of creation date (e.g., release date, date recorded, or date edited). |
| location | A tuple identifying a name or a set of geographic coordinates, in a given system, that describe where the resource has been created, developed, recorded, or otherwise authored. The name can be defined using either a URI (recommended best practice) or plain text. |

Table 2.4: Ontology for Media Resources - Technical properties

| Name | Description |
|---|---|
| frameSize | A tuple defining the frame size of the resource (e.g., width and height of 720 and 480 units, respectively). The units can be optionally specified; if the units are not specified, then the values MUST be interpreted as pixels. |
| compression | The compression type used. For container files (e.g., QuickTime, AVI), the compression is not defined by the format, as a container file can have several tracks that each use different encodings. In such a case, several compression instances should be used (see RFC 3986 and RFC 4281) for more details. |
| format | The MIME type of the resource (e.g., wrapper or bucket media types, container types), ideally including as much information as possible about the resource such as media type parameters, for example, using the "codecs" parameter - RFC 4281. |
| samplingRate | The audio sampling rate. The units are defined to be samples/second. |
| frameRate | The video frame rate. The units are defined to be frames/second. |
| averageBitRate | The average bit rate. The units are defined to be kbps. |
| numTracks | A tuple defining the number of tracks of a resource, optionally followed by the type of track (e.g., video, audio, or subtitle). |

http://www.ietf.org/rfc/rfc3986.txt
http://www.ietf.org/rfc/rfc4281.txt
http://www.w3.org/2003/01/geo/

Table 2.5: Ontology for Media Resources - Content description properties

| Name | Description |
|------|-------------|
| description | Free-form text describing the content of the resource. |
| keyword | A concept, descriptive phrase or keyword that specifies the topic of the resource, using either a URI (recommended best practice) or plain text. |
| genre | The category of the content of the resource, using either a URI (recommended best practice) or plain text. |
| rating | The rating value (e.g., customer rating, review, audience appreciation), specified by a tuple defining the rating value, an optional rating person or organization defined as either a URI (recommended best practice) or as plain text, and an optional voting range. |

### 2.2.5.5   Relational Properties

Relational properties are intended to convey a semantic relationship between a source content and other resources that sometimes are derivative. For example, one can express a semantic relationship between a movie and its trailer. This set of properties will be useful for typing the relationship between a seed video content and the suggested hyperlinked resources. In order to add provenance information about the relationship, or specify more details about the type of connection stablished, we will again study the use of the Open Annotation Model.

Table 2.6: Ontology for Media Resources - Relational properties

| Name | Description |
|------|-------------|
| relation | A tuple that identifies a resource that the current resource is related with (using either a URI -recommended best practice- or plain text), and optionally, specifies the nature of the relationship. |
| collection | The name of the collection (using either a URI or plain text) from which the resource originates or to which it belongs. We recommend to use a URI, as a best practice. |

### 2.2.5.6   Fragment Properties

This is one of the features inside The Ontology for Media Resources that brings more potential to the Web ecosystem. Being able to structure resources as Web addressable media fragments opens new possibilities for a fine grained description of multimedia content, providing application tailored anchors where other annotations can be attached to. In order to achieve this, the properties to describe media fragments are designed to have standard Media Fragments URI's as domain.

Table 2.7: Ontology for Media Resources - Fragment properties

| Name | Description |
|---|---|
| fragment | Materialized as "hasFragment" property, is a tuple containing a fragment identifier and optionally, its role. A fragment is a portion of a Resource, as defined by the MediaFragment Working Group. |
| namedFragment | A tuple containing a named fragment identifier and its label. |
| isFragmentOf | The inverse of the "hasFragment", annotates which media resource a fragment is belonging to. |

#### 2.2.5.7 Rights Properties

The Ontology for Media Resources contains simple properties to describe the rights to attach to a media resource. It consists mainly in a tuple containing the copyright statement associated with the resource and optionally, the identifier of the copyright holder. Issues related to Digital Rights Management are out of scope for this specification, apart from the metadata supported by the copyright and policy attributes.

#### 2.2.5.8 Distribution Properties

The Ontology for Media Resources contains properties to describe the publisher and the target audience of a media resource. This is however much simpler that what a standard such as TV Anytime can express. The properties considered are mainly the publisher of the resource, defined as either a URI or plain text, and the targetAudience for identifying the audience being addressed (demographic class, parental guidance group, or geographical region) and an optional classification system (e.g., a parental guidance issuing agency).

### 2.2.6 Event Metadata Models

#### 2.2.6.1 Event Ontology

The Event Ontology[36] is developed by Y.Raimond and S. Abdallah in the Centre for Digital Music in Queen Mary, University of London. The central concept of this ontology is the notion of event understood as the way by which cognitive agents classify arbitrary time/space regions. The Event ontology is inspired by the work of J. F. Allen and G. Fergusson who claim: *"events are primarily linguistic or cognitive in nature. That is, the world does not really contain events. Rather, events are the way by which agents classify certain useful and relevant patterns of change."* [2].

The Event ontology defines the classes: Event (see Figure 2.8), Factor, Product, and the properties Agent, agent_in, factor, factor_of, has Agent, hasFactor, hasLiter-

---

[36] http://motools.sourceforge.net/event/event.html
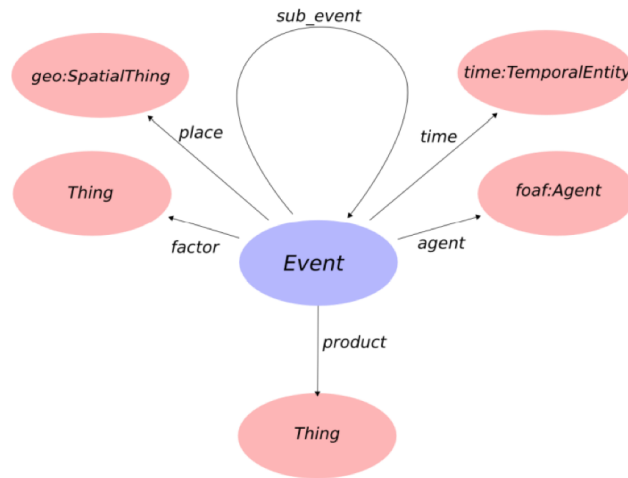
Figure 2.8: Event Ontology

alFactor, hasProduct, hasSubEvent, isAgentIn, isFactorOf, literal_factor, place, producedIn, produced_in, product, sub_event, time.

### 2.2.6.2 LODE Ontology

LODE [37] is a minimal model that encapsulates the most useful properties for describing events, enabling an interoperable modeling of the "factual" aspects of events, where these can be characterized in terms of the *four Ws*: *What* happened, *Where* did it happen, *When* did it happen, and *Who* was involved. "Factual" relations within and among events are intended to represent intersubjective "consensus reality" and thus are not necessarily associated with a particular perspective or interpretation. The LODE model thus allows to express characteristics about which a stable consensus has been reached, whether these are considered to be empirically given or rhetorically produced will depend on one's epistemological stance.

The LODE ontology contains numerous axioms that establish mappings with other event vocabularies such as Event, Dolce Ultra Light (DUL), Cyc, ABC, CIDOC-CRM, SEM. It consists of a single class `lode:Event` and a number of properties like *atPlace*, *atTime*, *illustrate*, *inSpace*, etc.

### 2.2.7 Annotation Models

The Open Annotation specification is being developed by the W3C Open Annotation Community Group[38]. The document aims at developing an open common specification for annotating digital resources in the Web, no matter their nature. Therefore this model is well appropriate for our purposes. The current model is made from

---

[37]http://linkedevents.org/ontology/
[38]http://www.w3.org/community/openannotation/

the reconciliation of two recent proposals: the Annotation Ontology[39] and the Open Annotation Model[40].

The Open Annotation Community Group has published two drafts:

- Core Open Annotation Specification[41].

- Open Annotation Extension Specification[42].

In the following, we describe how the features of the Open Annotation specification can be used for annotating multimedia documents in the Web. The general structure of an annotation in this proposal is depicted in Figure 2.9. In this model, an anno-
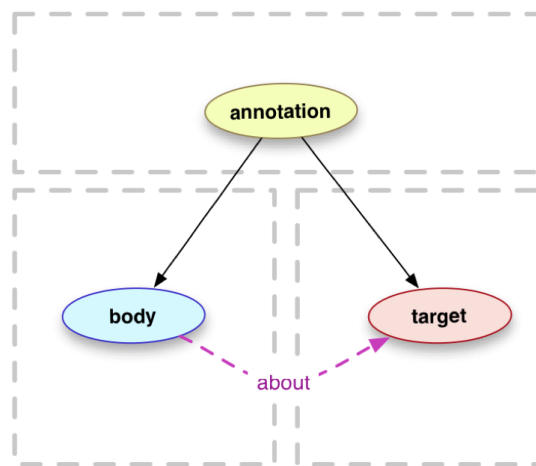


Figure 2.9: Open Annotation Core Model

tation consists of a set of connected resources referenced as body and target. The particularities of the relationship can be further specified through different properties in the oa:annotation instance, but a priory it is indicating that the body is saying something about the target.The core model of Open Annotation consists of one class and two relations:

- *oa:Annotation*: The class for making explicit the Annotation.

- *oa:hasBody*: The relationship between an Annotation and the Body of the Annotation

- *oa:hasTarget*: The relationship between an Annotation and the Target of the Annotation.

---

[39]http://code.google.com/p/annotation-ontology/
[40]http://www.openannotation.org/spec/beta/
[41]http://www.openannotation.org/spec/core/
[42]http://www.openannotation.org/spec/extension/

The Open Annotation model can include also tracking basic provenance information, in particular:

- *oa:annotator* - Relation - Identification of agent (human or software) responsible for annotation.

- *oa:annotated* - Property - Time of creation of annotation.

- *oa:generator* - Relation - Agent (software) responsible of generating serialization of annotation.

- *oa:generated* - Property - Time at which the software agent generated the serialization.

- *oa:modelVersion* - Relationship - The version of model of annotation.
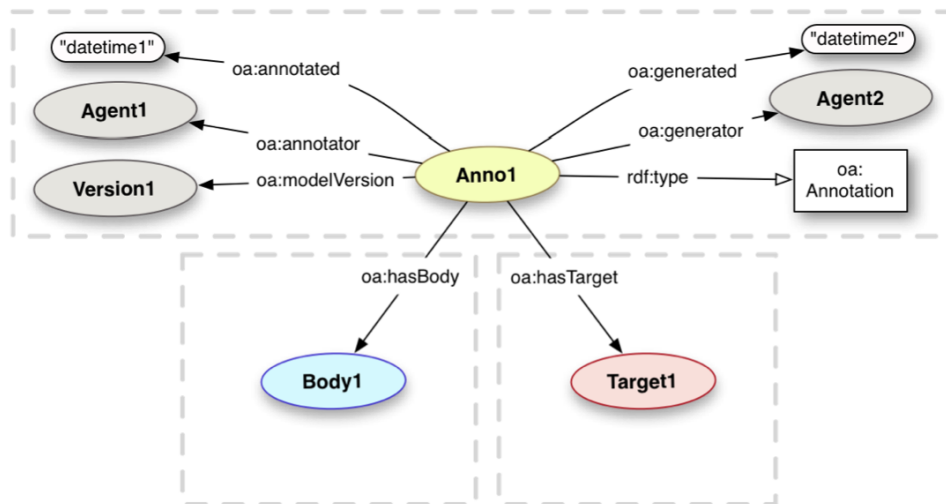


Figure 2.10: Open Annotation Provenance Model

However, as we will see later, provenance information can be further specified by the W3C Provenance Ontology.

### 2.2.7.1   Open Annotation Semantic Tags

Tagging a resource, either with a short text string or a with a URI, is a common use case for Annotation. Tags are typically keywords or labels, and used for organization, description or discovery of the resource being tagged. In the Semantic Web, URIs are used instead of strings to avoid the issue of polysemy where one word has multiple meanings so two different URIs referring to entities with same surface form are clearly distinguished. In the Open Annotation Core model, the tag is represented as the Body of the Annotation, and the resource being tagged is the Target. The

body resource is annotated as oa:Tag class in case the tag is a textual item, and oa:SemanticTag in case it is a URI with a tagging resource. For more details, check Figure 2.11.
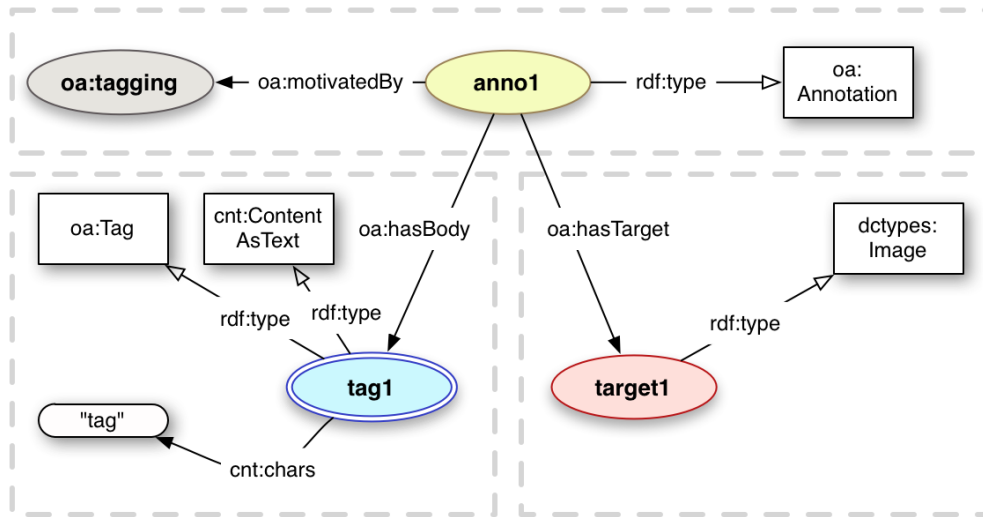


Figure 2.11: Open Annotation Semantic Tagging

#### 2.2.7.2   Open Annotation Motivation

In most of the cases it is important to understand the reasons why the Annotation has been created and not just the agents involved. For this reason the latest version of the Open Annotation Core offers the possibility of materializing such motivations as SKOS Concepts, so they can be inter-related between communities with more meaningful distinctions than a simple class/subclass tree. This frees up the use of subclassing for situations when it is desirable to be more explicit and prescriptive about the form an Annotation takes. In order to make explicit the motivation of an annotation, we will attach to it one oa:motivatedBy pointing to an instance of oa:Motivation, which is a subClass of skos:Concept. A list of high level Motivations is presented below.

oa:bookmarking, oa:classifying, oa:commenting, oa:describing, oa:editing, oa:highlighting, oa:identifying, oa:linking, oa:moderating, oa:questioning, oa:replying, oa:tagging,

As there could be many situations where more exact definitions of Motivation are required or desirable, it is possible to create a new oa:Motivation resource and relate it to one or more that already exist. New Motivations must be instances of oa:Motivation, and it is recommended to assert a skos:broader relationship between

the new motivation and at least one high level Motivation. Of course other relation-
ships, such as skos:relatedMatch, skos:exactMatch and skos:closeMatch, can be used
to compare it with concepts created by other communities.

### 2.2.8   Other Common Used Vocabularies

We conclude this overview of metadata models by surveying some of the most com-
mon used vocabularies in the Semantic Web.

#### 2.2.8.1   FOAF Ontology

The Friend of a Friend (FOAF) project started with the aim of creating a Web of
machine-readable pages describing people, the links between them and the things they
do, work on, create and like, with an emphasis on the on-line presence of people [43].
The FOAF project is well known in the Linked Data community and since 2004,
more than 1.5 million FOAF documents have been generated.

There is a number of sites that use the FOAF vocabulary as a standard for data
exchange: blogging sites[44] or content management systems such as Drupal 8 which
uses FOAF as one of the vocabularies for its RDF-based core[45].There are also sev-
eral FOAF extensions that have been applied to this vocabulary with a focus new
requirements discovered on the way [46].

#### 2.2.8.2   PROV-O Ontology

PROV-O[47] is an ontology (W3C Recommendation since April 2013) that provides a
set of classes, properties, and restrictions allowing users to represent and interchange
provenance information. It also aims at providing a common ground for exchange
of provenance information generated in heterogeneous systems. PROV-O is being
developed by the W3C Provenance Working Group. PROV is actually a family of
specifications consisting of the following documents:

- PROV-OVERVIEW , an overview of the PROV family of documents [PROV-
  OVERVIEW];

- PROV-PRIMER , a primer for the PROV data model [PROV-PRIMER];

- PROV-O the PROV ontology, an OWL2 ontology allowing the mapping of the
  PROV data model to RDF (this document);

---

[43]http://www.foaf-project.org
[44]http://www.livejournal.com/
[45]http://drupal.org/node/574624
[46]http://xmlns.com/foaf/spec/#sec-evolution
[47]http://www.w3.org/TR/prov-o/

- PROV-DM the PROV data model for provenance [PROV-DM];

- PROV-N a notation for provenance aimed at human consumption [PROV-N];

- PROV-CONSTRAINTS a set of constraints applying to the PROV data model [PROV-CONSTRAINTS];

- PROV-XML an XML schema for the PROV data model [PROV-XML];

- PROV-AQ mechanisms for accessing and querying provenance [PROV-AQ];

- PROV-DICTIONARY introduces a specific type of collection, consisting of key-entity pairs [PROV-DICTIONARY];

- PROV-DC provides a mapping between PROV-O and Dublin Core Terms [PROV-DC];

- PROV-SEM a declarative specification in terms of first-order logic of the PROV data model [PROV-SEM];

- PROV-LINKS

- introduces a mechanism to link across bundles [PROV-LINKS].

The core of the PROV-O model consists of the so called "Starting Point Terms", which are :

- *prov:Entity* - physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary.

- *prov:Activity* - something that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities.

- *prov:Agent* - is something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity.

Those classes are related via some well defined properties (such as prov:wasDerivedFrom or prov:wasAttributedTo) that build at the core of the PROV-O model depicted in Figure 2.12 PROV-O enables the extension of its core model with more detailed description of agents, concepts concerning activity and entities.

### 2.2.8.3 NERD Ontology

The NERD ontology[48] is a set of mappings established manually between different taxonomies of named entity types recognized by numerous Web APIs that perform

---

[48]http://nerd.eurecom.fr/ontology
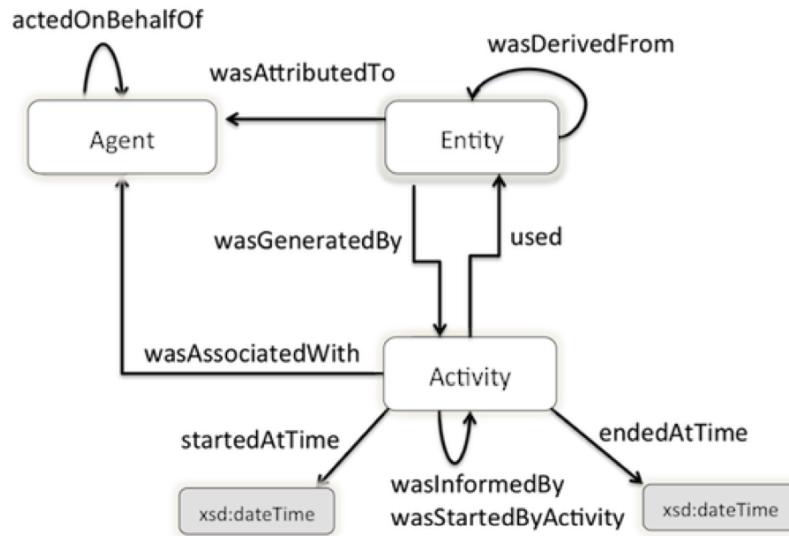
Figure 2.12: PROV-O ontology core model

**Named Entity extraction.** Concepts included in the NERD ontology are collected from different schema types: ontology (for DBpedia Spotlight, Lupedia, and Zemanta), lightweight taxonomy (for AlchemyAPI or Yahoo!) or simple flat type lists (for Extractiv, OpenCalais, Saplo, and Wikimeta). The selection of these concepts has been done considering the greatest common denominator among the taxonomies.

The NERD ontology becomes a good reference ontology for comparing the classification task of NE extractors. We show an example mapping among those extractors below: the `City` type is considered as being equivalent to `alchemy:City`, `dbpedia-owl:City`, `extractiv:CITY`, `opencalais:City`, `evri:City` while being more specific than `wikimeta:LOC` and `zemanta:location`.

```
nerd:City a rdfs:Class ;
  rdfs:subClassOf wikimeta:LOC ;
  rdfs:subClassOf zemanta:location ;
  owl:equivalentClass alchemy:City ;
  owl:equivalentClass dbpedia-owl:City ;
  owl:equivalentClass evri:City ;
  owl:equivalentClass extractiv:CITY ;
  owl:equivalentClass opencalais:City .
```

## 2.3 Requirements for Lightweight Web Models for Multimedia Annotation

After having reviewed the numerous multimedia metadata standards, we derive some general requirements (Section 2.3.1), some functional requirements (Section 2.3.2) and some other requirements dealing with intellectual property management (Section 2.3.3) that we have identified as important for our research. In top of all them we have always prioritized to keep the model as lightweight as possible and re-using already existing vocabularies.

### 2.3.1 General Requirements

In this subsection we list a set of general requirements and good design principles applicable to other software engineer fields and in particular to all metadata models that intend to be useful for information transferring, have wide coverage and remain applicable for a long time.

- **Extensibility.** Information systems, especially audio-visual and multimedia ones, evolve over time and keep being extended, connected, linked, combined or integrated. The lightweight model infrastructure needs to support such system evolution and enable implementation of new extensions to information systems that satisfy new functional requirements as they arise. Therefore a model infrastructure should be extensible enough to provide tools and development support for new unforeseen needs.

- **Modularity.** Some systems may require only certain parts of the entire model. Making the underlying infrastructure modular allows each information system to select only the desired modules without unnecessary increase in complexity.

- **Reusability.** The proposed metadata modeling infrastructure will be used by several information systems built for different tasks and different users. Sometimes they may even work in different domains. Life of involved systems may span long periods so the reusability of the developed metadata model is an important requirement.

- **Formal Precision.** The great majority of models available aim at establishing a common foundation for interoperability of many different information systems developed by authors with different backgrounds and working in different domains. The modeling infrastructure should be formally described to enable a common understanding and usage of the important concepts.

- **Machine Accessible Semantics.** With the advent of Semantic Web technologies, the requirement of machine accessible semantics has gained in popularity

and importance. Information consumption can no longer be available only to human but it should be directly accessible by information systems.

- **Standardization.** The developed model should rely on standards and commonly used vocabularies. Standardized solutions allow interoperability with other systems and future extensibility and reusability, so it is logically connected with those aforementioned requirements.

- **Seamlessness.** The adopted solution should allow seamless internal and external integration. The metadata modeling infrastructure must be internally designed to provide seamless and coherent interconnection of its structural components.

- **Unobtrusiveness.** The metadata model should not be undesirably noticeable or blatant. It should not be sticking out in an unwelcome way. There should be no obstacles or difficulties preventing a consumers and producers for effectively using the proposed model.

- **Multilingual Support.** The adopted solution should not be specific to a single language but should be applicable to the description of multilingual content.

- **Well Documented.** The classes and properties should be well documented. The ontology should be published following the best practices of the semantic Web community.

### 2.3.2   Functional Requirements

#### 2.3.2.1   Semantic Web and Linked Data Conformity

The Semantic Web is an extension of the current Web in which information is coming together with annotations making explicit the well-defined semantics expressed inside them [13]. The well-defined semantics is based on common knowledge representation formalisms defined and promoted by the World Wide Web Consortium. The Semantic Web initiative is based on three fundamental formal languages specifically designed for data: XML (Extensible Markup Language), a syntax for serializing information, RDF (Resource Description Framework), a simple data model that consists in representing knowledge in the forms of triples and OWL (Web Ontology Language), a description logic based language for defining schema knowledge.

The other important characteristics are the usage of URI for identifying any resources or entities and the usage of Unicode for encoding text. The Semantic Web Stack is depicted in the Figure 2.13. The conformity with the Semantic Web initiative is based on general requirements of machine accessible semantics, extensibility, reusability, standardization and seamlessness.
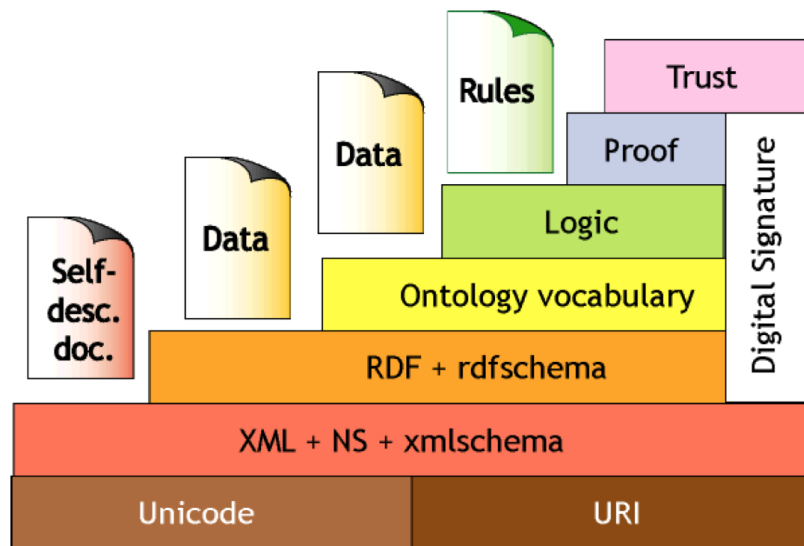
Figure 2.13: The Semantic Web Stack

The Linked Data initiative was started by Tim Berners-Lee as an architectural vision for the Semantic Web. It explores the idea of Semantic Web as putting emphasis on making links explicit, so that both people and machines can explore a semantically interconnected Web of data. If the data is linked, then "when you have some of it, you can find other, related, data"[49]. Just like in HTML where there are relationships and hypertext links between documents, the Linked Data initiative wants to encourage a similar approach in the case of general data content, represented in RDF. The key requirements for Linked Data are quite simple:

- Use URIs as names for things.

- Use HTTP URIs so people can look up those names.

- When someone looks up a URI, provide useful information, using standards (RDF, SPARQL).

- Include links to other URIs, so that they can discover more things.

Guidance provided by these general principles was later extended by technical documents[50] and papers [16, 15] by Bizer and Sauermann [161] among others. Linked Data can be crawled with appropriate browsers by following RDF links. A search engine can also search these information sources similarly to conventional Web sites. However, unlike HTML, which only provides a generic linking capability, links in Linked Data environment can have different types: we can e.g. specify that one

---

[49]http://www.w3.org/DesignIssues/LinkedData.html
[50]http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

person is an author of a paper, or that this person knows another one. General requirements of extensibility, reusability, standardization, seamlessness and unobtrusiveness imply also to conform with Linked Data principles.

### 2.3.2.2    Requirement for Media Broadcasting

Our metadata model has to be able to provide tools for precise broadcaster and media identification. This includes:

- Broadcaster identification: company or institution providing the content

- Broadcast service: individual service (e.g. logical channel)

- Medium / Channel: physical channel providing broadcast service

The BBC Programme ontology reviewed in the Section 2.2.1 fulfills this requirement by offering the adequate classes to represent such scenario.

### 2.3.2.3    Requirement for Structured Modeling of Broadcast Content

The proposed metadata model should allow to distinguish the different structural components of broadcasted programmes. Important concepts with respect to the structure of content like: Brand (e.g. Red Dwarf), Series (e.g. Season 3) and Episode (e.g. Episode 15) need to be taken into account. Then, more complex relationships can be stablished between these agents: episodes can have several versions, and one version can be broadcasted several times. Again, the BBC Programme ontology reviewed in the Section 2.2.1 fulfills this requirement.

### 2.3.2.4    Requirement for Media Resource and Fragments Modeling

The metadata model should contain properties for describing the multimedia documents as media resources in the Web. This includes the unique identification of the resource (preferably with a URI) and general properties such as the Title, Description, Creator, Date of creation, Language, Genre and Publisher. Finally, the model should enable to describe the content of the media resource at a lower level of granularity, in order to address particular fragments inside the media documents. The W3C Ontology for Media Resources reviewed in the Section 2.2.5, together with the Media Fragment URI's specification, fulfill completely this requirement. The media fragments instantiated using this ontology can be further specified by semantic typing using other taxonomies and class hierarchies depending on the specific scenario requirements.

### 2.3.2.5    Requirement for Annotation and Tagging Support

The metadata model should support advanced annotation of multimedia resources and fragments, explicitly materializing those annotations into instances where other information can be attached to, like provenance data or multiple semantic tagging. In particular, the main requirements for a flexible media annotation are:

- Support for Annotations of various types, leaving opened the possibility of using external vocabularies to further refine them

- Support for semantic tagging, not only over the content but also over the annotations themselves

- Tracking of annotator

- Tracking of annotation software

- Tracking of annotation date and time

The Open Annotation model reviewed in the Section 2.2.7 fulfills those requirement while keeping in mind the general prerequisites explained in Section 2.3.1.

### 2.3.2.6    Requirement for Representing Multimedia Analysis Results

The model to be designed has to deal with numerous multimedia analysis processes performed by pure video and audio processing algorithms such as shot or scene segmentation, concept detection, face detection and identification, automatic speech recognition, etc. In the context of the LinkedTV project, those results will be available in different formats that need to be incorporated to the semantic model, which should be able to represent this information. Candidate vocabularies are the Large Scale Concept Ontology for Multimedia (LSCOM [51] and their semantic alternatives [52]) for representing concepts detected in video frames. Other fragment oriented concepts such as shots or scenes have already been addressed via the W3C Ontology for Media Resources in previous requirements.

### 2.3.2.7    Requirement for Provenance Tracking

Managing provenance information is vital to ensure quality and reliability in data. The model proposed needs to provide means to describe how certain agents realized some activities concerning some data object at a particular date or time. The main of provenance tracking requirements are therefore the following:

---

[51]http://www.lscom.org/
[52]http://vocab.linkeddata.es/lscom/

- agent: human, software or other agent actively causing changes or transformations

- activity: description of activity, change or transformation that takes place

- object: entity that is changed, transformed or is object of activity

- datetime: time when activity takes place

The PROV-O model reviewed in the Section 2.2.8.2 fulfills this requirement.

### 2.3.2.8   Requirement for Entity Information Modeling

The model must be able to accommodate different entities that describe what is being identified on the media document. For example, persons detected in the video, such as politicians, athletes, artists... or places where the action is happening, or organizations involved in the plot of the facts being told. Our model should therefore use other standard and well-known vocabularies able to deal with such entities. The NERD ontology reviewed in Section 2.2.8, offering a top level classification of such entities harmonized among many different entity vocabularies, together with others like FOAF for persons or DBpedia ontology should be considered.

## 2.3.3   Intellectual Property Requirements

IP requirements have a huge importance in today's data ecosystem because they address critical legal and ethical issues related with the media being broadcasted. Even they will not target them as primary objective in this thesis because they require a deeper analysis in available legislation, we would like to emphasis their necessity and enumerate the most important ones:

- Support for rights management: type of copyrights, licensing terms, identification of copyright holder...

- Including certified provenance information: extra level of assessment on the source from which the original media was obtained and possible compliance checking with respect to licensing terms

- Dealing with personal information: preferences and user behavior collected during the interaction with the data should be managed carefully in order to maintain privacy and avoid unauthorized usage.

## 2.4 Specification of the LinkedTV Ontology

After having surveyed the numerous multimedia metadata models proposed by various communities and industries in Chapter 2.2 and derive a set of requirements for the model in the Chapter 2.3, we present in this chapter our proposed ontology. This ontology makes use of several widely used vocabularies, defining new items (classes and properties) only when necessary for ensuring reusability. This ontology has been developed under the scope of the LinkedTV project [53], hence its name. The LinkedTV ontology is available at http://data.linkedtv.eu/ontologies/core/.

We first describe this model in the Section 2.4.1. Then, we describe two LinkedTV scenarios in order to illustrate how the different legacy metadata, automatic analysis results and semantic annotations are converted into RDF using this model (Section 2.4.2 and Section 2.4.3).

### 2.4.1 Description of the LinkedTV Ontology

The following vocabularies have been selected as a basis for the LinkedTV ontology:

- BBC Program ontology for representing broadcast related metadata: series, episodes, brands, categories, subtitles, physical channel, audio format, video compression, etc.

- Ontology for Media Resources for representing general properties about the content itself such as the title, description, format, license, etc. Also, it contains the classes for representing media items and fragments of media items (*ma:MediaResource* and *ma:MediaFragment*).

- Ninsuna ontology for explicitly describing the media fragments boundaries.

- Open Annotation ontology for linking the analysis results (spatiotemporal segments, scene segmentation, shot segmentation, ASR, etc.) with media fragments. It could also be used for representing additional information such as ratings or user preferences. Finally, it offers support for representing annotations of various types and simple tagging.

- NERD ontology for representing the general types of the named entities recognized by a Named Entity extractor.

- LSCOM ontology for representing the visual concepts detected by multimedia analysis processes.

- FOAF ontology for representing the people recognized in video frames.

---

[53]http://www.linkedtv.eu/

- PROV-O ontology for representing provenance information.

- LODE Ontology for representing events.

- DBpedia ontology: http://dbpedia.org/ontology/

- WordNet 3.0: http://semanticweb.cs.vu.nl/lod/wn30/

These ontologies have been considered together to generate the LinkedTV data model. Some of them are already importing others internally, like for example, the BBC Programmes ontology uses FOAF for the descriptions of actors and makes use of the Event ontology for modeling a broadcast as an event, or the PROV-O ontology is used by the Core Annotation ontology to describe who has created an annotation and when this annotation has been generated. On top of those imported classes and properties, we have created some missing ones in order to give support to the different media consumption needs considered during the work done during this thesis, like Scene, Chapters, or Shots. We have also proposed a methodology in order to conveniently represent the content metadata using the available classes and properties.

In the following sections and in order to display the possibilities of this data model, we will illustrate how the data available for the different LinkedTV scenarios can be represented using this ontology. This includes:

- Legacy Data offered by the content providers. The seed video content offered by broadcasters comes together with some basic information like title, description and air times.

- Multimedia Analysis Data. The results from applying different visual and audio analysis techniques over the multimedia content have to be incorporated into the RDF graph.

- Semantic Annotations. After performing Named Entity Recognition (see Section 3.2.1.7) on texts associated with the seed video content (generally the program subtitles), we obtain entities that are attached to the data graph.

When converting metadata in RDF, we need generate new identifiers for the first class objects of the model. According to the Linked Data principles, those identifiers must be dereferencable URIs. We follow the best practices of the Linked Data community, by minting new URIs when necessary in the http://data.linkedtv.eu domain. The first class objects in LinkedTV are addressable using the following scheme:

- http://data.linkedtv.eu/episode/UUID for the resources of type `po:Episode`

- http://data.linkedtv.eu/brand/UUID for the resources of type `po:Brand`

- http://data.linkedtv.eu/broadcast/UUID for the resources of type `po:Broadcast`
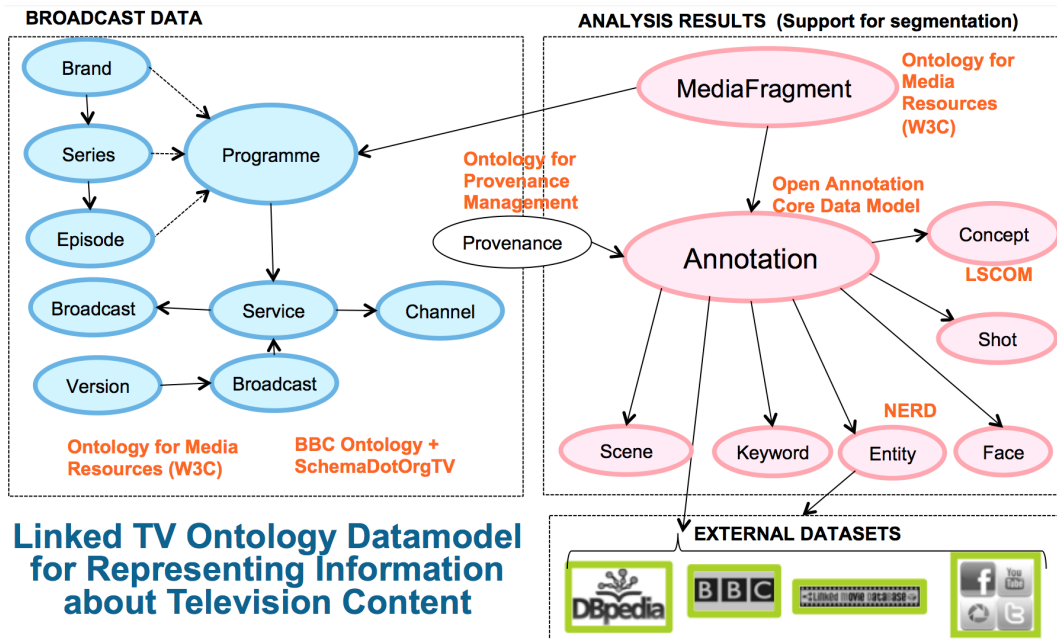
Figure 2.14: General LinkedTV metadata model

- [http://data.linkedtv.eu/version/UUID](http://data.linkedtv.eu/version/UUID) for the resources of type `po:Version`

- [http://data.linkedtv.eu/media/UUID](http://data.linkedtv.eu/media/UUID) for the resources of type `ma:MediaFragment`

- [http://data.linkedtv.eu/annotation/UUID](http://data.linkedtv.eu/annotation/UUID) for the resources of type `oa:Annotation`

- [http://data.linkedtv.eu/organization/UUID](http://data.linkedtv.eu/organization/UUID) for the resources of type `foaf:Organization`

- [http://data.linkedtv.eu/shot/UUID](http://data.linkedtv.eu/shot/UUID) for the resources of type `linked:Shot`

- [http://data.linkedtv.eu/person/UUID](http://data.linkedtv.eu/person/UUID) for the resources of type `foaf:Person`

- [http://data.linkedtv.eu/entity/UUID](http://data.linkedtv.eu/entity/UUID) for the resources of type `linkedtv:Concept` or `nerd:Concept`

- [http://data.linkedtv.eu/asr/UUID](http://data.linkedtv.eu/asr/UUID) for the resources of type `linkedtv:ASR`

Below we will illustrate through two different scenarios the adequacy of this model to accommodate different multimedia annotations in a Linked Data compliant way, that brings multimedia content to the Web at different levels of granularities and according to open and widely used domain vocabularies. There will remain one feature to be shown: the ability of representing also related content obtained via discovery and enrichment processes over the resulting annotation graph. However, those functionalities will be explained in a separate Section 4.2.5 and generated via a different Web service described in Section 4.2.3.

It is necessary to clarify that for the sake of simplicity, in the examples shown in next subsections we manually generate human friendly identifiers for all primary objects displayed. However, when those instances are generated by an automatic converter like the ones in Section 2.4.4, those identifiers will be replaced by machine readable UUID's.

### 2.4.2   Scenario 1: Cultural Heritage

In this example we analyze the video of the dutch TV program Tussen Kunst & Kitsch (Antiques Roadshow) which is offered by the public broadcaster AVRO [54]. In particular we have chosen an episode of the show from 8 December 2010 [55].

The Cultural Heritage scenario, proposed under the scope of the LinkedTV, project has been described in [97] and [96]. The general aim of this scenario is to study how the viewers of the Antiques Roadshow can have their information needs satisfied when consuming such a complex content, with so frequent mentions to historical and cultural facts. Via different non-intrusive but also more interaction demanding prototypes, it is possible to provide the user with data from external sources, such as Europeana[56], museum collections, and different encyclopedic datasets.



Figure 2.15: Ground truth metadata of automatic multimedia analysis

The legacy metadata for this program comes in the form of a spreadsheet. The automatic multimedia analysis results have been serialized in an EXMaRALDA[57] file. Finally, a In the following RDF excerpts, we show how both type of metadata is converted in RDF using the LinkedTV ontology. The general overview of the resulting conversion is depicted in the Figure 2.16.

---
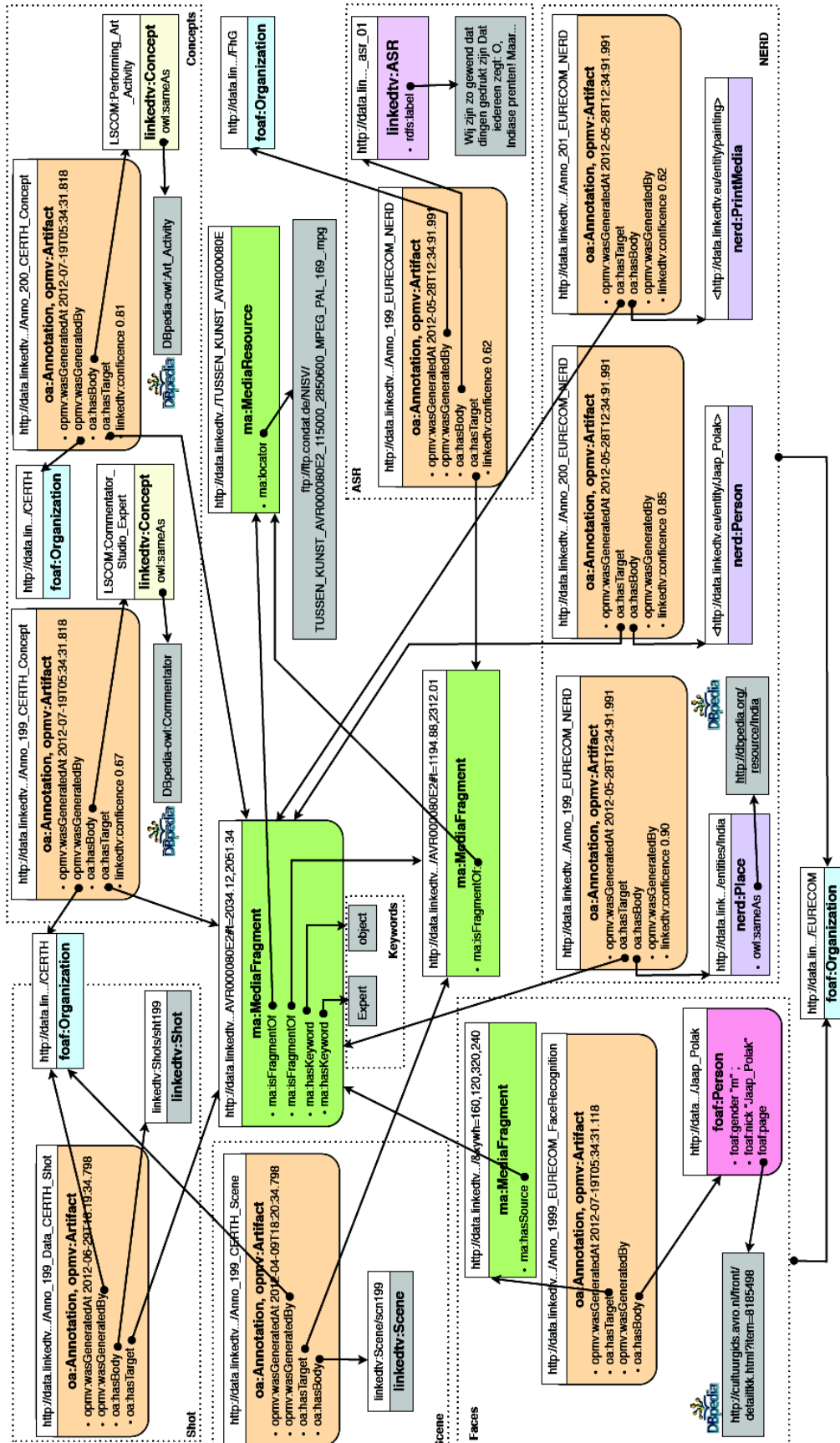
Figure 2.16: Instances involved in the Sound & Vision scenario

### 2.4.2.1   Legacy Metadata

In first place, an instance of the class `po:Episode` is created. This instance stores the title, the synopsis, the related subjects, and other basic attributes for the current material.

```
<http://data.linkedtv.eu/episode/TUSSEN_KUNST_AVR000080E2_115000_2850600>
 a po:Episode ;
 dc:title "Najaar" ;
 po:id "AVR000080E2_115000_2850600" ;
 po:microsites <http://cultuurgids.avro.nl/front/indextkk.html> ;
 po:short_synopsis "De nieuwe opnamedata en locaties van Tussen Kunst & Kitsch
     zijn weer bekend. Of je spulletjes nu waardevol zijn of niet, je mag drie
     voorwerpen meenemen naar de" ;
 po:subject "Tussen Kunst & Kitsch" , "Nelleke van der Krogt" , "Programma" ;
 po:version <http://data.linkedtv.eu/version/1_AVR000080E2_115000_2850600> .
```

At the same time, one instance of the class `po:Brand` stores information about the brand this episode belongs to.

```
<http://data.linkedtv.eu/brand/AVRO>
 a po:Brand ;
 dc:title "Algemene Vereniging Radio Omroep" ;
 po:episode <http://data.linkedtv.eu/episode/
     TUSSEN_KUNST_AVR000080E2_115000_2850600> ;
 po:microsites <http://avro.nl/> .
```

One instance of the class `po:Broadcast` establishes the relationship between a particular version of a program and the `po:Service` instance where this version is broadcasted on.

```
<http://data.linkedtv.eu/broadcast/1_7ffdb885-fcf4-44cd-80a7-7c137c8d457a>
 a po:Broadcast ;
 po:broadcast_of <http://data.linkedtv.eu/version/1_AVR000080E2_115000_2850600> ;
 po:broadcast_on <ftp://ftp.condat.de/NISV/> .
```

An instance of the class `po:Version` represents the appearance of a program at a particular date and hour and in a particular format.

```
<http://data.linkedtv.eu/version/1_AVR000080E2_115000_2850600>
 a po:Version ;
 po:aspect_ratio "urn:ard:tva:metadata:cs:ARDFormatCS:2008:90.3" ;
 po:time [ a event:Interval ;
 event:end "2010-12-08T20:35:23"^^xsd:dateTime ;
 event:start "2010-12-08T21:20:48"^^xsd:dateTime ];
 linkedtv:hasMediaResource <http://data.linkedtv.eu/media/
     TUSSEN_KUNST_AVR000080E2> .
```

### 2.4.2.2   Multimedia Analysis Metadata

This video program has been completely processed by the LinkedTV multimedia analysis tool chain, yielding numerous metadata results serialized in the EXMaR-ALDA file. In the following, we show how each layer composing the EXMaRALDA file are converted in RDF using the LinkedTV ontology.

First, we create an instance of the class `ma:MediaResource` that represents the particular media item and links it with its physical location in the LinkedTV platform.

```
<http://data.linkedtv.eu/media/TUSSEN_KUNST_AVR000080E2>
  a ma:MediaResource ;
  ma:locator <ftp://ftp.condat.de/NISV/
      TUSSEN_KUNST_AVR000080E2_115000_2850600_MPEG_PAL_169_.mpg> .
```

Instances of the class `ma:MediaFragment` represent the different spatio-temporal fragments that belong to a particular media resource. These media fragments could be related to other media fragments in a containment relationship (e.g. a scene contains shots). Keywords are also stored when the analyzed media fragment corresponds to a shot. The media fragments boundaries are explicitly serialized using the Ninsuna ontology.

```
<http://data.linkedtv.eu/media/TUSSEN_KUNST_AVR000080E2#t=2034.12,2051.34>
  a ma:MediaFragment, nsa:TemporalFragment ;
  ma:hasKeyword
    [ a  linkedtv:keyword ;
      rdf:label "Expert"
    ]:
    [ a  linkedtv:keyword ;
      rdf:label "Object"
    ];
  ma:isFragmentOf <http://data.linkedtv.eu/media/TUSSEN_KUNST_AVR000080E2> ;
  nsa:temporalStart 2034^^xsd:int;
  nsa:temporalEnd 2051^^xsd:int.
```

Instances of the class `oa:Annotation` attach the analysis results obtained from the different automatic processing tools. In this example, we can see an annotation that corresponds to a shot detected by the visual analysis algorithms in the media. The body of the annotation is an instance of a `linkedtv:Shot`, and the target is the media fragment this shot is related to. Provenance information is also included in this class through the use of the properties `opmv:wasGeneratedAt` and `opmv:wasGeneratedBy`.

```
<http://data.linkedtv.eu/annotation/Anno_199_CERTH_Shot>
  a oa:Annotation , opmv:Artifact ;
  opmv:wasGeneratedAt "2012-06-29T18:19:34.798Z"^^xsd:dateTime ;
  opmv:wasGeneratedBy
    [ a opmv:Process ;
    opmv:wasPerformedBy <http://data.linkedtv.eu/organization/CERTH>
    ];
  oa:hasTarget <http://data.linkedtv.eu/media/TUSSEN_KUNST_AVR000080E2#t
      =2034.12,2051.34> ;
  oa:hasBody <http://data.linkedtv.eu/shot/sht53> .
```

The instance of the class `linkedtv:Shot` that is being referred in the previous annotation is explicitly typed as follows.

```
<http://data.linkedtv.eu/shot/sht53>
  a linkedtv:Shot .
```

Shots are not the only temporal units that the video content can be divided into. For example, instances of the class `oa:Annotation` can relate a particular media fragments with instances of the class `linkedtv:Scene` recognized by specialized visual techniques.

```
<http://data.linkedtv.eu/scene/scn199>
  a linkedtv:Scene .
```

Instances of the class `oa:Annotation` can also help to associate LSCOM concepts detected on the media with the temporal fragment where the concept is being depicted on, including a level of confidence serialized using the `linkedtv:confidence` property.

```
<http://data.linkedtv.eu/annotation/Anno_199_CERTH_Concept>
  a oa:Annotation , opmv:Artifact ;
  opmv:wasGeneratedAt "2012-06-29T18:19:35.153Z"^^xsd:dateTime ;
  opmv:wasGeneratedBy
    [ a opmv:Process ;
      opmv:wasPerformedBy <http://data.linkedtv.eu/organization/CERTH>
    ];
  linkedtv:confidence "0.67"^^xsd:float ;
  oa:hasTarget <http://data.linkedtv.eu/media/TUSSEN_KUNST_AVR000080E2#t
      =2034.12,2051.34> ;
  oa:hasBody <lscom:Commentator_Studio_Expert> .
```

The instance `lscom:Commentator_Studio_Expert` that is being linked through the previous oa:annotation instance is also annotated as a `linkedtv:Concept` class.

```
<lscom:Commentator_Studio_Expert>
  a linkedtv:Concept .
```

Instances of the class `oa:Annotation` also relate particular media fragments with a face recognition results performed by EURECOM.

```
<http://data.linkedtv.eu/annotation/Anno_199_EURECOM_FaceRecognition>
  a oa:Annotation , opmv:Artifact ;
  opmv:wasGeneratedAt "2012-06-29T18:19:35.153Z"^^xsd:dateTime ; opmv:
      wasGeneratedBy
  [ a opmv:Process ;
    opmv:wasPerformedBy <http://data.linkedtv.eu/organization/EURECOM>
  ];
  oa:hasTarget <http://data.linkedtv.eu/media/TUSSEN_KUNST_AVR000080E2#t=2045&xywh
      =144,112,300,250> ;
  oa:hasBody <http://data.linkedtv.eu/person/5f1c5480-bdac-4c7a-881b-4e28476fd2f12
      > .
```

The instance of the class `foaf:Person` that is being referred in the previous annotation is also an instance of a `linkedtv:Person` that contains other attributes offering more information about the detected persona.

```
<http://data.linkedtv.eu/person/5f1c5480-bdac-4c7a-881b-4e28476fd2f12>
    a foaf:Person ;
  foaf:gender "m" ;
  foaf:nick "Jaap Polak" ;
```

```
   foaf:page <http://cultuurgids.avro.nl/front/detailtkk.html?item=8185498> .
<http://cultuurgids.avro.nl/front/detailtkk.html?item=8185498>
    a foaf:Document .
```

### 2.4.2.3  Semantic Annotations

Instances of the class `oa:Annotation` may also relate a particular media fragments with entities recognized in them. In this case, there are two target being annotated: the text block inside the subtitles where the entity has been spotted, and the corresponding temporal window when this subtitle block appears, as it is shown in the following RDF except under property `oa:hasTarget`.

```
<http://data.linkedtv.eu/annotation/c3a06786-6b69-4bd2-9a03-43cd663dbb7f>
        a                    prov:Entity , oa:Annotation ;
        oa:hasBody           <http://data.linkedtv.eu/entity/5f1c5480-bdac-4c7a
            -881b-4e2095fd2f60> ;
        oa:hasTarget         <http://data.linkedtv.eu/text/99335254-de81-42ec-806
            b-8b36f1b6d3c6#offset_2266_2270_Amsterdam> , <http://data.linkedtv.eu/
            media/2431e124-798a-11e5-b024-005056a40191#t=279.041,282.077> ;
        prov:startedAtTime   "2015-11-11T08:10:28.037Z"^^xsd:dateTime ;
        prov:wasAttributedTo <http://data.linkedtv.eu/organization/NERD> ;
        prov:wasDerivedFrom  <http://data.linkedtv.eu/text/99335254-de81-42ec-806
            b-8b36f1b6d3c6> .
```

The textual resource containing the surface form of the entity is serialized as a string offset (`str:OffsetBasedString`) according to the NIF 2.0 Core Ontology [58], specifying the char position inside the text where that textual fragment is appearing. At the same time, this surface form is part of a bigger textual unit, the subtitle block that is also represented in the RDF graph via the class `str:String`.

```
<http://data.linkedtv.eu/text/99335254-de81-42ec-806b-8b36f1b6d3c6#
    offset_2266_2270_Amsterdam>
        a               str:OffsetBasedString ;
        str:beginIndex  "2266"^^xsd:long ;
        str:endIndex    "2270"^^xsd:long ;
        str:subString   <http://data.linkedtv.eu/text/99335254-de81-42ec-806b-8
            b36f1b6d3c6> .

<http://data.linkedtv.eu/text/99335254-de81-42ec-806b-8b36f1b6d3c6>
        a          prov:Entity , str:String ;
        str:label  " Deze kocht ik op de veiling in Amsterdam. 15 jaar geleden.
            "^^xsd:string .
```

Finally, the instance of the class `linkedtv:Entity` that is being attached to the previous annotation has been typed as a `nerd:Location`, and disambiguated with a DBpedia resource (`http://dbpedia.org/resource/Amsterdam`).

```
<http://data.linkedtv.eu/entity/5f1c5480-bdac-4c7a-881b-4e2095fd2f60>
        a                     nerd:Location , linkedtv:Entity ;
        rdfs:label            "Amsterdam" ;
        linkedtv:hasConfidence "0.0637705"^^xsd:float ;
```

---

[58]http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html#

```
linkedtv:hasRelevance    "0.0730215"^^xsd:float ;
dc:identifier            "16564084" ;
dc:source                "nerdml" ;
owl:sameAs               <http://dbpedia.org/resource/Amsterdam> .
```

### 2.4.3   Scenario 2: News Items

RBB is the public broadcaster for the area of Berlin and Brandenburg in Germany. The idea behind the LinkedTV RBB's scenario is to enrich the local news program according to the needs and interests of the viewers. In some cases this requires to just highlight the important entities on daily news shows, in other cases the viewer may prefer to dig deeper into agents playing a role in the news stories, because there are missing details that need to be clarified, or he/she would like to be better informed about some specific aspects of the facts being depicted.

In order to illustrate this use case we have chosen as seed video content a number of episodes of its daily local news program "RBB Aktuell". This show is broadcasted four times a day, but we will focus on the night air (at 21:45). On the one hand, RBB provides legacy data in the form of TV-Anytime files. As in previous use case, LinkedTV has processed RBB videos in order to generate EXMaRALDA files containing visual analysis results. In particular the RDF excerpts used in this section correspond to the legacy information of the episode "Erlebe Deine Stadt" from the show "RBB Aktuell" broadcasted on 15 November 2011, 21:45h . The general overview of the resulting conversion is depicted in the Figure 2.17.
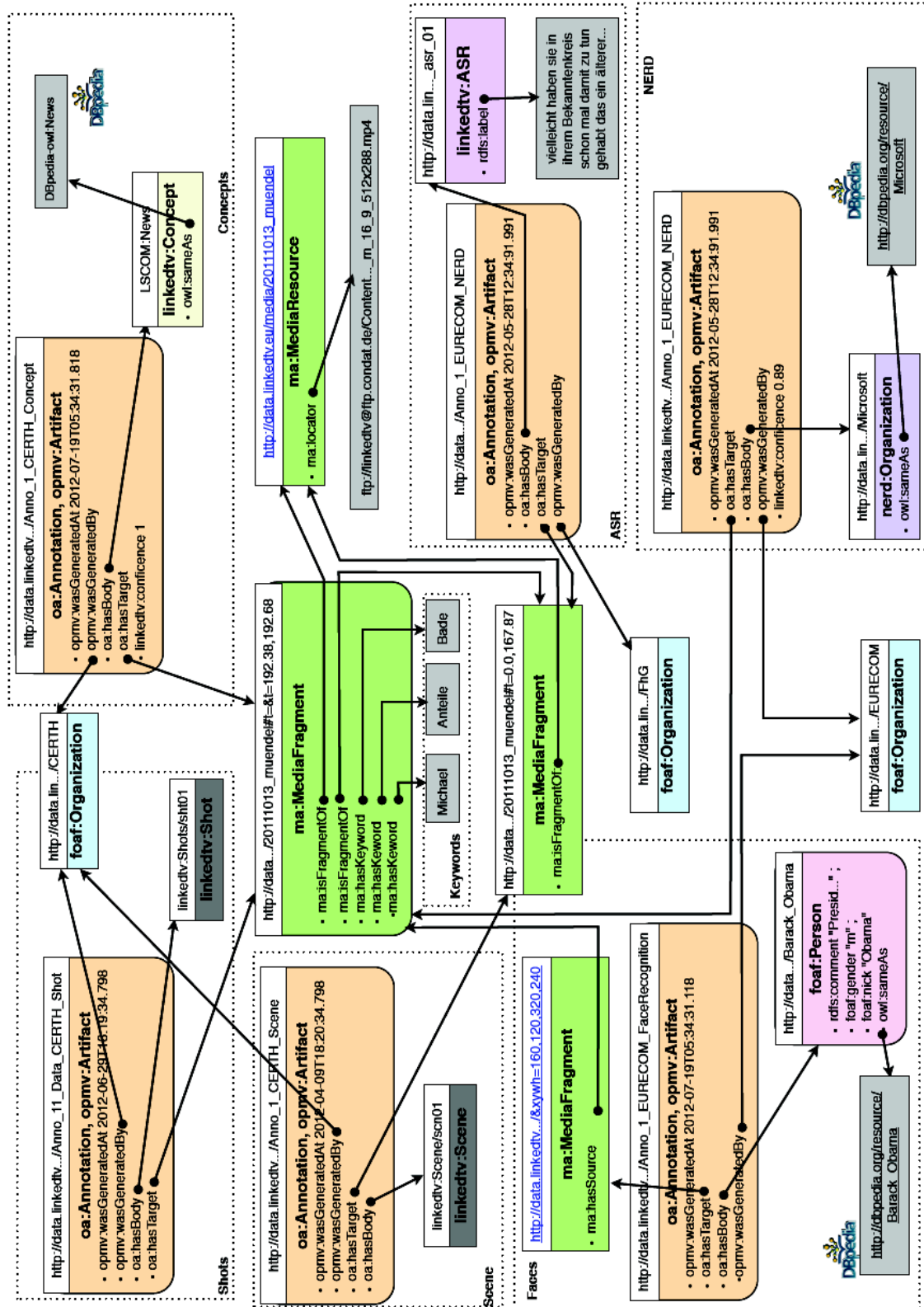
Figure 2.17: Instances involved in the RBB scenario

### 2.4.3.1   Legacy Metadata

The legacy files from RBB are expressed in a TV-Anytime format, which is converted to RDF according to the BBC Program Ontology. The instances created are described in more detail below.

There will be an instance of the class `po:Episode` that stores the title, the synopsis, the related subjects, and other basic attributes for the current material. Also, this individual has references to the different versions of the episode through the use of the `po:version` property.

```
<http://data.linkedtv.eu/episode/Erlebe_7ffdb885-fcf4-44cd-80a7-7c137c8d457a>
 a po:Episode ;
 dc:title "Hotel-Aktion \"Erlebe Deine Stadt\"" ;
 po:id "crid://rbb-online.de/rbbaktuell/7ffdb885-fcf4-44cd-80a7-7c137c8d457a" ;
 po:long_synopsis "Zu Jahresbeginn machen viele Hotels ein Berlin-Wochenende aus
     Touristensicht mglich: F?r 99 Euro knnen Berliner zu zweit in einem
     ausgewhlten Haus ?bernachten. Studiogast: Burkhard Kieker, visitBerlin. (
     Beitrag von Arndt Breitfeld)" ;
 po:masterbrand "Rundfunk Berlin Brandenburg" ;
 po:microsites <crid://ard.de/bewertbar> , <crid://rbb-online.de/rbbaktuell/0
     a566f0d- 27f4-9648-adf5-03a0cabf365a> ;
 po:short_synopsis "Zu Jahresbeginn machen viele Hotels ein Berlin-Wochenende aus
      Touristensicht mglich: F?r 99 Euro knnen Berliner zu zweit in einem
     ausgewhlten Haus ?bernachten. Studiogast: Burkhard Kieker, visitBerlin. (
     Beitrag von Arndt Breitfeld)" ;
 po:subject "Information" , "Politik" , "Kulturtipps" , "rbb AKTUELL" , "Neue
     Bundeslnder" , "Rundfunk Berlin-Brandenburg" , "Brandenburg" , "rbb online"
      , "rbb" , "Regionales" , "rbb Fernsehen" , "Berlin" , "TV" , "Nachrichten"
      ;
 po:version <http://data.linkedtv.eu/version/2_7ffdb885-fcf4-44cd-80a7-7
     c137c8d457a> , <http://data.linkedtv.eu/version/1_7ffdb885-fcf4-44cd-80a7-7
     c137c8d457a> , <http://data.linkedtv.eu/version/0_7ffdb885-fcf4-44cd-80a7-7
     c137c8d457a> .
```

An instance of the class `po:Brand` that stores information about the brand this episode belongs to, in this case "rbb AKTUELL".

```
<http://data.linkedtv.eu/brand/rbb_AKTUELL_0a566f0d-27f4-9648-adf5-03a0cabf365a>
  a po:Brand ;
  dc:title "rbb AKTUELL" ;
  po:episode <http://data.linkedtv.eu/episode/Erlebe_7ffdb885-fcf4-44cd-80a7-7
     c137c8d457a> ;
  po:id "crid://rbb-online.de/rbbaktuell/0a566f0d-27f4-9648-adf5-03a0cabf365a" ;
  po:microsites <crid://ard.de/sendung> , <crid://rbb-online.de/rbbaktuell> .
```

Instances of the class `po:Broadcast` establish a relationship between a particular version of a program and the `po:Service` instance where this version is broadcasted on.

```
<http://data.linkedtv.eu/broadcast/7ffdb885-fcf4-44cd-80a7-7c137c8d457a>
  a po:Broadcast ;
  po:broadcast_of <http://data.linkedtv.eu/version/7ffdb885-fcf4-44cd-80a7-7
     c137c8d457a> ;
  po:broadcast_on <http://data.linkedtv.eu/brand/rbb_AKTUELL_0a566f0d-27f4-9648-
     adf5- 03a0cabf365a> .
```

```
<http://data.linkedtv.eu/broadcast/7ffdb885-fcf4-44cd-80a7-7c137c8d457b>
  a po:Broadcast ;
  po:broadcast_of <http://data.linkedtv.eu/version/7ffdb885-fcf4-44cd-80a7-7
      c137c8d457b> ;
  po:broadcast_on <rtmp://ondemand.rbb-online.de/ondemand/mp4> .

<http://data.linkedtv.eu/broadcast/7ffdb885-fcf4-44cd-80a7-7c137c8d457c>
  a po:Broadcast ;
  po:broadcast_of <http://data.linkedtv.eu/version/7ffdb885-fcf4-44cd-80a7-7
      c137c8d457c> ;
  po:broadcast_on <ftp://linkedtv@ftp.condat.de/rbb/rbbaktuell/> .
```

Instances of the class po:Service are generated in order to specify the transmission channels used by broadcaster for making the content available, in this case an FTP server and a RTMP streaming service.

```
<ftp://linkedtv@ftp.condat.de/rbb/rbbaktuell/>
  a po:Service .
<rtmp://ondemand.rbb-online.de/ondemand/mp4>
  a po:Service .
```

Finally the instances of the class po:Version represent the appearance of a program at a particular date and hour and in a certain multimedia format.

```
<http://data.linkedtv.eu/version/7ffdb885-fcf4-44cd-80a7-7c137c8d457a>
  a po:Version ;
  po:time
    [ a event:Interval ;
      event:start "Tue Nov 15 22:45:00 CET 2011"^^xsd:dateTime
    ];
  linkedtv:hasMediaResource <http://data.linkedtv.eu/media/20111013_muendel> .

<http://data.linkedtv.eu/version/7ffdb885-fcf4-44cd-80a7-7c137c8d457b>
  a po:Version ;
  po:aspect_ratio "urn:ard:tva:metadata:cs:ARDFormatCS:2008:1.24" ;
  po:time
    [ a event:Interval ;
      event:end "2011-11-23T00:00:00"^^xsd:dateTime ;
      event:start "2011-11-15T21:45:00"^^xsd:dateTime
    ];
  linkedtv:hasMediaResource <http://data.linkedtv.eu/media/20111013_muendel> .

<http://data.linkedtv.eu/version/7ffdb885-fcf4-44cd-80a7-7c137c8d457c>
    a po:Version ;
  po:aspect_ratio "urn:ard:tva:metadata:cs:ARDFormatCS:2008:90.3" ;
  po:time
    [ a event:Interval ;
      event:end "2011-11-23T00:00:00"^^xsd:dateTime ;
      event:start "2011-11-15T21:45:00"^^xsd:dateTime
    ];
  linkedtv:hasMediaResource <http://data.linkedtv.eu/media/20111013_muendel> .
```

### 2.4.3.2   Multimedia Analysis Metadata

This "RBB Aktuel" episode has been completely processed by the LinkedTV multimedia analysis tool chain, yielding numerous metadata results serialized in the EXMaRALDA file. In the following examples, we showcase how each layer composing the EXMaRALDA file is converted in RDF using the LinkedTV ontology. First, we create one instance of the class `ma:MediaResource` that represents the particular media item and links it with its physical location.

```
<http://data.linkedtv.eu/media/20111013_muendel>
  a ma:MediaResource ;
  ma:locator<ftp://linkedtv@ftp.condat.de/Content\%20RBB\%20News\%20Scenario/
      RBB_AKTUELL_WEB_15_11_11/ kontraste_20111013_muendel_m_16_9_512x288.mp4> .
```

Instances of the class `ma:MediaFragment` represent the different spatio-temporal fragments that belong to a particular media resource. These media fragments store also keywords when they correspond to instances of the class `linkedtv:Shot`.

```
<http://data.linkedtv.eu/media/20111013_muendel#t=80,83>
  a ma:MediaFragment ;
  ma:hasKeyword
    [ a linkedtv:keyword ;
      rdf:label "Michael"
    ];
  ma:hasKeyword
    [ a linkedtv:keyword ;
      rdf:label "Baden"
    ];
  ma:hasKeyword
    [ a linkedtv:keyword ;
      rdf:label "Anteile"
    ];
  ma:isFragmentOf <http://data.linkedtv.eu/media/20111013_muendel> ;
  ma:isFragmentOf <http://data.linkedtv.eu/media/20111013_muendel#t=0.0,167.87> .
```

Instances of the class `oa:Annotation` are used to connect the analysis results obtained from the different automatic processing tools to the corresponding media fragments where those annotations are relevant. In this example, we can see an annotation indicating that a particular media fragment is indeed a shot detected by LinkedTV tools in the video content. Provenance information is also included in this instance through the use of the properties `opmv:wasGeneratedAt` and `opmv:wasGeneratedBy`.

```
<http://data.linkedtv.eu/annotation/Anno_1_CERTH_Shot>
  a oa:Annotation , opmv:Artifact ;
  opmv:wasGeneratedAt "2012-06-29T18:19:34.798Z"^^xsd:dateTime ;
  opmv:wasGeneratedBy
    [ a opmv:Process ;
      opmv:wasPerformedBy <http://data.linkedtv.eu/organization/CERTH>
    ];
  oa:hasTarget <http://data.linkedtv.eu/media/20111013_muendel#t=80,83> ;
  oa:hasBody <http://data.linkedtv.eu/shot/sht01> .
```

The instance of the class `linkedtv:Shot` that is being referred in the previous annotation is explicitly serialized like follows:

```
<http://data.linkedtv.eu/shot/sht01>
  a linkedtv:Shot .
```

Same occurs in the case of fragments with higher level of granularity, in this case the `linkedtv:Scene` instances that are also detected by LinkedTV analysis tools.

```
<http://data.linkedtv.eu/scene/scn01>
  a linkedtv:Scene .
```

Instances of the class `oa:Annotation` can also link to LSCOM concepts detected in the media with a level of confidence represented by the `linkedtv:confidence` property. The instance `lscom:News` that is being used as body of the annotation is also serialized as `linkedtv:Concept` class for being easily identifiable inside the model.

```
<http://data.linkedtv.eu/annotation/Anno_1_CERTH_Concept>
  a oa:Annotation , opmv:Artifact ;
  opmv:wasGeneratedAt "2012-06-29T18:19:35.153Z"^^xsd:dateTime ;
  opmv:wasGeneratedBy
    [ a opmv:Process ;
      opmv:wasPerformedBy <http://data.linkedtv.eu/organization/CERTH>
    ];
  linkedtv:confidence "1.0"^^xsd:float ;
  oa:hasTarget <http://data.linkedtv.eu/media/20111013_muendel#t=80,83> ;
  oa:hasBody <lscom:News> .

  <lscom:News>  a linkedtv:Concept .
```

Face recognition results are incorporated to the RDF model in the same way, in this case linked to spatial media fragment indicating the region inside the video where that person has been detected and recognized.

```
<http://data.linkedtv.eu/annotation/Anno_1_EURECOM_FaceRecognition>
  a oa:Annotation , opmv:Artifact ;
  opmv:wasGeneratedAt "2012-06-29T18:19:35.153Z"^^xsd:dateTime ;
  opmv:wasGeneratedBy
    [ a opmv:Process ;
      opmv:wasPerformedBy <http://data.linkedtv.eu/organization/EURECOM>
    ];
  oa:hasTarget <http://data.linkedtv.eu/media/20111013_muendel#t=80&xywh
      =160,120,320,240> ;
  oa:hasBody <http://data.linkedtv.eu/person/person98032> .
```

### 2.4.3.3 Semantic Annotations

We will proceed here in the same way that Named Entity results were converted to RDF in the cultural heritage scenario presented in previous section. Via instances of the class `oa:Annotation` we relate the text block inside the subtitles where the entity has been spotted, and the corresponding temporal window when this subtitle

block appears, with the spotted named entity serialized via a linkedtv:Entity
class which is attached through the property oa:hasTarget.

```
<http://data.linkedtv.eu/annotation/1a10a34a-ac56-497f-a214-943072fa04f5>
        a                     prov:Entity , oa:Annotation ;
        oa:hasBody            <http://data.linkedtv.eu/entity/87ab666e-9f9d-401a
            -90a7-576d27cf8ecc> ;
        oa:hasTarget          <http://data.linkedtv.eu/text/edee1bdc-0f89-4fb0-9
            ab1-0c9ce94d2e27#offset_2588_2592_Alzheimer> , <http://data.linkedtv.
            eu/media/43ccc924-36b2-11e5-b040-005056a40191#t=1389.6,1393.491> ;
        prov:startedAtTime    "2015-08-04T08:03:17.614Z"^^xsd:dateTime ;
        prov:wasAttributedTo  <http://data.linkedtv.eu/organization/NERD> ;
        prov:wasDerivedFrom   <http://data.linkedtv.eu/text/edee1bdc-0f89-4fb0-9
            ab1-0c9ce94d2e27> .
```

The textual resource containing the surface form of the entity is serialized as a
string offset (str:OffsetBasedString) according to the NIF 2.0 Core Ontol-
ogy [59], specifying the starting and ending char position inside the text where that
textual fragment is appearing. At the same time, this surface form is part of a bigger
textual unit, the subtitle block that is also represented in the RDF graph via the
class str:String.

```
<http://data.linkedtv.eu/text/edee1bdc-0f89-4fb0-9ab1-0c9ce94d2e27#
    offset_2588_2592_Alzheimer>
        a               str:OffsetBasedString ;
        str:beginIndex  "2588"^^xsd:long ;
        str:endIndex    "2592"^^xsd:long ;
        str:subString   <http://data.linkedtv.eu/text/edee1bdc-0f89-4fb0-9ab1-0
            c9ce94d2e27> .

<http://data.linkedtv.eu/text/edee1bdc-0f89-4fb0-9ab1-0c9ce94d2e27>
        a          prov:Entity , str:String ;
        str:label  "vielleicht haben sie in ihrem Bekanntenkreis schon mal damit
            zu tun gehabt dass ein Łlterer Mensch GeschŁfts unfŁhig wird etwa auf
            Grund einer schweren Krankheit wie Alzheimer oder Demenz"^^xsd:string
            .
```

Finally, the instance of the class linkedtv:Entity that is being referred in the
previous annotation has been typed as a nerd:Thing, and disambiguated with a
DBpedia resource (http://dbpedia.org/resource/Alois_Alzheimer).

```
<http://data.linkedtv.eu/entity/87ab666e-9f9d-401a-90a7-576d27cf8ecc>
        a                     nerd:Thing , linkedtv:Entity ;
        rdfs:label            "Alzheimer" ;
        linkedtv:hasConfidence "0.0637705"^^xsd:float ;
        linkedtv:hasRelevance  "0.0730215"^^xsd:float ;
        dc:identifier         "16564084" ;
        dc:source             "nerdml" ;
        owl:sameAs            <http://dbpedia.org/resource/Alois_Alzheimerr> .
```

---

[59] http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html#

### 2.4.4 TV2RDF REST Service

The LinkedTV model proposed in previous section is already available and documented at http://data.linkedtv.eu/ontologies/core/, for everybody interested in analyzing and exploiting it. But in order to ease and promote the adoption of this model and all the related Semantic Web technologies and principles that this ontology promotes, we have developed the conversion tool *TV2RDF*[60]. This tool is offered as a REST API Web service that serializes the information about a certain multimedia content into RDF according to the LinkedTV core ontology. This tool has been integrated inside the LinkedTV processing workflow where it has been further tested over different video content from the associated partners S&V and RBB.

In a nutshell, this service takes as input the identifier (UUID[61]) of a `MediaResource` and its corresponding data files, and produces a RDF representation of the provided information by using the classes and properties considered in the LinkedTV core ontology. The resulting serialization includes mainly (1) legacy information from the content providers, (2) subtitles and extracted name entities via the NERD framework [149, 150, 151, 152], and (3) data obtained after the execution of certain analysis techniques like shot segmentation, concept detection, or face recognition as serialized in an EXMaRALDA format, coming from LinkedTV workflow execution.

#### 2.4.4.1 Implementation

This service has been developed considering media resources as the main citizen of the media conversion. Consequently, at the data storage level, we maintain the list of media resources that have been created inside a particular TV2RDF instance. For each resource, there are two types of items included: *metadata files* and *serialization files*. Both kind of documents are directly uploaded/accessed via the REST API methods.

The logic for converting metadata files into LinkedTV compliant RDF documents is encapsulated inside three different core components. First, all the results from the analysis algorithms generated by LinkedTV workflow and available in EXMaRALDA format are processed inside the module *EXMaRALDA_Serialization*. Second, subtitles and the entities extracted over them are converted into RDF within the component *Subtitle_Entity_Serialization*. The format accepted for this kind of metadata files is SRT[62]. This component is in charge of invoking NERD for extracting the named entities over the transcripts. Finally, the legacy metadata information is processed in the module *Legacy_Serialization*. The information managed in this step always refer to the entire media resource, but not at the finer granularity of

---

[60]http://linkedtv.eurecom.fr/tv2rdf/api/
[61]https://en.wikipedia.org/wiki/UUID
[62]http://en.wikipedia.org/wiki/SubRip
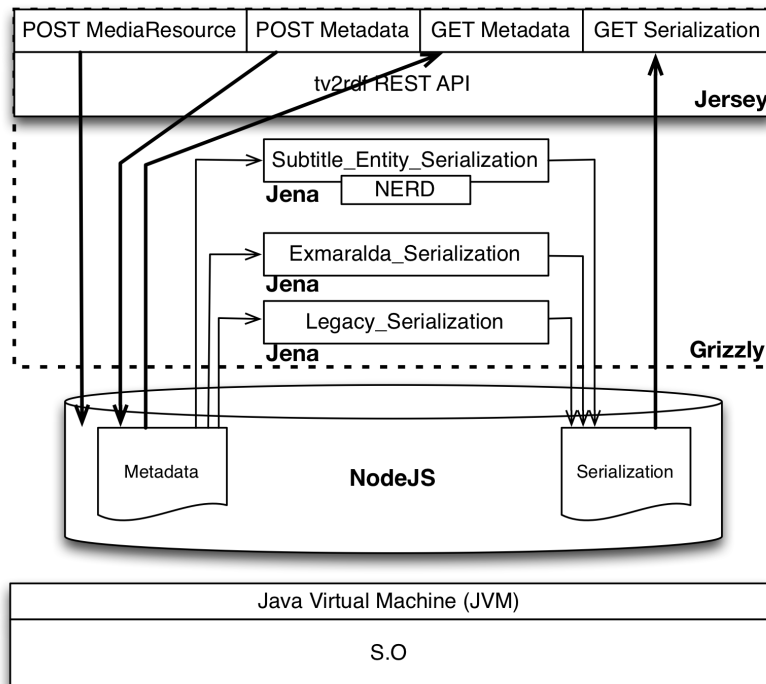
Figure 2.18: TV2RDF implementation details

fragments.

Those three core components produce separated serialization files. However, they populate a unique graph when is stored in a RDF triplestore. The Figure 2.18 details this workflow.

fsf The list of supported formats is planned to be extended in future versions of the TV2RDF service in order to broaden the scope of the service and to enable third parties to make their data available under the same ontology model.

Concerning the storage of the data, the system chosen is Mongodb[63], a cross-platform document oriented database solution classified as a "NoSQL" alternative. The reason behind this decision is that MongoDB performs really well with JSON-like files with dynamic schemas and makes the integration of data in certain types of applications (like the case of this Web Service) easier and faster. Regarding other details about the software used, we have relied on two different initiatives that make easier to write scalable server applications: Grizzly[64], which supports Java New I/O API (NIO) and manages threads in order to allow a server to scale to thousands of users. The Grizzly NIO framework has been designed to help developers to build robust servers using NIO as well as offering extended framework components like Web Framework (HTTP/S), WebSocket, or Comet. In addition, in order to build a

---

[63]http://www.mongodb.org/
[64]https://grizzly.java.net/

RESTful Web service and seamlessly expose data in a variety of representation media types while abstracting the low-level details of the client-server communication, we need a toolkit like Jersey[65]. Jersey framework is an open source, production quality framework for developing REST services in Java that provides support for JAX-RS APIs[66] and serves as a JAX-RS (JSR 311 & JSR 339).

Finally, there are two other libraries used in TV2RDF. The first one is nerd4java[67], a java library providing a programmable interface to NERD for conveniently launching named entity extractions with different parameters. The second one is Apache Jena[68], a free and open source Java framework for building Semantic Web and Linked Data applications that makes easy to create and read Resource Description Framework (RDF) graphs, work with RDFS and OWL models to add extra semantics to RDF data, and serialize triples using well-known formats such as RDF/XML or Turtle.

### 2.4.4.2 TV2RDF Integration in the LinkedTV Platform

The TV2RDF service has been integrated in the general LinkedTV workflow. In a nutshell, the REST API service interacts with three main actors taking part in the LinkedTV scenario (see Figure 2.19 for further details):

- LinkedTV Platform, for obtaining the video UUID, the locator of the media resource, and the namespace to be used for generating the instances URLs.

- Dataset containing metadata files from the providers. Those files include the subtitles and legacy metadata that need to be serialized into RDF.

- Modules containing the results from the analysis processes. The corresponding Exmaralda file generated by those modules has to be serialized as well so it is also used by TV2RDF when generating the RDF graph.

Once all those LinkedTV components have provided TV2RDF with the necessary information, the internal serialization engine reads those particular formats, generates the corresponding media fragments, annotates them at different level of granularities, and links them with resources from the Web of Data cloud via NERD's entities. The final set of triples is serialized in the Turtle format and pushed to the LinkedTV platform.

As illustrated in Figure 2.20, everything starts with the ingestion of a television programme by the LinkedTV platform. When all the associated attributes (UUID, locator, namespace) have been generated and visual analysis algorithms have finished,

---

[65]https://jersey.java.net/
[66]https://jax-rs-spec.java.net/
[67]https://github.com/giusepperizzo/nerd4java
[68]http://jena.apache.org/
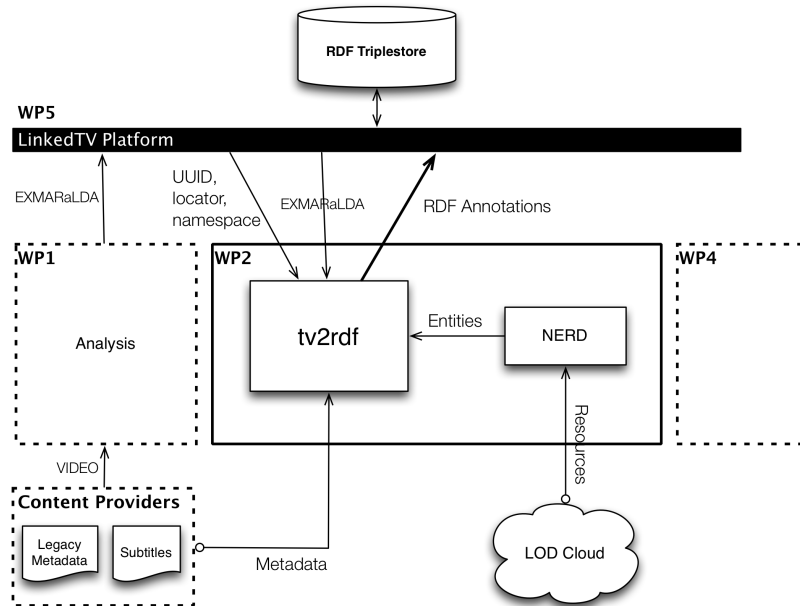
Figure 2.19: Integration of tv2rdf inside the LinkedTV workflow

the platform sends a POST request to TV2RDF in order to create the corresponding media resource. The next step consists of uploading the different metadata files associated to this particular television programme. On the TV2RDF side, this will automatically trigger the execution of the serialization processes that could imply a request to NERD in the case of processing subtitles. At the end of the execution, the results are ready to be retrieved by the LinkedTV platform, which can actually perform the three GET requests to the REST service in order to download the corresponding Turtle files. In a last step, the Turtle files are loaded into the LinkedTV triplestore within the default graph http://data.linkedtv.eu/graph/linkedtv.

### 2.4.4.3   REST API Calls Supported in TV2RDF

**Creating a Media Resource** This request creates a new media resource in TV2RDF to be serialized to RDF. It is necessary to provide the *UUID* (Universal Unique Identifier) of the media resource for uniquely identifying the television content to be processed. Optionally, it is also possible to specify the *locator* of the video (the logical address at which the resource can be accessed, a URL) and the *namespace* for building up the URI's of the instances generated during the serialization.

```
curl -X POST http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/UUID\_media\
    _resource --header "Content-Type:text/xml" -v
```

For specifying properties such as the locator and the namespace, one can use the query parameters as follows:

```
curl -X POST "http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/UUID\_media\
```
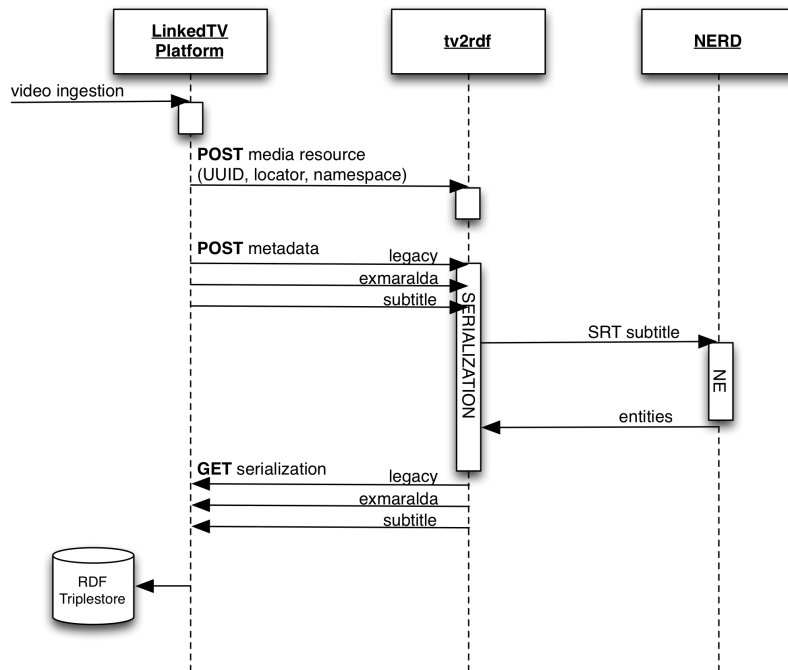
Figure 2.20: Sequence diagram of the serialization of a Media Resource by TV2RDF

```
_resource?locator=URL\_locator&namespace=http://data.linkedtv.eu/" --header "
    Content-Type:text/xml" -v
```

**Uploading Metadata Files for a Media Resource** This request allows to specify the metadata files describing a particular media resource: the legacy file, the subtitles file and the analysis results file. Immediately after the storage of the files in the server, the corresponding serialization process is automatically launched, so the RDF results will be available as soon as possible by performing one of the REST requests shown in Section 2.4.4.3. If these POST requests are executed multiple times, the files uploaded in the past are substituted by the ones specified in the current calls and the serialization processes are re-started conveniently.

**Legacy Metadata.** This request upload the file that contains the information provided by the broadcasters for a particular media resource, and launch the process of converting it into RDF. The format supported by TV2RDF is TVAnytime.

```
curl -X POST --data-binary @LEGACY_file.tva  http://linkedtv.eurecom.fr/tv2rdf/api
    /mediaresource/UUID\_media\_resource/metadata?metadataType=legacy --header "
    Content-Type:text/xml" -v
```

**Subtitles.** This request uploads the subtitle file for a particular Media Resource, launch the process of extracting Named Entities from the time text and serialize them into RDF. Up to now, the format supported by TV2RDF is SRT.

```
curl -X POST --data-binary @SUBTITLES_file.srt  http://linkedtv.eurecom.fr/tv2rdf/
    api/mediaresource/UUID\_media\_resource/metadata?metadataType=subtitle --
    header "Content-Type:text/xml" -v
```

**Analysis Results.** This request upload the file with the results from the execution of various analysis techniques over a particular media resource, in EXMaRALDA format.

```
curl -X POST --data-binary @EXMARALDA\_file.exb  http://linkedtv.eurecom.fr/tv2rdf
    /api/mediaresource/UUID\_media\_resource/metadata?metadataType=exmaralda --
    header "Content-Type:text/xml"
```

**Getting Metadata Files** It is possible to retrieve the original metadata files that were uploaded to TV2RDF for a certain media resource at any time, by performing a GET request instead of the corresponding POST calls, for each of the three data files that can be uploaded [legacy, subtitle,analysis].

```
curl -X GET  http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/UUID\_media\
    _resource/metadata?metadataType=legacy --header "Content-Type:text/xml" -v
```

```
curl -X GET  http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/UUID\_media\
    _resource/metadata?metadataType=subtitle --header "Content-Type:text/xml" -v
```

```
curl -X GET  http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/UUID\_media\
    _resource/metadata?metadataType=analysis --header "Content-Type:text/xml" -v
```

**Getting Serialization Results** This request allows to retrieve the RDF graph generated after the serialization of the uploaded metadata files. Those results will be available after the corresponding core component has finished its processing, by just performing the corresponding GET request to the TV2RDF service. Hence, a client can repeatedly make those GET calls to the REST service until the resource is available (the 404 responses turn into a 200 OK and the file is downloaded from the server). It is also possible to see if a serialization result is available by checking the status attributes of the media resource via the REST calls explained in Section 2.4.4.3.

```
curl -X GET http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/UUID\_media\
    _resource/serialization?metadataType=legacy --header "Content-Type:text/xml" -
    v
```

```
curl -X GET http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/UUID\_media\
    _resource/serialization?metadataType=subtitle --header "Content-Type:text/xml"
     -v
```

```
curl -X GET http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/UUID\_media\
    _resource/serialization?metadataType=exmaralda --header "Content-Type:text/xml
    " -v
```

**Complete Serialization.** This request allows to retrieve in a single serialization file the complete RDF graph about a certain Media Resource.

```
curl -X GET http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/UUID\_media\
    _resource/serialization --header "Content-Type:text/xml" -v
```

### Other Requests

**Get MediaResource's description.**   With this request, it is possible to obtain a JSON serialization of the data available in the TV2RDF about a particular Media Resource: id, metadata files that have been uploaded, serialization files available, and base URL used for generating the different data instances of the graph.

```
curl -X GET http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/19a73f0a-d023-49f8
    -9203-cbd721053c55 --header "Content-Type:text/xml" -v
```

**Modify MediaResource's parameters.**   If some of the parameters (locator or namespace) need to be modified, we can use the same REST call as for creating a Media Resource for the first time. The parameters will be overwritten on the server side, and the serialization processes will be automatically re-launched (if the corresponding metadata files are available).

```
curl -X POST "http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/19a73f0a-d023-49
    f8-9203-cbd721053c55?locator=http://stream6.noterik.com/progressive/stream6/
    domain/linkedtv/user/rbb/video/59/rawvideo/2/raw.mp4&namespace=http://data.
    linkedtv.eu/" --header "Content-Type:text/xml" -v
```

**Get a List of Media Resources.**   This operation returns the list of Media Resource's that have been processed inside TV2RDF REST service.

```
curl -X GET http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/list --header "
    Content-Type:text/xml" -v
```

#### 2.4.4.4   TV2RDF Example

In this section, we provide the REST API calls needed for serializing content from RBB partner in LinkedTV, corresponding with the daily news aired on 2013-06-19 in RBB Aktuell program.

We consider this content has already been pre-ingested in the LinkedTV platform: there is already an abstract UUID, synchronized with the physical locator of the corresponding video. We assume that the different analysis algorithms have already been performed, resulting in a set of EXMaRALDA files that will be used as input for TV2RDF. As depicted in the sequence diagram (Figure 2.20), the procedure can start by first creating the corresponding media resource into TV2RDF:

```
curl -X POST "http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/b82fb032-d95e-11
    e2-951c-f8bdfd0abfbd?locator=http://stream17.noterik.com/progressive/stream17/
    domain/linkedtv/user/rbb/video/249/&namespace=http://data.linkedtv.eu" -v --
    header "Content-Type:text/xml"
```

The next step consist of uploading the corresponding metadata files, three per
review video (subtitles, exmaralda files, and legacy files). Immediately after every
file has been uploaded to TV2RDF, the serialization starts in the background making
the results available through the REST API as soon as the corresponding modules
have finished the processing.

```
curl -X POST --data-binary @BERLIN-2013-06-19-22-00-37-06192145.srt http://
    linkedtv.eurecom.fr/tv2rdf/api/mediaresource/b82fb032-d95e-11e2-951c-
    f8bdfd0abfbd/metadata?metadataType=subtitle -v --header "Content-Type:text/xml
    "
```

```
curl -X POST --data-binary @rbbaktuell_20130619_sdg_m_16_9_512x288.exb http://
    linkedtv.eurecom.fr/tv2rdf/api/mediaresource/b82fb032-d95e-11e2-951c-
    f8bdfd0abfbd/metadata?metadataType=exmaralda -v --header "Content-Type:text/
    xml"
```

```
curl -X POST --data-binary @rbbaktuell_20130619.html http://linkedtv.eurecom.fr/
    tv2rdf/api/mediaresource/b82fb032-d95e-11e2-951c-f8bdfd0abfbd/metadata?
    metadataType=legacy -v --header "Content-Type:text/xml"
```

Once, the metadata files have been uploaded and serialized, the corresponding
Turtle files obtained as results can be accessed. In the following script, we show how
to make the GET call and store the corresponding triples in a local file for further
processing or re-loading into a RDF triplestore.

```
curl http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/b82fb032-d95e-11e2-951c-
    f8bdfd0abfbd/serialization?metadataType=subtitle >
    rbbaktuell_20130619_entities_subtitles.ttl;
```

```
curl http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/b82fb032-d95e-11e2-951c-
    f8bdfd0abfbd/serialization?metadataType=exmaralda >
    rbbaktuell_20130619_exmaralda.ttl;
```

```
//getting legacy
curl http://linkedtv.eurecom.fr/tv2rdf/api/mediaresource/b82fb032-d95e-11e2-951c-
    f8bdfd0abfbd/serialization?metadataType=legacy > rbbaktuell_20130619_legacy.
    ttl;
```

### 2.4.4.5   Future Work

The RESTful TV2RDF application is currently under improvement cycles that aim to
increase the reliability of the service and include new features that could be interesting
in a multimedia and television scenario. We enumerate below some of the main
features that are planned for the future

- More formats supported. The television ecosystem is not limited to formats
  such as TVAnytime or SRT, but considers a bigger set of metadata schemas
  that currently play a role in the audiovisual market and should be supported
  in TV2RDF.
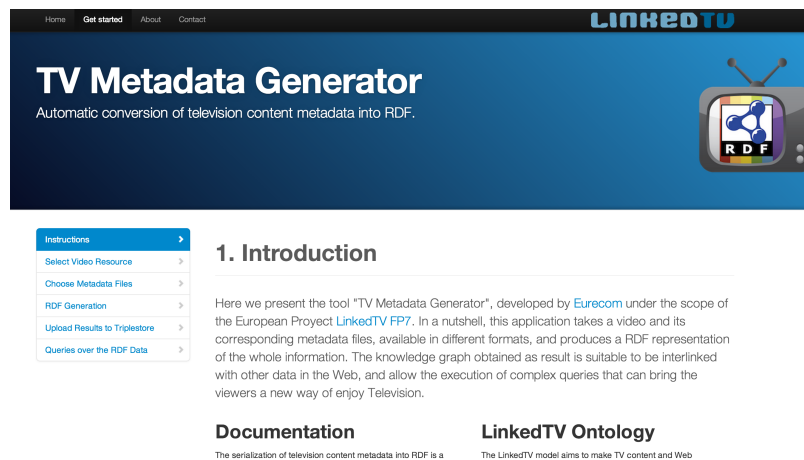
Figure 2.21: Front-end user interface for serializing media resources in TV2RDF (under development)

- User interface. TV2RDF is primarily a REST Web service but we have also developed a more human-friendly interface (see Figure 2.21).

- LinkedTV core ontology evolution. The LinkedTV core ontology needs to evolve and since it depends on a number of third-party ontologies such as the Open Annotation ontology which itself is evolving. The TV2RDF service needs to be always compliant with those latest specifications.

## 2.5   Summary

In this Chapter we have proposed an ontology model for representing multimedia content in the Web, inspired in the previous work done in the field through different formats developed by industrial and broadcasting partners. Following the Linked Data publishing principles, it creates URI's for different involved agents including fragments inside the content, and reuses vocabularies in the domain like the Ontology for Media Annotation or the Open Annotation Model.

We have illustrated the use of the LinkedTV vocabulary through two different use cases dealing with cultural heritage and news items respectively, probing the adequacy of the ontology for representing different multimedia features at different level of granularities and bringing them to the Web. In addition we have described the service TV2RDF, which helps to serialize traditional multimedia format into the LinkedTV model in an automatic way, making easy for publishers to provide their content annotated in a semantic, Web compliant way. This knowledge representation and publishing techniques open the room for new advanced annotation techniques that will be explained in next Chapter 3, as well as advanced operations for interlinking the media fragments with other relevant content as we will cover in Chapter 4. The results of this additional processed can be also represented according to this same model as we will cover in Section 4.2.5 .

# Generating Video Annotations

## 3.1 Introduction

In Chapter 2 we presented an ontology model for multimedia content that aims to bring video documents to the Web in a more flexible and better annotated way. But this model needs to be somehow populated, and given the huge amount of content being generated every minute[1], this task is unfeasible for human annotators in terms of time and efforts. Visual analysis techniques for annotating videos have been out there for some time, but unfortunately they do not take advantage of working over the Web ecosystem, therefore lacking some important features this scenario can offer, such as better interoperability through common standards and vocabularies, and easily exploration of the vast amount of additional knowledge already available to be used.

In this Chapter we introduce some semantic Web techniques that tackle this annotation process in a more automatic and autonomous way, relying not only in the information that can be found in the multimedia content itself but also other knowledge present on the Web. Those techniques will leverage on different parts of the media content, mainly: (1) the text obtained from subtitles, automatic speech recognition, and OCR. and (2) the results of visual analysis executed over the images of the video that bring different information such as: visual concepts, object recognized in images, object tracking, and face detection and recognition. In addition, we analyze the efforts made in combining both textual and visual annotations in the so called multimodal approaches, which aim to get the best from both worlds and provide the more rich and detailed information possible to work over.

## 3.2 Textual Annotations

In this first section we will describe different techniques to produce semantic annotations relying on text. Although the content we are dealing with is video and audio files, a considerable part of the information coming inside those documents can be collected an expressed in words so tools annotating text can be applied over them. In order to proceed this way, we look mainly those two video dimensions:

---

[1] http://media.fb.com/2015/01/07/what-the-shift-to-video-means-for-creators/

- **Text in Audio Track**: the ideal situation is to have access to subtitles produced by humans and normally offered by the content providers together with the video itself. In cases where no curated subtitles are available, we can rely in automatic transcription techniques like Automatic Speech Recognition described in Section 3.3.6 that are not as good as the manual alternatives, but are getting closer in quality.

- **Text in Video**: It refers to the text that appears embedded in the video, like banners, headers, or simply written text that is recorded by the camera. This information is normally not available by providers, and automatic techniques have more problems for automatically detecting what is being shown in there. However in particular situations and for particular kinds of multimedia content they can be applied so the resulting text can be exploited to produce additional annotations.

### 3.2.1   Named Entity Recognition and Disambiguation

Originally, Named Entity Recognition (NER) is an information extraction task that seeks to locate atomic elements in text. The NER and disambiguation problems have been addressed in different research fields such as NLP, Web mining and Semantic Web communities. All of them agree on the definition of a Named Entity, which was coined by Grishman et al. as an information unit described by the name of a person or an organization, a location, a brand, a product, a numeric expression including time, date, money and percent found in a sentence [64].

Initially, these NER techniques focused on identifying atomic information unit in a text (the named entities), later on classified into predefined categories (also called context types) by classification techniques, and linked to real world objects using Web identifiers. Such a task is called Named Entity Disambiguation. The NER task is strongly dependent on the knowledge base used to train the NE extraction algorithm. Leveraging on the use of DBpedia[2], Freebase[3] and YAGO[4], recent methods coming from the Semantic Web community have been introduced to map entities to relational facts exploiting these fine-grained ontologies. In addition to detect a Named Entity (NE) and its type, efforts have been spent to develop methods for disambiguating information unit with a URI. Disambiguation is one of the key challenges in this scenario and its foundation stands on the fact that terms taken in isolation are naturally ambiguous. Hence, a text containing the term London may refer to the city London in UK or to the city London in Minnesota, USA, depending on the surrounding context. Similarly, people, organizations and companies can have multiple names and

---

[2]http://wiki.dbpedia.org/
[3]https://www.freebase.com/
[4]https://www.mpi-inf.mpg.de/yago-naga/yago/

nicknames. These methods generally try to find in the surrounding text some clues for contextualizing the ambiguous term and refine its intended meaning. Therefore, a NE extraction workflow consists in analyzing some input content for detecting named entities, assigning them a type weighted by a confidence score and by providing a list of URIs for disambiguation. The problem of word sense disambiguation is defined as the task of automatically assigning the most appropriate meaning to a polysemous word within a given context. Such a word sense disambiguation facilitates more accurate information filtering and enables enhanced text browsing. In multimedia context, named entity recognition helps in retrieval of additional related content and locating related videos [10].

The named entity recognition and disambiguation process consists generally in the following steps:

- Named Entity Recognition – Identification of named entities in a given text.

- Candidate Generation – Finding possible word senses or identifiers of concrete candidate entities that can occur under the recognized surface form.

- Disambiguation – Selecting the most appropriate meaning (concrete category or resource identifier from a knowledge base) within a given context.

In the following sections, we describe the two main approaches for performing named entity recognition:

- Statistical approaches grounded in computational linguistics that often use some representation of an entity context to classify it in a predefined or open set of categories (section 3.2.1.1).

- Knowledge based approaches that aim at mapping recognized entities to concrete records in a backing knowledge base[5] (section 3.2.1.2). An advantage of such a detailed disambiguation is the possibility to enrich unstructured text with additional structured data from the knowledge base beyond just the type of an entity.

We conclude this section by describing Web APIs that offer named entities and disambiguation functionalities (Section 3.2.1.3) and a comparison of those APIs (section 3.2.1.4).

### 3.2.1.1 Statistical Approaches Grounded in Computational Linguistics

Early studies were mostly based on hand crafted rules, but most recent ones use supervised machine learning as a way to automatically induce rule-based systems or

---

[5]Such a knowledge base can include a proprietary data source like social networks for names of people or a general data source such as Wikipedia or DBpedia.

sequence labeling algorithms starting from a collection of training examples. However, when training examples are not available, even recent approaches stick with some kind of hand crafted rules often backed by a knowledge base [163]. Statistical approaches to named entity recognition can be divided into three groups: Supervised Learning Approaches, Semi-Supervised Learning Approaches and Unsupervised Learning Approaches.

**Supervised Learning** The idea of supervised learning is to study the features of positive and negative examples of named entities over a large collection of annotated documents and design (learn) rules that capture instances of a given type. Supervised machine learning techniques include Hidden Markov Models [14], Decision Trees [162], Maximum Entropy Models [18], Support Vector Machines [9] and Conditional Random Fields [53]. In [128], the LEXAS system is described as using a wide range of features that can be used to train the disambiguation algorithm. These include Part of Speech (POS) tags of surrounding words, POS tag of the disambiguated word, surrounding words in their basic form, collocations (words or phrases often co-occurring with the given sense), verb-object syntactic relations. LEXAS determines the correct meaning of the word by looking for the nearest meaning in terms of the features. In [137], bigrams occurring nearby the disambiguated word are used as features. Weka [211] implementations of the C4.5 decision tree learner, the decision stump and the Naive Bayesian classifier are used.

**Semi-Supervised Learning** As opposed to supervised learning methods, semi-supervised methods require only a limited set of examples or initial seeds in order to start the learning process. For example, the system may ask for a limited number of names of sought entities. They are then located in a text and the system tries to identify some contextual features characteristic for all the located entities. The results are then used to locate additional entities found in similar contexts. The learning process is then repeated.

In [125] a named entity extractor exploits the HTML markup of Web pages in order to locate named entities. It is reported to outperform baseline supervised approaches but it is still not competitive with more complex supervised systems. In [20] semi-supervised learning is used to extract names of books and their authors. At the beginning example pairs of `author name -- book name` are given. They are used to learn patterns that model the context of these pairs. A limited class of regular expressions is used for the patterns. Such derived patterns are then used to extract new names.

A Web scale fact extraction is performed in [135]. The recall of fact extraction is increased by pattern generalization - words from the same class are replaced by the same placeholder. The authors report a precision of about 88% by 1 million extracted facts from 100 million Web documents. In [63] a similar task of word sense disambiguation is supported by semantic resources obtained from large corpora where

terms are mapped to domains. This domain model is constructed in the completely unsupervised way using clustering based on Latent Semantic Analysis. The authors report that such a domain model contributes to better results even with limited amount of training data that are often difficult to gather.

**Unsupervised Learning** An example of unsupervised named entity recognition using WordNet is given in [1]. The aim is to assign a known concept from Word-Net to an unknown concept in a text. It is achieved by analyzing words that often co-occur with each known concept. Certain language patterns (e.g. such as, like, or other) are exploited in [47]. The Google search engine is used to locate additional hypernyms. The sets of hypernyms are then clustered in an attempt to find general types of named entities. An observation that a Named Entity is likely to appear synchronously in several news articles, whereas a common noun is less likely exploited as proposed in [164]. Authors report they successfully obtained rare Named Entities with 90% accuracy just by comparing time series distributions of a word in two newspapers. KnowItAll [46] uses the redundancy of the Web to perform a bootstrapped information extraction process.

**Discussion** Statistical-based approaches often do not disambiguate entities into many diverse categories. Hence, the standard types used are: people, locations, organizations and others. From this point of view, knowledge-based approaches are more suitable for the need of future algorithms that want to rely in ever-growing data sources available on the Web. Hence the task becomes finding unique identifiers that disambiguate named entities and obtaining additional information for these named entities. However, statistical approaches provide very good results in the process of named entity recognition in texts. The de facto standard state-of-the-art solution in this area is the Stanford Named Entity Recognizer [53] which exploits conditional random fields (CRF) models [108].

CRF belongs to the group of supervised learning algorithms and, as such, needs a comprehensive training data set. This could be an issue since in real use cases we sometimes have to deal with at least three different languages (English, German and Dutch). The authors provide models for English texts. Trained models for German can be found in [48]. Fortunately, the CoNLL 2003 shared task[6] [186] provides a comprehensive annotated corpus for various languages including Dutch.

Additional semi-supervised and unsupervised techniques can be used in later stages of the project in order to improve the named entity recognition process. The approaches to extraction of further information about named entities [20] and exploiting HTML structure of Web pages [125] can be used to enhance indexing and retrieval of additional content. This is the subject of our further evaluation.

---

[6] http://www.cnts.ua.ac.be/conll2003/ner/

### 3.2.1.2    Knowledge Based Approaches

Apart from statistical approaches to named entity recognition, the recognition and disambiguation may be supported by a knowledge base. The knowledge base serves on one hand as the white list of names that are located in a text. On the other hand, many services supported by a knowledge base assign concrete identifiers to recognized entities and thus can be mapped to additional information describing the recognized entities. Many general purpose named entity recognition tools use DBpedia [17] as their knowledge base (e.g. DBpedia Spotlight [110], Wikify [113]) or map recognized named entities directly to Wikipedia articles [21].

A big advantage of Wikipedia is that links created in articles by Wikipedia contributors can be used as manual annotations. Each link to a Wikipedia article represents a mention of an entity represented by the target article. In Figure 3.1, we show an example of links in a Wikipedia article and the representation of their anchor texts in the source of this article. We can see that the entity British Empire[7] has the same anchor text, whereas the entity American Revolutionary War[8] has the anchor text American Revolution, which is an alternative surface form for this entity.



Figure 3.1: A sample of links in a Wikipedia article together with their representation in the source of a Wikipedia article.

One important feature of an entity is its commonness [109] (i.e. prior probability of a particular sense of a given surface form). In the case of Wikipedia, this is usually measured as the count of incoming links having a given anchor text (i.e. surface form) leading to a corresponding Wikipedia article. At least, when we do not have access to any context of the entity (e.g. when we just see USA), the most common meaning of that shortcut is probably the most meaningful match. In [175], the authors claim that disambiguation based purely on the commonness of meanings outperforms some of the state of the art methods dealing with the context of entities. However, the most popular or most common meaning is not always the best match and the proper model of an entity context is very important. We can divide the approaches used for named entity disambiguation into two groups: either textual features of a context are compared in order to disambiguate a meaning, or structural relations between entities mentioned in a text are considered.

**Textual Disambiguation** Textual representation of an entity context is used in [21]. Links in Wikipedia articles are used as annotations and their surroundings

---

[7]http://en.wikipedia.org/wiki/British_Empire
[8]http://en.wikipedia.org/wiki/American_Revolutionary_War

(words within a fixed size window around the annotation) are collected and indexed. They are then compared against the context of a disambiguated entity in new texts. When the context of an entity is not sufficiently big, the taxonomy of Wikipedia categories is taken into account for the disambiguation. For comparison of textual context vectors, the cosine similarity and TF-IDF [155] weight are used.

Wikify [113] and Spotlight [110] use the textual representation of entities described in Wikipedia articles too. Wikify attempts to identify the most likely meaning for a word in a given context based on a measure of contextual overlap between the dictionary definitions of the ambiguous word – here approximated with the corresponding Wikipedia pages, and the context where the ambiguous word occurs (the current paragraph is used as a representation of the context). The approach is inspired by [94].

Spotlight represents the context of an entity in a knowledge base by the set of its mentions in individual paragraphs in Wikipedia articles. DBpedia resource occurrences are modeled in a Vector Space Model [160] where each DBpedia resource is a point in a multidimensional space of words. The representation of a DBpedia resource thus forms a meta document containing the aggregation of all paragraphs mentioning that concept in Wikipedia.

The meta document context representation of each candidate entity for an ambiguous surface form is compared to the target paragraph (containing disambiguated entity). The closest candidate in terms of cosine similarity in the vector space model is selected. For weighting individual terms, the TF-ICF weight [110] is introduced. The TF-ICF measure is an adaptation of the TF-IDF [155] measure. The only difference is that the IDF part is counted among concrete selected candidates and not over the entire knowledge base. Thus, the discriminator terms specific for the concrete candidate selection are weighted higher.

In more recent work [86], a weakly semi-supervised hierarchical topic model is used for named entity disambiguation. It leverages Wikipedia annotations to appropriately bias the assignment of entity labels to annotated words (and un-annotated words co-occurring with them). In other words the frequency of occurrence of the concrete form of the word in annotations of particular entities in Wikipedia is taken into account, when selecting the correct entity. The Wikipedia category hierarchy is leveraged to capture entity context and co-occurrence patterns in a single unified disambiguation framework.

**Structural Disambiguation** In [118], the structure of links to Wikipedia articles corresponding to disambiguated entities is analyzed. Each entity is represented by a Wikipedia article. The most similar entities to entities which are not ambiguous in the texts get higher score. The similarity [109] between two entities represented by Wikipedia articles depends on the number of Wikipedia articles that link to both of them. The score computed this way is then combined with an overall entity

commonness for a particular surface form using a C4.5 classifier.

A very similar approach to word sense disambiguation was proposed in [127]. WordNet [117] is used as the knowledge base. The disambiguation starts with non-ambiguous words in the text and searches for senses that are connected to these non-ambiguous words. The grammar for this kind of disambiguation is proposed.

A more general approach to structural disambiguation of word senses is introduced in [112]. Distance between candidate labels or senses is counted and a graph is constructed consisting of labels as vertices and distances as weights of edges. The Random Walk adaptation in the form of PageRank algorithm is used to determine scores for individual labels. For each word, its label with the best score is selected. Various representation of distance measures are proposed. For the evaluation, the definition overlap of individual label definitions in a dictionary is used. This sense similarity measure is inspired by the definition of the Lesk algorithm [94]. Word senses and definitions are obtained from the WordNet sense inventory [117].

The work presented in [118] was further improved in [90]. An annotation is scored based on two types of features: one set is local to the occurrence of the surface form of the mentioned entity and the other set of features is global to the text fragment. The annotation process is modeled as a search for the mapping that maximizes the sum of the local and global scores of the selected annotations. Experiments over a manually annotated dataset showed that the approach presented in [90] yields a precision comparable to [118] but outperforms it in terms of recall.

**Disambiguation Discussion** As the model proposed in Chapter 2 focuses on disambiguating named entities in order to retrieve additional content and to obtain background knowledge about those named entities, we will favor the approaches using Wikipedia or DBpedia [113, 110, 118, 109] since their knowledge base seem to be ideal for this purpose. Wikipedia is one of the biggest freely available knowledge bases on the Web. It is also relatively up-to-date, as new concepts (e.g. new products, celebrities, companies) appear relatively early in Wikipedia. Wikipedia is also a general knowledge base which fits into the wide variety of multimedia scenarios. URLs of Wikipedia articles can be easily translated to URIs of entities in DBpedia [17] which provides another valuable source of information about identified entities – in this case in a structured form of RDF documents. Last but not least, Wikipedia is available in the comprehensive extent in many language variations.

For this reason in this thesis we will consider mainly the combination of representative approaches from both groups – namely the approach of DBpedia Spotlight [110] for textual representation of entity context and the structural representation of entity context proposed in [118] together with overall popularity measure [109, 175]. Our preliminary experiments show that these methods do not overlap and can provide complementary results. The proposal of concrete combination of these methods and the evaluation is subject of our future work.

### 3.2.1.3 NER Web API's

Nowadays different Web APIs for named entitiy recognition and disambiguation are available online, such as: AlchemyAPI[9], DBpedia Spotlight[10], Evri[11], Extractiv[12], Lupedia[13], OpenCalais[14], Saplo[15], Wikimeta[16], Yahoo! Content Analysis (YCA)[17], TextRazor[18]and Zemanta[19].

They represent a clear opportunity for the Linked Data community to increase the volume of interconnected data. Although these tools share the same purpose, extracting semantic units from text, they make use of different algorithms and training data. They generally provide a similar output composed of a set of extracted named entities, their type and potentially a URI disambiguating each named entity. The output varies in terms of data model used by the extractors. These services have their own strengths and shortcomings but, to the best of our knowledge, few scientific evaluations have been conducted to understand the conditions under which a tool is the most appropriate one. This section attempts to fill this gap. A comparative study involving them has been published in [152].

The NE recognition and disambiguation tools vary in terms of response granularity and technology used. As granularity, we define the way how the extraction algorithm works: One Entity per Name (OEN) where the algorithm tokenizes the document in a list of exclusive sentences, recognizing the full stop as a terminator character, and for each sentence, detects named entities; and One Entity per Document (OED) where the algorithm considers the bag of words from the entire document and then detects named entities, removing duplicates for the same output record *(NE, type, URI)*. Therefore, the result set differs from the two approaches.

### 3.2.1.4 Benchmarking Initiatives and NER Comparison attempts

### 3.2.1.5 NER Web APIs Comparison

The creators of the DBpedia Spotlight service have compared their service with a number of other NER extractors (OpenCalais, Zemanta, Ontos Semantic API[20], The Wiki Machine[21], AlchemyAPI and M&W's wikifier [119]) according to a particular

---

[9]http://www.alchemyapi.com
[10]http://dbpedia.org/spotlight
[11]http://www.evri.com/developer/index.html
[12]http://extractiv.com
[13]http://lupedia.ontotext.com
[14]http://www.opencalais.com
[15]http://saplo.com
[16]http://www.wikimeta.com
[17]http://developer.yahoo.com/search/content/V2/contentAnalysis.html
[18]https://www.textrazor.com/
[19]http://www.zemanta.com
[20]http://www.ontos.com
[21]http://thewikimachine.fbk.eu/

annotation task [111]. The experiment consisted in evaluating 35 paragraphs from 10 articles in 8 categories selected from the "The New York Times" and has been performed by 4 human raters. The final goal was to create wiki links. The experiment showed how DBpedia Spotlight overcomes the performance of other services to complete this task. The "gold standard" does not adhere to our requirement because it annotates unit information with just Wikipedia resource and it does not link the annotation to the NE and their type. For this reason, we differentiate from this work by building a proposal for a "golden standard" where we combine NE, type and URI as well as a relevance score of this pattern for the text.

Nathan Rixham[22] and Benjamin Nowack[23] have both reported in their blogs their experiences in developing a prototype using Zemanta and OpenCalais. They observe that Zemanta aims at recommending "tags" for the analyzed content while Open-Calais focuses on the extraction of named entities with their corresponding types. They argue that Zemanta tends to have a higher precision for real things while the performance goes down for less popular topics. When OpenCalais provides a Linked Data identifier or more information about the named entity, it rarely makes a mistake. OpenCalais mints new URIs for all named entities and sometimes provides `owl:sameAs` links with other linked data identifiers. In contrast, Zemanta does not generate new URIs but suggests (multiple) links that represent the best named entity in a particular context. In another report, Robert Di Ciuccio[24] notices on a simple benchmarking test of five NER APIs (OpenCalais, Zemanta, AlchemyAPI, Evri, OpenAmplify and Yahoo! Term Extraction) over three video transcripts in the context of ViewChange.org. The author argues that Zemanta was the clear leader of the NLP API field for his tests, observing that OpenCalais was returning highly relevant terms but was lacking disambiguation features and that AlchemyAPI was returning disambiguated results but that the quantity of entities returned was low. Finally, Veeeb provides a simple tool enabling to visualize the raw JSON results of AlchemyAPI, OpenCalais and Evri[25]. Bartosz Malocha developed in EURECOM a similar tool for Zemanta, AlchemyAPI and OpenCalais[26]. We conclude that to the best of our knowledge, there have been very few research efforts that aim to compare systematically and scientifically Linked Data NER services. In this thesis we have contributed to the development of a framework enabling the human validation of NER Web services that is also capable to generate an analysis report under different conditions (see Section 3.2.1.7).

---

[22]http://webr3.org/blog/experiments/linked-data-extractor-prototype-details/
[23]http://bnode.org/blog/2010/07/28/linked-data-entity-extraction-with-zemanta-and-opencalais
[24]http://blog.viewchange.org/2010/05/entity-extraction-content-api-evaluation/
[25]http://www.veeeb.com/examples/flex/nlpapicompare/nlpCompare.html
[26]http://entityextraction.appspot.com/

#### 3.2.1.6 NER Benchmark Initiatives

The Natural Language Processing (NLP) community has been addressing the NER task for the past few decades, with two major guidelines: establishing standard for various tasks, and metrics to evaluate the performances of algorithms. Scientific evaluation campaigns, starting in 2003 with CoNLL, ACE (2005, 2007), TAC (2009, 2010, 2011, 2012), and ETAPE in 2012 were proposed to involve and compare the performance of various systems in a rigorous and reproducible manner. Various techniques have been proposed along this period to recognize entities mentioned in text and to classify them according to a small set of entity types. We will show how we have used those benchmarks in order to evaluate the NERD platform presented in the section 3.2.1.7.

#### 3.2.1.7 Entity Recognition and Disambiguation on Media Resources using NERD

NERD is a Web framework plugged on top of various NER extractors, including some of the API's reviewed in the section 3.2.1.3. Its architecture follows the REST principles [51] and includes an HTML front-end for humans and an API for computers to exchange content in JSON. Both interfaces are powered by the NERD REST engine. NERD has been previously used for evaluating the quality of the extraction results collected by the different integrated extractors [149]. In [150] the authors offered statistics about precision measures for each tool, with the goal to highlight strengths and weaknesses and to compare them.

The primary sources for performing Named Entity Recognition and Disambiguation are the subtitles of the seed videos content. Alternatively, another textual source can be the ASR transcripts. By nature, those transcripts will be more noisy, often grammatically incorrect depending on the performance of the ASR engine. However, as we will see in the section 3.2.1.9, the performance of NER on ASR transcripts is similar than on perfect subtitles using our proposed named entity framework.

**NERD Data Model** We propose the following data model that encapsulates the common properties for representing NERD extraction results. It is composed of a list of entities for which a label, a type and a URI is provided, together with the mapped type in the NERD taxonomy, the position of the named entity, the confidence and relevance scores as they are provided by the NER tools. The example below shows this data model (for the sake of brevity, we use the JSON syntax):

```
"entities": [{
    "entity": "Kalifornien",
    "type": "StateOrCounty",
    "nerdType": "http://nerd.eurecom.fr/ontology#Location",
    "uri": "http://de.dbpedia.org/resource/Kalifornien",
    "startChar": 346,
    "endChar": 357,
```

```
"confidence": 0.288741,
"source": "alchemyapi",
"startNPT": 79622.9,
"endNPT": 79627.3
}]
```

which indicates that "Kalifornien" is a named entity of type `StateOrCounty` for the extractor `AlchemyAPI`, which has been mapped to the type `nerd:Location` and disambiguated with the German DBpedia URI http://de.dbpedia.org/resource/Kalifornien. It also indicates that the source of this extraction is `AlchemyAPI` with a confidence score of 0.288741, and that this named entity has been spotted in the transcript of a video in the time range [79622.9, 79627.3] in seconds.

**NERD REST API** The REST engine runs on Jersey and Grizzly technologies already presented in Section 2.4.4. Their extensible frameworks enable to develop several components. NERD is composed of 7 modules namely authentication, scraping, extraction, ontology mapping, store, statistics and Web. The authentication takes as input a FOAF profile of a user and links the evaluations with the user who performs them. The scraping module takes as input the URI of an article and extracts all its raw text. Extraction is the module designed to invoke the external service APIs and collect the results. Each service provides its own taxonomy of named entity types it can recognize. We therefore designed the NERD ontology which provides a set of mappings between these various classifications. The ontology mapping is the module in charge to map the classification type retrieved to our ontology. The store module saves all evaluations according to the schema model we defined in the NERD database. The statistic module enables to extract data patterns from the user interactions stored in the database and to compute statistical scores such as the Fleiss Kappa score and the precision measure. Finally, the Web module manages the client requests, the Web cache and generates HTML pages.

Plugged on the top of this engine, there is an API interface[27]. It is developed following the REST principles and it has been implemented to enable programmatic access to the NERD framework. It follows the following URI scheme (the base URI is http://nerd.eurecom.fr/api):

**/document** : GET, POST, PUT methods enable to fetch, submit or modify a document parsed by the NERD framework;

**/user** : GET, POST methods enable to insert a new user to the NERD framework and to fetch account details;

**/annotation/{extractor}** : POST method drives the annotation of a document. The parametric URI allows to pilot the extractors supported by NERD;

---

[27]http://nerd.eurecom.fr/api/

**/extraction** : GET method allows to fetch the output described as described at the
    beginning of this Section;

**/evaluation** : GET method allows to retrieve a statistic interpretation of the extrac-
    tor behaviors.

**NERD Ontology** Although these tools share the same goal, they use different
algorithms and different dictionaries which makes their comparison hard. We have
developed the NERD ontology, a set of mappings established manually between the
taxonomies of NE types. Concepts included in the NERD ontology are collected from
different schema types: ontology (for DBpedia Spotlight, Lupedia, and Zemanta),
lightweight taxonomy (for AlchemyAPI, Evri, and Yahoo!) or simple flat type lists
(for Extractiv, OpenCalais, Saplo, and Wikimeta).

The NERD ontology tries to merge the linguistic community needs and the logician
community ones: we developed a core set of axioms based on the Quaero schema [55]
and we mapped similar concepts described in the other scheme. The selection of
these concepts has been done considering the greatest common denominator among
them. The concepts that do not appear in the NERD namespace are sub-classes
of parents that end-up in the NERD ontology. This ontology is available at http:
//nerd.eurecom.fr/ontology (Figure 3.2).

To summarize, a concept is included in the NERD ontology as soon as there
are at least two extractors that use it. The NERD ontology becomes a refer-
ence ontology for comparing the classification task of NE extractors. We show
an example mapping among those extractors below: the `City` type is considered
as being equivalent to `alchemy:City`, `dbpedia-owl:City`, `extractiv:CITY`,
`opencalais:City`, `evri:City` while being more specific than `wikimeta:LOC`
and `zemanta:location`.

```
 nerd:City a rdfs:Class ;
   rdfs:subClassOf wikimeta:LOC ;
   rdfs:subClassOf zemanta:location ;
   owl:equivalentClass alchemy:City ;
   owl:equivalentClass dbpedia-owl:City ;
   owl:equivalentClass evri:City ;
   owl:equivalentClass extractiv:CITY ;
   owl:equivalentClass opencalais:City .
```

**NERD User Interface** The user interface[28] is developed in HTML/Javascript.
Its goal is to provide a portal where researchers can find information about the NERD
project, the NERD ontology, and common statistics of the supported extractors.
Moreover, it provides a personalized space where a user can navigate through a
dashboard, see his profile details, browse some personal usage statistics and get a
programmatic access to the NERD API via a NERD key. The simple user account

---

[28]http://nerd.eurecom.fr

| NERD type | Occurrence |
|-----------|-----------:|
| Person | 10 |
| Organization | 10 |
| Country | 6 |
| Company | 6 |
| Location | 6 |
| Continent | 5 |
| City | 5 |
| RadioStation | 5 |
| Album | 5 |
| Product | 5 |
| ... | ... |
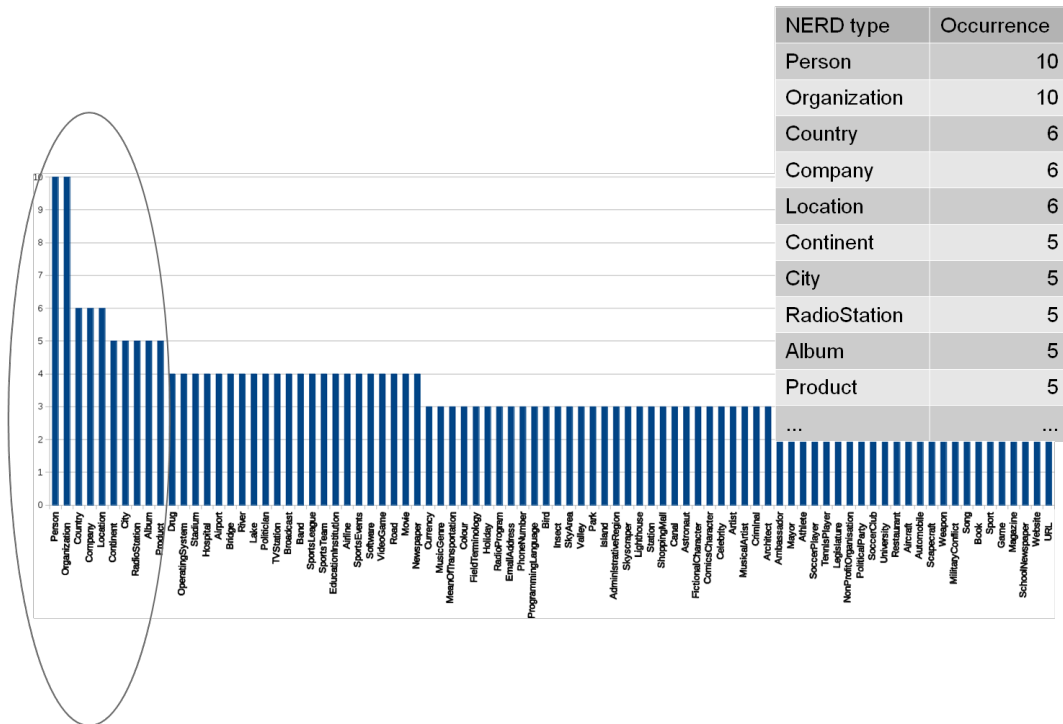
Figure 3.2: NERD ontology: the long tail of common denominator between NER extractors taxonomies

enables to annotate any Web documents via its URI. The raw text is first extracted from the Web source and a user can select a particular extractor in order to annotate the submitted document.

### 3.2.1.8   Other Named Entity Recognition Services

Under the scope of LinkedTV we have been also leveraging in other extraction and disambiguation tools that have been integrated under the umbrella of the NERD framework, but are worth to be explained in more deep since they bring particular functionalities and tackle the same problem from different perspectives.

**SemiTags** SemiTags is a Web service for named entity recognition and disambiguation. It is intended to recognize named entities in unstructured texts and discover links to Web based knowledge basis (namely Wikipedia and DBpedia). SemiTags works in two phases:

- Named Entity Recognition – The phrases corresponding to named entities are located in the text.

- Link Discovery – Local version of Wikipedia (corresponding to selected language) is queried for retrieving a suitable article describing entities located in

the previous phase. The link to Wikipedia is then used to map the entity to the corresponding DBpedia resource (if available).

For named entity recognition in English and German SemiTags uses the state of the art Stanford Named Entity Recognizer [53]. For Dutch they use OpenNLP[29] library trained on the CONLL-2002 [186] datasets. It has been found that Stanford Named Entity Recognizer trained on the same dataset performs significantly better.

For the second phase – Link Discovery – Semitags considers the combination of the textual based approach introduced in [110] and structural based approach introduced in [118] together with a structural based co-occurrence disambiguation. They generate the set possible candidates $C$ to surface forms of named entities discovered in the text. If there is more than one candidate for a given surface form a disambiguation is performed.

Contrary to the approach presented in [118] the model does not compare similarities of individual entities. Instead the objective is searching for the the best combination of candidates for individual surface forms in the analyzed text, being the entire document the context of the search task. Consider for example the following sentence: *Michael Bloomberg is the mayor of New York*. Simple observation shows that the entity Michael Bloomberg (mayor of New York) co-occurs in the same paragraph in Wikipedia together with the correct entity New York City in United States much more often (88 times) than with the New York in England (0 times).

Because generating all candidate combinations is a very demanding task, a heuristic that quantifies the impact of co-occurrences in the same paragraph has been used. Starting from an incidence matrix $I$ of the size $|C| \times |C|$ (where $|C|$ is the number of candidates), which represents a weighted graph, weights are obtained from the co-occurrence and assigned according to Equation 3.1.

$$
d_{e_{i,s},e_{j,t}} = \begin{cases} 0 & \text{if } s = t \\ 0 & \text{if } i = j \\ |P_{e_{i,s},e_{j,t}}| & \text{if } i \neq j \text{AND } s \neq t \end{cases} \tag{3.1}
$$

The weight $|P_{e_{i,s},e_{j,t}}|$ (count of paragraphs, where $e_i$ and $e_j$ were mentioned together) is used only in the case that the candidates represent a different entity $i \neq j$ and belong to a different surface form $s \neq t$, otherwise it is 0. Then they compute a score $e_{i,s}$ for each candidate as a sum of lines of the matrix representing the candidate (Equation 3.2).

$$
e_{i,s} = \sum_{j=1}^{|C|} e_{i,j} \tag{3.2}
$$

**Targeted Hypernym Discovery (THD)** The Targeted Hypernym Discovery

---

[29]http://opennlp.apache.org/

(THD) approach implemented in this extraction tool is based on the application of hand-crafted lexico-syntactic patterns. Although lexico-syntactic patterns for hypernym discovery have been extensively studied since the seminal work [74] was published in 1992, most research focused on the extraction of all word-hypernym pairs from the given generic free-text corpus.

Lexico-syntactic patterns were in the past primarily used on larger text corpora with the intent to discover all word-hypernym pairs in the collection. The extracted pairs were then used e.g. for taxonomy induction [173] or ontology learning [28]. This effort was undermined by the relatively poor performance of lexico-syntactic patterns in the task of extracting *all* relations from a *generic* corpus. On this task, the state-of-the-art algorithm of Snow [172] achieves an F-measure of 36 %. However, applying lexico-syntactic patterns on a *suitable document* with the intent to extract *one hypernym* at a time can achieve F1 measure of 0.851 with precision 0.969 [104]. In [104], the suitable documents were Wikipedia entries for persons and the target of the discovery was the hypernym for the person covered by the article. The algorithm implemented in THD is based on similar principles as [104], but without being limited to a certain entity type.

The design and evaluation of the THD algorithm has been done so far in English. Of course, using Wikipedia of the particular language has its benefits, even for named entities. Local versions are smaller, but they are not subsets of English Wikipedia. Many named entities of local importance not present in the English Wikipedia are covered. However, use of non-English Wikipedia for THD would require the design of the extraction grammar for the particular language as well as the availability of other resources and processing tools. Also, an issue with mapping the non-English hypernyms to the English DBpedia may arise.

With the aim of achieving interoperability between THD and other NLP tools, the processed results are exposed in the NIF format [76]. The results from the *entity* and *hypernym extraction* together with information about their *resource representations* in DBpedia are translated into the NIF format and published as Linked Data.

**Soft Entity Classification** So far, the description of the annotation task has been focused on a sharp or crisp classification, meaning an input entity is typically assigned one class as type. This standard approach implies some limitations, with some specific problems for the multimedia oriented use:

- the NER systems are sometimes unsure of which of the types is correct. However according to traditional approaches, just one type needs to be picked.

- in some cases, multiple types can be correct simultaneously. For example, the Obama entity can be simultaneously classified as nerd:Politician and nerd:Celebrity.

- the result of NER in multimedia use cases is used also for personalization: the type(s) of the entity present in the shot are aggregated to one feature vector,

```
1  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  @prefix str: <http://nlp2rdf.lod2.eu/schema/string/>
3  @prefix dbpedia: <http://dbpedia.org/resource/>
4  @prefix sso: <http://nlp2rdf.lod2.eu/schema/sso/>
5  @prefix : <http://example.org/>
6  :offset_0_80_Diego+Armando+Maradona+Franco+is+from+Argentina.+Argentina+is+next+to
       +Chile.
7    rdf:type str:Context ;
8    str:isString "Diego Armando Maradona Franco is from Argentina. Argentina is next
         to Chile." ;
9  :offset_0_29_Diego+Armando+Maradona+Franco
10   rdf:type str:String ;
11   str:referenceContext :offset_0_80_Diego+Armando+Maradona+Franco+is+from+
         Argentina.+Argentina+is+next+to+Chile. ;
12   sso:oen dbpedia:Diego_Maradona ;
13   str:beginIndex "0" ;
14   str:endIndex "29" .
15   str:isString "Diego Armando Maradona Franco" ;
16 dbpedia:Diego_Maradona rdf:type dbpedia:Manager .
17
18 :offset_38_47_Argentina
19   rdf:type str:String ;
20   str:referenceContext :offset_0_80_Diego+Armando+Maradona+Franco+is+from+
         Argentina.+Argentina+is+next+to+Chile. ;
21   sso:oen dbpedia:Argentina_national_football_team ;
22   sso:oen dbpedia:Argentina ;
23   str:beginIndex "38" ;
24   str:endIndex "47" .
25   str:isString "Argentina" ;
26 dbpedia:Argentina rdf:type dbpedia:Country .
27
28 :offset_70_75_Chile
29   rdf:type str:String ;
30   str:referenceContext :offset_0_80_Diego+Armando+Maradona+Franco+is+from+
         Argentina.+Argentina+is+next+to+Chile. ;
31   sso:oen dbpedia:Chilean_peso ;
32   sso:oen dbpedia:Chile ;
33   str:beginIndex "70" ;
34   str:endIndex "75" .
35   str:isString "Chile" ;
36 dbpedia:Chilean_peso rdf:type dbpedia:Currency .
```

Figure 3.3: The excerpt of a NIF export.

which is very useful for those kind of algorithms. For this purpose, it is better to have a more robust entity representation (multiple types, some of them with lower confidence), rather than a single type with non-negligible likelihood of being incorrect.

The above mentioned points can be addressed by providing soft (or sometimes referred to as "fuzzy") entity classification. Some tools described so far have the option to provide soft output. These systems include:

- DBpedia spotlight, included in the NERD platform, can be configured to return n-best candidates along with confidence levels.

- SCM algorithm [88], which uses THD algorithm to map entities to WordNet concepts, and then uses WordNet similarity measures to compute the similarity with each of the target classes. Target classes (concepts) are WordNet concepts.

- BOA algorithm [89] is based on the Rocchio classifier applied on Wikipedia articles. Target classes (concepts) are Wikipedia articles.

The advantages provided by soft classification have not been deeply probed in this thesis, but the experiments performed suggest they work better for our current use cases where very broad knowledge from different domains can be considered.

### 3.2.1.9   NER Evaluation: the ETAPE Campaign

ETAPE is a project targeting the organization of evaluation campaigns in the field of automatic speech processing and natural language processing. Partially funded by the French National Research Agency (ANR), the project brings together national experts in the organization of such campaigns under the scientific leadership of the AFCP, the French-speaking Speech Communication Association, a regional branch of ISCA.

In order to evaluate NERD in an audiovisual oriented corpora, the framework was presented to the 2012 ETAPE campaign. The evaluation focused on TV material with various level of spontaneous speech and multiple speaker speech. Apart from spontaneous speech, one of the originality of the ETAPE 2012 campaign is that it does not target any particular type of shows such as news, thus fostering the development of general purpose transcription systems for professional quality multimedia material. More precisely, the ETAPE 2012 data consists of 30 hours of radio and TV data from TV news, TV debates, TV amusements and Radio shows.

Several tasks are evaluated independently on the same dataset. Four tasks are considered in the ETAPE 2012 benchmark. For historical reasons, tasks belong to one of the three following categories: segmentation (S), transcription (T) and information extraction (E). The named entity task (E) consists in detecting all direct

mentions of named entities and in categorizing the entity type. The taxonomy follows the LIMSI Quaero definition as per the version 1.22 of the guide. Two conditions will be evaluated, detection on manual transcriptions and detection on ASR. Entity types are organized in a hierarchical way (7 types and 32 sub-types):

1. Person: pers.ind (invidual person), pers.coll (collectivity of persons);

2. Location: administrative (loc.adm.town loc.adm.reg loc.adm.nat loc.adm.sup), or physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro);

3. Organization: org.ent (services), org.adm (administration);

4. Amount: quantity (with unit or general object), duration;

5. Time: date time.date.abs (absolute date), time.date.rel (date relative to the discourse), or hour time.hour.abs, time.hour.rel

6. Production: prod.object (manufactury object), prod.art, prod.media, prod.fin (financial products), prod.soft (software), prod.award, prod.serv (transportation route), prod.doctr (doctrine), prod.rule (law);

7. Functions: func.ind (individual function), func.coll (collectivity of functions).

In order to participate in the campaign, we first built 426 axioms in the NERD ontology to the 32 concepts in the Quaero schema. The dataset being composed of French documents, we only consider the extractors Wikimeta, AlchemyAPI, Lupedia and OpenCalais. We developed a combined strategy of these 4 extractors which outperforms the performance of each individual extractor (Table 3.1).

|  | SLR | precision | recall | F-measure | %correct |
|---|---|---|---|---|---|
| AlchemyAPI | 37,71% | 47,95% | 5,45% | 9,68% | 5,45% |
| Lupedia | 39,49% | 22,87% | 1,56% | 2,91% | 1,56% |
| OpenCalais | 37,47% | 41,69% | 3,53% | 6,49% | 3,53% |
| Wikimeta | 36,67% | 19,40% | 4,25% | 6,95% | 4,25% |
| NERD combined | 86,85% | 35,31% | 17,69% | 23,44% | 17,69% |

Table 3.1: Performance comparison of the combined strategy of NERD with each individual extractor in the ETAPE campaign

The analysis per-type class highlights contrasted results: the class Person is generally well-detected while other categories show a very low recall. Interestingly, the NERD approach performs equally on perfect transcriptions than on automatically transcribed texts which are generally noisy and grammatically incorrect. This proves that the approach is robust to non grammatically correct text since it is much less dependent on a specific learning corpora as traditionally performed by the other participants in this campaign. This is positive since most of the multimedia content that we are interested in process do not come together with curated transcripts.

### 3.2.2   Keywords Extraction

Keyword extraction has also been considered in order to annotate videos out of the transcription files. We have experimented with the results obtained from the algorithm as presented at [196]. It is mainly based on the inverse document frequency (TF-IDF, see [154]) paradigm and employs Snowball[30] as stemmer.

Results have been converted to RDF according to the LinkedTV model via the class *linkedtv:Keyword*. However in this research we have prioritized entities againts keywords for two main reasons: 1) entities provide links to resources on the Web therefore promoting the exploitation of other available knowledge, and 2) keywords are textual units that by definition already include the notion of relevancy, while we are interested in generating more flexible annotations that allow to delegate the relevancy criteria to the particular system consuming them (what is relevant in some domain and particular application can be just meaningless for another). Keywords will also be used for derivate visual concept annotations as explained in Section 3.3.9.

### 3.2.3   Named Entity Expansion

The second important technique applied over textual information is the Named Entity Expansion algorithm. This algorithm has been implemented under the scope of this thesis and has been used in various scenarios from the LinkedTV project, and has motivated the different experiments included in Part II about news annotation.

Within the Linked Data community, a first objective is to increase the volume of interconnected data. Tools and frameworks like the one described in Section 3.2.1.7 contribute to generate new bridges between different documents and the knowledge available on the Web. However, from an exploitation point of view, those important techniques still do not tackle some issues. On the one hand, subtitles are not always complete enough to be the only textual source to rely on. The context around a particular event is broader than what is said in a video, and some important information pieces can be missing. On the other hand, a flat list of name entities fails to characterize what is described in the multimedia content: sometimes, one also needs to know how important those entities are with respect to an event or how those entities relate to each other.

In this section, we present an approach that generates extended annotations for the story happening in a video by alleviating the lack of textual resources that limits the application of semantic extraction techniques. We extend the initial set of descriptions about an event via Google searches and entity clustering. Applying this workflow allows to discover relevant resources and context-sensitive filtering resources. The general implementation of the algorithm described here will be further

---

[30]http://snowball.tartarus.org/

refined, tuned and exhaustively used in Part II to annotate News videos and generate their semantic context.

### 3.2.4 The Named Entity Expansion Pipeline

In order to build up the semantic context associated with one particular news document, we extract the main concepts and entities from the subtitles and explain how they are related to each other. The complete processing workflow takes as input the textual transcript of a multimedia resource illustrating an event, as well as the start and end date for which that particular event is considered.

We assume that this event has a considerable presence and coverage on the Web to ensure that the subsequent data mining techniques can collect sufficient data to reconstruct the event's context. The output of the algorithm is a list of named entities together with a numeric relevance score ($\varepsilon = \{E \times \mathbb{R}\}$, $E$ being a set of named entities classified using the NERD ontology[31]).

Our hypothesis states that this representation of stories provides a sufficient source of information for satisfying the viewer's information needs and better supports complex operations such as search and hyperlinking.

For each news item, we perform named-entity recognition over the corresponding subtitles using the NERD framework [151]. In our experiments, the language of the videos is English but NERD supports other languages. The output of this phase is a collection of entities annotated using the NERD Ontology, that comes with a first relevance score obtained from the extractors which have been used. This set includes a list of ranked entities that are explicitly mentioned during the video. Other entity based video annotation tools [99] stop at this point even when entities that can be relevant for the viewer in the context of the event are still missing. We tackle this problem by extending this first list of concepts via the entity expansion component.

The set of entities obtained from a traditional named entity extraction operation is normally insufficient and incomplete for expressing the context of a news event. Sometimes, some entities spotted over a particular document are not disambiguated because the textual clues surrounding the entity are not precise enough for the name entity extractor, while in other cases, they are simply not mentioned in the transcripts while being relevant for understanding the story. This is an inherent problem in information retrieval tasks: a single description about the same resource does not necessarily summarize the whole picture.

The named entity expansion operation relies on the idea of retrieving and analyzing additional documents from the Web where the same event is also described. By increasing the size of set of documents to analyze, we increase the completeness of the context and the representativeness of the list of entities, reinforcing relevant

---

[31]http://nerd.eurecom.fr/ontology/nerd-v0.5.n3

entities and finding new ones that are potentially interesting inside the context of that news item.

The entire logic will further be described in the following subsections and mainly consist of (1) building an appropriate search query from the original set of entities, (2) retrieving additional documents about the same news event, and (3) analyzing them for providing a more complete and better ranked set of final entities, as illustrated in Figure 3.4.
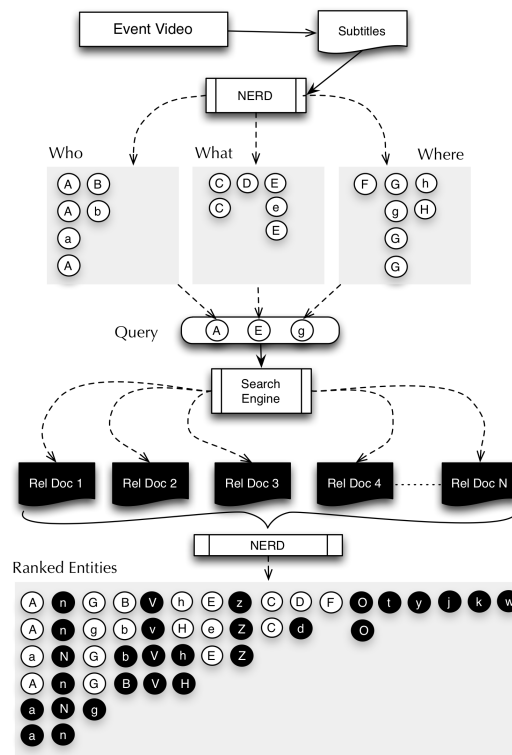


Figure 3.4: Schema of Named Entity Expansion Algorithm.

### 3.2.4.1    Query Generation

The *Five W's* is a popular concept of information gathering in journalistic reporting. It captures the main aspects of a story: who, when, what, where, and why [98]. We try to represent the news item in terms of four of those five W's (who is involved in the event, where the event is taking place, what the event is about, and when it has happened) in order to generate a query that retrieves documents associated to the same event.

First, the original entities are mapped to the NERD Core ontology, which considers 10 main classes: Thing, Amount, Animal, Event, Function, Organization, Location, Person, Product and Time. From those ten different categories, we generalize to

three classes: the *Who* from nerd:Person and nerd:Organization, the *Where* from nerd:Location, and the *What* from the rest of NERD types after discarding nerd:Time and nerd:Amount. The *When* or so-called temporal dimension does not need to be computed since it is considered to be provided by the video publisher.

After generating the three sets of entities, the next step consists in ranking them in relevance according to a weighted sum of two different dimensions: their frequency in the transcripts and their former relevance scores coming from the named entity extractors. We have defined the function *filterEntities(S)* for selecting the $n$ entities inside the set of entities $S$ whose relative relevance

$$R_{rel}\left(e_i, S\right) = R\left(e_i\right)/Avg\left(R\left(e_i\right)\right) \tag{3.3}$$

falls into the upper quarter of the interval

$$\left[max\left(R_{rel}\left(e_i, S\right)\right) - min\left(R_{rel}\left(e_i, S\right)\right)\right] \tag{3.4}$$

The final query is a pair

$$\text{Query}_{Event} = [\text{textQuery}, t] \tag{3.5}$$

where *textQuery* is the result of concatenating the labels of the most relevant entities in the sets Who, What, Where in that particular order, and $t$ the time period dimension. This query generation is depicted in the upper part of Figure 3.4.

### 3.2.4.2 Document Retrieval

Once $\text{Query}_{Event}$ is built out of the original set of named entities, it will be ready to be injected into a document search engine where additional descriptions about the news event can be found. In this situation, the kind of query generated in the previous step and the search engine chosen should be closely tied in order to maximize the quality of the obtained results. The different behavior of search engines make some alternatives more suitable than others for certain kinds of events. The way the resulting documents change in the search engines for a particular kind of event is a research question that will not be studied in this thesis.

For the first experiments described in this section we leverage on the Google Search REST API service[32] by launching a query with the text *textQuery*. Due to quota restrictions imposed by Google, the maximum number of retrieved documents is set to 30 by default. However, as we will show during the experiments performed in Part II, this is enough for significantly extending the initial set of entities directly spotted by NERD.

---

[32]http://ajax.googleapis.com/ajax/services/search/web?v=1.0

Concerning the temporal dimension, we only keep the documents published in the time period $t + t_e$. We increase the original event period in $t_e$ because documents concerning a news event are not always published during the time of the action is taking place but some hours or days after. The value of $t_e$ depends on many factors such as the nature of the event itself (whether it is a brief appearance in a media, or part of a longer story with more repercussion) or the kind of documents the search engine is indexing (from very deep and elaborated documents that take time to be published, to short and fresh posts quickly generated by users). Based on the simple assumption that that the longer an event takes, the bigger the buzz it generates in form of Web document published around those dates, we approximate $t_e = t$ which means that we look at a temporal window that is double the size the time the event was ongoing.

The middle part of Figure 3.4 shows this process. The query is input in the search engine in order to retrieve other documents that report on the same event discussed in the original video. Those documents (colored in black in the Figure 3.4) will be further processed to increase the size of the collection and get additional insights about the news item.

### 3.2.4.3   Entity Clustering

In this phase, the additional documents which have just been retrieved are now processed and analyzed in order to extend and re-rank the original set of entities and consequently get a better insight about the event. Since most of the resources retrieved are Web pages, HTML tags and other annotations are removed, keeping only the main textual information. This plain text is then analyzed by the NERD framework in order to extract more named entities.

In order to calculate the frequency of a particular resource within the entire corpora, we group the different appearances of the same instance and check their cardinality. This is not a trivial task since the same entity can appear under different surface forms, contain typos or have different disambiguation URL's pointing to the same resource. We performed a centroid-based clustering operation over the instances of the entities. We considered the centroid of a cluster as the entity with the most frequent disambiguation URL's that also have the most repeated labels. As distance metric for comparing pairs of entities, we applied strict string similarity over the URL's, and in case of mismatch, the Jaro-Winkler string distance [210] over the labels. The output of this phase is a list of clusters containing different instances of the same entity.

### 3.2.4.4 Entity Ranking

The final step of the expansion consists of ranking the different named entities obtained so far. To create this ordered list, we assigned a score to every entity according to the following features: (1) relative frequency in the transcripts of the event video; (2) relative frequency over the additional document; and (3) average relevance according to the named entity extractors. The three dimensions are combined via a weighted sum where the frequency in the video subtitles has a bigger impact, followed by the frequency on the searched documents and the relevance from the extractors.

The final output of the entity expansion operation is a list of entities together with their ranking score and the frequency in both the main video and in the collected documents retrieved from the search engine. Entities with a higher $relScore_i$ in the final classification are considered more representative for describing the context than the original entities. Furthermore, we observe that:

- The bigger the sample size, and the better the ranking becomes. Entities appearing repeatedly in the additional documents will be promoted while those appearing rarely will be pushed back to the end of the list.

- Entities that originally have not been disambiguated can now have their corresponding URL if any of the similar instances appearing in the same cluster but coming from different documents can be used to provide a link to a Web resource. The same occurs with incomplete or misspelled labels: cleaner surface forms from the same entity spotted in the related documents can alleviate the problem of having to rely on one single and therefore error-prone instance.

- Finally, some entities not spotted in the original transcripts but important in the context of the event are now included in the list of relevant items since they have been extracted from the collected documents. The seed document being analyzed can be biased or lack important information for the described facts, so having a wider-scope semantic annotations can help in many other operations leveraging in them like content recommendation or personalization.

Finally, the ranking mechanism explained in this section, which is primarily based on frequency measures relying on entity mentions on both the transcripts and related documents, will be extended and improved in Part II of this thesis with more sophisticated heuristics, in order to promote other entities that are also highly relevant for the story being told, but are barely mentioned on related documents.

## 3.3 Visual-Based Annotations: a Multimodal Approach

In this section we will introduce a set of different visual techniques that directly rely in the audiovisual information of the multimedia document to detect different visual

cues that further characterize the content. Those techniques have been studied in the literature since much before the Linked Data Web philosophy irrupted in the media scenario, and in this thesis we have worked on adequating them to a new Web scenario where they can be globally used and exploited at high scale, polishing their semantic and reinforcing their outcomes by complementing them with other information already available on the Web. In Chapter 4.5 we make use of those visual annotations, together with other textual based approaches and semantic techniques, to offer the media consumers a new set of features and advanced operations that were not widely available before.

### 3.3.1 Concepts Detection

Visual concept detection is one of the techniques we have applied in the context of the LinkedTV project. In particular we have followed the approach presented in [165], using a sub-set of 10 base (key-frame) detectors. The algorithm is applied on the key-frames of the video, aiming to detect objects out of the 151 different semantic concepts, both static and dynamic ones, selected from the list of concepts defined in the TRECVID 2012 SIN task [133]. The 10 used classification modules are derived from different combinations of the employed interest point detector, descriptor and visual word assignment method. Specifically, the considered interest point detectors are the Harris-Laplace corner detector [70] and a dense pixel sampling strategy, while the employed descriptors are the well known SIFT [105] and two colored variations of it, named RGB-SIFT and Opponent-SIFT [200]. Then, the low-level descriptors are assigned to visual words from two vocabularies that were created off-line through K-means clustering, employing hard- and soft-assignment [201], respectively.

For each one of the employed classification modules, one vector per key-frame is finally extracted and used as the actual input to the utilized SVM classifier. In order to increase the computational efficiency, linear SVM classifiers are employed instead of kernel SVMs that are typically used for this task, while another boost in time performance is obtained by using only one range file for the classification of the overall set of concepts (not one range file per concept, as when the algorithm runs for a small collection of videos). The latter results in a slightly lower detection accuracy, however reducing by 145 times the needed processing time, which is crucial when the algorithm is applied on large collections of videos, such as the MediaEval dataset (see Section 4.5.2). The output of each of the employed classifiers is a Degree of Confidence (DoC) score for the corresponding concept, which expresses the classifier's confidence in this concept being suitable for annotating the current shot. This process is iterated for each considered concept and for all used modules, and the extracted DoC scores are averaged to generate the final concept detection score. Finally, a vector of such scores, where each element of the vector corresponds to a different concept, is the

system's output.

This output is converted to RDF and integrated into the LinkedTV model via the LSCOM vocabulary presented in Section 2.4.1. The concepts detected and serialized in RDF are attached to the corresponding media fragments (*ma:MediaFragment* instances) through the use *oa:Annotations*.

### 3.3.2 Shot Segmentation

The temporal segmentation of the videos into shots is performed using the algorithm proposed in [194]. This technique extracts visual features, namely color coherence, Macbeth color histogram and luminance center of gravity, and forms an appropriate feature vector per frame. Then, given a pair of neighboring (either successive or non-successive) frames of the video, the distances between their vectors are computed, composing distance vectors, that are finally evaluated using one or more SVM classifiers, resulting to the detection of both abrupt and gradual transitions between the shots of the video. Shot detection accuracy of this techniques over some LinkedTV's material has reached a 98.5% accuracy [5]. The resulting Shots are incorporated to the LinkedTV knowledge graph by serializing them as instances of *ma:MediaFragment* classes and further annotating them as *linkedtv:Shot*, where the temporal references are both present in the URL's via Media Fragment URI's specification and explicitly encoded via *nsa:temporalStart* and *nsa:temporalEnd* attributes.

### 3.3.3 Scene Segmentation

We have considered also scene segmentation annotations based on the algorithm introduced in [166]. While shots are considered as temporal units where there is visual irruption between frames, scenes are generally longer segments concerning a particular location and action taking place, so they are significantly more difficult to spot from a pure visual point of view. The method we have considered groups the shots of the video (either automatically detected, or predefined as for the MediaEval 2013 Search and Hyperlinking task described in Section 4.5.2) into sets that correspond to individual scenes of the video, based on the visual similarity and the temporal consistency among them. Specifically, one representative key-frame is extracted from each shot of the video and the visual similarity between pairs of key-frames is estimated via HSV histogram comparison. The grouping of shots into scenes is then performed, by utilizing the two proposed extensions of the well-known Scene Transition Graph (STG) method [213], which clusters shots into scenes by examining whether a link, between two shots, exists.

The first extension, called Fast STG, reduces the computational cost of shot grouping, by considering shot linking transitivity and the fact that scenes are by definition convex sets of shots, thus limiting the number of shot pairs whose possible linking

needs to be evaluated. The latter allows for faster detection of the scene boundaries, while maintaining the same performance with the original STG algorithm. The second extension, called Generalized STG, builds on the former in order to construct a probabilistic framework, towards multiple STGs combination, alleviating the need for manual STG parameter selection. As described in [166], this probabilistic framework can also be used for the realization of a multi-modal approach for scene segmentation, allowing the fusion of STGs built by considering different forms of information extracted from the video, such as low level audio or visual features, visual concepts and audio events. As already happened with Shots, Scenes are incorporated to the LinkedTV knowledge graph by serializing them as instances of *ma:MediaFragment* classes and further annotating them as *linkedtv:Scene*.

### 3.3.4   Optical Character Recognition (OCR)

For text localization in pictures and video frames, we employ the algorithm presented at [179]. The detection is based on color segmentation, using statistical region merging [130]. For a refined text separation, a Gaussian model is computed based on uniform colored connected components. For Optical Character Recognition (OCR), we employ the widely used tesseract[33] engine. Results have been serialized as *str:String* and attached to particular media fragments via *oa:Annotation* instances. They have not been deeply used in our experiments due to the lack of context that those spare strings sometime have inside the main facts being told in the video, but given its importance in some cases they are definitely a subject of study for the future.

### 3.3.5   Face Detection and Tracking

Face analysis starts with face detection, where all frames of the videos are processed to extract faces. For this task, we use the well-known Viola and Jones' cascade face detector [204], or more precisely its implementation in the C++ openCV library as improved by Lienhart and Maydt [102]. Detection is combined with a skin color detector [122] for filtering out detections that are not likely to be faces.

The tested framework performs well on images. However, it has to be adapted to videos and to create face tracks. Under the scope of the LinkedTV project, we have used the spatio-temporal information in order to smooth the results. Detected faces are linked with shots using a spatio-temporal matching of faces: if two faces in adjacent frames are in a similar position, we assume we can match them. Linear interpolation of missing faces also relies on matching similar bounding boxes in close but none adjacent frames through a shot. This process enables to smooth the tracking results and to reject some false positive.

---

[33]http://code.google.com/p/tesseract-ocr/

Results have been converted to RDF according to the LinkedTV model as explained in Section 2.4.2.2. Faces detected and recognized can be used not only to support search operations and linking content by individuals being appeared, but also to reinforce other annotations that are temporally close to the spotted face, like explained in [180].

### 3.3.6  ASR on Spontaneous Speech

Spoken content is one of the main sources for information extraction on most audiovisual documents. However, manual transcripts of that spoken information is not always available, so we have to rely on other machine driven techniques like ASR. In [177], the authors performed a manual ASR transcript evaluation which performed good on planned speech segments, but rather poor on spontaneous parts which were quite common in interview situations in the news show scenarios.

Under the scope of the LinkedTV project this algorithm was improved by extending the training material with new data and adopt new settings for the clarification phase. Focusing on german as language, the authors collected and manually transcribed a huge new training corpus of broadcast video material, with a volume of approx. 400 h and containing roughly 225 h of clean speech. The new corpus is segmented into utterances with a mean duration of 10 seconds and is transcribed manually on word level. The recorded data covered a broad selection of news, interviews, talk shows and documentaries, both from television and radio content across several stations. Special care was taken in order to ensure the material contains large parts of spontaneous speech. The effort in acquiring this new training data ended up making this dataset of the largest corpora of German TV and radio broadcast material.

This new training material made a revisit of the free speech decoder parameters necessary, to guarantee optimality. In the literature, these parameters are often either set empirically using cross-validation on a test set, which is a rather tedious task, or the default values of toolkits are retained. Few publications analyze the parameter adaption with automatic methods; among them are [37], using gradient descent, [106], using large-margin iterative linear programming, or [85], using evolutional strategies. Since we aim at facilitating the optimization process by employing a fast approach and therefore enable this step for a wide range of applications, we employ Simultaneous Perturbation Stochastic Approximation (SPSA) [174] for optimizing the free decoding parameters and show in [178] that it leads to stable and fast results.

In a nutshell, the algorithm works as follows. For a tuple of free parameters in each iteration, SPSA perturbates the given values simultaneously, both adding and subtracting a random perturbation vector for a total of two new tuples. The gradient at the current iteration is estimated by the difference of the performance

Table 3.2: WER results on the test corpora, for the SPSA iterations and their respective loss functions. Each optimization on a given loss function has been executed two times from scratch with 18 iterations to check for convergence.

| parameter set | WER planned | WER spontaneous |
|---|---|---|
| baseline | 27.0 | 52.5 |
| larger training data | 26.4 | 50.0 |
| SPSA 1st run | 24.6 | 45.7 |
| SPSA 2nd run | 24.5 | 45.6 |

(here measured as word error rate, WER) between these two new tuples, and a new tuple is then computed by adapting the old tuple towards the gradient using a steadily decreasing step function. We refer to [178] for further implementation details. For developing and optimizing those free parameters, we used the aforementioned corpus from German broadcast shows, which contains a mix of planned (i.e., read news) and spontaneous (i.e., interviews) speech, for a total of 2,348 utterances (33,744 words).

For evaluation, we test the decoding performance on the news show content, separated into a planned set (1:08h, 787 utterances) and a spontaneous set (0:44h, 596 utterances). The results are listed in Figure 3.2. Here, it can be seen that while the performance for planned speech improved by 2.5% absolute (9.3% relative) in terms of WER, spontaneous speech segments now have a WER of almost 7% lower (13.3% relative) than the original baseline, which is quite a nice advance in the ASR quality.

Such promising results, combined with the power of NERD to analyze prune error texts, have turn ASR results in a crucial technique for having semantically annotated raw videos that have not been manually transcribed or annotated by any means.

### 3.3.7   Fast Object Re-detection

We have also considered a semi-automatic annotation of the video based on the re-detection of specific objects of interest selected by a video editor so that, e.g., instances of the same painting in a culture heritage show can be identified and tracked throughout the movie, allowing to automatically show to the viewer timely descriptions or related information.

We detect instances of a manually pre-defined object of interest $O$ in a video $V$ by evaluating its similarity against the frames of this video, based on the extraction and matching of of SURF (Speeded UP Robust Features) descriptors [11]. The time performance of the method is a crucial requirement, since the object-based video annotation will be handled by the editor. A faster than real-time processing is achieved by combining two different strategies: (a) exploit the processing power of the modern Graphic Processing Units (GPUs) and (b) introduce a video-structure-

based frame sampling strategy that aims to reduce the number of frames that have to be checked.

The algorithm utilizes the analysis results of the shot segmentation method of [195], which can be interpreted as a matrix $S$ where its $i$-th row $S_{i,j}, j = 1, ...5$ contains the information about the $i$-th shot of the video. Specifically, $S_{i,1}$ and $S_{i,2}$ are the shot boundaries, i.e. the indices of the starting and ending frames of the shot and $S_{i,3}$, $S_{i,4}$, $S_{i,5}$ are the indices of three representative key-frames of this shot. By using this data, the algorithm initially tries to match the object $O$ with the 5 frames of the $i$-th shot that are identified in matrix S (i.e. $S_{i,j}, j = 1, ...5$), and only if the matching is successful for at least one of these frames it proceeds with comparing $O$ against all the frames of that shot. It then continues with the key-frames of the next shot, until all shots have been checked. Following this approach the algorithm analyses in full only the parts (i.e. the shots) of the video where the object appears (being visible in at least one of the key-frames of these shots) and quickly rejects all remaining parts by performing a small number of comparisons, thus leading to a remarkable acceleration of the overall procedure. More details on our object re-detection approach can be found in [4].

Our experiments on the object re-detection technique, using objects and videos from the LinkedTV dataset, show that the algorithm achieves 99.9% Precision and 87.2% Recall scores, identifying successfully the object for a range of different scales and orientations and when it is partially visible or partially occluded (see for example Fig. 3.5), while the needed processing time using a modest modern PC (e.g. having an Intel i7 processor, 8GB RAM memory and a CUDA-enabled GPU) is about 10% of the video's actual duration, thus making the implemented technique an efficient tool for fast and accurate instance-based annotation of videos.

### 3.3.8 Towards Localized Person Identification

One of the issues of developing a good face recognition approach is to have an adequate database containing the instances to be recognized. This requires an high editorial effort as described in [177], where authors highlight the challenge of obtaining a reasonable person identification database for a local context. To overcome this issue in the news domain we have experimented with the following hypothesis: for most news shows, banner information is shown whenever a specific person is interviewed. Manually checking videos of one show over the course of two months, seems reasonable to assume that (a) the banner is only shown when the person is speaking, and (b) mostly (but not always) only this single person is seen in these shots. We can thus use this information for speaker identification and face recognition (see Figure 3.6 for a graphical representation of this workflow).

Figure 3.5: Object of interest (top row) and in green bounding boxes the detected appearances of it, after zoom in/out (middle row) and occlusion-rotation (bottom row).

We tested this approach over 50 episodes of the show "Brandenburg aktuell"[34], 30 minutes length each. Every single show contains on average around seven interviewed persons with their name contained in the banner. Since the banner will be always at a certain position, Optical Character Recognition (OCR) heuristic using tesseract [169] was applied: we check each screen-shot made every half second and decide that a name is found whenever the Levenshtein distance over three consecutive screen-shots is below 2. On manually annotated 137 screen-shots, the character accuracy is at convenient 97.4%, which further improves to 98.4% when optimizing tesseract on the shows font, using a distinct training set of 120 screen-shots. The results of this visual techniques is not materialized in form of annotations about the video, but are intended to improve the quality of the face recognition results already covered in

---

[34]http://www.rbb-online.de/brandenburgaktuell/
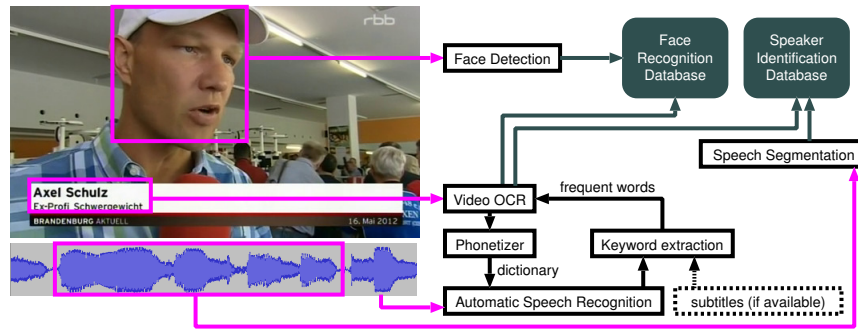
Figure 3.6: Workflow for a an automatically crawled person identification database, using news show banner information

previous subsections.

### 3.3.9 From Visual Cues to Detected Concepts

Different visual annotations such as concept detected in shots has been probed to be valuable for different operations over the video, such as hyperlinking (see Section 4.5.2). However, running those techniques over the content is expensive in terms of time and processing resources, making this task impossible to perform when the corpora is too big or we need results in real time with no preprocessing phase allowed. Also, sometimes we need to describe pure textual resources in term of the same visual cues so we can map them with relevant videos annotated with concepts. Even this situation is expected to be alleviated in the future, thanks to better performing hardware and more advanced techniques in the field, we have tackled this issue by generating those visual concepts by starting from more lightweight text-based techniques like keyword extraction. The objective is to link certain tags to visual concepts via the semantic knowledge available in certain knowledge bases. Taking as input raw text accompanying the analyzed video, we have run keyword extraction operations using Alchemy API[35]. We have then aligned every spotted keywords with a subset of LSCOM concepts (more than one hundred items from TRECVID 2012 SIN task) by using a semantic word distance based on Wordnet synsets [103]. This way, certain keywords like "car" or "bicycle" can be used to derive visual concepts like "lscom:vehicle". The output of this algorithm includes two items per concept detected: the confidence score, which helps to decide how close a visual concept is related to a keyword, and the relevance, which gives an idea of the importance of a certain concept inside the scope of the text submitted. Manual evaluations lead us to discard concepts when the confidence score was below 0.7 [159]. The usefulness of such approach has been initially probed by the good results obtained in MediaEval 2013 explained in Section 4.5.2, where the initial anchors raising the corresponded queries did not bring any visual cues that allow us to adequately bring in related

---

[35]http://www.alchemyapi.com/

video results.

## 3.4   Summary

In this Chapter we have reviewed different automatic annotation techniques for populating the multimedia knowledge model presented in previous Chapter 2. We have focused mainly in text based approaches working over subtitles, including Named Entity Extraction and Named Entity Expansion. But we have also considered different visual analysis algorithms, whose results are equally incorporated into the model for providing a more adequate description of the multimedia content.

Those annotations are attached to the corresponding media fragments so other applications consuming the multimedia content can rely on them for providing advanced features such as the ones we will describe in next Chapter 4. Named Entities are disambiguated in resources on the Web always that possible, so they become anchors from the fragments to knowledge bases like DBpedia where we can discover implicit relations between those entities and even the reasons behind those relations, so what is being told in the video can be better understood. Visual annotations can complement and refine such text based annotations in what are called multimodal approaches. Those hybrid techniques are still a domain to be further explore but our initial experiments in Section 4.5 have probed that they have a great potential when we apply them in operations like hyperlinking.

Finally, the annotations covered in this Chapter are mainly focusing in what is being explicitly described in the video, but the research direction taken with Named Entity Expansion exemplifies how the approaches solely relying on the analyzed document are often not enough for describing particular aspects requiring a contextualization. News is one of those scenarios where the lack of a valid background information prevent us to precisely interpret what has been told in the items. In Part II we will study how, using at starting point the results from Entity Expansion we can get to recreate the big picture of the events being described and significantly improve the way viewers consume those items.

# Exploiting Annotated Media Fragments

## 4.1 Introduction

Having the multimedia content represented according to the Web of Data compliant model presented in Chapter 2, segmented in fragments with different levels of granularity by relying on the metadata offered by the content publishers or in visual techniques introduced in Section 3.3, and having those fragments annotated according to different techniques leveraging on text and visual dimensions as explained in Chapter 3, the next step is to exploit this multimedia descriptions in order to implement advanced operations leveraging on them, which can make a difference in the way we consume audiovisual information.

Having the content annotated according to widely used vocabularies makes easier o fetch other relevant resources from the Web to also rely on, opening the window to a new set of possibilities: more precise reasoning over the knowledge, bringing more information to the context of the video by enriching it with other resources (see Section 4.2), further promote and refine the most prominent fragments and annotations (see Section 4.3), classify video in an automatic fashion (see Section 4.4), or finding hyperlinks inside a collection of fragments in a multimodal way (see Section 4.5).

## 4.2 Media Fragment Enrichment

Media fragments annotated with different techniques described in Chapter 3 can be associated with other media content that further illustrates what is being told in them. The nature of the related content that can be attached to the seed fragments can be diverse: social items being shared in different online platforms, blog posts generated by users about the same matter being discussed in the original video, curated journal articles giving a more professional point of view about the facts, or other pictures and media fragments belonging to different sources. In this section we show different approaches to perform this enrichment process.

### 4.2.1   Enriching Fragments with Social Media Content

The widespread availability of mobile phones with higher resolution cameras has transformed citizens into media publishers and witnesses, who feel keen to comment and share event-related media on social networks. Some examples with global impact include the shootings in Utøya, which first appeared on Twitter, the capture and arrest of Muammar Gaddafi, which first appeared on YouTube, or the emergency ditching of a plane in the Hudson river, which first appeared on Twitpic. Some news agencies[1] have even specialized in aggregating and brokering this user-generated content. In this section, we illustrate an approach for retrieving all those event-related media items that are being published by users on several social networks, that has been published in [148, 116, 181].

#### 4.2.1.1   Social Networks

A social network is an online service or media platform that focuses on building and reflecting social relationships among people who share interests and/or activities. The boundary between social networks and media platforms is rather blurry. Several media sharing platforms, such as YouTube, enable people to upload content and optionally allow other people to react to this content in the form of comments, likes or dislikes. On other social networks (*e.g.,* Facebook), users can update their statuses, post links to stories, upload media content and also give readers the option to react. Finally, there are hybrid clients (*e.g.,* TweetDeck for Twitter using Twitpic) where social networks integrate with media platforms typically via third party applications. Therefore, we consider three types of support of media items with social networks:

- *First-order support*: The social network is centered on media items and posting requires the inclusion of a media item (*e.g.* YouTube, Flickr);

- *Second-order support*: The social network lets users upload media items but it is also possible to post only textual messages (*e.g.* Facebook);

- *Third-order support*: The social network has no direct support for media items but relies on third party application to host media items, which are linked to the status update (*e.g.* Twitter before the introduction of native photo support).

We consider 12 different social networks that all have powerful and stable APIs and, together, represent the majority of the market. The criteria for including media sharing platforms follow a study performed by the company Sysomos, specialized in social media monitoring and analytics [95]. Table 4.1 lists these platforms according to the categorization defined above.

---

[1]*e.g.* Citizenside (http://www.citizenside.com)

| Social Network | URL | Category | Comment |
|---|---|---|---|
| Google+ | http://google.com/+ | second-order | Links to media items are returned via the Google+ API. |
| MySpace | http://myspace.com | second-order | Links to media items are returned via the MySpace API. |
| Facebook | http://facebook.com | second-order | Links to media items are returned via the Facebook API. |
| Twitter | http://twitter.com | second-/third-order | In second order mode, links to media items are returned via the Twitter API. In third order mode, Web scraping or media platform API usage are necessary to retrieve links to media items. Many people use Twitter in third order mode with other media platforms. |
| Instagram | http://instagram.com | first-order | Links to media items are returned via the Instagram API. |
| YouTube | http://youtube.com | first-order | Links to media items are returned via the YouTube API. |
| Flickr | http://flickr.com | first-order | Links to media items are returned via the Flickr API. |
| MobyPicture | http://mobypicture.com | first-order | Media platform for Twitter. Links to media items are returned via the MobyPicture API. |
| Twitpic | http://twitpic.com | first-order | Media platform for Twitter. Links to media items must be retrieved via Web scraping. |
| img.ly | http://img.ly | first-order | Media platform for Twitter. Links to media items must be retrieved via Web scraping. |
| Lockerz | https://lockerz.com/ | first-order | Media platform for Twitter. Links to media items must be retrieved via Web scraping. |
| yfrog | http://yfrog.com | first-order | Media platform for Twitter. Links to media items must be retrieved via Web scraping. |

Table 4.1: Social networks with different support levels for media items and techniques needed to retrieve them

### 4.2.1.2  Collecting Items from Social Netwoks: MediaCollector module

We have developed a collection module called **MediaCollector** composed of media item extractors for all the media sharing networks listed in Table 4.1. The media collector takes as input a search term, *e.g.,* "obama" and performs a parallel keysearch in all the social networks. Each platform has a 30 second timeout window to deliver its results. When the timeout has expired, or when all social networks have responded, a unified output is delivered.

The Media Collector was originally developed by Thomas Steiner at https://github.com/tomayac/media-server and forked twice for the purpose of LinkedTV experiments and this thesis at https://github.com/vuknje/media-server (in order to enable temporal search across social media platforms) and at https://github.com/MathildeS/media-server (in order to integrate a dedicate LinkedTV enriching module, Unstructured Search Module). It is based on NodeJS[2]

It proposes a common alignment schema for all social networks in order to be agnostic of a particular social network. The resulting metadata for a media item are detailed below (URI examples for the search by "io12" keyword, shortened for legibility):

**Media URL** Deep link to the media item (*e.g.,* http://goo.gl/zl2Tg).

**Type** Type of the media item (photo or video).

**Story URL** URL of the micropost where the media item appeared (*e.g.,* http://goo.gl/R4lv8).

**Message** Text Description of the micropost in raw format.

**Clean** Cleaned text description of the micropost where some characters are removed.

**User** URL of the micropost author (*e.g.,* http://goo.gl/zl2Tg).

**Timestamp** Reference time when the micropost was authored or the media item was uploaded.

**Implementation** Twitter and its ecosystem (TwitPic, TwitterNative, MobyPicture, Lockerz or yfrog), GooglePlus and YouTube, Facebook and Instagram, Flickr and FlickrVideos, MySpace, all offer search APIs over the content they host. Those search functions, however, provide results that vary according to the time the query has been triggered, covering a window of time that ranges from only the recent past to many years ago. In addition, they offer different parameters that enable to customize search queries (e.g. filtering by location). The MediaCollector module is composed of media item extractors for these 12 media sharing platforms. It takes as input

---

[2]http://nodejs.org/
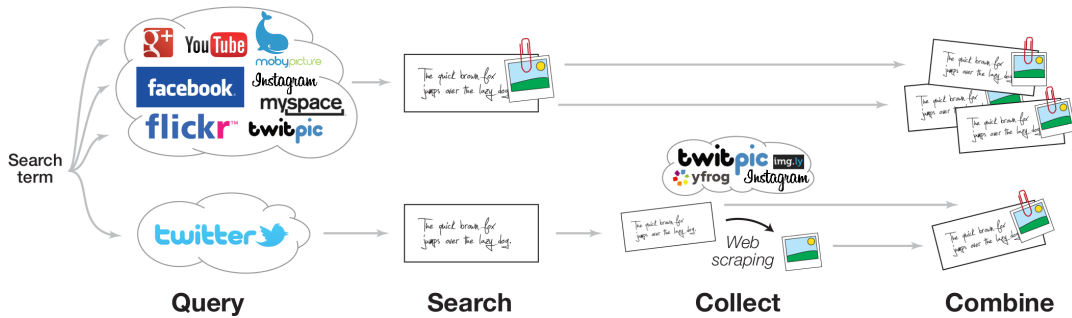
Figure 4.1: The media collector architecture: it proposes a hybrid approach for the media item extraction process using a combination of API access and Web scraping.

a search term and a parallel key-search is then performed to these social networks. Each platform has a 30 second timeout window to deliver its results. When the timeout has expired, or when all social networks have responded, a unified output is delivered [148, 116, 181]. Figure 4.1 depicts the overall architecture of the media collector.

The metadata attached to the microposts retrieved varies in terms of schemas, data types and serialization formats. We harmonize these results and project them to a common schema described below. This component performs also a cleansing process, discarding items which are older than seven days ago, in order to keep only fresh media items. Besides this abstraction layer on top of the native data formats of those social networks, we perform a similar task for the social interactions (Table 4.2 [181]).

| Likes | Shares | Comments | Views |
|---|---|---|---|
| Facebook Like | Facebook Share | Facebook Comments | YouTube Views |
| Google+ +1 | Google+ Share | Google+ Comments | Flickr Views |
| Instagram Like | Twitter ReTweet | Instagram Comments | Twitpic Views |
| Flickr Favorite | | Twitter RT, @Replies | MobyPicture Views |
| YouTube Like | | Twitpic Comments | |
| YouTube Favorite | | MobyPicture Comments | |
| Twitter Favorite | | Flickr Comments | |

Table 4.2: Abstract social network interaction paradigms and their underlying native counterparts [181]

Media Collector provides not only a way to capture a snapshot at a particular instant of what has been shared in social media platforms, but enables also to monitor the results of a search query over a longer period of time, by automatically re-issuing the same query at a regular frequency and by cumulating the results as explained in [114].

**Query and Response Format** Different settings are possible depending on

the type of search performed. On the one hand, we can perform focused searches, based on white lists made available via dedicated extractors, like LinkedTV's one (see Section 4.2.2) that allows to search customized content for a S&V or RBB providers. On the other, we can query fresh media items on social networks through different extractors working over their APIs. Those functionalities are exposed via the query pattern: http://linkedtv.eurecom.fr/api/mediacollector/search/TYPE/TERM where:

- TYPE is one of [RBB, SV, freshMedia, combined] and

- TERM is the search term.

The resulting behavior is:

- **RBB**: returns results from the RBB white list. The IRAPI module (see Section4.2.2) provides media items crawled from numerous white-listed Web sites, while we also query some APIs offering white-listed content namely the YouTube API (on specific channels) and the Arte replay programs API (arte+7).

- **SV**: returns results from the Sound and Vision white list. Similarly as RBB, we query the unstructured search module and YouTube white-listed channels.

- **freshMedia**: returns fresh media items from the 12 social platforms.

- **combined**: combines results from all possible sources.

The Media Collector produces results into a unified JSON structure that first contains *sources* and then *items* for each source:

```
{
  www.ard.de: [],
  Arte+7: [],
  TwitterNative: [],
  Instagram: [
      {...},
      {...},
      {...},
      {...},
      {...},
      {...}
      {...},
  ],
  YouTube: [],
  FlickrVideos: [],
  Flickr: [],
```

```
  MobyPicture: [],
  TwitPic: [],
  Lockerz: [ ],
}
```

An *item* is described using the following attributes:

- **mediaUrl**: deep link to the media item itself.

- **posterUrl**: URL of a thumbnail (miniature) of the media item.

- **micropostUrl**: URL of the Web document where the item is embedded (permalink), e.g. a YouTube page.

- **micropost**:

    - **html**: text description of the micropost in the original marked-up format.

    - **plainText**: cleaned text description of the micropost where the markup and some characters are removed.

- **userProfileUrl**: URL of the user that has published the micropost.

- **type**: type of the media item (photo or video).

- **timestamp**: reference time when the micropost was authored or the media item was uploaded (date in UNIX timestamp format).

- **publicationDate**: publication date using human-friendly syntax ("yyyy-MM-dd'T'HH:mm:ss'Z'").

- **socialInteractions**:

    - **likes**: total number of likes, +1 or hearts for this media item.

    - **shares**: total number of re-shares or re-postings of this media item.

    - **comments**: total number of comments made about this media item.

    - **views**: number of times this item has been viewed.

### 4.2.2  Customized Collection Services: IRAPI

The IRAPI Module (also called Unstructured Search Module or USM) has been designed for the retrieval of enrichment content from a curated list of Web sites. This collection component performs the crawling and indexing of Web sites that have been previously selected as a "white list", which ensures that only results from credible Web sites are returned.

**Hand-crafted Wrappers** Apache Nutch ([nutch.apache.org](nutch.apache.org)) is used as a highly extensible and scalable Web crawler. Apache Nutch can run on a single machine, but it can also run in a cluster (Apache Hadoop). Apache Hadoop allows for the distributed processing of large data sets across clusters of computers. Hadoop, Nutch core and all their modules and plugins are written in Java.

Nutch, by default, does not offer features for multimedia parsing. To the best of our knowledge, there is no plugin that would allow an easy integration and retrieval of multimedia content. As a consequence, a dedicated extension was developed for the purposes of enriching media with other media content. This leads to substantial changes in the index structure, because Nutch follows a single Web page - single index entry paradigm, which no longer holds if images, podcasts, and videos featured on a page become objects of standalone importance.

The main class of the developed plugin is *MediaParseFilter*. This class overrides the default Nutch extension point *ParseFilter*. This can be considered as entry point for every Web page to be parsed. In its overridden method called "filter", this class searches for multimedia on the Web page. Additionally, it provides additional meta data about the Web page. At the beginning of every parse step for a single Web page, lists of Document Object Model (DOM) managers and taggers are created.

The result of the method is a set of multimedia objects. Each of these objects is saved as a separate document to the index (during indexing phase), and assigned with a document representing the HTML page within which the document was located. This plugin performs also a first filtering of some multimedia objects such as logos and banners. The current implementation uses a blacklist of names, which, if present in the multimedia object's file name, cause the object to be skipped.

**Metadata Extraction Service (MES)** The other main component behind those steps is the Metadata Extraction Service (MES), which extracts supplementary metadata information for an identified media object. The relevant media objects include videos, podcasts and images. The textual metadata includes text in Web pages containing the titles or descriptions of the metadata objects.

To extract metadata describing media objects in Web pages, IRAPI adopted the approach of *extraction ontologies* (EO). The method was first introduced by Embley [39] and it consists in augmenting a domain ontology with extraction knowledge that enables automatic identification and extraction of references to ontology concepts in text.

In their basic form, extraction ontologies define the concepts, the instances of which are to be extracted, in the sense of various attributes, their allowed values as well as higher level (e.g. cardinality or mutual dependency) constraints. Extraction ontologies are assumed to be hand-crafted based on observation of a sample of resources but are often suitable for intra-domain reuse. They have been primarily applied to the extraction of records consisting of textual attribute-values like product

descriptions from heterogeneous HTML Web pages, but are also applicable to other formats of text documents.

A key benefit is that extraction ontologies provide immediate semantics to the extracted data, alleviating the need for subsequent mapping of extracted data to a domain ontology. At the same time, they allow for rapid start of the actual extraction process, as even a very simple extraction ontology is likely to cover a sensible part of target data and generate meaningful feedback for its own redesign; several iterations are of course needed to obtain results in sufficient quality.

The model consists of a single class titled "MediaRecord" which encapsulates a single occurrence of a *media* object, exactly one occurrence of its textual *title* found nearby, and optional multiple occurrences of further textual *descriptions* and mentions of *dates* occurring near the extracted media object. The media object is extracted in the form of a fragment (subtree) of the analyzed HTML document (e.g. the corresponding <img >, <video >or <embed >tag along with its contents).

**Indexed Web Sources** To take advantage of MES and the hand-crafted wrappers simultaneously, the following approach has been implemented. The metadata identified by the hand-crafted wrappers are saved to the original index fields (title and description), the output of the new MES module is saved into a separate field. All of the fields are used for retrieval. While MES supports also other media types, the initial release focuses on video, which is the most significant enrichment media type and at the same time the most difficult one to extract metadata from due to the variety of ways videos are embedded into Web pages. All this logic has been incorporated into the so called focused video crawler module. The purpose of the focused video crawler is to index documents that are relevant to queries issued to IRAPI. The focused video crawler is based on the assumption that for each scenario, there are several high priority Web sites which can be assumed to contain multiple relevant results for a significant portion of enrichment queries.

The Web sites covered by the focused video crawler are RBB Mediathek and ARD Mediathek (RBB use case) and avro.nl (SV usecase). These Web sites are too large to crawl exhaustively, which results in IRAPI retrieving only a portion of relevant content. Additionally, the search results can omit recently added items. To address this issue, the focused crawler wraps the video facet search facility that these large Web sites offer. Using the on site search, the crawler identifies Web pages embedding video that are relevant to the query issued to IRAPI. These are crawled in a priority queue, which indexes them typically within minutes of the original user query.

The focused video crawler is a separate service, which is invoked each time the IRAPI component of IRAPI receives a query. The system first checks the query against the history: if the same query was issued in a predefined history window, it is believed that up-to-date results are already in the index. Otherwise, the supported on-site search interfaces are queried using the video facet, and the top N results

(where N is pre-specified parameter) are saved to the index along with the MES extraction results.

**REST API Queries** IRAPI is available via a Web Service interface at https://ir. lmcloud.vse.cz/irapi/media-server/. The URI scheme for calling the API is composed of:

- query argument *q* for Lucene query,

- query argument *row* defining a number of items to be retrieved (by default 10).

Wildcards *?* and *\** replace one or more characters in parameter *q*. The common boolean operators *AND*, *OR*, *NOT* are also available. It is possible to use *[ from TO to ]* syntax to form range queries. The symbol *\** in ranges may be used for either or both endpoints to specify an open-ended range query.

- *field:[\* TO 100]* matches all field values less than or equal to 100

- *field:[100 TO \*]* matches all field values greater than or equal to 100

- *field:[\* TO \*]* matches all documents with the field present

Range queries are also applied on timestamp* fields. The complete *ISO 8601* date syntax that fields support, or the *DateMath* Syntax to get relative dates. For example to get all documents crawled since August 22, 2013:

```
parse_time:[2013-08-22T07:33:21.699Z TO *]
```

Other features exposed are: 1) specifying that queried words should be found within a specific distance on the text (*content:"Berlin parliament"~2*), 2) considering similar terms via string edit distances (*media_title:meine~0.7*), or 3) boosting certain fields when performing the search (*(media_title:\*Berlin\*)^5 (media_description:\*Parlament\*)^0.5*).

**Rest API Responses** The output of IRAPI follows the same JSON format implemented by MediaCollector (Section 4.2.1.2), with the following exception: the *type* field inside each item can be set to a wider list of values: [photo — video — audio — webpage]. One example coming from the crawled Web site www.ard.de:

```
www.ard.de: [
{
        micropostUrl: "http://www.ard.de/",
        micropost: {
                html: "Woller und Schwester Hanna (Bild: ARD/Barbara Bauriedl)",
                plainText: "Woller und Schwester Hanna (Bild: ARD/Barbara Bauriedl
                    )"
        },
        mediaUrl:
            "http://static.daserste.de/cmspix/tvtipp/thumb_128_72_05022013187955.
                jpg",
        type: "photo",
        timestamp: 1398772232014,
```

```
        publicationDate: "2013-08-29T13:50:32Z"
  }
  ]
```

### 4.2.3   Enriching Television Content with TVEnricher

In this section, we will present how television content ingested and serialized according to the LinkedTV ontology as explained in Section 2.4 is enriched with other media resources extracted from external platforms, either coming from white-listed Web sites or social networks. The corresponding logic has been published as separate Web service called *TVEnricher*[3] Web service, developed under the scope of this thesis. In a nutshell, TVEnricher detects suitable anchors within media resources for the enrichment based on volume and frequency of named entities detected in media fragments. The named entities become query terms that are used by the Media Collector[4].

Before TVEnricher enters in action, the different description files about a particular media document are converted into RDF and represented according to the LinkedTV Ontology via the REST API service *tv2rdf*[5]. The instances of the *MediaFragment* have attached various entities describing what is being told in that part of the video. From this initial RDF, TVEnricher starts its processing following a workflow that can be summarized like this:

- The enrichment process is launched by providing a valid media resource identifier (UUID) to TVEnricher. A pre-requisite is that the annotation process of the corresponding media resource is completed in TV2RDF, and the resulting RDF data has been already pushed to a particular the triple-store.

- Given the provided media resource identifier, TVEnricher accesses the triple-store for retrieving all the entities identified by a linked data URI (e.g. the ones spotted by NERD) for this television program. From the set of entities retrieved from the graph, TVEnricher performs a rank and filter operation in order to promote a smaller set of relevant entities to be enriched. In the current version, the ranking operation is based on a pure TF approach consisting in giving more importance to the entities who are mentioned the most inside that particular media fragment

- We iterate over the set of top ranked entities like follows: a search operation on the MediaCollector is triggered by using as input the label of the entity. By default, the so-called *combined* strategy is launched in order to involve as

---

[3]http://linkedtv.eurecom.fr/tvenricher/api/
[4]http://linkedtv.eurecom.fr/api/mediacollector/
[5]http://linkedtv.eurecom.fr/tv2rdf
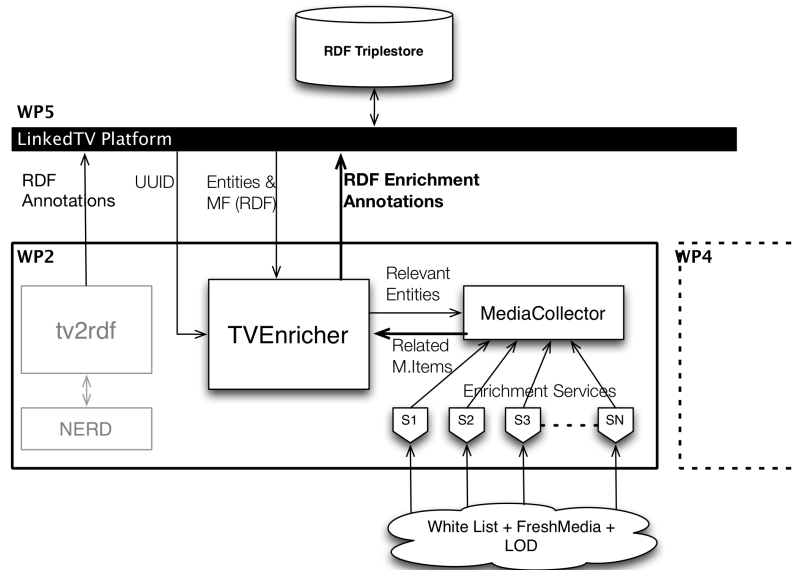
Figure 4.2: Diagram showing the role of the TVEnricher service within the LinkedTV Platform

many media sources as possible, including items from white-listed Web sites and fresh media resources from social networks. For example being the entity "S-Bahn" one candidate for enrichment, TVEnricher generates the query: http://linkedtv.eurecom.fr/api/mediacollector/search/combined/S-Bahn in MediaCollector, obtaining as result a list of media resources (photos and videos) grouped by source.

- We serialize the results into RDF (see Section 4.2.5). When serializing the information, every item returned by MediaCollector is represented as a new 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, as proposed in the Open Annotation Ontology. At the same time we align the results with a media fragment of a certain granularity (`linkedtv:Chapter`, `linkedtv:Scene`, `linkedtv:Shot`) inside the seed video according to the temporal proximity of the media fragment with the entity that triggered the enrichment. All the enrichment results are stored into a single Turtle file that can be pushed back to the original triplestore.

The Figure 4.2 illustrates this workflow while Figure 4.2 summarizes all the main services involved in the processing chain.

The sequence diagram in Figure 4.3 illustrates how the different components play their role over time for the LinkedTV's enrichment workflow. The conversation is initiated by providing the UUID of a media resource to be enriched. Immediately after, TVEnricher answers back for requesting an excerpt for the RDF data containing
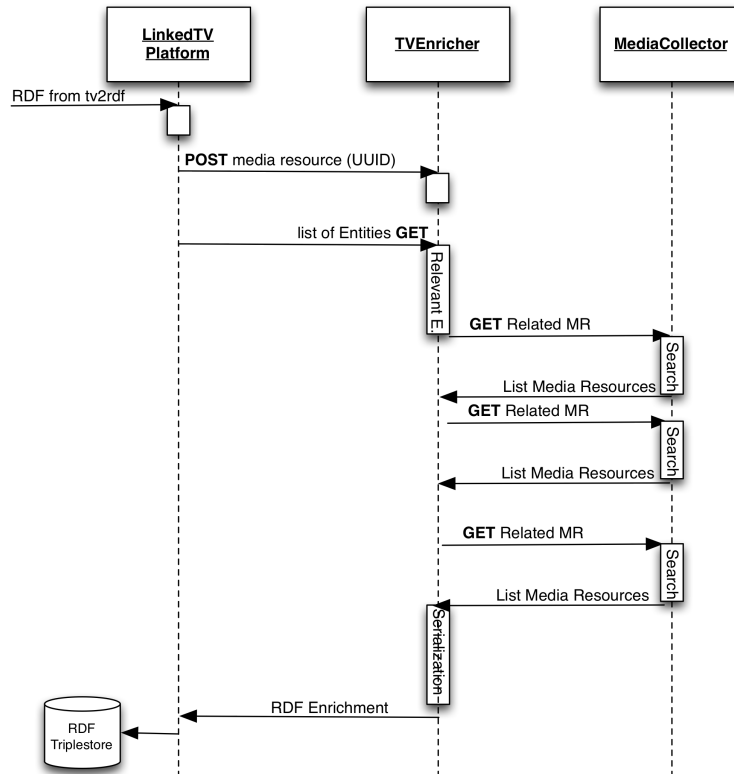
Figure 4.3: Sequence diagram of the enrichment serialization of a Media Resource by TVEnricher

the entities to be used as anchors and the media fragments to align related media items to. As it is illustrated in the aforementioned figure, the main interaction occurs between the TVEnricher and the Media Collector for collecting media items related to an entity. The latency can be more or less long depending on the duration of the seed video and the number of relevant entities detected.

Once the metadata about a particular content has been gathered, serialized into RDF, and interlinked with other resources in the Web, it is ready to be used in the subsequent consumption phases like the editorial review or data display. The creation of a MediaFragments hierarchy with different levels of granularity provides a very flexible model to (1) easily incorporate new data describing the media resource and (2) allowing different interpretations of the available information depending on the final user and the particular context.

In future work we have planned to extend TVEnricher by better leveraging on other media annotations available when launching the enrichment. Up to now, the only anchors that trigger the operation are the labels of the top N relevant entities. In the future, it could be interesting to use other available metadata such as LSCOM visual concepts, faces, or keywords, separately or in aggregated manner, for obtaining
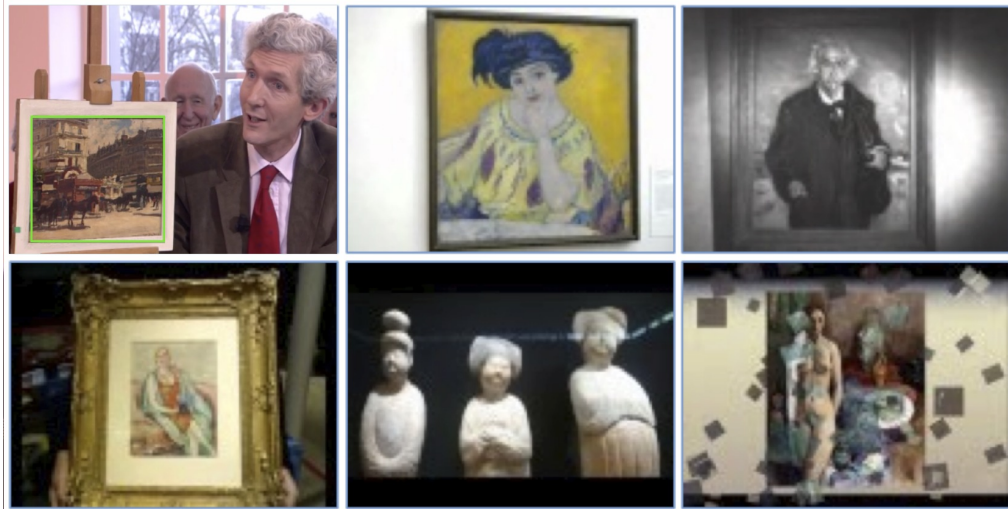
Figure 4.4: List of media items retrieved from MediaCollector service for the search term "Jan Sluijters".

more precise results. In the same way, the results from the enrichment services can be combined with the already existing annotations like spatial objects and named entities for obtaining new insights about what is happening in the video. The upper left image in Figure 4.4 illustrates a painting, detected by the object re-detection algorithm and highlighted with a green bounding box, that appears in the Tussen Kunst en Kitsch show, between the 1492nd and 1504th second. Looking for information attached to temporarily similar MediaFragments in the model, there is an entity about the artist "Jan Sluijters" that is mentioned from the second 1495 to 1502. Media items retrieved for the entity "Jan Sluijters" in TVEnricher can be aligned with the object detected by visual analysis techniques and mark all them as paintings created by this author. Similar deductions can be done by relying in other items in the model like keywords and LSCOM concepts.

**REST API**. In the current version of TVEnricher the main REST supported call is the one for triggering the enrichment of a media resource given its UUID inside ta particular RDF knowledge base. The call includes two query parameters: (1) the granularity of the media fragments with which enrichments will be associated to, and (2) the namespace used in the instances of the dataset where the RDF representation of media resource to enrich is stored.

```
curl "http://linkedtv.eurecom.fr/tvenricher/api/mediaresource/UUID\_media\
    _resource/enrichment?granularity=Granularity&namespace=http://data.linkedtv.eu
    "  --header "Content-Type:text/xml" -v;
```

- UUID = ID of a valid media resource inside ENDPOINT // [mandatory]

- endpoint = URL of the SPARQL endpoint where the media fragments annotations are available. // [optional, by default http://data.linkedtv.eu/sparql]

- namespace = base URI for creating the enrichment RDF instances // [optional, by default http://data.linkedtv.eu/]

- broadcaster = token indicating the company authoring the multimedia content // [optional, values: [SV, RBB], by default SV]

- granularity = the level of granularity of the media fragment where the enrichment content will be attached to. // [optional, values: [Chapter, Shot], by default Shot]

- graph = the RDF graph inside ENDPOINT where the media resource is located. [optional, e.g. http://data.linkedtv.eu/graph/linkedtv, by default ALL graphs in the ENDPOINT]

If the UUID corresponds to a media resource that does not exist in ENDPOINT, the enrichment will be registered in TVEnricher with a status value of `Error`. Every time a media resource parameter is changed, the enrichment is automatically reprocessed again.

This API call shows the status information for a particular media resource that has been or is been processed by TVEnricher. Apart from the parameters specified during the creation, it is also possible to check the "state" flag value, and some details about the serialization file in the case the enrichment process has already finished.

```
GET \url{http://linkedtv.eurecom.fr/tvenricher/api/mediaresource/UUID}
```

- UUID = ID of a valid media resource inside ENDPOINT // [mandatory]

Some possible values of the "state" flag value depending on the current state of the enrichment are:

- Processing: The media resource has been already uploaded on the REST service, but the enrichment process is still under execution. A later call will be needed to assess when the enrichment is finished and available to be downloaded.

- Processed: The media resource has been successfully enriched, the property STATE changes to processed and a link to the RDF file is included in the JSON serialization under the field `enrichment`.

- Error: In case of non-existent media resource, problems during enrichment process, or exceptions during the RDF conversion.

Once the media resource enrichment gets to state "processed" we are in situation of getting the enrichments which have been previously computed for a particular media resource via the following call:

```
GET \url{http://linkedtv.eurecom.fr/tvenricher/api/mediaresource/UUID/enrichment}
```

- UUID = ID of a valid media resource inside ENDPOINT // [mandatory]

If a client tries to retrieve the enrichment of a media resource where "state" is equal to `Error` or `Processing` , a 404 Not Found error code is returned as a response.

Finally, TVEnricher can enrich not only complete media resources but also be invoked on-demand using some particular textual search terms. The following REST call generates a set of enrichments based on a term or a list of query terms (comma separated), and returns results serialized in JSON according to MediaCollector's output format described in previous section.

```
GET \url{http://linkedtv.eurecom.fr/tvenricher/api/entity/enrichment/RBB?q=Obama}
```

- q = query string

- strategy = allows to select between different sets of enrichment sources [combined, freshMedia, Europeana, RBB, SV]

### 4.2.4    Enriching Fragments with News Documents

In previous sections we have shown how media fragments annotated with with different visual and textual techniques can be enriched with other media content that further describe, complement, or refine what is being said in the original resource. In this section we will illustrate how certain fragments, particularly in the news domain, can be also contextualized and enriched with textual documents talking about similar matters.

The way this enriching approach has been tackled is different than the ones previously covered because it intends to be more targeted to a particular kind of content (in this case, news items) and it is organized around different enrichment dimensions identified by experts in the domain (such as same news in other journalistic sources, opinion articles, and other categories explained in Section 4.2.4.1). This higher specialization enrichment allows to better address particular needs of the domain, at the cost of higher efforts for setting up and properly configure the enrichment workflow. However, in different studies [140] have revealed that in certain scenarios such as the news domain, we need a better targeted, editorial reviewed content. In the approach we have developed we ensure this customization level via two mechanisms: (1) using particular types of named entities from the seed video for triggering relevant searches, (2) perform searches over pre-defined white lists of Web sources covering the enrichment dimensions specified by user studies.

In order to index and expose query functionalities over those white list, we have relied on the search tools provided by Google via their Custom Search Engine[6] (CSE).

---

[6]https://cse.google.com/cse/

This service allows to specify a list of Web sources that are automatically indexed by Google, that can be queried afterwards via a dedicated API[7]. Apart of the query terms to launch the search, we can leverage on different parameters for further tuning the retrieval process, such as considering results from the whole Web when the white list sources do not provide relevant enough results, selecting a date interval to filter results, or restricting the Web documents indexed to those including particular semantic annotations according to Schema.org vocabulary. For example in the case of the news domain and apart of other configurations considered in Section 4.2.4.1, we have set up a CSE over a list of ten worldwide famous international news papers in English[8] that allow us to obtain better curated documents from trustable sources as enrichment.

The input of this enrichment service will therefore be, a list of concatenated entities' surface forms generated from the seed news video's annotations, a date indicating the time period where the displayed facts were relevant and a particular enrichment dimension selected from the set of white lists specified by domain experts. The output will be a set of Web documents from the specified sources, ordered by relevance and ready to be exploited by different applications assisting viewers in consuming the content.

In the following subsections we will further describe the way we can set up an enrichment workflow for the news scenario (Section 4.2.4.1), and how the corresponding logic has been wrapped into a Web Service called TVNewsEnricher, which can programmatically provide enrichment to the different identified dimensions described in Section 4.2.4.2.

### 4.2.4.1   Use Case: Enriching International News

In the paper [144] we proposed a second screen application for assisting the users in consuming international news. The prototype was designed according to different guidelines and requirements derived from user studies on the domain [139]. In order to feed the GUI with the adequate enrichments in a timely matter and particularly the so called "active" mode (see Figure 4.5), we have relied on the document based enrichment approach explained in this section. The active mode of the application acts as a hub where the viewers can access extra documents for complementing what is being told in the main news video. A logic for generating the appropriate query terms per each video is applied over the main entities coming from the expansion process (see Section 3.2.3) and afterwards injected into the adequate Google CSE engine. This logic works relaxing or promoting some particular entity types according to the *W's* of the journalism [98] and depending on the desired dimension. In contrast to

---

[7]https://developers.google.com/custom-search/json-api/v1/reference/
[8]http://www.4imn.com/top200/

traditional news aggregators that simply gather related documents, the results are organized around five different axes that intent to fulfill the viewer' needs.
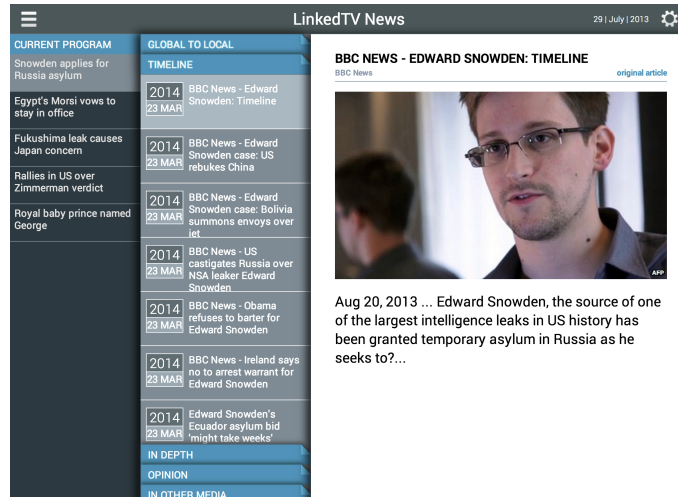


Figure 4.5: Active mode for news consumption as implemented in LinkedTV demo

**Timeline.** Follows the news story throughout time by confectioning an list of ordered documents that includes the main antecedents of the present facts. For getting those document we rely on a query without any prior time constraint, which is created when including the most relevant entity from the Who, Was, Where inside the pattern "The" + entity + "case".

**In other sources.** This section of the interface is dedicated to showing the selected news as it was reported in other newspapers, radio, or TV programs. We launch a query generated from the set of expanded entities by following exactly the same logic used during the entity expansion (see Section 3.2.4.1), over the curated list of resources including the top english journals introduced before.

**Opinion.** This section is devoted to gathering opinions regarding the selected news item from different authors with a certain presupposed knowledge about the matter. The list of documents is obtained by executing the same query generated for dimension "In other Sources", but operating over a different list of curated resources that considers only subdomains specialized in opinion documents, like[9].

**Geo-localized information.** This section includes live feeds from Twitter API expressing people's remarks, comments and feelings filtered by subject and geo-location. We use the same textual query than in "In other Sources" dimension, but reducing the temporal dimension $t$ to the last 7 days from the current time, in order to see what people are thinking about the particular news item.

**In depth.** Includes in depth coverage articles that offer a more extensive view about the seed news item. The documents under this dimension are obtained by

---

[9]http://www.nytimes.com/pages/opinion/index.html

combining the most relevant entity from the Who, Was, Where with the keyword "in depth" and removing any temporal restriction and extending the search domain to the entire Web. In our example the textual will be "Edward Snowden in depth";

### 4.2.4.2   REST API Service: TVNewsEnricher

TVNewsEnricher is a public REST API service for enriching news items with online articles and media from the Web in five different dimensions, following the process described in previous section.  This collection process is performed along the five aforementioned axes:  Opinion, OtherMedia, TimeLine, InDepth and geolocalized data from Twitter. The results of each dimension can be invoked via separate REST API calls as it is further explained below. The service is available at: http://linkedtv.eurecom.fr/newsenricher/api/.

This service makes use of the search capabilities from Google CSE and Twitter API[10]. It is intended to be plugged over the results obtained from the Name Entity Expansion service.  A live demo displaying the related documents found by this service is available at http://linkedtv.project.cwi.nl/news/auto/.  The four CSE based dimensions are available via the following REST API calls:

#### Opinion Dimension

```
API call: GET \url{http://linkedtv.eurecom.fr/newsenricher/api/opinion?query=TERMS
    &startdate=START&enddate=END&cse=CSE&limit=50}
```

#### Other Media Dimension

```
API call: GET \url{http://linkedtv.eurecom.fr/newsenricher/api/othermedia?query=
    TERMS&startdate=START&enddate=END&cse=CSE&limit=50}
```

#### Timeline Dimension

```
API call: GET \url{http://linkedtv.eurecom.fr/newsenricher/api/timeline?query=
    TERMS&startdate=START&enddate=END&cse=CSE&limit=50}
```

#### In Depth Dimension

```
API call: GET \url{http://linkedtv.eurecom.fr/newsenricher/api/indepth?query=TERMS
    &startdate=START&enddate=END&cse=CSE&limit=50}
```

where:

- TERMS: query string with + separated entity labels

- START: start date in the format YYYYMMDD, mandatory

- END: end date in the format YYYYMMDD, optional, by default, the current date

- limit: maximum number of documents to be retrieved, optional, by default 10

---

[10]https://dev.twitter.com/rest/public/search

- CSE: ID of the Google custom search engine to be used for collecting related documents

The last dimension **Related Tweet Dimension**, based on Twitter API, can be reached at the following URL:

```
API call: GET \url{http://linkedtv.eurecom.fr/newsenricher/api/tweets?query=TERMS&
    startdate=START&enddate=END&lat=LAT&lon=LON&rad=RAD&limit=50}
```

where:

- TERMS: query string with + separated entity labels

- START: start date in the format YYYYMMDD, mandatory

- END: end date in the format YYYYMMDD, optional, by default, the current date

- LAT: latitude in degrees for geolocalized tweets, optional

- LON: longitude in degrees for geolocalized tweets, optional

- RAD: radius in km, optional, by default 5 km

- limit: maximum number of documents to be retrieved, optional, by default 10

### 4.2.5   LinkedTV Ontology for Representing Enrichment Results

In this section we show how the ontology model for describing multimedia content presented in Section 2.4 can be also used as means for representing the enrichment results such as the TVEnricher output in Subsection 4.2.3. Having in mind to keep the solution as simple as possible, those are the main RDF modeling decisions made:

- The list of media resources retrieved after each Media Collector search operation are modeled as instances of the *ma:MediaResource* class from the Ontology for Media Resources. As the ontology by itself does not cover entirely the set of attributes considered in the MediaCollector JSON schema, we have used additional properties coming from the Dublin Core Metadata Element Set[11] and the LinkedTV Ontology. The Table 4.3 summarizes the mappings between both serialization formats.

- Those media resources considered as enrichment results are attached to the particular temporal fragment where the seed entities that triggered the enrichment query have been taken from. In order to do so we identify the instance of the *ma:MediaFragment* class (annotated with a desired level of granularity:

---

[11]http://dublincore.org/documents/dces/

*linkedtv:Chapter*, *linkedtv:Shot*, or *linkedtv:Scene*) for which the entity's temporal boundaries are closer to the interval defined by *nsa:temporalStart* and *nsa:temporalEnd*.

- An instance of the class *oa:Annotation* is generated for explicitly linking the collected media resources with the enriched media fragment. Apart from various provenance information according to the PROV[12] ontology, those annotations include two properties that are worth to be mentioned: (1) *prov:wasDerivedFrom* that indicates the instance of the class *linkedtv:Entity*, which triggered the Media Collector search operation, and *oa:motivatedBy*, which specifies the reasons why the annotation was created (in our case its value is set to `oa:linking` by default).

Table 4.3: Properties inside `ma:MediaResource` instances for representing Media Collector's JSON attributes

| JSON attribute | Ontology property inside LinkedTV |
|---|---|
| mediaUrl | ma:locator |
| posterUrl | linkedtv:hasPoster |
| micropostUrl | dc:isPartOf |
| plainText | dc:description |
| userProfileUrl | dc:creator |
| type | dc:type |
| timestamp | dc:date |
| socialInteractions | linkedtv:hasSocialInteraction |
| likes | linkedtv:likes* |
| shares | linkedtv:shares* |
| comments | linkedtv:comments* |
| views | linkedtv:views* |

The Figure 4.6 illustrates this process with an example of an enrichment serialization operation.

In order to access enrichment information serialized according to the LinkedTV model, we can perform SPARQL queries over a triple-store where the generated annotations are published. In order to know more about how to consume this graph, you can see the queries included as example in Appendix 5.

#### 4.2.5.1 Scenario 1: Enriching Cultural Heritage

In this Section we show the RDF generated out of the enrichment results obtained for the one episode of the dutch TV program Tussen Kunst & Kitsch (Antiques Roadshow) which is offered by the public broadcaster AVRO, available at http://web.

---

[12]http://www.w3.org/TR/prov-o/
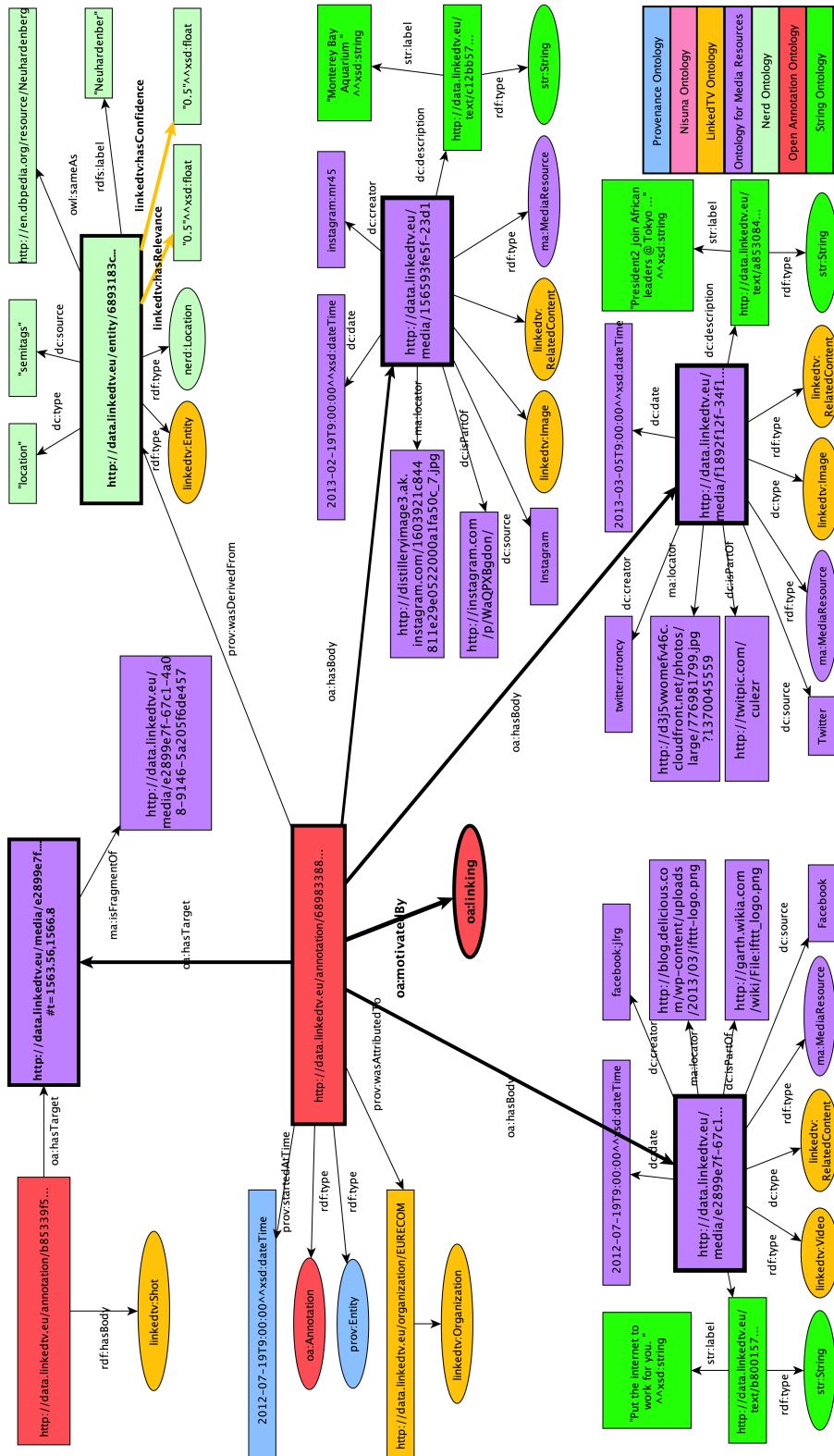
Figure 4.6: Instances involved in the RDF serialization of a media resource's enrichment

avrotros.nl/tussenkunstenkitsch/player/POMS_AVRO_677886/. Considering the media
resource has been already serialized in TV2RDF as explained in Section 2.4.2 and we
have access to its unique identifier (UUID), in this case `8a8187f2-3fc8-cb54-0140-7dd151100003`,
we will issue the following request on TVEnricher:

```
curl "http://linkedtv.eurecom.fr/tvenricher/api/mediaresource/8a8187f2-3fc8-cb54
    -0140-7dd151100003/enrichment?granularity=Shot&namespace=http://data.linkedtv.
    eu"  --header "Content-Type:text/xml" -v;
```

The service first retrieves the list of entities spotted by NERD for that particular
video. TVEnricher will list those entities in order of relevance by considering the
number of times they have appeared over the entire show. In our example, those are
the labels of the top 8 ranked entities are: Amsterdam, Jan Toorop, Nederland, Parijs,
Dat, Leuk and Tholen. We will take as example the RDF enrichment generated for
the named entity **Jan Toorop**. The first step to be done is to inject a search
query in Media Collector in order to start the collection of media relevant resources
for that particular search term: http://linkedtv.eurecom.fr/api/mediacollector/search/
SV/Toorop.

- **Named entity**: Toorop

- **Query**: http://linkedtv.eurecom.fr/api/mediacollector/search/SV/Toorop

- **Media content returned from API**:

  - source: YouTube
    type: video
    url : http://www.youtube.com/watch?v=csVdymOUyNA
  - source: www.geschiedeniszeeland.nl
    type: photo
    url: http://www.geschiedeniszeeland.nl/topics/kunstindomburg_18.jpg
  - source: www.geschiedeniszeeland.nl
    type: photo
    url: http://www.geschiedeniszeeland.nl/topics/kunstindomburg_22.jpg
  - source: www.boijmans.nl
    type: photo
    url: http://www.boijmans.nl/images/pages/slides/72/2090_MK_De_Theems_
    bij_Londen_van_Jan_Toorop_web.jpg

Once the JSON response containing all the retrieved items is available, we start
the RDF serialization of all the aspects concerning that set of media resources. First,
an instance of the class *oa:annotation* is generated for that particular entity. This
annotation has for *oa:hasBody* the list of media resources found by the Media Col-
lector and different provenance information such as the date of enrichment creation,
the tool generating the enrichment, etc.

```
<http://data.linkedtv.eu/annotation/a2664015-3b62-47aa-b999-b3069a9054c0>
      a        oa:Annotation , prov:Entity ;
      oa:Motivation oa:linking ;
      oa:hasBody <http://data.linkedtv.eu/media/5ee11dc8-38cc-4034-807e-5
          eabdfb5bc84> , <http://data.linkedtv.eu/media/8b97a931-f7ea-4370-914a
          -73019eed62df> , <http://data.linkedtv.eu/media/f1cc03fe-be2e-41c3-bb9f
          -3bf9b77683c7> , <http://data.linkedtv.eu/media/263fddda-c7a7-47a0-bd6f
          -99be810874b1> , <http://data.linkedtv.eu/media/6415bd27-b025-4924-8e84-
          a256212bfc85> , <http://data.linkedtv.eu/media/4ea709f1-f11e-4af4-be85
          -7610b3b50633> , <http://data.linkedtv.eu/media/e3d59d8a-0a29-4fc5-80b0
          -34a31c4308af> , <http://data.linkedtv.eu/media/de89c3f1-e6f8-4d08-aabf
          -2a7a1a7f9e15> , <http://data.linkedtv.eu/media/b35fa154-e8b2-4b0b-8a21-
          aebc264072d6> , <http://data.linkedtv.eu/media/fb24909d-b765-4040-bacf
          -85b6345dc353> ;
      oa:hasTarget <http://data.linkedtv.eu/media/a8187f2-3fc8-cb54-0140-7
          dd151100003#t=785.56,790.56> ;
      prov:startedAtTime "2013-09-14T00:33:30.788Z"^^xsd:dateTime ;
      prov:wasAttributedTo
            <http://data.linkedtv.eu/organization/EURECOM> ;
      prov:wasDerivedFrom "http://data.linkedtv.eu/entity/226ba0ff-537a-485d-bbfa-
          f356eba105a9" .
```

As in the RBB use case, the property *prov:wasDerivedTo* points to the *linkedtv:Entity* instance which has triggered the enrichment process so we have a direct access to other possible information such as the text where that entity was retrieved from. The property *oa:motivatedBy* is set to *oa:linking* for specifying the nature of the relationship, again a resource that is potentially related to the target.

```
<http://data.linkedtv.eu/media/a8187f2-3fc8-cb54-0140-7dd151100003#t
    =785.56,790.56>
      a        nsa:TemporalFragment , ma:MediaFragment ;
      nsa:temporalEnd "790.56"^^xsd:float ;
      nsa:temporalStart "785.56"^^xsd:float ;
      nsa:temporalUnit "npt" ;
      ma:duration "5.0"^^xsd:float ;
      ma:isFragmentOf <http://data.linkedtv.eu/media/8a8187f2-3fc8-cb54-0140-7
          dd151100003> .
```

An example of the serialization of the metadata corresponding to a video that is relevant to the entity "Jan Toorop" is shown below:

```
<http://data.linkedtv.eu/media/8b97a931-f7ea-4370-914a-73019eed62df>
      a        ma:MediaResource , linkedtv:RelatedContent ;
      linkedtv:hasPoster <https://i1.ytimg.com/vi/QLLRMEF9Bt4/default.jpg> ;
      linkedtv:hasSocialInteraction
            [ linkedtv:comments "5"^^xsd:int ;
              linkedtv:likes "24"^^xsd:int ;
              linkedtv:shares "0"^^xsd:int ;
              linkedtv:views "1180"^^xsd:int
            ] ;
      dc:creator <https://www.youtube.com/channel/UCKmKCcwzzFTL8HgsOhPu5mg> ;
      dc:date "2012-03-13T17:10:28Z"^^xsd:dateTime ;
      dc:description <http://data.linkedtv.eu/text/e08ff4a0-4bcc-44b1-bba4-8
          bdd7cffdf3b> ;
      dc:isPartOf <http://www.youtube.com/watch?v=QLLRMEF9Bt4> ;
      dc:source <http://data.linkedtv.eu/socialplatform/YouTube> ;
```
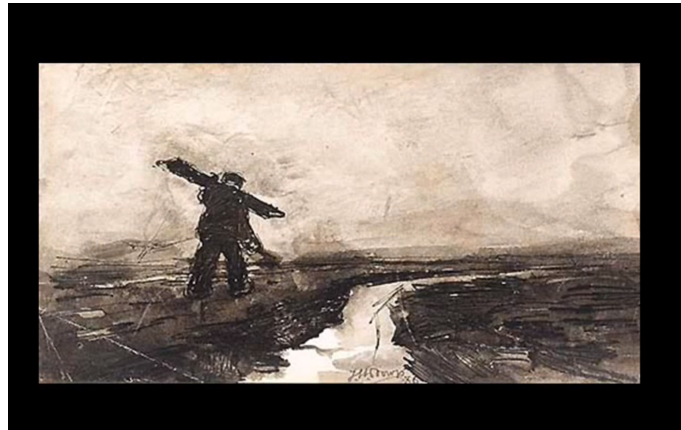
Figure 4.7: Painting of Johan and Mies Drabbe by Jan Toorop, 1898 (collection H.F. Elout), relevant to the named entity `Toorop` in the Sound and Vision scenario. Source: http://www.geschiedeniszeeland.nl/topics/kunstindomburg_22.jpg

```
dc:type linkedtv:Video ;
ma:locator <https://www.youtube.com/embed/QLLRMEF9Bt4> .
```

The Figure 4.7 depicts one of the media resources retrieved as an enrichment for the "Toorop" entity. This media item corresponds to a painting from this famous Dutch painter, who worked on various different styles such as Realism, Impressionism Post-Impressionism, Symbolist or Art Nouveau.

### 4.2.5.2 Scenario 2: Enriching News Items

In this Section we will show the resulting RDF of enriching one episode of the news show "RBB Aktuell" from the RBB german broadcaster, aired on the 19th of June of 2013. Considering the media resource has been already serialized in TV2RDF as explained in Section 2.4.3 and we have access to its unique identifier (UUID), in this case `b82fb032-d95e-11e2-951c-f8bdfd0abfbd`, we will issue the following request on TVEnricher:

```
curl "http://linkedtv.eurecom.fr/tvenricher/api/mediaresource/b82fb032-d95e-11e2
    -951c-f8bdfd0abfbd/enrichment?granularity=Shot&namespace=http://data.linkedtv.
    eu" --header "Content-Type:text/xml" -v;
```

In first place, the service retrieves the list of entities spotted by NERD for that particular video. TVEnricher will list those entities in order of relevance by considering to the number of times they have appeared over the entire show. In our example, the labels of the top 12 ranked entities are: Berlin, Insekten, Mücken, Barack Obama, Deutschland, Matthias Platzeck, S-Bahn, Maschine, Bornstedt, Freiheit, Krongut.

We will take as example the RDF enrichment generated for the named entity **S-Bahn**. The first step to be done is to issue a search query in Media Collector in order to start the collection of media resources in the different considered platforms: http:

//linkedtv.eurecom.fr/api/mediacollector/search/RBB/S-Bahn. Below we enumerate some items obtained from MediaCollector given the aforementioned query:

- **Named entity**: S-Bahn

- **Query**: http://linkedtv.eurecom.fr/api/mediacollector/search/RBB/S-Bahn

- **Media content returned from API**:

  - source: YouTube
    type: video
    url : http://www.youtube.com/watch?v=jHdpP2X9tE8

  - source: www.s-bahn-berlin.de
    type: webpage
    url: http://www.s-bahn-berlin.de/unternehmen/firmenprofil/historie.html

  - source: www.mdr.de
    type: photo
    url: http://www.mdr.de/sachsen-anhalt/halle/halleipzig100_v-standard43_zc-698fff06.jpg?version=54569

Once the JSON response containing all the retrieved items is available, we start the RDF serialization of all the aspects concerning that set of media resources. First, an instance of the class *oa:annotation* is generated for that particular entity. This annotation has for *oa:hasBody* the list of media resources found by the Media Collector and different provenance information such as the date of enrichment creation, the tool that has triggered the enrichment, etc.

```
<http://data.linkedtv.eu/annotation/2c60da24-d2fc-451a-888b-48f49bf0c75d>
    a        oa:Annotation , prov:Entity ;
    oa:hasBody <http://data.linkedtv.eu/media/656b2d1c-ded8-4e1f-a888-889
        beabeaff1> ,
             <http://data.linkedtv.eu/media/808c5ec3-4fd6-428d-8dbf-9
                 d7266184328> ,
             <http://data.linkedtv.eu/media/1d115295-87a5-40b5-a512-62
                 fbe1f60a98> ,
             <http://data.linkedtv.eu/media/809a7b79-c02e-4db6-ae5f-6
                 bbafee52a7c> ,
             <http://data.linkedtv.eu/media/b4d7cadb-d33b-495f-9aae-4526
                 cec3ba69> ,
             <http://data.linkedtv.eu/media/38f4aefc-b176-4d64-900f-8
                 ada4dc1a3c9> ,
             <http://data.linkedtv.eu/media/e54ff382-6247-47ef-9446-
                 ac222503dfc9> ,
             <http://data.linkedtv.eu/media/023211d5-3275-44a9-bbd4-
                 bd4b41e02a76> ,
             <http://data.linkedtv.eu/media/3f072470-4981-4cbd-9339-595
                 c501990ab> ,
             <http://data.linkedtv.eu/media/5876ff1d-9ce0-40f1-98f6-05
                 d7c2d9efb9> ,
```

```
              <http://data.linkedtv.eu/media/aa7bdf54-2853-410a-99a4-1
                  a9805c447ef> ,
              <http://data.linkedtv.eu/media/0e12a03b-5910-458f-b278-20
                  dd62097fd3> ;
    oa:hasTarget <http://data.linkedtv.eu/media/b82fb032-d95e-11e2-951c-
        f8bdfd0abfbd#t=962.36,964.84> ;
    oa:motivatedBy oa:linking ;
    prov:startedAtTime "2013-09-23T09:13:45.858Z"^^xsd:dateTime ;
    prov:wasAttributedTo <http://data.linkedtv.eu/organization/EURECOM/
        TVEnricher> ;
    prov:wasDerivedFrom <http://data.linkedtv.eu/entity/e81bb960-9e4e-4f6e-b084-
        e4bfbfca27ce> .
```

The property *prov:wasDerivedTo* points to the *linkedtv:Entity* instance which has triggered the enrichment process in order to have a direct access to other possible information such as the subtitle block where this entity has been spotted. The property *oa:motivatedBy* is set to *oa:linking* for specifying the nature of the relationship: an untyped link to a resource related to the target[13]. The target of the annotation (*hasTarget*) refers to the *ma:MediaFragment* instance to which those items will be attached to. The corresponding Turtle code for that instance is:

```
<http://data.linkedtv.eu/media/b82fb032-d95e-11e2-951c-f8bdfd0abfbd#t
    =962.36,964.84>
      a       nsa:TemporalFragment , ma:MediaFragment ;
      nsa:temporalEnd "964.84"^^xsd:float ;
      nsa:temporalStart "962.36"^^xsd:float ;
      nsa:temporalUnit "npt" ;
      ma:duration "2.4800415"^^xsd:float ;
      ma:isFragmentOf <http://data.linkedtv.eu/media/b82fb032-d95e-11e2-951c-
          f8bdfd0abfbd> .
```

All the media resources coming from the Media Collector are described in RDF as well. We use The Ontology for Media Resources for that purpose, as described in Section 4.2.5. An example for the entity "S-Bahn" is shown below:

```
<http://data.linkedtv.eu/media/e54ff382-6247-47ef-9446-ac222503dfc9>
      a       ma:MediaResource , linkedtv:RelatedContent ;
      linkedtv:hasPoster <https://i1.ytimg.com/vi/29aSPW6xltM/default.jpg> ;
      linkedtv:hasSocialInteraction
            [ linkedtv:comments "22"^^xsd:int ;
              linkedtv:likes "35"^^xsd:int ;
              linkedtv:shares "0"^^xsd:int ;
              linkedtv:views "41172"^^xsd:int
            ] ;
      dc:creator <https://www.youtube.com/channel/UC_3uchSRXVGSOPszleKUEBw> ;
      dc:date "2012-01-14T19:53:51Z"^^xsd:dateTime ;
      dc:description <http://data.linkedtv.eu/text/ea8283ec-5771-4149-9075-66
          b7db364ab3> ;
      dc:isPartOf <http://www.youtube.com/watch?v=29aSPW6xltM> ;
      dc:source <http://data.linkedtv.eu/organization/YouTube> ;
      dc:type linkedtv:Video ;
      ma:locator <https://www.youtube.com/embed/29aSPW6xltM> .
```

---

[13]http://www.openannotation.org/spec/core/appendices.html#ExtendingMotivations provides the list of possible values for this property

Figure 4.8: Screenshot of the media resource (video) at https://www.youtube.com/embed/29aSPW6xltM, relevant to the named entity "S-Bahn" - RBB scenario

If we further analyze the different properties available for this media resource, we can see that it corresponds to a video, from the social platform Youtube, published on 2012-01-14. The video comes also within a Web page accessible at the URL via the *dc:isPartOf* property, which at the same time contains a textual description. The rest of media items from the Media Collector have been serialized in the same way. The Figure 4.8 depicts one media resource retrieved as an enrichment for the "S-Bahn" entity. This media resource shows one of the trains of this suburban metro-like railway system that serves city centre traffic as well as surroundings and nearby towns.

## 4.3 Refining and Promoting Media Fragments

In the second section of this Chapter we will present a set of techiques that rely on the information attached to the previously annotated media fragments in order to reinforce or re-arrange their temporal boundaries (subsection 4.3.1) or identify certain parts of the video that seem semantically more prominent and therefore should be prioritized or highlighted in different consumption operations (subsection 4.3.2).

### 4.3.1 Using Annotations to Redefine Visual Fragment Boundaries

The temporal boundaries of the video segments are crucial for a proper implementation of meaningful operations over them, like linking, recommendation, or summarization. Visual shots detected via different analysis techniques like the one explained in Section 3.3.2 are sometimes too artificial and meaningless from a user point of view to be considered as unit to work with in the successive steps of the consuming workflow. In this section we present our approach for, starting from the shot detected in raw news broadcast video material and relying on some of the textual and visual annotation techniques included in Chapter 3 (such as automatic speech recognition, keyword extraction or named entity recognition), identifying the boundaries of bigger temporal segments corresponding to important topics in the content. Those topic segments correspond to the different subjects of discussion, dominant ideas, or themes that therefore viewers would be more interested to consume. We offer a first evaluation of the approach probing our precision when exploiting annotations inside automatically generated fragments to find other temporal boundaries which are better suited for certain multimedia tasks.

#### 4.3.1.1 Related work in Topic Segmentation

In topic segmentation one popular, commonly used aspect is the lexical cohesion (e.g. [75, 183]). The general idea is to segment text into fixed-length blocks or into blocks containing sentences or paragraphs. In a next step cohesion (by word repetitions, synonyms, related words, etc.) between these blocks is determined and areas with low cohesive strength are considered to indicate topic changes.

Fuller et al. [54] present a topic segmentation for pod-casts based on lexical cohesion calculated on Automatic Speech Recognition (ASR) transcripts. Instead of punctuation and paragraph information, which are not present in ASR transcripts, the authors use speech related information like time information from ASR and low energy points for segmentation into smaller blocks. Guinaudeau et al. [65] introduce an approach for topic segmentation incorporating confidences and likelihoods from the ASR process to calculate lexical cohesion for segmentation.

Still, most approaches are limited to one modality like text or speech only. Here we

consider an additional modality by using visual cues from shot segmentation replacing sentence and paragraph structures to determine meaningful blocks. Furthermore, we consider lexical cohesion aspects beyond word repetitions by including linking information from Wikipedia in our approach.

### 4.3.1.2   Processing Chain and Dataset

We work with videos from the local German news show "Brandenburg Aktuell", taken from Public Service Broadcaster Rundfunk Berlin-Brandenburg (RBB). The videos were collected over a time period of 5 month, and each day was already segmented by human editors from RBB, for an average of 6.8 segments per day. Note that these segments were not mono-thematic; roughly one-third contained short news summaries consisting of multiple topics.

For each video, we apply automatic speech recognition (as shown in Section 3.3.6) and, on top, keyword extraction (according to method in Section 3.2.2)and named entity extraction (Section 3.2.1.7). Using visual shot segmentation as a fine-granular temporal segmentation to create the original media fragments, we attached the keywords and the NERs to these time segments and compute a distance score on them using the approach that will be described in next subsection.

### 4.3.1.3   Refining Video Segmentation based on Semantic Annotations

For tuning and evaluating our approach, we selected ASR transcripts of any two segments, with the goal of retrieving the seam shot (shot where the disruption between topics is happening) of these segments by relying on annotations inside them in an automatic fashion. We restricted these sets by only allowing segments taken from the same day, in order to ensure that the topics of these segments are distinct (especially important for topics with a high longevity on time, such as the delay in Berlin's airport construction highly covered in many episodes). Keyword extraction has been re-run on each of these double-segments. With an average of seven segments per day, all possible combinations of distinct segments gave 6830 testing instances. Of these, we used 5800 combinations as development set, and reserved 1030 combinations as test set.

Our approach for topic segmentation combines visual cues from Shot Segmentation with the annotations obtained from ASR. On the one hand, video production cuts, dissolves and wipes are used to visualize shifts in time or space including topic shifts in news productions. On the other, ASR provides the notion to determine lexical and semantic cohesion. First, we consider repetitions of extracted keywords in different shots. Second, we indirectly include associated words (e.g. generalizations, statistical associations, etc.) in a similar way by analyzing Wikipedia links and extract associated words for each of the extracted named entities. After combining

both sources of information shot boundaries in areas of low cohesive strength are considered to indicate topic transitions.

The best result on the dev set – 5-best recall showed 62% of the segmentation point to retrieve the exact seam point – was achieved by employing the following scoring:

- Compute keyword splitting penalty (cohesive strength) per shot and per keyword, by taking the minimum number of keywords cooccurrences between consecutive shots, weighted with the TF-IDF relevance of each keyword.

- Compute a Wikipedia splitting penalty (cohesive strength) per shot and wikipedia entry, by taking the minimum number of interlinked Wikipedia entry cooccurrences between consecutive shots, weighted with the (black-box) TextRazor relevance. In order to penalize generic links, we further divide this score by the tenth-part of the number of outgoing links of each Wiki entry based on different experiments done.

- add the penalties with an empirically set weight of the Wikipedia splitting penalty to 0.475.

- Smooth the joined penalty by applying a 5-window triangular filter bank, to emphasize concise penalty shifts.

- Compute the first-derivative.

- Determine the shot which has the highest absolute value indicating low cohesive strength.

#### 4.3.1.4 Evaluation

A qualitative analysis of the shots which were preferred as split points rather than the seam transition produced interesting results, being able to distinguish between: (a) split points whenever an interview started, because the interviewee used different wording, (b) split points when a news point was commented on a deeper level, and (c) many split points in between brief news summary sessions that were merged as one segments. One recurring peculiarity were weather reports, which normally were split well if the weather remained constant but received a split point in between if the forecast predicted a change in the upcoming weather (e.g. "tomorrow the sun finally shines again, bringing warm temperatures to the north...").

**Segmentation Results.** The splitting points for (a) and (b) are quite meaningful in the context of the news consumption. For a better grasp of the segmentation quality, we hand-labelled "weak" topic segmentation shots within a segment that correspond to (a) and (b), in the 1030 test combinations. This introduced on average

Table 4.4: Evaluation of predicted break points, with $\pm$ 1 shot accuracy.

|        | With short news | | Without short news | |
|--------|-----------|--------|-----------|--------|
|        | precision | recall | precision | recall |
| 1-best | 69.7      | 47.4   | 69.9      | 53.1   |
| 2-best | 60.3      | 51.8   | 61.0      | 57.6   |
| 3-best | 44.0      | 59.5   | 44.4      | 64.3   |
| 4-best | 37.2      | 64.6   | 35.0      | 68.8   |
| 5-best | 34.7      | 69.2   | 32.1      | 73.1   |

1.4 new split points for each file. Overall, the first-best guess has a precision of 69.7% to hit a correct topic break point within $\pm$ 1 shot accuracy, and a 47.4% recall to hit the actual seam point. See Table 4.4 for all results, and Figure 4.9 for an example. Precision indicates that the current break point candidate is indeed a valid topic shift, recall indicates that the seam break point was found either by this shot or by one of the higher ranked shots. The second set of experiments was conducted by leaving out all poly-thematic news summary segments.
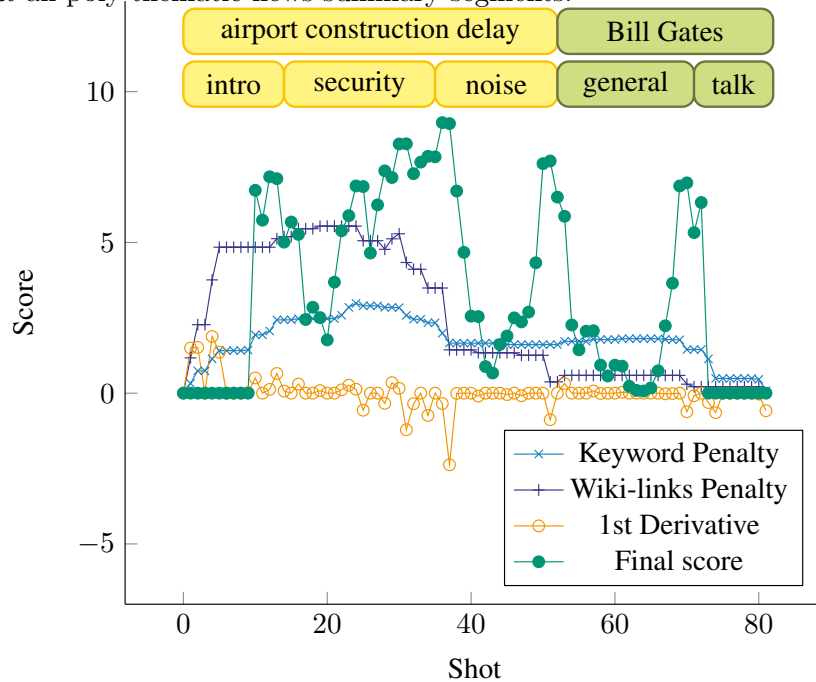


Figure 4.9: Example of the various scores for two joined segments, the first being on Berlin's airport construction delay, the second being on Bill Gates visiting Germany. The weaker topic changes (2 for Berlin, 1 for Gates) are marked accordingly.

In addition, a closer inspection of the results reveals that segment combination which do not include short news summaries roughly have the same precision but quite an improved recall (4–6% absolute), since the topic changes at the seam break points are less confusing. Knowing that a segment is poly-thematic when you see it thus seems like quite a valuable information. News summaries make up substantial

33.2% of the segments. For each file, we extract the following features: the sum of the penalty in each shot, the sum of the active keywords in each shot, and the number of sign changes in the penalty derivate, each divided by the number of shots per file, respectively. Using a support vector machine with a linear kernel, we can classify these short news segments with an accuracy of 83.1% (average over 4-fold cross-validation), which can potentially lead to a better topic boundaries detection that has been left for future experiments on the same line.

### 4.3.2 Promoting Key Fragments: Hotspots

In this section we present an approach that leverages on visual analysis techniques and the knowledge present on the Web for identifying relevant fragments (called Hot Spots) inside educational online videos, in order to get a good overview about what is being told and promote the consumption of media clips at a higher level of granularity. Our approach performs a first segmentation by combining visual features and semantic units in transcripts (paragraph). The resultant video fragments are semantically annotated via Entity Extraction and Topic detection. By identifying consecutive chapters talking about similar topics and entities, we merge the initial segments into bigger and independent semantic media units. Finally we rank them, filter out the lower scored candidates, and propose a summary that illustrates the fragment and can be visualized on the dedicate media player. The algorithm has been applied over a set of educational TED talks. An online demo of the proposed solution is available at http://linkedtv.eurecom.fr/Hyperted/.

#### 4.3.2.1 Motivating Fragment Selection and Promotion

Today people consume all kind of audiovisual content on a daily basic. From breaking news to satiric videos passing through a tutorial on how to cook that wonderful dinner you love, we are constantly bombarded with all kind of multimedia documents. In this media-overloaded scenario it becomes complicated for us to decide if a candidate video is really worth to be watch, or which are the fragments/s that can be potentially interesting without having to watch the entire video.

This phenomena is equally evident when it comes to curated, editorial and educational Web videos. Some studies made over media entertainment streaming services [214] reveal that the majority of partial content views (52.55%) are terminated by the user within the first 10 minutes, and about a 37% of these sessions do not last past the first five minutes. In practice, it is difficult and time consuming to manually gather video insights that (1) give the viewers a fair understanding about what the video is talking about and (2) allow to easily visualize which fragments in particular are illustrating the main topics. Our research tackles this inconvenience by proposing a set of automatically annotated media fragments called Hot Spots, which intend to

highlight the main ideas of the video and make easier for the user to decide which fragment can be relevant for him to watch or share.

The challenge of video segmentation has been addressed by many previous research approaches. Some of them rely exclusively on visual and low-level features like color histograms or visual concept detection clustering operations [171]. On the other hand, there are some pure textbased implementations which leverage on the transcripts and written annotations that goes together with the video, like for example [24]. A special variety of the latter tries to go further and study the semantic behind the text by identifying relevant concepts and linking them to a taxonomy, like in [33]. Finally, there are some initiatives that combine different kinds of techniques [25] in order to keep the best of each. Our demo fits into this last category, with the added value of leveraging on the Web: it is applicable over online videos and it relies on the Web knowledge in order to analyze and annotate the content itself.

### 4.3.2.2   Generating and Displaying Hot Spots in Web Videos

We have implemented a multimodal algorithm for detecting key fragments in a set of 1681 TED talks and annotating them in order to have a quick overview of which are the main topics involved and watch or share the specific parts of the media content which talk about those main ideas. In this section we unveil the details of this approach, specially in what concerns the segmentation of the video, the annotation of the obtained fragments, the selection of the Hot Spots, and the summarization of the main topics inside them.

**Video Segmentation** In first place we perform a video segmentation based only on pure visual features for detecting shots, following the technique described in Section 3.3.2. However, those shots are too small when it comes to semantic consistency. A visual change in the frame flow does not necessary reveal a disruption in what is being told at that particular time of the video. We introduce then the notion of Chapters for naming chunks which illustrate particular topics inside the entire video context. In order to obtain such fragments we have leveraged in some marks embedded in the available video transcripts and not currently exploited by the TED portal, which indicate the start of a new paragraph. According to their definition [14], paragraphs are self-contained units of a discourse dealing with a particular point or idea, which is exactly what we are looking for at this point. Mapping those textual boundaries to the temporal references of the corresponding starting and ending subtitle blocks, we obtain the desired chapters.

In a last step semantic fragments are combined with visual shots for keeping the best of both approaches. In particular we extend chapters back and forward in time in order to include entire shots. This way we end up having semantically independent

---

[14]http://en.wikipedia.org/wiki/Paragraph

segments with visually consistent borders at the same time.

**Media Fragment Annotation** Once the video has been segmented the corresponding fragments are analyzed and annotated. We will rely on the textual information (subtitles) available for the 1681 TED talks in order to detect two kinds of semantic clues: topics (prominent matters the video is talking about) and Named Entities (resources taking part in the story). For the formers we have used TextRazor[15], while for the latter we have used the NERD framework (see Section 3.2.1.7).

Both entities and topics come accompanied by a relevance score which indicates the importance of that particular semantic unit inside the whole context of the video story. Every item detected is attached to the media fragment (in our case, chapters) where that annotation falls into. The output of this phase is a list of chapters which are individually annotated with topics according to TexRazor taxonomy and entities classified according to the NERD Ontology.

**Hot Spots Generation** At this point chapters are accompanied with their own semantic annotations. However sometimes the semantic descriptions between two temporally close chunks are similar enough to consider both fragments as a single unit, either because they talk about the same topics or because they mention the same named entities. In order to tackle this phenomena we apply a clustering algorithm over the chapters, which accumulatively merges consecutive similar fragments. In order to perform this operation we apply a similarity function between consecutive pairs of chapters until no new merges are possible. Being $Rel_i$ the relevance score provided by the topic detection and named entity recognition tools, this comparison leverages on the annotations attached to each segment by analyzing the number of coincidences between the three more prominent topics $T = max_3 \left\{ \sum_{topic_i} Rel_i \right\}$ and the five entities selected according to the five W's principles of the journalism [98], $E = max_{5W's} \left\{ \sum_{entity_i} Rel_i \right\}$. Denoting the two consecutive chapters to be compared as $Ch_1$ and $Ch_2$, and defining the weights $w_{topic}$ and $w_{entity}$ in order to combine the importance of topics a entities into a single score, we define the distance between two chapters as:

$$d\left(Ch_1, Ch_2\right) = w_{topic} \cdot \left( \frac{|T_1 \bigcap T_2|}{max\left\{|T_1|, |T_2|\right\}} \right) + w_{entity} \cdot \left( \frac{|E_1 \bigcap E_2|}{max\left\{|E_1|, |E_2|\right\}} \right) \quad (4.1)$$

After clustering process is finished the chapters have grown in length and decreased in number, but there are still too many candidates which should not be proposed as Hot Spots. Therefore we filter out those fragments which contain potentially less decisive topic and entities. We define the a function for measuring the relevance of a video segment, which directly depends on the relevance and frequency of its main

---

[15]http://en.wikipedia.org/wiki/Paragraph

annotations and is inversely proportional to its length:

$$Relevance\,(Fragment) = \frac{w_{topic} \cdot \sum_{t \in T} Rel_t + w_{entity} \cdot \sum_{e \in E} Rel_e}{Duration(Ch)} \qquad (4.2)$$

In our current approach, the Hot Spots are those fragments whose relative relevance falls under the first quarter of the final score distribution. In a last step, for each Hot Spot we also generate a summarization to be shown in the dedicated media player. Again, we take advantage of the previously calculated main topics $T$ and main entities $E$, which are distributed along 5 dimensions corresponding to the Five W's [98] journalist concept.

### 4.3.2.3   Displaying Hot Spots from TED Talks: HyperTED Prototype

The obtained Hot Spots and their summaries are visualized in a user friendly MediaFragment URI compliant Web media player. The workflow to get Hot Spots available for a certain Ted talk goes like follows: after introducing a valid URL, we land in the video page from where the hot spot detection can be launched for the first time (see Figure 4.10a). When results are available, the corresponding fragments get highlighted on the timeline together with the label of the most relevant chapter annotation. This brief description can be extended to the broader set of main entities and topics in order to get a more exhaustive summary (see Figure 4.10c). Finally as shown in Figure 4.10d) you can always relive that part of the talk you like the most or you would like to share with others by just using its URL.

Regarding upcoming research efforts in this line, we plan to carry out an exhaustive evaluation of our solution involving feedback from real users, in order to optimize the results of our Hot Spot generation algorithm and to improve the usability and efficiency of the developed interface.
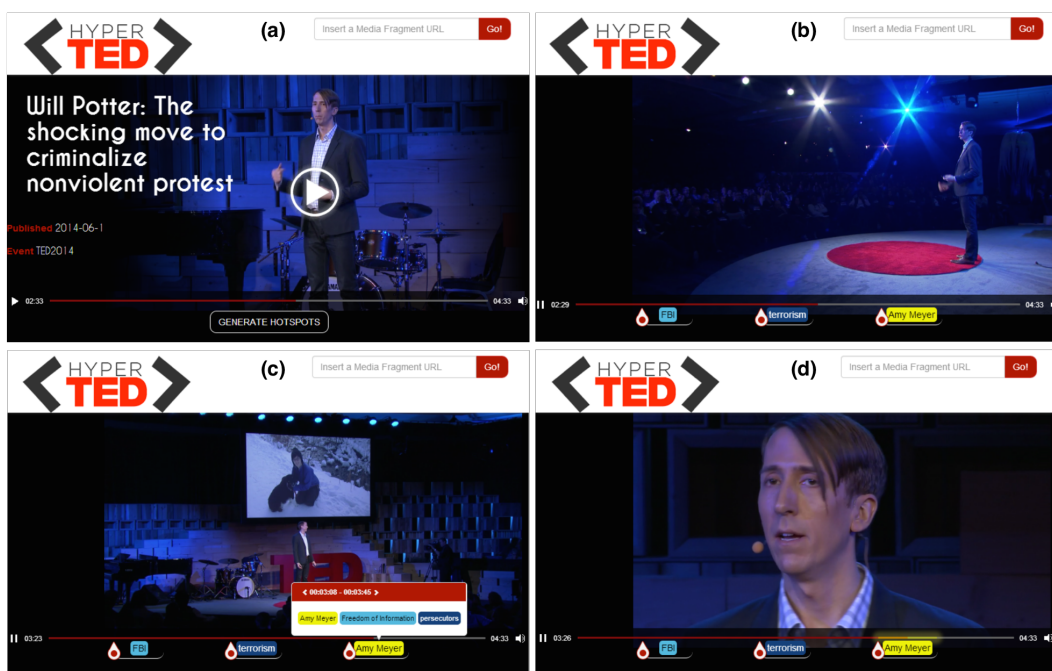
Figure 4.10: Visualizing the Hot Spots of a TED Talk (available at http://linkedtv.eurecom.fr/mediafragmentplayer/video/bdc4a8ba-b092-4b55-8981-69e938208c4d)

## 4.4 Video Classification using Media Annotations

In this section we will propose an approach for exploiting the semantic annotations spotted on textual features in media fragments, in particular named entities extracted following the approach explained in 3.2.1.7. The objective this time is classifying video items. We study the temporal distribution patterns of named entities extracted from 805 Dailymotion videos. We present and evaluate a video classification approach exploiting temporal and textual features used as inputs of four well-known machine learning algorithms. The work has been published at **??**.

We rely on the general system described in [100] for handling media fragments, where named entities from video subtitles are conveniently attached. We design an empirical experiment that considers the number of named entities extracted from video subtitles and their type as features for classifying videos into a subset of Dailymotion channels. With these experiments we aim to address two research questions: *(i)* are there any correlation between the number of named entities per NERD type, or the total number of named entities across the video duration, or the number of named entities per NERD type and temporal distribution for categorizing the video into a channel? *(ii)* which machine learning algorithm(s), among Logistic Regression (LG), K-Nearest Neighbour (KNN), Naive Bayes (NB) and Support Vector Machine (SVM), can best find those correlations in order to best predict what is the category of a particular video.

### 4.4.1 Brief State of the Art in Video Classification

With the steady increase of video items published on media sharing platforms such as Dailymotion and YouTube, more and more efforts are spent to optimize the video handling process for improving video classification. In previous research attempts, text, audio and visual have been the most widely used features for dealing with such a task [19]. Video sharing platforms usually provide rich data that can be used for video classification, such as co-watch data [215], user-generated comments [52] and subtitles [79]. Katsiouli *et al.* [87] have pointed out that video segments can be used together with named entities extracted from subtitles to improve the classification results. However, to the best of our knowledge, no attempts have yet been made to analyze the temporal distribution of named entities in videos, and its implication for video classification. As already introduced in Chapter 2, the Media Fragment URI 1.0 standard and the Ontology for Media Resource open the room for linking video segments to structured annotations, therefore opening new possibilities of innovative video classification services based on semantic descriptions of fragments of video content.

Table 4.5: Video metadata statistics per channel (top). Number of named entities per channel grouped according to the main entity type (bottom).

| channel | fun | tech | sport | news | creat | life | film | music | other | total |
|---|---|---|---|---|---|---|---|---|---|---|
| id | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | - |
| video | 96 | 44 | **163** | 66 | 55 | **194** | 81 | 42 | 64 | 805 |
| ne | 1026 | 4071 | 2794 | 4921 | 1966 | 6996 | 16806 | 1617 | 4279 | 44476 |
| length | 30.2k | 24.2k | 35.9k | 28.4k | 24.2k | 62.5k | 231.7k | 17.4k | 29.8k | 484.4k |
| ne/vid | 10.67 | 92.57 | 17.14 | 74.58 | 35.75 | 36.09 | **207.64** | 38.52 | 66.88 | 55.28 |
| Thing | 274 | 1514 | 618 | 1018 | 581 | 2175 | 1511 | 337 | 933 | **8961** |
| Amount | 106 | 689 | 544 | 810 | 274 | 2010 | 1729 | 201 | 686 | **7049** |
| Animal | 0 | 92 | 2 | 3 | 11 | 5 | 14 | 2 | 49 | 178 |
| Event | 4 | 5 | 20 | 8 | 4 | 6 | 63 | 3 | 11 | 124 |
| Func | 11 | 66 | 55 | 138 | 60 | 107 | 492 | 45 | 126 | 1100 |
| Loc | 103 | 269 | 362 | 827 | 194 | 328 | 1369 | 206 | 604 | 4262 |
| Org | 125 | 233 | 197 | 554 | 132 | 550 | 1705 | 163 | 371 | 4030 |
| Person | 182 | 571 | 462 | 789 | 379 | 867 | **7532** | 403 | 791 | **11976** |
| Prod | 151 | 358 | 184 | 374 | 189 | 589 | 1233 | 136 | 381 | 3595 |
| Time | 70 | 274 | 350 | 400 | 142 | 359 | 1158 | 121 | 327 | 3201 |

## 4.4.2 A Dataset of Categorized Video Items

We collect[16] a set of 805 videos from Dailymotion with subtitles and their basic metadata such as the channel and the video duration using the Dailymotion API. The whole dataset has been processed using NERD [151], where named entities are automatically extracted from the video subtitles and aligned with the corresponding media fragments according to a start time and an end time. All the subtitles are written in English, but some include special characters from other languages. The duration of the videos ranges from 17 to 7654 seconds. There are 9 different channels in this video collection and the distribution of video per channel is: *fun* (96), *tech* (44), *sport* (163), *news* (66), *creation* (55), *lifestyle* (194), *shortfilms* (81), *music* (42) and *other* (64)[17]. Videos are assigned to channels by the video owner. The number of videos per channel and the total number of named entities extracted per channel are depicted in Table 4.5.

Table 4.5 illustrates the number of entities in each of the 10 NERD types per channel. Generally, Thing, Amount and Person have the largest number of entities, while Animal and Event have the smallest amount of entities. The *shortfilms* channel has a large amount of entities in terms of Person and Function, and more than one third of the named entities in Product and Time. Most Animal named entities are extracted from the *tech* channel.

Based on the pair *(entity, media fragment)*, we can further group the named entities by the temporal position the named entity is extracted from. Since the duration of the videos vary, we need to normalize the measure of temporal positions.

---

[16]This dataset has been provided by Dailymotion and the authors had no influence in the video selection process.

[17]Here after, we use the following abbreviations: *creat* for *creation*, *life* for *lifestyle* and *fil* for *shortfilms*
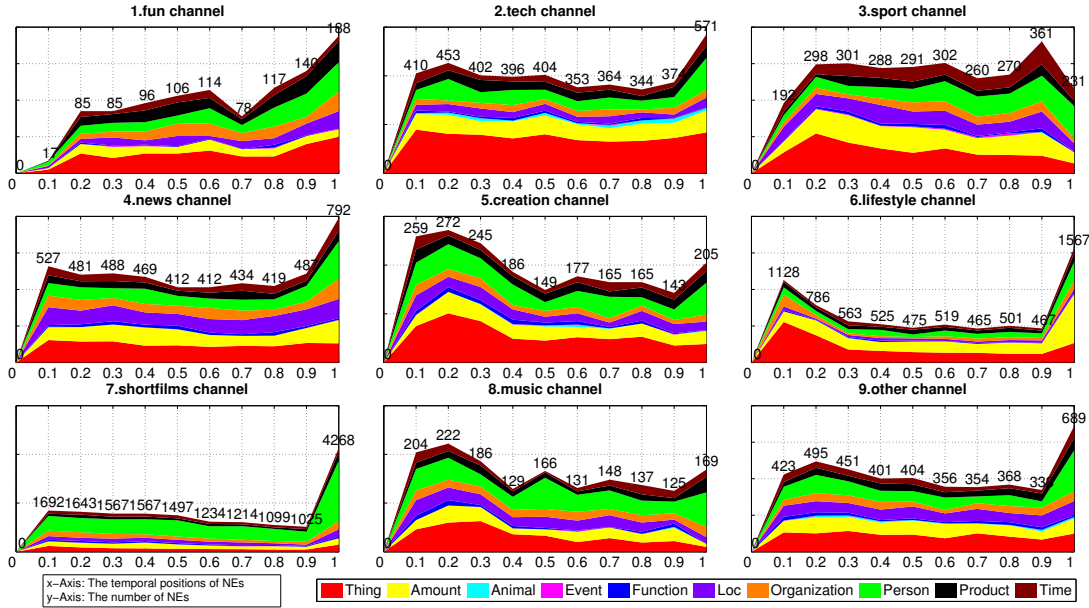
Figure 4.11: Distribution of named entities extracted from subtitles per channel and the summary of their temporal positions in the videos.

We define a temporal position variable $tp$ as $0 \leqq tp = \frac{st+et}{2 \times dur} \leqq 1$, where $st$ and $et$ are the start and end time the named entity is extracted from, and $dur$ is the duration of the video. When grouping the named entities according to their $tp$, each video is equally divided into $N$ fragments, and any entity belongs to one fragment selected according to $tp$. Figure 4.11 demonstrates the $tp$ distribution of different types of named entities for each channel. The different colors in the figure represent the different NERD types.

In Figure 4.11, we observe that for $shortfilms$, a large amount of named entities (4268) is extracted from the end of the video ($tp \in (0.9, 1]$), and a large proportion of them is Person. The $lifestyle$ has two spikes at the beginning and at the end. The $fun$ channel has a very low number of named entities at the beginning, and the named entities in $tech$, $news$ and $other$ are slightly higher at the end. The $sport$ channel has low numbers at both the beginning and the end, while it is difficult to see the characters for $creation$ and $music$. If we suppose that the named entities extracted are all correct, these characters mentioned above imply some important information that could be useful for video classification and retrieval based on temporal features.

### 4.4.3    Video Classification Methodology

We conducted a multi-class classification experiment to categorize videos into 9 different Dailymotion channels. We ran three different experiments, later on named $Exp1$, $Exp2$ and $Exp3$. In each experiment, we used different features which are commonly

denoted by the vector $\vec{x}$. The channel information retrieved from the Dailymotion API is considered as the outcome label.

We assign an *id* to each channel *c*, and $c \in \{1, 2, ..., 9\}$ (Table 4.5). In Exp1, we consider the number of named entity types as features. As there are 10 NERD types, each observation is a feature vector whose $|\vec{x}| = 10$. For Exp2, we weight the number of named entities type with their temporal position values *tp* and group them into *N* groups. The size of the feature vector is $|\vec{x}| = N$. The choice of *N* may affect the prediction results. For Exp3, we combine Exp1 and Exp2 together and use the temporal distribution of named entities in each NERD type as features. Therefore, there are $10 \times N$ features in Exp3 and if $N = 20$, then $|\vec{x}| = 200$.

We applied four basic classification algorithms in each experiment: LG (Logistic), KNN (K Nearest Neighbors), NB (Naive Bayes) and SVM (Support Vector Machine). First, we applied a 10-fold cross validation, where all the 805 videos are divided into 10 equal-sized groups. In each fold, we use 9 portions as the training set and 1 part as the test set and make sure each observation in the dataset appears only once in the test set. Then, when applying different algorithms into each fold, the results can be generically defined as:

$$\hat{\mathbf{R}} = predict(\mathbf{X}^e, \mathbf{Y}^e, \mathbf{X}^r, \mathbf{Y}^r, params) \tag{4.3}$$

where $\mathbf{X}^r$ and $\mathbf{X}^e$ are the matrix of training and testing data respectively and where each row in the matrix is an observation of the feature set. $\mathbf{Y}^r$, $\mathbf{Y}^e$ and $\hat{\mathbf{R}}$ are single column matrix. Each entry in $\mathbf{Y}^r$ and $\mathbf{Y}^r$ is the labeled channel id, named *c*, for training and testing respectively, while each entry in $\hat{\mathbf{R}}$ is the predicted channel $\hat{c}$ corresponding to the observation of the same row in $\mathbf{X}^e$. The actual definition of *predict* function in Equation 4.3 changes accordingly with different algorithms. The *params* represents a set of parameters that we use to tune each algorithm so that the best results can be obtained and will be further specified in next paragraph.

For the experiments, we adopted the multinomial LG so that the classification result for each video using LG is a vector $\vec{r} = [r_1, r_2 \cdots r_c]$, where $r_c$ is the probability that this video belongs to the channel *c*. We use the same setting for NB. The output channel is the one with the largest possibility as the final prediction result. To reduce the over-fitting problem, we applied to the logistic regression the L2-Regularization. Given our settings, we empirically assessed that $\lambda = 0.0001$ has the best bias-variance tradeoff. NB has many choices to model the data distribution, so we chose the multi-variate multinomial distribution, which best fits our problem. In KNN, the main tuning *param* is the choice of *k*, but there is still lack of principled way to define it. Hence, we empirically assessed that with $k = 20$ the algorithm has the best convergence given our settings. SVM cannot be directly applied for

multi-class classification problems, so we use LIBSVM[18] to implement a 1-vs-1 SVM algorithm, choosing the linear kernels as the kernel function for all the experiments. Finally, to measure the accuracy of each experiment and algorithm, we define the precision $P$, recall $R$ and F1-score $F1$ for each channel $c$ as:

$$P_c = \frac{\sum_{f=1}^{10}, |\hat{R}_f(c) \bigcap Y_f^e(c)|}{\sum_{f=1}^{10} \hat{R}_f(c)} \tag{4.4}$$

$$R_c = \frac{\sum_{f=1}^{10} |\hat{R}_f(c) \bigcap Y_f^e(c)|}{\sum_{f=1}^{10} Y_f^e(c)} \tag{4.5}$$

$$F1_c = 2 \times \frac{P_c \times R_c}{P_c + R_c} \tag{4.6}$$

$\hat{R}_f(c)$ is the set of videos that have been predicted belonging to channel $c$ in $f$th fold of cross validation, while $\hat{Y}_f(c)$ are the videos that have been labeled in channel $c$. So $|\hat{R}_f(c) \bigcap Y_f^e(c)|$ is the number of videos that is correctly categorized in channel $c$ in a cross validation fold. There is a possibility that $\sum_{f=1}^{10} \hat{R}_f(c) = 0$ if no video has been categorized to the channel $c$. In this case, the value of $P_c$ is $NaN$. Our dataset has videos in each channel, so $\sum_{f=1}^{10} Y_f^e(c) \neq 0$. To evaluate the overall accuracy $acc$ of the algorithm in each experiment on the entire dataset, we define:

$$acc = \frac{\sum_{c=1}^{9} \sum_{f=1}^{10}, |\hat{R}_f(c) \bigcap Y_f^e(c)|}{805} \tag{4.7}$$

The overall accuracy is the total number of videos that have been correctly classified divided by the total number of the video since every video appears exactly once in the test set.

### 4.4.4　Experiments and Discussion

Figure 4.12 shows the overall accuracy for each experiment detailed in Section 4.4.3. The best accuracy is obtained with KNN-Exp1 (46.58%) and the worst one is LG-Exp3 (33.54%). Generally, there are not major differences between each algorithms using different sets of features for the overall accuracy. The features chosen in Exp1 perform better than the other two feature sets using LG and KNN. For NB, the $acc$ for Exp1 and Exp3 are close and they are all better than Exp2. SVM-Exp3 outperforms Exp1 and Exp2, and it is also the best accuracy in Exp3 compared with other algorithms. Accuracy in Exp1 and Exp3 are usually better than Exp2. From this point of view, it is possible to infer that the number of named entities and their type is an indicator to be taken into account for improving the video classification

---

[18]http://www.csie.ntu.edu.tw/~cjlin/libsvm/
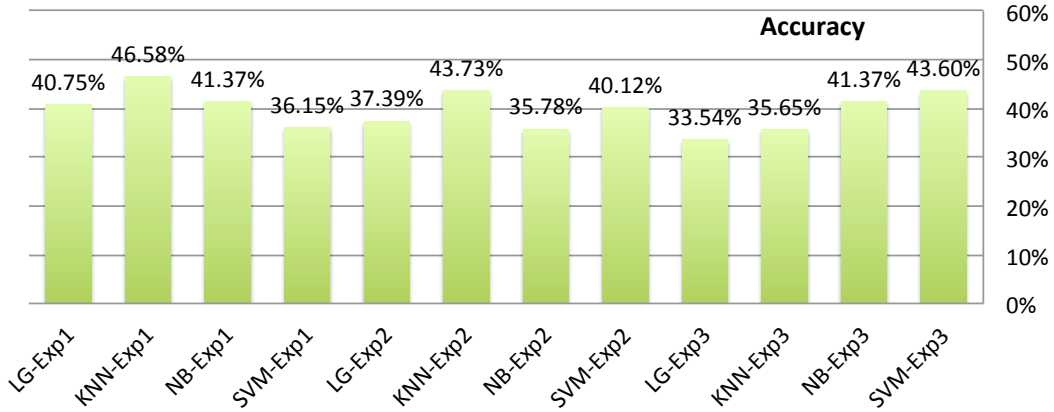
Figure 4.12: Accuracy comparison for each algorithm-experiment pair.

algorithm, assuming that there is a sufficient number of named entities detected for each NERD type.

Table 4.6 reports the breakdown scores per channel and experiment considering precision, recall and $F1$. The 3 largest numbers for each measurement are highlighted in bold. If we use $F1$ as the general measure of the accuracy, $sport$, $life$ and $shortfilms$ usually obtain best accuracy in the different experiments and using different regression algorithms, while the $F1$ of $news$, $creation$, $music$ and $other$ are usually below 20%. This behaviour makes sense since the number of samples available for that first set of channels is bigger than for the second group. Therefore, the algorithm is able to better define the classification model. Using LG, $lifestyle$ and $shortfilms$ consistently gain high accuracy in all the three experiments. All $P$, $R$ and $F1$ scores are high for $shortfilms$ in Exp1. For KNN, the $P$, $R$ and $F1$ are all above 70% for $shortfilms$ in Exp1, which is the overall best. Compared with Exp1, Exp2 and Exp3 obtained worse results nearly in every channel. Using NB, $F1$ for $sport$, $lifestyle$ and $shortfilms$ are good in both Exp1 and Exp3. SVM performs better when dealing with multi-dimensional data, so the best result for SVM is in Exp3, where 200 features are used for the classification. SVM also generates the best $F1$ for $lifestyle$ (61.5%) among all the other algorithms, while the accuracy for $shortfilms$ channel in SVM-Exp1 is very low. This is due to the fact that SVM relies more on the size of the samples when the size of the features are small.

Generally, channels with large sample size, such as $sport$ and $lifestyle$, are likely to obtain high accuracy in most of the algorithms. However, even though the sample size of $shortfilms$ is not big, the $NEs/Video$ value is much larger than others (Table 4.5). This is because the average video length in $shortfilms$ is longer and more named entities can be extracted from their subtitles. Considering the use of media fragments in this experiment, the characteristics of temporal fragments and

Table 4.6: Precision (P), recall (R) and F-measure (F1) on various channels for the experiments using LG ($\lambda = 0.0001$), KNN ($k = 20$), NB and SVM (%)

| | | fun | tech | sport | news | creat | life | films | music | other |
|---|---|---|---|---|---|---|---|---|---|---|
| LG-Exp1 | P | 28.87 | 33.33 | **35.69** | 32.26 | 8.33 | **49.78** | **73.13** | NaN | 16 |
| | R | 29.17 | 15.91 | **71.17** | 15.15 | 1.82 | **58.25** | **60.49** | 0 | 6.25 |
| | F1 | 29.02 | 21.54 | **47.54** | 20.62 | 2.99 | **53.68** | **66.22** | 0 | 8.99 |
| KNN-Exp1 | P | 23.91 | 45 | **50** | **54.17** | 28.57 | 48.01 | **72.29** | 20 | 23.08 |
| | R | 22.92 | 20.46 | **66.87** | 19.7 | 18.18 | **74.74** | **74.07** | 2.38 | 9.38 |
| | F1 | 23.4 | 28.13 | **57.22** | 28.89 | 22.22 | **58.47** | **73.17** | 4.26 | 13.33 |
| NB-Exp1 | P | 31.82 | 40.74 | **44.87** | 29.83 | 26.32 | 44.06 | **55.77** | 12 | 0 |
| | R | 29.17 | 25 | **42.95** | 25.76 | 9.09 | **72.68** | **71.61** | 7.14 | 0 |
| | F1 | 30.43 | 30.99 | **43.89** | 27.64 | 13.51 | 54.86 | **62.7** | 8.96 | 0 |
| SVM-Exp1 | P | 33.33 | NaN | **50** | 50 | 26.67 | 31.37 | **36.36** | NaN | 0 |
| | R | **8.33** | 0 | **62.58** | 4.55 | 7.27 | **87.63** | 4.94 | 0 | 0 |
| | F1 | **13.33** | 0 | **55.59** | 8.33 | 11.43 | **46.2** | 8.7 | 0 | 0 |
| LG-Exp2 | P | 35.71 | 24 | **32.18** | 30.77 | 5.26 | **50.23** | **66.67** | 5.88 | 0 |
| | R | 31.25 | 13.64 | **68.71** | 12.12 | 1.82 | **56.19** | **41.98** | 2.38 | 0 |
| | F1 | 33.33 | 17.39 | **43.84** | 17.39 | 2.7 | **53.04** | **51.52** | 3.39 | 0 |
| KNN-Exp2 | P | **47.69** | 37.93 | 42.48 | 18.75 | 6.25 | 44.04 | **86** | 20 | 19.05 |
| | R | 32.29 | 25 | **58.9** | 9.09 | 1.82 | **81.96** | **53.09** | 2.38 | 6.25 |
| | F1 | 38.51 | 30.14 | **49.36** | 12.24 | 2.82 | **57.3** | **65.65** | 4.26 | 9.41 |
| NB-Exp2 | P | 18.75 | 30.77 | 32.89 | **38.46** | 9.09 | 46.28 | **62.9** | 3.7 | 8.33 |
| | R | 12.5 | 9.09 | **60.74** | 15.15 | 5.45 | **60.83** | **48.15** | 2.38 | 3.13 |
| | F1 | 15 | 14.04 | **42.67** | 21.74 | 6.82 | **52.56** | **54.55** | 2.9 | 4.55 |
| SVM-Exp2 | P | **45.71** | NaN | **43.48** | 0 | 37.5 | 49.1 | 26.75 | NaN | NaN |
| | R | 16.67 | 0 | **61.35** | 0 | 10.91 | **70.1** | **80.25** | 0 | 0 |
| | F1 | 24.43 | 0 | **50.89** | 0 | 16.9 | **57.75** | **40.12** | 0 | 0 |
| LG-Exp3 | P | 18.87 | 17.54 | **38.82** | 15.39 | 7.02 | **57.9** | **54.76** | 10 | 12.9 |
| | R | 20.83 | 22.73 | **36.2** | 18.18 | 7.27 | **56.7** | **56.79** | 11.91 | 6.25 |
| | F1 | 19.8 | 19.8 | **37.46** | 16.67 | 7.14 | **57.29** | **55.76** | 10.87 | 8.42 |
| KNN-Exp3 | P | 21.05 | **42.86** | 29.86 | **42.86** | 33.33 | 34.36 | **80.65** | NaN | 0 |
| | R | 20.83 | 6.82 | **26.38** | 4.55 | 1.82 | **86.08** | **61.73** | 0 | 0 |
| | F1 | 20.94 | 11.76 | **28.01** | 8.22 | 3.45 | **49.12** | **69.93** | 0 | 0 |
| NB-Exp3 | P | 22.68 | 28.26 | **47.4** | 33.33 | 13.51 | **52.56** | **61.18** | 19.36 | 12.5 |
| | R | 22.92 | 29.55 | **55.83** | 25.76 | 9.09 | **63.4** | **64.2** | 14.29 | 6.25 |
| | F1 | 22.8 | 28.89 | **51.27** | 29.06 | 10.87 | **57.48** | **62.65** | 16.44 | 8.33 |
| SVM-Exp3 | P | **52.63** | 26.92 | 34.78 | 25.81 | 0 | **66.47** | **49.17** | NaN | 40 |
| | R | 20.83 | 15.91 | **88.34** | 12.12 | 0 | **57.22** | **72.84** | 0 | 3.13 |
| | F1 | 29.85 | 20 | **49.91** | 16.49 | 0 | **61.5** | **58.71** | 0 | 5.8 |

NERD type distribution of named entities for *shortfilms* are also outstanding: large number of named entities are associated with the end of the videos and most of them are Person. So considering those two factors and the sample size of *shortfilms* is not very low, it is possible to understand why the accuracy of this channel is higher for most of experiments. But as mentioned earlier, sample size is still the key factor for SVM regression in this context.

Some experiments achieve very high score of $R$ but very low $P$, such as *lifestyle* in KNN-Exp3 and SVM-Exp1, and *sport* in SVM-Exp3. There are many videos in *lifestyle* and *sports*, but in the results of the classification, many more instances have been marked as belonging to these channels, most of which are wrongly categorized. To the contrary, in some cases, $P$ is much higher than $R$, such as the *news* channel in KNN-Exp3. This is because not enough instances are predicted to belong to this channel and the classification accuracy of such channels is usually very low and not

stable. If we put these two phenomenons together with Table 4.5 and Figure 4.11, it is possible to find out that channels with very strong and clear patterns in the entity distribution or with large sample size are easily to have high $R$ but low $P$. Channels with small amount of samples are likely to be badly classified as there are not many samples provided for training the algorithms.

**Wrapping up Results**. The results obtained for the three proposed experiments indicate that the implemented fully automatic method is very promising in the context of online videos. In detail, there is no dominant algorithm that outperforms the rest of them for the 3 experiments in terms of the overall accuracy, but the number of named entities and their types are very useful to improve the video classification algorithm. KNN in Exp1 obtains the best overall accuracy among all experiments and algorithms, and when using named entities and media fragment features together, SVM gets the best overall performance. For each individual channel, $sport$, $lifestyle$ and $shortfilms$ have the highest prediction accuracy.

## 4.5   Multimodal Media Fragment Hyperlinking

Managing multimedia content and offering worthy operations over it is key in today's consumption scenario. As one may be overwhelmed by the huge amount of available information being created every day, features like searching relevant content are crucial. But also other procedures such as media fragment enrichment (see Section 4.2) or recommendation further improve user experience, as they provide the viewers with more relevant content about the topic. This advanced navigation from one video to another at different levels of granularity is similar to the browsing activity when users follow classic html hyperlinks to move from one textual document to another on the Internet.

In order to bring this browsing capabilities to the visual ecosystem, the network of media hyperlinks has to be created beforehand and/or adjusted on the fly in order to better support the different operations. In this section we describe our efforts in hyperlinking a closed collection of video items at the level of fragments, by relying both in semantic techniques applied on textual information attached to video chunks (Section 3.2), and visual based annotations as covered in Section 3.3, in what we call a multimodal hyperlinking approach. This multimodality of the approach has been proven to improve results when compared to other algorithms only considering some of the elements of the media annotation spectrum.

Through the MediaEval participation reported in Section 4.5.2 and successive campaigns, we study how much improvement this visual and textual combination can provide to the search, and how we can tune this approach to get better results. The aim is to explore the cross modality between textual and visual features: text has traditionally been able to give valuable results, but lacks the specificity of the visual information, while visual features exploit this visual part but are not descriptive enough by themselves. We argue that improved information retrieval operations can be achieved by combining textual and visual information to create a multimodal query as already proposed before in the literature [171]. The originality of our work lies in the fact that we can start from a pure text query to perform a visual-based search. Hence we attempt to overcome the semantic gap by automatically mapping input text to semantic concepts.

To summarize, in this section we will be showing how to implement and evaluate a system relying on low level visual concepts and high-level semantic entities in combination with pure text-based searches in order to improve video retrieval tasks. We investigate the following two questions: i) to which extent visual concepts can bring value when retrieving media fragments and resources, and ii) how we can combine it with the textual based annotations to bring the desired improvements.

### 4.5.1 State of the Art in Exploiting Visual Cues: Bridging the Semantic Gap

Popular search engines retrieve documents on the basis of textual information. This is especially the case for text documents, but also is valid for images and videos, as they are often accompanied with textual metadata. Several research works attempt to include visual information based on input images and/or on relevance feedback [168, 176, 141, 185].

The work of Hauptmann *et al.* [72] analyses the use of visual concepts only for video retrieval in the scenario of a news collection. The authors study the impact of different factors: the number, the type and the accuracy of concept detectors. They conclude that it is possible to reach valuable results within a collection with fewer than 5000 concepts of modest quality. In their evaluation, they start from a query directly constituted of concepts, while we propose to automate the concept mapping from a text query. Nevertheless, they suggest the use of semi-automated methods for creating concepts-based queries.

Such work inspired the study of [67], although their focus is slightly different: they want to represent *events*. They aim at creating a concept detectors vocabulary for event recognition in videos. In order to derive useful concepts, they study the words used to describe videos and events. The resulting recommendations on the concepts are the following: concepts should be diverse, both specific and general. They also have results on the number of concepts to be used: vocabularies should have more than 200 concepts, and it is better to increase the number of concepts than the accuracy of the detectors.

Hamadi *et al.* [69] proposed a method, denoted as 'conceptual feedback', to improve the overall detection performance that implicitly takes into account the relations between concepts. The descriptor of normalized detection scores was added to the pool of available descriptors, then a classification step was applied on this descriptor. The resulting detection scores are finally fused with the already available scores obtained with the original descriptors. They have concluded that significant improvement on the indexing system's performance can be achieved, when merging the classification scores of the conceptual feedback with their original descriptors. However, they have evaluated their approach on TRECVID 2012 semantic indexing task, which is based only on detecting semantic visual concepts, and no text-based queries was used.

How much can different features (textual, low-level descriptors and visual concepts) contribute to multimedia retrieval? The authors in [26] have addressed this question by studying the impact of different descriptors, both textual and visual ones, for video hyperlinking. They concluded that the textual features (in this case transcripts) perform the best for this task, while visual features by themselves (both

low level and high level) cannot predict reliable hyperlinks, due to a great variability in the results. Nevertheless, they suggest that using visual features for reranking results obtained from a text search slightly improves the performance. In this section we endeavor to estimate how visual concepts can improve a search, depending on the way they are used.

Several works achieve the automatic linking of a textual query to visual concepts through a semantic mapping by exploiting ontologies. In [170], the authors developed an OWL ontology of concept detectors that they have aligned with Word-Net [49]. They question whether semantically enriched detectors help in multimedia retrieval tasks. Similarly, an ontology based on LSCOM taxonomy [126] has been developped[19], and has been aligned with ontologies such as DBpedia[20].

In our work we will focus on the use of visual information to improve content retrieval in a video collection. Videos are visually very rich and it is not straight-forward to exploit such data when searching for specific content. This phenomena is commonly called *semantic gap*: there is no direct or easy match between the meaning of a situation or an object, a concept, and the representation that can be made of it, in particular by a computer [167]: a gap between the low-level features extracted from an image, and the high-level semantics that can be understood from it.

### 4.5.2   MediaEval 2013 Participation

MediaEval is a benchmarking initiative for multimedia evaluation. It offers several tasks, among which the Search and Hyperlinking is the one that better fits our research objectives since it tackles the issue of information seeking in a video dataset [41]. The scenario proposed is that of a user watching a particular video segment and therefore could be interested in watching related content proposed by the system. Hence, this represents the perfect scenario for testing the different annotation techniques and technologies covered in this thesis in real-world conditions.

The MediaEval 2013 workshop took place in Barcelona, Catalunya, Spain, on Friday-Saturday 18-19 October 2013[21], just preceding the ACM Multimedia 2013[22] conference. The dataset for this task contains 2323 videos from the BBC, amounting to 1697 hours of television content of all sorts: news shows, talk shows, series, documentaries, etc. The collection contains not only the videos and audio tracks, but also some additional information:

- Subtitles (manually transcribed).

- Two types of ASR transcripts (LIMSI [92] and LIUM/Vocapia [156]).

---

[19]http://vocab.linkeddata.es/lscom/
[20]http://www.eurecom.fr/~atemezin/def/lscom/lscom-mappings.ttl
[21]http://www.multimediaeval.org/mediaeval2013/
[22]http://acmmm13.org/

- Metadata giving show title, description, date of airing, format, etc.

- Additional metadata: synopsis and cast from the BBC Web site.

- Shot boundaries and keyframes.

- Face detection and face similarity information.

- Concept detection.

First, we applied different processing techniques over the dataset, in order to have as much information as possible. We summarized the pre-processing performances in table 4.7: the processing time is fairly important due to the size of the dataset (1697 hours of video). The data extracted will be further described below.

Table 4.7: Performances of the different analysis techniques on the whole dataset

| Concepts Detection | 20 days on 25 4-cores computers |
|---|---|
| Scene Segmentation | 2 days on 6 cores |
| Keywords Extraction | 5 hours |
| OCR | 1 day on 10 cores |
| Face Detection and Tracking | 4 days on 40 4-cores computers |
| Named Entities Extraction | 4 days |

#### 4.5.2.1 Description of the Search and Hyperlinking Task

**Search Task**. The goal of the search task is to identify relevant video segments in the dataset using a textual query provided by a user. The query is constituted of two parts: the first part gives information for a text search while the other gives cues on visual information in the searched segments, using words. For example, the query composed by the terms "Little Britian Fat Fighters problem of gypsies in the area" and the visual clues "fat club comedy". 29 users defined 50 search queries related to video segments watched among this dataset.

The evaluation of the search task is based on the following measures:

- the Mean Reciprocal Rank (MRR) assesses the rank of the relevant segments returned by the queries. It averages the multiplicative inverse of the ranks corresponding to correct answers (within a given time window, here 60 seconds).

- the Mean Generalized Average Precision (mGAP) is a variation of the previous that takes into account the distance to the actual relevant jump-in point. Hence, this measure also takes into account the start time of the segment returned.

- the Mean Average Segment Precision (MASP) assesses of the search in terms of both precision of the retrieved segments and the length of the segments

that should be watched before reaching the relevant content [44]. It takes into account the length of overlap between the returned segments and the relevant segment. It hence favors segments whose boundaries are close to the expected ones.

A window size of 10, 30 and 60 seconds was originally planned to be reported, but for the sake of simplicity in the final results delivered only the 60s time window has been considered.

**Hyperlinking task**. The Hyperlinking task aimed at offering to the viewer content from some particular sources that could be potentially related to what (s)he is watching. The user defines an anchor (a video segment, identified by the video name, the begin and end times), which is the basis for seeking related content in the collection. Condition "LA" requires to use only this segment for the hyperlinking, while condition "LC" also allowed to use the context of this segment (i.e. a broader segment from the video).

Examples of an hyperlinking query:

```
<anchor>
  <anchorId>anchor_1</anchorId>
  <startTime>13.07</startTime>
  <endTime>13.22</endTime>
  <item>
   <fileName>v20080511_203000_bbcthree_little_britain</fileName>
   <startTime>13.07</startTime>
   <endTime>14.03</endTime>
  </item>
</anchor>
```

A manual evaluation by users will follow by performing crowd-sourcing evaluation through Mechanical Turk. It was carried out for the top 10 ranks of the runs. Mean Average Precision will be reported, as well as precision at different ranks (i.e. how many relevant targets were retrieved in the top n ranks).

#### 4.5.2.2  Media Content Indexing

**Solr Indexing** We have used the search platform Solr[23] based on Lucene[24] for indexing the different media fragments and their attached annotations. Media fragments are included at different granularities: whole video, scene level, shot level, subtitle block level, and speech segments from transcripts. Hence, searches can be performed at a chosen granularity, the next step being to design an appropriate query.

Different indexes were created with different information, as described below. Documents were represented by both textual fields (for a text search) and floating point fields to represent concepts calculated using the technique described in Section 3.3.1.

---

[23]http://lucene.apache.org/solr/
[24]https://lucene.apache.org/core/

In order to search over those visual cues, we will execute range queries, like for example: concept `Animal` is between 0.63 and 1.

- **Video Index**. Videos were defined using their name. Different text fields enabled to store and index diverse information taken from the metadata: title of the series the show is part of, episode title, channel the video was aired on, synopsis, short synopsis, description, cast. We also indexed subtitle and keywords extracted, plus the number of shot in each video.

- **Scene Index**. Scenes (created using scene segmentation, see 3.3.3) were defined using the id of the video they are part of, with begin and end times. Each scene was aligned with the corresponding subtitles. Hence, we indexed each scene with associated textual information: subtitle and various metadata (title, cast, synopsis, etc). Extra fields representing concepts were also introduced. For each concept, we stored the value of the highest score of the concept across shots that constitute the scene.

- **Shot Index**. Similarly to scenes, shots were defined using the id of the video they are part of, with begin and end times. We indexed subtitles (aligned to each shot) as well as OCR when available. The number of faces per shot was also indexed but not used in the search and hyperlinking algorithms.

- **Speech Segments Index**. We create speech segments by merging adjacent speech segments when they were uttered by the same speaker. Speech segments don't overlap and are connected to each other. It was prepared only for LIMSI transcripts using information about speakers. Hence, speech segment are defined using video id, begin time and end time. The indexed text was the corresponding transcript. This algorithm produces short video segments, whose length can vary from 1 to 45 seconds, and it is not suitable for videos with one person speaking, because it will produce an extreme situation with a single long segment. In the final runs, we didn't submit any results based on speech segments because of poor results obtained.

- **Sliding Windows Index**. Sliding window algorithm [42] was used with different parameters and different schema in Apache Solr. First, each document is divided into sentence segments. Because of different format of data, the sentence segments have to be prepared differently. For LIUM transcripts, a fix length "sentence" of about 17 words was used, as it is considered as the average length of English sentence [209]. For LIMSI transcripts, splitting is done on punctuation. Finally, for subtitles, each line in `<span>` tag was extracted as one sentence. Sentences are processed using POS (part-of-speech) tagging (for each sentence segment) to obtain information about the number of open words

they contain. Open words are those words that carry the content or the mean-
ing of a sentence - verbs, nouns, adjectives, adverbs and interjections. For open
class words recognition, MaxtenTagger[25] was used. Last, we index sentences as
follows: we insert each sentence into sliding window and count open words for
the whole actual window (a window can contain many sentences). When the
sliding window is full, i.e. when the count of open class words is bigger than the
parameter specified at startup, we merge sentences into one text and index it as
one segment (start time is start time of first sentence and end time is end time
of last sentence). We then remove the first sentence and repeat this process
until we reach the end of the document. The last segment in the document can
have a smaller count of open class words than other. This algorithm produces
many overlapping segments and the density of coverage document is high.

### 4.5.2.3   The Search Approach

The search task required to divide videos into smaller segments, between 1 and 15
minutes. We submitted 9 different runs by adopting two different strategies: (1) we
either used pre-constructed segments that were indexed in the Lucene engine, or (2)
we performed queries creating segments on the fly, by merging video segments based
on their score.

For most of the searches, we concatenated textual and visual part of the query
(content of *queryText* and *visualQueues* on the given data), as it yields better results
than using the textual dimension only.

After indexing the entire collection of videos, we found out that performing a text
query on the video index returned the accurate video in the top of the list. Hence,
for some runs, we could restrict the pool of videos that are going to be searched to
a small number. The query then has to be made in two steps: first, we query the
overall video index, extracting the 20 first videos, and then we perform additional
queries for smaller segments restricted to this smaller dataset.

**Search Using with Visual Cues**. Text search is straightforward searchable
in Lucene/Solr: the search engine provides a default text search based on TF-IDF
values. For this reason, incorporating visual information in the search task requires
to design a new query that combines the provided visual cues and the output of
the visual concept detection algorithm. Therefore, we initially created a mapping
between keywords (i.e. terms) extracted from the visual cues query via text analysis
in Section 3.3.9, and the visual concepts that are considered by the applied concept
detection algorithm in Section 3.3.1. Based on this mapping, we were able to integrate
the output of this algorithm (i.e. presence of visual concepts in the searched segment)
into the search task that is not bringing those kind of annotations in an explicit way,

---

[25]http://nlp.stanford.edu/software/tagger.shtml

resulting in an enriched query that includes both textual and visual information, therefore obtaining the desired *multimodality*.

**Normalizing Visual Concept Scores in Query**. For each of the considered visual concepts we defined two values: its highest score across the entire group of key-frames/shots of the MediaEval data-set and a "valid detection" rate, calculated by examining its presence/absence within the key-frames/ shots of the MediaEval dataset that correspond to the top-100 highest scores for this concept. Built on this, we performed a filtering step using the "valid detection" rate, that occurred from the top-100 highest scores, as a new confidence score about this concept and we then discarded the concepts with a rate lower than 0.5. Afterwards, we linearly normalized these rates between 0 and 1, by mapping the highest score among them to 1.

**Concept detector scores for each scene:** $v$

The concept scores extracted from the videos express the confidence that the corresponding concepts appears in the main frame of each shot. By extension, we assume that they represent the confidence of appearance for the entire shot.

We first normalize all the visual scores on a scale from 0 to 1 by a min-max normalization function. This function aims to scale the scores for each concept, so that they all fall in a range of $l$ *to* $u$ bounds. Thus, the visual scores values are normalized by subtracting the minimum and maximum score for each concept and then applying the following equation on each bin value:

$$v'_{ij} = l + \frac{(u - l) \times (v_{ij} - min_j)}{max_j - min_j} \tag{4.8}$$

where $v_{ij}$ is the score of the $j^{th}$ concept for the $i^{th}$ frame, $min_j$ and $max_j$ are respectively the minimum and maximum score of the $j^{th}$ concept, and $u$ and $l$ are the new dimension space. Results in $v'$ are often normalized to the $[0, 1]$ range. Then, the visual score $v$ of each scene is obtained by the mean average of its shots' scores.

**Building Search Query**. Finally, when a visual concept was derived from the visual cues query and if it hadn't been rejected by the previous filtering, we also added a range query to the Lucene query, beside the textual query: in order for the concept to be present, its normalized score had to lie above a threshold value, i.e. between this threshold and 1. We empirically set the threshold at 0.7. Hence, we considered that all the normalized scores above this threshold truly contain the given concept. The range query was similar to:

```
Animal:[0.7 TO 1]
```

**Submitted Runs**. We submitted 9 runs in total:

- *scenes-C*: Scene search using textual and visual cues. No filtering by video.

- *scenes-noC*: Scene search using textual cues only. For comparison purposes, this run has the same settings as the previous one, except that it makes no use of the visual concepts.

- *part-scenes-C*: Partial scenes search from shot boundary using textual and visual cues. This search was made in three steps: first, we filtered the list of videos as explained earlier; then we queried for shots inside each video and ordered them by score. As a shot is a unit that is too small to be returned to a viewer, we completed the segment using the scenes boundaries: for each shot, we created a segment that begins with the shot and ends at the end of the scene this shot is part of.

- *part-scenes-noC*: Partial scenes search from shot boundary using textual cues only. Same settings as the previous except for the visual concepts.

- *clustering10-C*: Temporal clustering shots within a video using text and visual cues. The clustering was done as following: First, we filtered out the set of videos to search as explained earlier. Second, we computed scores of every shot in the video, and clustered together shots that were closer than 10 seconds apart. We added the score of shots together to form the score of the segment.

- *clustering10-noC*: Temporal clustering shots within a video using text search only. Same settings as the previous except from the visual concepts.

- *scenes-S or scenes-U or scenes-I*: Scene search using only textual cue from transcript (scenes-I for LIMSI, scenes-U for LIUM) or subtitle (scenes-S). No metadata was used.

- *slidingWindows-60-I or slidingWindows-60-S*: Search over segments created by the sliding window algorithm [42] for LIMSI transcripts (slidingWindows-60-I) and subtitles (slidingWindows-60-S). The size of the sliding windows is 60.

- *slidingWindows-40-U*: The same as above for LIUM transcript with sliding window size of 40.

### 4.5.2.4   The Hyperlinking Approach

For the hyperlinking task, we designed two kind of aproaches:

1. *Hyperlinking using the Search Component.* Our first approach aims to reuse the search logic previously introduced into the hyperlinking task. The challenge was to create from the given temporal anchor a query that looks like the ones considered in the search task. For the first condition (LA, considering

only anchor), the text query was made by extracting keywords from the subtitle (aligned at start time and end time), using Alchemy API default keyword extraction. The visual cues were directly extracted from the shots contained in the anchor. If the anchor was constituted by more than one shot, we took for each concept the highest score among all shots. Then, we performed queries using the scene and shot indexes. For the second configuration (LC, considering context), we used the following methodology: for the textual query, we used both keywords extracted from subtitles aligned to the anchor and from the context containing the anchor, and gave a higher weight to words coming from the anchor. For the visual query, we took the highest score of all concepts across the anchor and its context with no distinction.

2. *MoreLikeThis Component.* Our second approach amounts to use the MoreLikeThis feature of the Solr component combined with the semantic annotations produced by THD (Targeted Hypernym Discovery, see Section 3.2.1.8). This feature from Apache Solr is aimed at finding documents similar to the provided one[26], based on the premise that similar indexed documents are those that have a similar distribution of elements inside. If a user is watching a video, e.g. on the theme "drugs and legalization of cannabis", we suppose that he is also interested in other videos with this subject. We experimented with THD and MoreLikeThis components.

The hyperlinking framework works in 5 main steps:

- Make segments from the LIMSI transcript with sliding window algorithm.

- Annotate text with THD annotations and index them.

- Create, annotate and index temporary document from the anchor for query.

- Ask with MorelikeThis for similar documents.

- Delete temporary document (to exclude contamination of index).

### 4.5.2.5   Results in Search and Hyperlinking

In this section, we report the results of our approaches given the measures described in 4.5.2.3 and 4.5.2.4.

**Search Task.**The results of the search task are listed in table 4.8. Overall, runs using scenes and the sliding window with the size 60 approach have the best performance. First, we notice than under the same conditions, subtitles perform significantly better than any of the transcripts. This was expected since subtitles are

---

[26]http://wiki.apache.org/solr/MoreLikeThis

human-generated and are thus more accurate than transcripts. Hence, the runs that got the best results for each measure were generated using subtitles. If the subtitles were not available, the best option would be to use the scenes with LIMSI transcripts (scenes-I) approach as it yields the higher results over all measures when using transcripts.

It is also interesting to note that **using the visual concepts in the query increases the results for all measures** (e.g., clustering10-C vs clustering10-noC). This improvement could be even bigger if more visual concepts would have been taken into account (we used a subset of 151 concepts), probing the value of multimodal approaches.

Table 4.8: Results of the Search task

| Run | MRR | mGAP | MASP |
|---|---|---|---|
| scenes-C | **0.30946346604** | **0.176992668643** | 0.195100503655 |
| scenes-noC | 0.309135738072 | 0.176714986483 | 0.194691217552 |
| scenes-S | **0.315224269655** | 0.163526098095 | **0.202065496585** |
| scenes-I | 0.261337609154 | 0.144401090929 | 0.158159589898 |
| scenes-U | 0.245763345432 | 0.134440892192 | 0.152835325636 |
| part-scenes-C | 0.22839916764 | 0.12414594745 | 0.102360579399 |
| part-scenes-noC | 0.228115789255 | 0.12399354915 | 0.102092742224 |
| clustering10-C | 0.292888150402 | 0.152518978931 | 0.181357372131 |
| clustering10-noC | 0.284924316767 | 0.147870395447 | 0.171263250592 |
| slidingWindows-60-S | 0.283311232393 | **0.192486941209** | **0.202706179669** |
| slidingWindows-60-I | 0.196479921556 | 0.120561880191 | 0.120417180025 |
| slidingWindows-40-U | 0.236816510313 | 0.134195367345 | 0.150069623509 |

In the following, we further look into the measures and compare the three runs that have the best performances: scenes-C, scenes-S and slidingWindows-60-S. MRR favors the approaches using scenes, scenes-S having slightly higher results. Hence, retrieving segments as scenes returns a better list of results in terms of "corresponding" segments, within a time window of 60s. When looking at mGAP, the sliding window approach is better: as this measure takes into account the distance of the jump-in point to the actual start of the video, it shows that using sliding windows is more precise when it comes to a more fine-grained analysis. Retrieved segments start points are closer to the actual ones, to an extent that enables them to catch up with other approaches with better MRR ranking. Last, MASP measures are very similar for scenes-S and slidingWindows approaches, the latter being higher by very little. This measure takes into account the boundaries of the segment watched: once again, such results hint that the sliding windows approach produces segments closer to the expected ones in term of start time and end time. A next step of this work could then be to refine the scenes in order to preserve the ranking but redefine the

start and end times.

In Figure 4.13 we show the performance of our searching approach (labeled as LinkedTV13 and colored in orange) compared with the rest of results submitted by other participants in MediaEval 2013. In particular the diagram is focusing on MRR when temporal window is set to 10 seconds. Our run "scenes-S" has scored the best one among the candidates, while other configurations like "scenes-C" are falling also on first positions on the ranking, revealing the great potential of our multimodal technique in this task.
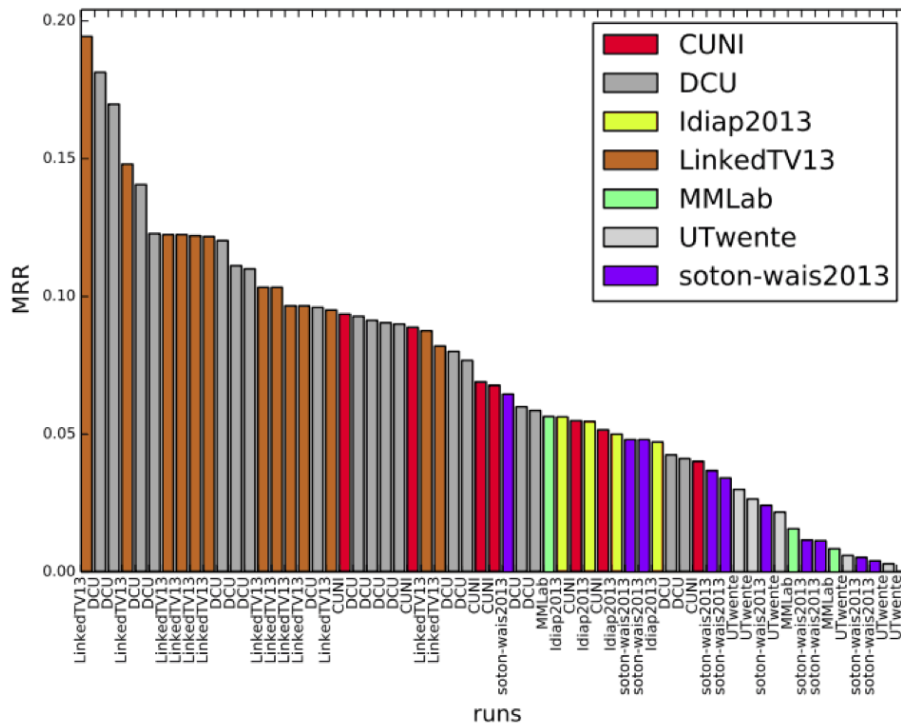


Figure 4.13: MediaEval 2013 Search Task MMR Results (10s window)

**Hyperlinking Task**. The results are provided in table 4.9, where MAP is the mean average precision, and P-5, P-10 and P-20 are the precisions at rank 5, 10 and 20:

For the hyperlinking task, there were originally 98 anchors. Due to time and financial constraints, the organizers have chosen a representative subset of 30 anchors to be evaluated via crowdsourcing. For both LA and LC conditions, runs using scenes outperform other runs for all metrics, while the THD/MoreLikeThis approach comes second. As expected, using the context increases the precision when hyperlinking video segments. It is also notable that the precision at rank n decreases with n. From this observation, we can conclude that those approaches will suit a scenario of

Table 4.9: Results of the Hyperlinking task

| Run | MAP | P-5 | P-10 | P-20 |
|---|---|---|---|---|
| LA clustering10 | 0.2764 | 0.3733 | 0.2967 | 0.2200 |
| LA THDMoreLikeThis | 0.4760 | 0.4667 | 0.4533 | 0.3533 |
| LA scenes | **0.5848** | **0.7467** | **0.6633** | **0.5233** |
| LC clustering10 | 0.3392 | 0.4467 | 0.3733 | 0.2417 |
| LC THDMoreLikeThis | 0.5457 | 0.5733 | 0.5433 | 0.4483 |
| LC scenes | **0.7042** | **0.8533** | **0.8000** | **0.6683** |

a search engine, where mainly the top results will raise the user's attention.

In Figure 4.14 we show the performance of our hyperlinking approach (labeled as LinkedTV13 and colored in orange) compared with the rest of results submitted by other participants in MediaEval 2013. In particular the diagram is focusing on Precision at position 10. Our run "LC scenes" has scored the best one among the candidates, while other configurations like "LA Scenes" are falling also on first positions on the ranking, revealing the great potential of our multimodal technique in this task.
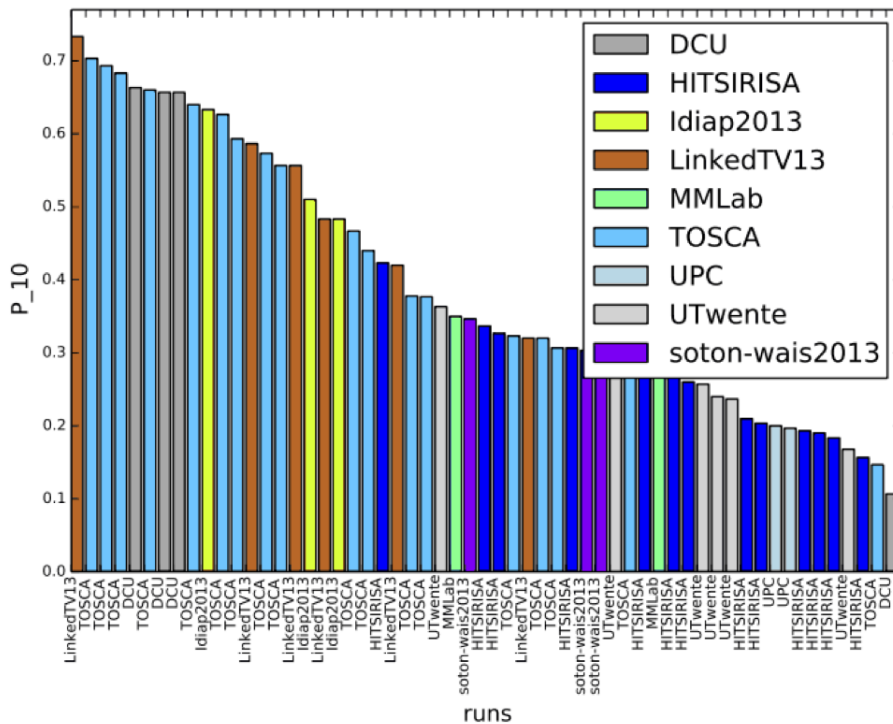


Figure 4.14: MediaEval 2013 Hyperlink Task  Precision Results @ 10

**The Importance of Visual Concepts**. As we described earlier in this docu-

ment, keywords spotted in the text query through a mapping were linked to visual
concepts. We performed a filtering over concepts at two levels: first, we discarded
visual concepts where the "valid detection" score among the top 100 scores was under
0.5. Second, we also applied a threshold of 0.7 on the mapping between text and
concepts on the 50 search queries.

Table 4.10: Number of concepts at various stages

| Filtering level | # concepts before f. | # concepts after f. | # concepts final |
|---|---|---|---|
| visual level | 151 | 47 | 41 |
| mapping level | 976 | 144 | over 50 queries |

As can be seen in table 4.10, our filtering stages drastically reduces the number
of concepts that are actually taken into account in the queries. 976 concepts were
discovered in the 50 queries, and we only take 41 of them into account in the final
queries. Those 41 concepts are spread over 23 queries, the remaining 27 queries hav-
ing no visual concept attached after filtering. However, search results indicate that
using concepts globally provides an improvement over textual query. This finding is
very promising and should lead to greater research work in exploring how to com-
bine different annotations from textual and visual nature into a single multimodal
approach.

#### 4.5.2.6   Lessons Learned

This challenge enabled us to test our approaches with an entire dataset of diverse
shows from a TV provider. Real users evaluated the results in the hyperlinking task,
which is of great interest for testing the different algorithms presented in this thesis.

The results evidence that approaches using scenes outperform others. In the
hyperlinking task, using scenes along with context of the anchor returns results with
a high precision (0.85 and 0.80) at rank 5 and 10: hence, proposing the first items to
a viewer is very likely to reach the expected goal, and emphasizing the importance
of the context also in such kind of media driven task. Hence, improving on scene
segmentation and using features such as speaker clustering, faces or semantics may
be a future direction to study in order to get further in the multimodal philosophy.

We also observe that using visual concepts in the queries improve the results: we
filtered both concepts detected in the videos and concepts mapped from the query,
and still got better results than without using those visual cues. Those outcomes
show that there is room for improvement and the use of visual concepts is another
direction that should be better explored. We see two main processes that are worth
being looked at: first, at the text query level, the mapping between text and concepts;
second, at the video analysis level, the confidence in visual concepts from the video.

### 4.5.3    MediaEval 2014 Participation

After the first class performance submission to the Search and Hyperlinking task at the MediaEval 2013 benchmarking campaign, the LinkedTV consortium participated also in the 2014 edition achieving great results. In this occasion, two alternative methods were proposed to identify which visual concepts are relevant to each query: using WordNet similarity or Google Image analysis. For Hyperlinking, relevant visual concepts were identified by analyzing the video anchor. As one of 9 participants, the LinkedTV submission obtained the fourth best result for the Search sub-task and achieved second best for the Hyperlinking sub-task according to the final results which were made public at the MediaEval Workshop in October 2014 in Barcelona, Spain.

#### 4.5.3.1    Novelties in Task Definition

In 2014, the framework of the Search sub-task has been changed in favor of large scale experiments evaluation, i.e. the queries became more general to allow the ad-hoc search that implies more than one relevant document within the collection [40]. At the same time the visual cues were no longer available for these new queries, thus our investigation into the semantic gap problem was more difficult to carry out.

This absence of visual cues is indeed the most significant change compare to 2013 edition. In MediaEval 2014 Search sub-task, queries are composed of a few keywords only (visual-cues are not provided). Hence, the identification of relevant visual concepts was more complex than last year [40]. The results of both sub-tasks within the ad-hoc scenario were evaluated using the same procedure: pooling of the top 10 results across all participants submissions, relevance assessment of those search results and anchor/target pairs using crowdsourcing, i.e. workers at the Amazon Mechanical Turk platform[27]. In this framework, precision at rank 10 is the most suitable metric to analyze the results, and we use the binned version of it as defined by the task organisers in [3]

#### 4.5.3.2    Updated Processing Chain

For the pre-processing of the data, we followed a very similar workflow to previous campaign but keeping in mind the objective now is to search for media segments inside a video collection given a pure textual query. Videos are pre-segmented into *scenes* and we extract textual and visual features (visual concepts) in order to give grounds to the search. The indexing phase is achieved by extracting the same descriptors on the test set, and using the learned model (on each descriptor) to predict the presence

---

[27]www.mturk.com

of the learned concept in these samples. Then, for each sample per concept, the system assigns a predicted score by fusing its scores from all the different models.

Each candidate scene was ranked according to two scores: the text-based scores computation: $T$ using using Lucene's default text search based on TF-IDF representation and cosine similarity, and the visual-based scores computation: $V$ as we explain below.

As introduced before, the main difference when calculating $V$ comes when generating the candidate visual concepts to be used during our multimodal search approach. In the absence of textual visual cues present in 2013, we tried to generate them directly from the purely text-based query provided this time in two ways:

1. Following the same algorithm used in Section 3.3.9, keywords were extracted from text using the Alchemy API[28], and then mapped with concepts for which a detector is available.

2. The query terms are used to perform a Google Image search. Visual concept detection (using 151 concepts from the TRECVID SIN task [132]) is performed on the first 100 returned images and concepts obtaining the highest average score are selected.

The impact of the threshold applied over the *confidence score $\beta$* of the set of concepts $C^q$ associated to each query $q$ was studied in more deep. We computed the performance of the system with different thresholds $\theta$ that will determine the set of visual concepts which should be included with each $q$. Given the set of concepts $C^q$ for query $q$ and a threshold $\theta$, the selected concepts $C'^q$ are those having $\beta \geq \theta$.

For each query $q$, we compute the visual score $v_i^q$ associated to every scene $i$ as the following:

$$v_i^q = \sum_{c \in C'^q} w_c \times v_i^c, \tag{4.9}$$

where $w_c$ is the valid detection rate of concept $c$, which is used as a weight for the corresponding concept detection score. $v_i^c$ is the score of scene $i$ to contain the concept $c$. The sum is made over the selected concepts $C'^q$.

Notice that when $\theta = 0$, all the set of $C^q$ is included. Therefore, evaluating the threshold $\theta$ was one the main objective of this research and this was compared with two baselines: i) using only text-based search and ii) using text-based search with all available visual concepts $C$ (e.g. the 151 visual concepts).

Once we have computed scores $(V)$ based on visual attributes and considering also scores of the scenes $(T)$ based on the text features we fusion both features in order to obtain the final multimodal ranking of items. The score of each scene is therefore created according to its $t_i$ and $v_i$ scores. Many alternative fusion methods

---

[28]http://www.alchemyapi.com/

are applicable to such situation [45, 12], but in this case a simple weighting fusion
function was chosen:

$$s_i = t_i^{\alpha} + v_i^{1-\alpha} \tag{4.10}$$

where $\alpha$ is a parameter in a range of [0,1] that controls the "strength" of the
fusion method. There are two critical values of $\alpha$: $\alpha = 0$ and $\alpha = 1$. $\alpha = 1$ gives the
baseline (i), which corresponds to the initial text-based scores only. $\alpha = 0$ uses the
visual scores of the corresponding concepts only, which are expected to be very low
on the considered task.

### 4.5.3.3   Dataset and Experiments

The work was conducted on the datasets offered by the MediaEval 2013-2014 Search
and Hyperlinking task, where the test set of 2013 edition became development set
for 2014 experiments. The dataset contains 2323 and 3520 videos from the BBC
(amounting to 1697 hours and 2686 hours) for development and test sets respec-
tively. This represents television content of all sorts: news shows, talk shows, series,
documentaries, etc. The collection contains not only the videos and audio tracks,
but also some additional information such as subtitles, transcripts or metadata.

The queries and anchors for both 2013 and 2014 task editions were created by
users at the premises of BBC. They defined 50 and 30 search queries for the develop-
ment and test sets accordingly, that are related to video segments inside the whole
collection; and 30 and 30 anchors for development and test sets for the input for the
Hyperlinking sub-task. In case of the known-item search, each query is associated
with the video segment seen by the user, described by the name of the video, the be-
ginning and end time of the segment inside the video. In case of an ad-hoc scenario,
these relevant segments were defined after the run submission.

**Search Task**. In this approach, relevant video segments are searched using Solr
using text ($TXT$) only. Two strategies are compared: one where search is performed
at the text segment level directly ($S$) and one where the first 50 videos are retrieved
at the video level and then the relevant video segment is locate using the scene-level
index. The scene-level index granularity is either the Visual-Scene ($VS$) or the Topic-
Scene ($TS$). Scenes at both granularities are characterized by textual information
only (either the subtitle ($M$) or one of the 3 ASR transcripts ( ($U$) LIUM [157], ($I$)
LIMSI [62], ($S$) NST/Sheffield [68])).

Motivated by [158], visual concept scores are fused with text-based results from
Solr to perform re-ranking. Relevant visual concepts, out of the 151 available, for
individual queries are identified using either the WordNet ($WN$) or the GoogleImage
($GI$) strategy. For those multi-modal ($MM$) runs only visual scene ($VS$) segmenta-
tion is evaluated.

**Hyperlinking Task**. Pivotal to the hyperlinking task is the ability to automatically craft an effective query from the video anchor under consideration, to search within the annotated set of media. We submitted two alternative approaches, one using the MoreLikeThis (*MLT*) Solr extension, and the other using Solr's query engine. *MLT* is used in combination with the sentence segments (*S*), using either text (*MLT1*) or text and annotations [36] (*MLT2*). When Solr is used directly, we consider text only (*TXT*) or with visual concept scores of anchors (*MM*) to formulate queries. Keywords appearing within the query anchor's subtitles compose the textual part of the query. Visual concepts whose scores within the query anchor exceed the 0.7 threshold are identified as relevant to the video anchor and added to the Solr query. Both visual (*VS*) and topic scenes (*TS*) granularities are evaluated in this approach.

#### 4.5.3.4  Results

In this campaign the performance of the multimodal approach dropped a little bit compared with other pure text-based methodologies followed by other participants. However considering that there were not visual cues provided on the queries, the results were still remarkable and revealed a good performance even in less advantageous situations like this one. At the same time, we performed extra efforts to quantify the influence of visual content in this kind of combined approached as we will show below.

**Evaluating Influence of Visual Concepts in Multimodal Approach**. The goal of this experiment is to study the influence of the visual concept mapping to text-based queries based on WordNet. In order to achieve this, we have evaluated the proposed method to find the best combination of visual concepts scores with text-based scores, in function of the confidence threshold ($\theta$). We have fixed the values of the $\alpha$ parameter as shown in Section 4.4 in [158]), for $w = Score(c)$ and $w = 1$. The experiment has been performed over subset of the 50 MediaEval search task queries that includes only those that triggered at least one visual concept with high confidence mapping $\beta \geq 0.9$. This results on 21 queries for which the visual information is important, and where the textual description maps to visual concept detectors with a high probability.

Figure 4.15 shows the system performance (with MRR measure) when combining the visual content (selected using threshold $\theta$) with the text-based search approach when $w = 1$. When $\theta = 0$, all mapped concepts (using the WordNet-based mapping) are selected, and as the $\theta$ value increases, the number of selected concepts decreases. In other words, the $\theta$ values perform as a noise remover in the concept mapping, and as it increases the number of mapped concepts decreases. The system performance with the evaluation of $\theta$ is compared to the two aforementioned baselines: i) using
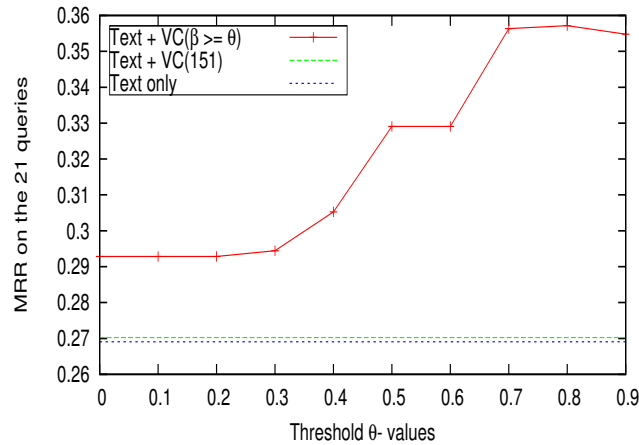
Figure 4.15: MRR values on only 21 queries that have minimum one concept with high confidence ($\beta \geq 0.9$) from WordNet, with different $\theta$-values using concepts validation rate $w = 1$

the text-based scores only and ii) combining the text-based scores with the visual scores of the 151 visual concepts.

As we can see, concept mapping improves significantly the performance of the text-based search task on these queries. Moreover, the best performance was achieved at $\theta \geq 0.7$, with gain of about $32-33\%$ comparing to the text-based search system. This concludes that mapping text-based queries to concepts improves the performance of the search system. The same performance was observed with both mGAP and the MASP measures, but for simplicity we report only the results with the MRR measure.

**Search Results**. Table 4.11 shows the performance of our search runs. Our best performing approach (*TXT_VS_M*), according to MAP, relies on manual transcript only segmented according to visual scenes. Looking at the precision scores at 5, 10 and 20, one can notice that multi-modal approaches using WordNet (*MM_VS_WN_M*) and Google images (*MM_VS_GI_M*) boost the performance of text only approaches. There is a clear performance drop whenever ASR (*I, U or S*) are employed, instead of subtitles (*M*). Same difference between ASR and manual transcript based runs was observed across submissions of the other participants.

Figure 4.16 shows in details the performance of all participants runs in terms of precision at rank 10 as evaluation score. LinkedTV was the only participant that addressed the visual aspect of the task and achieved high results that are present in the top. The other runs that achieved the top performance (e.g. DCU, CUNI) based their techniques on manual transcripts and use of metadata and prosodic features.

**Hyperlinking Results**. Table 4.12 shows the performance of our hyperlinking runs. Again, the approach based on subtitle only (*TXT_VS_M*) performed best (MAP = 0,25) followed by the approach using MoreLikeThis (*TXT_S_MLT1_M*).

Table 4.11: Results of the 2014 Search sub-task

| Run | map | P_5 | P_10 | P_20 |
|---|---|---|---|---|
| TXT_TS_I | 0,4664 | 0,6533 | 0,6167 | 0,5317 |
| TXT_TS_M | 0,4871 | 0,6733 | 0,6333 | 0,545 |
| TXT_TS_S | 0,4435 | 0,66 | 0,6367 | 0,54 |
| TXT_TS_U | 0,4205 | 0,6467 | 0,6 | 0,5133 |
| TXT_S_I | 0,2784 | 0,6467 | 0,57 | 0,4133 |
| TXT_S_M | 0,3456 | 0,6333 | 0,5933 | 0,48 |
| TXT_S_S | 0,1672 | 0,3926 | 0,3815 | 0,3019 |
| TXT_S_U | 0,3144 | 0,66 | 0,6233 | 0,48 |
| TXT_VS_I | 0,4672 | 0,66 | 0,62 | 0,53 |
| TXT_VS_M | **0,5172** | 0,68 | 0,6733 | **0,5933** |
| TXT_VS_S | 0,465 | 0,6933 | 0,6367 | 0,5317 |
| TXT_VS_U | 0,4208 | 0,6267 | 0,6067 | 0,53 |
| MM_VS_WN_M | 0,5096 | **0,7** | **0,6967** | 0,5833 |
| MM_VS_GI_M | 0,509 | 0,6667 | 0,68 | **0,5933** |
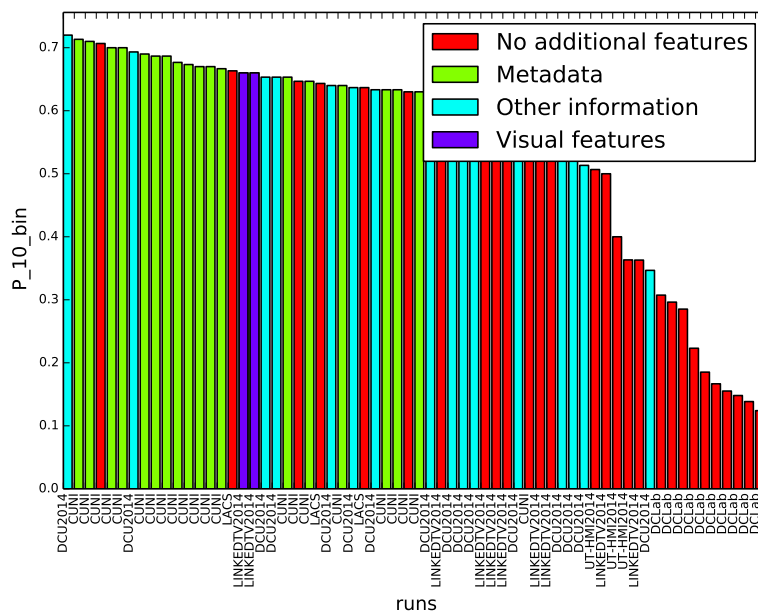


Figure 4.16: Mediaeval 2014 Search Performance (All Participant's Runs)

Multi-modal approaches did not produce the expected performance improvement. this is due to the significant duration reduction of anchors compared with last year which meant that less visual and audio context was available for processing and feature extraction.

Figure 4.17 brings the Hyperlinking sub-task results in context of comparison with the other participants submissions. The runs that use visual features achieve lower scores, however LinkedTV approach using visual features is still better than the other

Table 4.12: Results of the Hyperlinking sub-task

| Run | map | P_5 | P_10 | P_20 |
|---|---|---|---|---|
| TXT_S_MLT2_I | 0,0502 | 0,2333 | 0,1833 | 0,1117 |
| TXT_S_MLT2_M | 0,1201 | 0,3667 | 0,3267 | 0,2217 |
| TXT_S_MLT2_S | 0,0855 | 0,2067 | 0,2233 | 0,1717 |
| TXT_VS_M | **0,2524** | **0,504** | **0,448** | **0,328** |
| TXT_S_MLT1_I | 0,0798 | 0,3 | 0,2462 | 0,1635 |
| TXT_S_MLT1_M | 0,1511 | 0,4167 | 0,375 | 0,2687 |
| TXT_S_MLT1_S | 0,1118 | 0,3 | 0,2857 | 0,2143 |
| TXT_S_MLT1_U | 0,1068 | 0,2692 | 0,2577 | 0,2038 |
| MM_VS_M | 0,1201 | 0,3 | 0,2885 | 0,1923 |
| MM_TS_M | 0,1048 | 0,3538 | 0,2654 | 0,1692 |

runs submitted. As the anchors became shorter this year, removing completely the notion of the **context**, so the metadata proved to become important for the task performance.
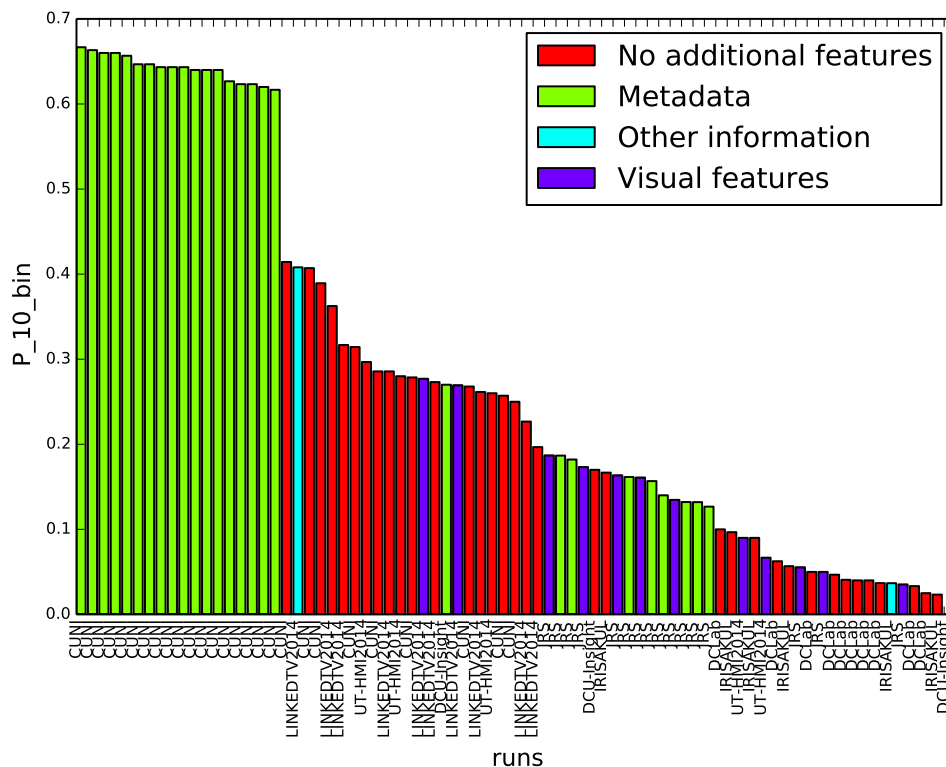


Figure 4.17: Mediaeval 2014 Hyperlinking Performance (All Participant's Runs)

The results of LinkedTV's approaches on the 2014 MediaEval S&H task show that it is difficult to improve over text based approaches when no visual cues are

provided. But overall, our investigation into the use of visual cues found in the data and its temporal structure has shown that our technique is competitive and manages to achieve high score results. The work was distinguished with a "Distinctive Mention Award" during the closing session of the MediaEval 2014 workshop which indicates that the techniques being covered in this thesis are able to provide leading results for Multimedia Search and Hyperlinking.

### 4.5.4 MediaEval 2015 Participation

For 2015 the Search and Hyperlinking task in MediaEval workshop has been splitted into two different events: SAVA2015[29] inside MediaEval 2015, specialized the search aspect and this time also considering the automatic selection of convenient anchors for a given set of videos, being those anchors media fragments for which users could require additional information. Eurecom has submitted a contribution to the workshop [43], this time with a different approach that, even inspired in the multimodal hyperlinking techniques covered in this thesis, does not consider visual features as critical for performing the search operation and only rely on them for a late re-ranking of media fragments previously promoted using pure textual features. Due to the low level of participation in this year edition, the organizers have not been able to produce a fair comparison between the few runs provided by the contenders.

The hyperlinking sub-task has became part of the TRECVID evaluation campaign [30]. The task has gotten contributions from 10 different international teams, showing the good acceptance of the task inside the venue and high demand on the development of these techniques. For the benchmarking, the data from 2014 has been used this time putting special emphasis on the anchoring generation via a editor tool that has been used by journalists, employees of British Film Institute, and students in journalism, and relying on Amazon Mechanical Turk for the ground truth generation. The approach submitted to this track has been explained in [129], and is based on the indexing and ranking workflow presented to SAVA2015 but adapted to be trigged by the provided video anchors. From a total of 100 runs submitted (being 67 multimodal) Eurecom has performed on third place in the list of participants in terms of Mean Average Precision (MAP), as shown in Figure 4.18.

---

[29]http://www.multimediaeval.org/mediaeval2015/searchandanchor2015/
[30]http://www-nlpir.nist.gov/projects/tv2015/tv2015.html
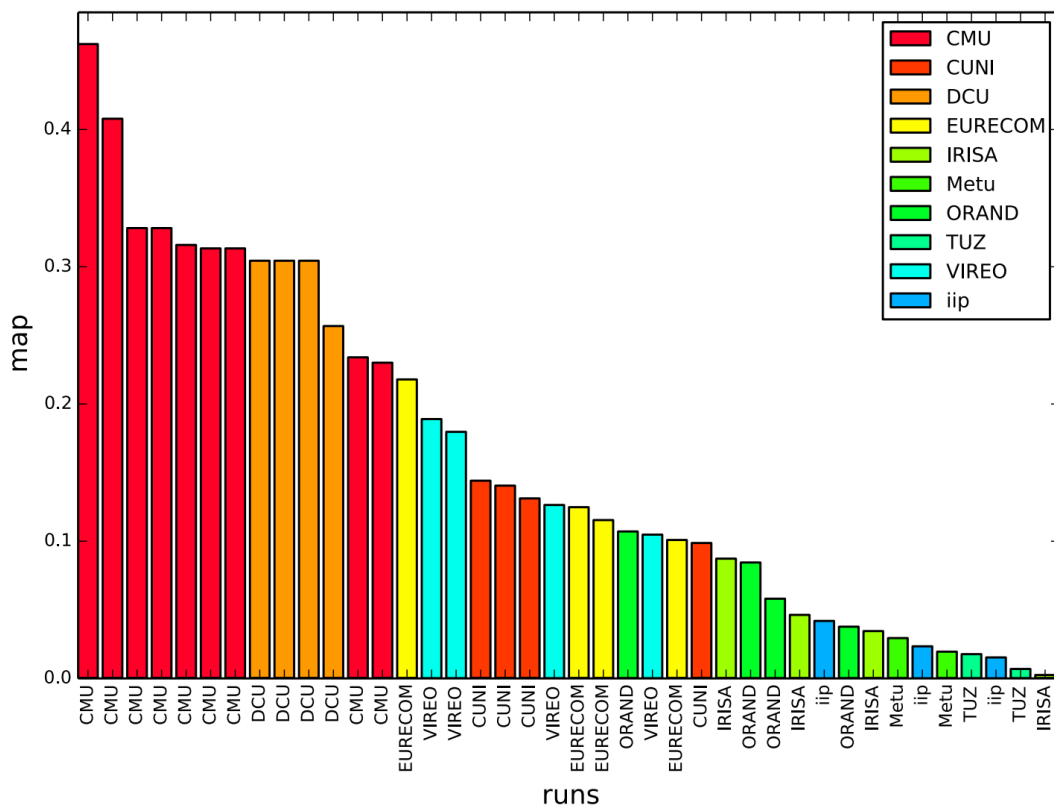
Figure 4.18: TRECVID 2015 Video Hyperlinking Performance MAP (All Participant's Runs)

## 4.6 Summary

In this Chapter we have showcased different techniques leveraging on annotations produced according to methods in Chapter 3 for being able to offer advanced features over multimedia content at a fragment level. In particular, we have elaborated on four kinds of media operations: fragment enriching, temporal fragments refining and promoting, video classification, and multimodal fragment hyperlinking.

Concerning media fragment enrichment, we have showcased how is possible to retrieve relevant content from different sources (mainly social networks, white-listed Web sites, and other online resources) in order to further illustrate, complement or emphasize what is happening inside a media fragment. The obtained resources can be explicitly attached to the seed media fragments who originated the enrichment operations by using the same ontology-based methodology introduced in Section 2.4 in order to make them available on the Web in a Linked Data fashion.

Also, we have developed a prototype for automatically discovering Hot Spots in online and educational videos. We leverage on a visual analysis of the multimedia content and in the knowledge available on Web for detecting the fragments better illustrating the main topics being told in the video. The possibility of having the video annotated with those Hot Spots allows the viewer to quickly decide if the video is interesting for him and incentives a more fine-grained consumption of the resource in the form of media fragments.

Semantic annotations can also support innovative means of video classification by applying named entity recognition and media fragment generation techniques over video subtitles. First results of this techniques are promising, proving that the temporal distribution of the entity types is a distinctive feature for recognizing video categories. Enlarging the sample size of the dataset may produce a better model and defining clearer classification outcome. This kind of unsupervised techniques are suited to be applied over video items and metadata from others video sharing platforms in order to improve the performance and reduce the temporal and human cost of those categorization systems. Relying not only in names entities but also in other more visual features is a clear step forward in improving video classification and retrieval that needs to be better explored.

We have also studied how multimodal approaches combining textual based features with visual techniques perform in the task of hyperlinking of multimedia content at the level of fragments. Through the different MediaEval campaigns like the 2013, we have been able to investigate the performance of multimodal searches over a collection of media fragments, emphasizing on the difficulty to accurately interpret the user's query, with or without explicit visual cues. The proposed strategy, which semantically maps visual cues inferred from textual queries to LSCOM visual concepts based on WordNet similarity, improves significantly the performance of the video search

system (up to 41% MRR and 40% mGAP). Even on cases where visual cues are not available like 2014 campaign, we attempted to overcome the so-called semantic gap problem by automatically mapping text from the query to visual concepts, resulting in an improvement against the pure text based searches. At the same time we have highlighted the importance of considering different levels of granularity in both the video anchors triggering the hyperlinking and in the obtained results in order to better adequate to the users' needs.

As final remark, from the results obtained along the different approaches covered in this Chapter we have identified an increasing need of specialization in order to keep improving their results. General algorithms can reach adequate scores in performance but in order to go further and provide high quality features, it is needed to better focus on a particular kind of content and better identify the use case to be targeted.

# Conclusion of Part I

In the first part of this thesis, "Towards a Semantic Multimedia Web", we have investigated on how turn multimedia content into a first citizen of the Web, with all the advantages that this can bring to the different actors in the scenario: from TV broadcasters, to people uploading videos to YouTube showing their expertise, passing through students who want to learn and explore in a more interactive way or just some relatives reunited around a screen for the sake of entertainment.

We have proposed an ontological model able to describe a wide variety of multimedia content at the level of fragment, succeeding in bringing into a single ecosystem many different media descriptors covered in the literature. This way the media information becomes available on the Web in a Linked Data fashion that promotes a better reusing, more effective interlinking with other data in the Web, and easier exploitation by third parties.

However content does not always come annotated. In most cases, only raw videos with not metadata attached are provided. Given the huge amount of multimedia documents being uploaded every day in this conditions, we need to rely on automatic annotation techniques that alleviate the amount of time and efforts human annotators spend on such tasks. On the one hand, we have revealed the huge potential of semantic approaches working over the textual dimension of the videos: they can spot certain anchors on the text corresponding to entities mentioned in the video, which are automatically linked with other resources on the Web where more information can be fetched. Those anchors are characterized according to different vocabularies that further shape the semantic context of the information being transmitted on the video. On the other hand, the visual part of the multimedia documents cannot be left aside: techniques working over this features have been traditionally developed outside the Web ecosystem. Thanks to the LinkedTV model we have showcased how they can effectively be brought in, to empower a new generation of multimodal algorithms.

In the last Chapter of this Part we have focused on algorithms that rely on the advantages of having the content represented and annotated as described before, making the most of working over standard data formats and having available the huge amount of information the Web offers. This leads to new ways of enriching content with very diverse information (from the fresher social network to more curated Web sources for editorial purposes), better selecting meaningful temporal boundaries for fragments, identifying parts of the video that are potentially more relevant, automatically classify clips so they can be faster ingested and searchable, and generate hyperlinks between different videos inside a collection in order to explicitly relate the fragments the user is watching with other chunks that could be worth for him/her

to watched.

This new multimedia ecosystem opens the window to a new scenario of possibilities, not only for content providers that can make their information available in a better manner, but also for viewers who want to enjoy a richer, better contextualized, multimedia consuming experience. Content being published is well structured in the form of interlinked Web resources, annotated with less efforts, in ways that were not imaginable years back. Providers and third parties can leverage on those annotations for creating new operations that traverse the content in more intuitive and powerful ways, ultimately benefiting the users.

Convinced that the techniques covered during this first part of the thesis will lay the foundations for a better suited and more innovative approaches, in next Part II we will deeper elaborate on the annotation of international news items, a very particular kind of video documents where satisfying viewers' information needs is particularly critical in order to properly understand the story being told.

# Part II

# Advanced Semantic Annotation
of News

# Overview of Part II

In Part II, we apply and extend the methodologies and techniques described during Part I to semantically annotate news items. We identify the various open challenges in building up a proper context that summarizes the plot of the stories and can assist the users in consuming news. We will iterate over the initial context generation algorithm in order to improve it in an incremental manner, while justifying the decisions taken.

In Chapter 5 **The Semantic Snapshot of a News item**, we motivate the need of recreating the semantic context of a news item, in order to allow humans to better consume the different stories happening in the world, and machines to perform advanced operations over them. We propose the so-called News Semantic Snapshot (NSS) of a news items, a semantic data structure that explicitly represents this context in terms of relevant entities involved. We describe the methodology for creating of the Gold Standard that will be used in subsequent chapters in this part, and introduce a first approach to programmatically generate the NSS of a news story. This logic has been published as a Web service.

In Chapter 6 **The Multidimensionality of the News Entity Relevance**, we highlight the importance of exploiting different relevancy dimensions to properly generate the NSS. Story-related entities can become part of the ideal NSS because of very different reasons, and frequency based techniques introduced in previous Chapter are not enough to not to spot them. We formalize the different parameters involved in each phase of the News Entity Expansion algorithm, propose new ranking functions working on different relevancy dimensions, and execute different experiments in order to identify which configuration and ranking method works better when recreating the NSS of news stories.

In Chapter 7 **The Concentric Nature of the News Semantic Snapshot**, we highlight the difficulties of approaching the NSS generation by combining different relevancy functions. Integrating them into a single relevancy score makes very hard to achieve significant improvements over the experiments performed in previous chapter. For this reason, we propose a revolutionary concentric model approach for reconstructing the NSS of news, based on two main layers: the *Core*, composed by entities which are highly representative for the story, but sometimes too evident for the users; and the *Crust*, where other relevant entities, semantically attached to the *Core* for very different reasons, are included. This new conceptual model, together with some new metrics prioritizing Recall and size of the sample against more application-dependent rankings focused only in top positions, led us to take the NSS generation process to a much better level of performance.

In Chapter 8 **The NSS in the News Consumption Paradigm**, we analyze the versatility and powerfulness of the NSS in supporting applications and prototypes

assisting users in consuming news. Through different concrete examples implemented during the period of this thesis, we show how the entities organized in a concentric model and the duality between *Core* and *Crust* layers can lead to a better conceptualization of the news story that has a positive impact in the way those applications are designed and implemented.

# The Semantic Snapshot of a News item

## 5.1 Introduction

Consuming information is an undoubtedly complex task, both for humans and machines. Data comes in very different manners and formats, and the process of interpreting and making sense out of it is never a trivial task. Sometimes, the documents to be consumed are preceded or followed by other excerpts of information that are necessary to properly understand the message inside them. Those pieces around, influencing the meaning of the initial document and refining its effects, is what we call the *context*. The importance of this concept has been well noted by other research works in other areas, from multimedia annotation and semantic technologies [203] to the information retrieval techniques in medical domains[35], for example.

In general, the more complex an information unit is, the bigger the context we need to successfully consume it. In this regard, international news are deeply complicated by nature. News items in textual and audiovisual form come full of connotations, codes and conventions, related topics, implicit paradigms, unwritten referents, and very different ideologies [71]. The plots of the stories they tell contain a considerable number of agents involved, evolve in a variety of places, and can be related with previous events that explain the causes behind the current facts. Looking at the television domain in particular, with the emergence of both citizen-based and social media traditional information channels are re-thinking their production and distribution workflow processes which are now much more complex. TV newscasts need to quickly report about the latest event-related facts occurring in the world. They often deliver partial information thus neglecting the whole picture of the event that is often assumed as known. For all those reasons, news items would benefit from a contextual information who further complements the stories they tell.

In this chapter we define a semantic structure called "Newscast Semantic Snapshot" (NSS) in order to explicitly represent the context of a particular news item. This knowledge abstraction intends to summarize into a single unit the necessary information for making sense of the news story being told, providing humans and machines with details that were not explicitly included in the original news docu-

ment. In this Chapter we illustrate this need of a context via some examples, and propose a first set of technologies to automatically generate the NSS of a news item and evaluate it against a Gold Standard. This research has been published at [143].

## 5.2    The Need of Contextualizing News Items

In this section we highlight the need of contextualizing international news stories through two different examples from the printed and audiovisual media catalog. After presenting them we will formulate our hypothesis on how the News Semantic Snapshot of a news item can alleviate this problem, and how the semantic techniques explained in PartI can help in this mission.

### 5.2.1    Augmenting News Items: Angela Merkel

On the 3rd of October 2015, Reuters news agency has published a news article [1] where the German chancellor Angela Merkel calls the European Union to protect its external frontiers as the continent is facing the greatest influx of refugees since World War II. This news item is understandable by reading the text and checking out the accompanying pictures, and the core details of the story seem to be there (see Figure 5.1).



Figure 5.1: News Item reporting Angela Merkel declarations on refugee crisis, Oct 3rd 2015

---

[1]http://www.reuters.com/article/2015/10/03/us-europe-migrants-germany-merkel-idUSKCN0RX0A020151003

However even if the news item can be interpreted with the available information, the story is lacking a little bit of the big picture about the explained fact. Why are those refugees coming so numerous in this particular period of time? A little bit more of context about this massive immigration flux would be very beneficial for understanding the current article. In this regard, entities like Syria, a country which is currently facing a war that is forcing people to scape the country, or another german politic figure Wolfgang Schäuble [2], who previously talked about the economic cost for the country of hosting the refugees do help to contextualize story being told. These and other entities would strengthen the current information available in the news item, providing users with a more comprehensive view of the issue (see Figure 5.2).
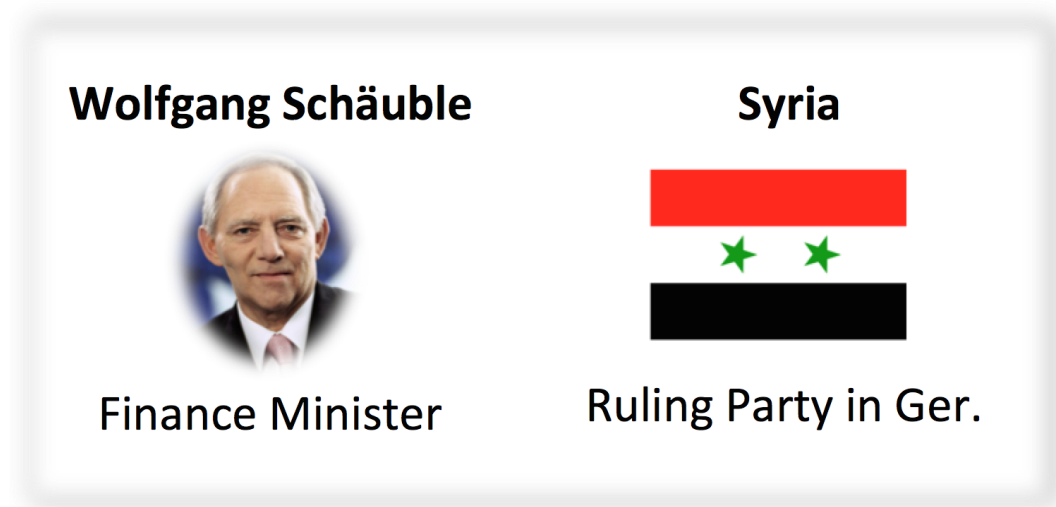


Figure 5.2: Entities augmenting Angela Merkel declarations on refugee crisis, potential candidates for becoming part of this article's NSS

### 5.2.2 Reinforcing News Items: Snowden's Asylum

Sometimes the news stories are not only missing pure contextual information augmenting the scope of the described fact, but they also present incomplete or scarce information. On the 17th July, 2013 Edward Snowden gave a press conference in a Russian airport asking for asylum. He was unable to travel to Latin America, where he was granted the right to stay. The BBC published a video item describing those an certain details around this interview (see Figure 5.3).

But some of the details mentioned and depicted in the video are vague and would need further clarification for certain viewers who want to dig more into the details. For example, it is being said that the interview is being held in an airport in Moscow, but there are two in the capital. In addition, in the images we can see a lady sitting

---

[2]http://panteres.com/2015/09/08/wolfgang-schauble-refugee-crisis-has-priority-in-budget/

Figure 5.3: Edward Snowden asking for asylum in a Russian Airport, Jul 17th 2013

next to him, but neither in the banners not in the narration it is explained who she is. These and other entities would reinforce and fix missing information in the current news item in order to offer the viewers a more clear descriptions about the facts (see Figure 5.3)



Figure 5.4: Entities reinforcing Edward Snowden press conference on asking asylum to Russia, potential candidates for becoming part of this article's NSS

To wrap up we can see that we live in a globalized world, a vast playing field where events happening are the result of complex interactions between many diverse agents along time. The interpretation of those news is problematic because of two issues: *i)* the *need of background*: viewers probably need to be aware of other facts happened

in a different temporal or geographic dimension, and *ii)* the *need of completeness*: a single representation of an event is not enough to capture the whole picture, because it is normally incomplete, it can be biased, or partially wrong.

### 5.2.3   Hypothesis: The News Semantic Snapshot

The examples shown above make emphasis on the importance of additional information for further complementing and refining news stories, and reinforces our hypothesis about news consumption: a proper context bringing in the big picture of the facts being addressed is necessary. For this reason, we define what we have called the "News Semantic Snapshot", a semantic knowledge structure representing the context of a news item.
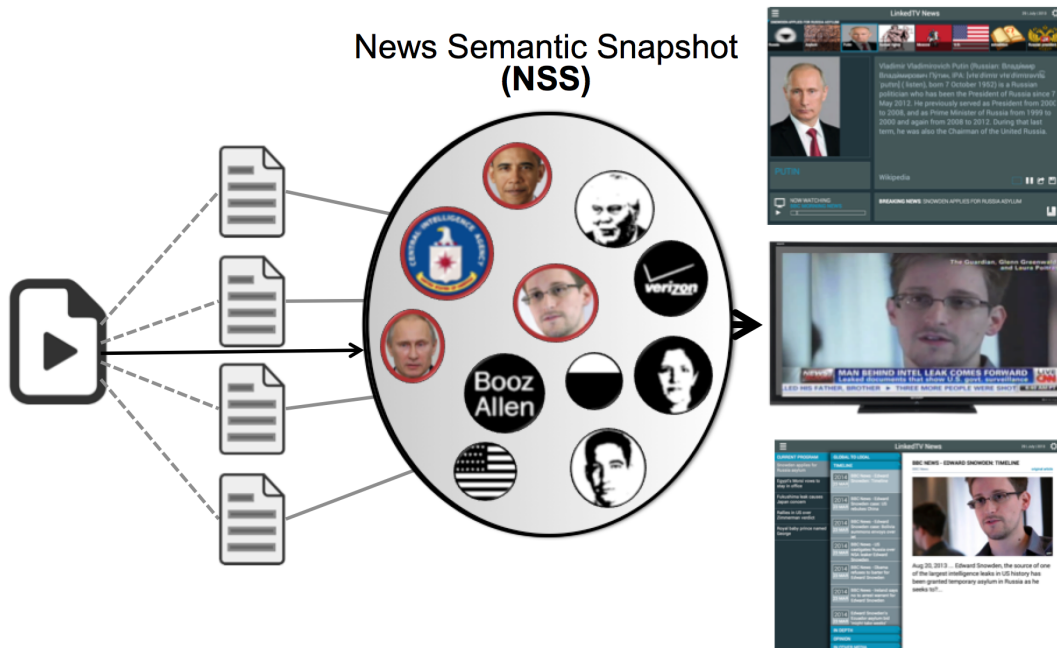


Figure 5.5: The News Semantic Snapshot (NSS): a set of relevant entities describing the big picture of a news story, which cannot be generated out of a single news testimony

As depicted in Figure 5.5, in our research this $NSS$ is composed by a set of semantic entities $e_i$ that are relevant for the facts addressed on the news item and play a role in the story being told.

$$NSS_{story} = e_{1_{d_i}}, ..., e_{n_{d_i}} \tag{5.1}$$

Given the nature of the news articles published on media, it is not possible to reconstruct the whole context of a news story by analyzing a single news item talking

about it. This semantic context is instead an upper-level unit integrating a wider set of view points, which requires a fair amount of documents talking about the same fact in order to be properly created.

The NSS has as primary objective: to be able to serve machines and humans in advanced operations over the news items. Viewers can better consume the story being told via applications leveraging on data contained in the NSS, while machines can exploit certain entities to trigger enrichments, propose recommendations, or link to other related documents that the users would like to explore.

**NSS for feeding News Applications**. Recent gadgets and applications, such as second screen devices, have recently irrupt as a good way of assisting the viewer in the challenge of becoming aware of that bigger picture of the event and complement the missing information problems reported before. In [31], the authors tracked 260 tablet users, concluding that even though there is a modest uptake and interest in using secondary screens to digitally share opinions, the use of second screen interaction with television content is something the viewer qualitatively appreciate.

At the end of the day those second screen applications need to be fed with meaningful details concerning a newscast. The way they offer the information to the viewers is very important, but the real challenge is in how we get this information ready in the first place. Such kind of applications would clearly benefit from having an explicit data representation of the news story in the form of a NSS, so they can directly leveraging on it without having to worry about implementing complicated information retrieval and selection techniques. In Chapter 8 we will analyze how adequate the NSS data structure is for supporting different kind of applications assisting users during news consumption.

## 5.3   State of the Art on News Stories's Context Generation

The need of a NSS for feeding certain applications is concept we have investigated in previous research works and prototypes like [144], which have probed the benefit for users of browsing the "surrounding context" of the newscasts. The same concept[3] has been presented to the forum Iberoamerican Biennial of Design (BID).[4] with great feedback from users and experts.

Research efforts have also underlined the importance of automatic collection algorithms, data processing, and social science methods for reporting and storytelling in professional journalism, under the term of computational exploration in journalism [66] (CEJ). CEJ demands precise event descriptions for supporting the journalistic co-creation of quantitative news projects that transcend geographical, disciplinary,

---

[3]https://vimeo.com/119107849
[4]http://www.bid-dimad.org

and linguistic boundaries. Projects like NEWs have studied how to disambiguate
named entity in the news domain by continuously learning while processing news
streams [50]. In the domain of Social Networks, named entities are also used for
identifying and modeling events and detecting breaking news. In [182], the authors
emphasize the importance of spotting news entities in short user generated post in
order to obtain a better understanding about what they are talking about.

The need for contextualizing news documents has been also tackled in some re-
search works such as [27], where the authors describe the challenge of generating
textual summaries of news items based on an entity-centric process of traversing a
graph. This generates a sub-graph containing the most relevant entities and the
structural relationships among them. In the particular case of this paper, those re-
lationships are used as the means for building human readable summaries instead of
aiming to create general and shareable data structures like the NSS is.

In order to build a comprehensive NSS that contextualizes the knowledge ex-
pressed in the new video, the entities spotted on subtitles have to be extended with
others that were not explicitly mentioned in the video item, but are still relevant. As
we will show in subsequent sections in this chapter, our approach performs an entity
expansion process that allows to collect on the fly event-related documents from the
Web. In the literature there are already some approaches relying on similar expansion
techniques, even if the driving force to decide about its relevance is not a news event
or story, but its belonging to a more general and usually easy to identify category.
They transform the feature space from a small number of named entities with the
same type to a more complete named entity set. One of them was Google Sets, which
is not longer available[5]. This expansion using the Web is also closely related to the
problem of unsupervised relation learning [22], and set-expansion-like techniques have
been used to derive features for concept-learning [29], to construct "pseudo-users"
for collaborative filtering [30], and to compute similarity between attribute values in
autonomous databases [212]. In [207], authors proposed the Set Expander for Any
Language (SEAL) approach. SEAL works by automatically finding semi-structured
web pages that contain lists of items and aggregating those lists in a way that the
most promising items are ranked higher. SEAL is a language-independent system
and it has shown good performance against previously published results like the al-
ready mentioned Google Sets. By using particular seeds and the top one hundred
documents returned by Google, SEAL achieves 93% in average precision in dataset
from various languages. The same authors published an improved version of the
algorithm [208], increasing the performance by handling unlimited number of super-
vised seeds. In each iteration, it expands a couple of randomly selected seeds while
accumulating statistics from one iteration to another. Our approach does not rely

---

[5]http://googlesystem.blogspot.fr/2012/11/google-sets-still-available.html not longer available since
2014.

on such kind of iterative mechanism, and it focuses on maximizing the quality of
the search query for obtaining the most appropriate set of related documents to be
analyzed. Another approach in extending set of entities is [187], which combines the
power of semantic relations between language terms like synonymy and hyponymy
and grammar rules in order to find additional entities on the Web sharing the same
category that the ones provides as input. Relying on Google they analyze documents
for parsing semi-structured text elements like tables and rank the final candidates
using different ranking algorithms like PageRank. Numerous approaches have dealt
with a set expansion method using free text rather than semi-structured Web docu-
ments; for instance authors in [184] presented a method for automatically selecting
trigger words to mark the beginning of a pattern, which is then used for bootstrap-
ping from free text. But again, this approach gathers category-related entities while
in our case the criteria for judging the relevancy of an entity is more its relatedness
with the story told in the news item.

In some other approaches such as [189], the contextualization process is not based
on entities but on other elements like related documents complementing the origi-
nal, or text snippets. However, a machine readable representation of the context of
the news items is still needed so it can be exploited by machines to offer operations
like searches and recommendations. In [188], the context is built up from different
keywords directly spotted in the news document without leveraging on other exter-
nal sources. Another difference is the existence of human intervention during the
workflow since the keywords have to be manually highlighted by the users.

Other examples of news contextualization efforts are found in [206], where the
authors aim to complement the original news document with the most interesting
reactions from different online platforms. In order to identify relevant candidates,
they propose to look at different dimensions including interestingness, informative-
ness, opinionatedness, and popularity, whose importance will be emphasized during
chapter Chapter 6 of this thesis. The difficulty of dealing with those dimensions
inspired also the implementation of the concentric approach presented in Chapter 7
that tries to ease the way relevance of the entities is managed by integrating them
under a common knowledge model that is easier to populate and consume.

To the best of our knowledge, there is no related work in the news domain that has
tackled the annotation of news stories by grounding in the power of Web expansion
algorithms over the initial set of initial entities retrieved in subtitles. In the rest of
Part II we will describe the research work done in this area during the period of this
thesis, which has been published in [143], [145], [57], and [58]. In Section 5.5 inside
this same chapter, we will present a first NSS generation implementation based on
a naive document collection strategy for bringing in related documents, and a pure
frequency-based algorithm ranking function for promoting entities. The next chapters
will elaborate on improving the initial approach in several directions: more specific

| Newscast Title | Date | Person | Organization | Location | Total |
|---|---|---|---|---|---|
| Fugitive Edward Snowden applies for asylum in Russia | 2013-07-03 | 11 | 7 | 10 | 28 |
| Egypt's Morsi Vows to Stay in Power | 2013-07-23 | 4 | 5 | 4 | 17 |
| Fukushima leak causes Japan concern | 2013-07-24 | 7 | 5 | 5 | 13 |
| Rallies in US after Zimmerman Verdict | 2013-07-17 | 9 | 2 | 8 | 19 |
| Royal Baby Prince Named George | 2013-07-15 | 15 | 1 | 6 | 22 |
| **Total** | | 46 | 20 | 33 | 99 |

Table 5.1: Breakdown entity figures per type and per newscast.

document retrieval mechanism, better ranking based on different dimensions, and advanced methodologies for organizing the knowledge inside the NSS.

## 5.4 Gold Standard for Evaluating Newscast Semantic Snapshot

We are interested in evaluating ranking strategies for generating semantic snapshots of newscasts, where each snapshot is characterized by a set of named entities. To the best of our knowledge, there are not datasets of news articles annotated with entities belonging to their corresponding context, but only those which are explicitly mentioned in them. For this reason we have created our own dataset, following the methodology described below.

We narrowed down the selection of named entity types to Person, Organization, and Location, since they can be directly translated in *who*, *what*, *when*, a subset of the fundamental questions in journalism known as the 5Ws. The title of the newscasts and the breakdown figures per entity type are shown in Table 5.1. The dataset is freely available[6].

### 5.4.1 Newscast Selection

We randomly selected 5 newscasts from the BBC One Minute World News website[7]. Each newscast lasted from 1 to 3 minutes. The selection covered a wide range of subjects specifically: politics, armed conflicts, environmental events, legal disputes, and social news. Subtitles of the videos were not available; therefore, a member of the team manually transcribed the speech in the newscasts. After obtaining the transcriptions, the following steps were performed in order to obtain a set of unbiased candidate entities.

---

[6]https://github.com/jluisred/NewsEntities
[7]http://www.bbc.com/news/video_and_audio/

### 5.4.2   Newscast Semantic Annotation

The annotation process involved two human participants: an annotator and a journalist (expert of the domain). No system bias affected the the annotation process, since each annotator performed the task without any help from automatic systems. The output of this stage is a list of entity candidates. The annotators worked in parallel.

The annotator of the domain was asked to detect for each newscast entities from:

**subtitle** : the newscast subtitle.

**image** : every time a recognizable person, organization or location was portrayed in the newscast, the entity was added to the list.

**image captions** : the named entities appearing in such tags, such as nametag overlays, were added to the candidate set.

**external documents** : the annotator was allowed to use Google Custom Search to look for articles related to the video. The query followed the pattern: title of the newscast, date. The sources were considered: The Guardian, New York Times, and Al Jazeera online (English). The results were filtered out by keeping only those falling inside the seven days period around the day when event took place.

The journalist, with more than 6 years of experience as a writer and editor for important American newspapers and websites, acted as the expert of the domain. He was asked to watch the newscasts and to identify for each the entities either mentioned or not that better serve the objective of showing interesting additional information a final reader. He was completely free to suggest any named entity he wanted.

### 5.4.3   Quality control and Ranking

A quality control, performed by another expert of the domain, refined the set of entities coming from the previous stage, eliminating all named entity duplicates and standardizing names. We then conducted a crowdsourcing survey with the objective to gather information about the degree of interestingness of the entities for each newscast. Based on [205] we define interestingness whether an entity is interesting, useful or compelling enough to tear the user away from the main thread of the document. Fifty international subjects participated in this online study. They responded an online call distributed via email and social networks. Their age range was between 25 and 54 years with an average age of 30.3 (standard deviation 7.3 years). 18 participants were female and 32 were male. Most of the participants were highly educated

and 48 of them had either a university bachelor degree or a postgraduate degree. The main requisite for participation was that they were interested in the news and followed the news regularly, preferably through means that include newscasts. During the interview participants were asked to choose at least 3 out of 5 videos according to their preferences. Then they were shown each one of the newscasts. Then they were asked to rate whether they would be interested in receiving more information about the named entities in the context of the news video and on a second screen or similar application. All the named entities from the candidate set related to the last seen video were shown in a list with ratio buttons arranged in a similar way to a three-point Likert-scale. The possible answers were "Yes" "Maybe" and "No".

### 5.4.4 Results from Online Survey

The number of respondents per video were: Snowden 49, Morsi 34, Fukushima 42, Zimmerman 27, and Royal baby 15. In order to calculate the interestingness scores from the users responses, we gave every answer a numerical value: Yes = 1, Maybe = 0 and No = −1. We then obtained an average score for each entity using the number of participants that rated an entity and the score that each participant gave to that particular entity. This average was used to obtain a score that ranked all the entities in the candidate set according to the user's responses. The list of ranked entities per video obtained from the online survey results is available at Appendix 6.

Finally, with the objective of finding out users preferences regarding the three analyzed entity types: person, organization and location, we calculated an average rating for each one of the entity types. The results in descending order were: organization = −0, 05, person = −0, 24 and location = −0, 52. These results suggest a preference from users for the entities of the type organization and person over those of the type location when getting informed about international news.

## 5.5 A first Approach for Generating NSS: the Snowden Asylum Case

Today users have access to multiple news portals, different services for commenting and debating on the news, and social media that instantaneously spread news information. However, this results in large amount of unreliable and repeated information, leaving to the user the burden of processing the large amount of potentially related data to build an understanding of the event. The generation of the NSS of a news story is a very challenging task that currently requires a high efforts by humans trying to interpret the news, and would be benefit from more automatic strategies that this problem with a much lower human intervention and in a shorter time.

One strategy reported in the literature for tackling such objective is to perform

named entity extraction over the newscast transcript [101]. However, the set of named entities obtained from such an operation is insufficient and incomplete for expressing the context of a news event [77]. Sometimes entities spotted over a particular document are not disambiguated because the textual clues surrounding the entity are not precise enough for the name entity extractor. While in some others, entities are simply not mentioned in the transcripts while being relevant for understanding the story. This is also an inherent problem in information retrieval tasks: a single description about the same resource does not necessarily summarize the whole picture. In this paper we automatically retrieve and analyze additional documents from the Web where the same event is also described, in a process called Newscast Named Entity Expansion. By increasing the size of the document set to analyze, we increase the completeness of the context and the representativeness of the list of entities, reinforcing relevant entities and finding new ones that are potentially interesting inside the context of that news item. This approach is able to produce a ranked list of entities called Newscast Semantic Snapshot (NSS), which includes the initial set of detected entities in subtitles with other event-related entities captured from the seed documents.

In order to recreate the context of a news item, we have applied the Named Entity Expansion algorithm described in Section 3.2.3, Part I. This entity annotation technique does not only rely on the metadata of the video itself, but analyzes other relevant Web documents that are collected by using entities spotted in subtitles via named entity extraction algorithm as seeds to crawl the Web. Our initial hypothesis states that this mechanism is able to bring to the table entities that, despite they were not explicitly mentioned in the original news video, are important to understand the context of the described story. As the process is also introducing noise on the form of non relevant entities being retrieved by the expansion method, we perform a selection over the list of candidates relying mainly on frequency measures. The big picture of this two step procedure is depicted on Figure 5.6.

The results in terms of recall when comparing with the Gold Standard presented in Section 5.4 are very promising and reveal that this first step of the News Semantic Snapshot generation process (component on top in Figure 5.6) is bringing in many relevant entities that were missing. In particular, average recall of the entities spotted by named entity extraction over the subtitles of the videos in the ground truth is 0.42 while the average recall using Named Entity Expansion goes up to 0.91, proving the benefit of relying on the related document collected.

We will also conduct a first evaluation of the results obtained after performing the selection phase depicted at the bottom of Figure 5.6, which at the moment is purely based on frequency measures as described in Section 3.2.4.4. In particular, we will analyze the results obtained after applying the workflow over one of the videos[8]

---

[8]http://www.bbc.co.uk/news/world-europe-23339199
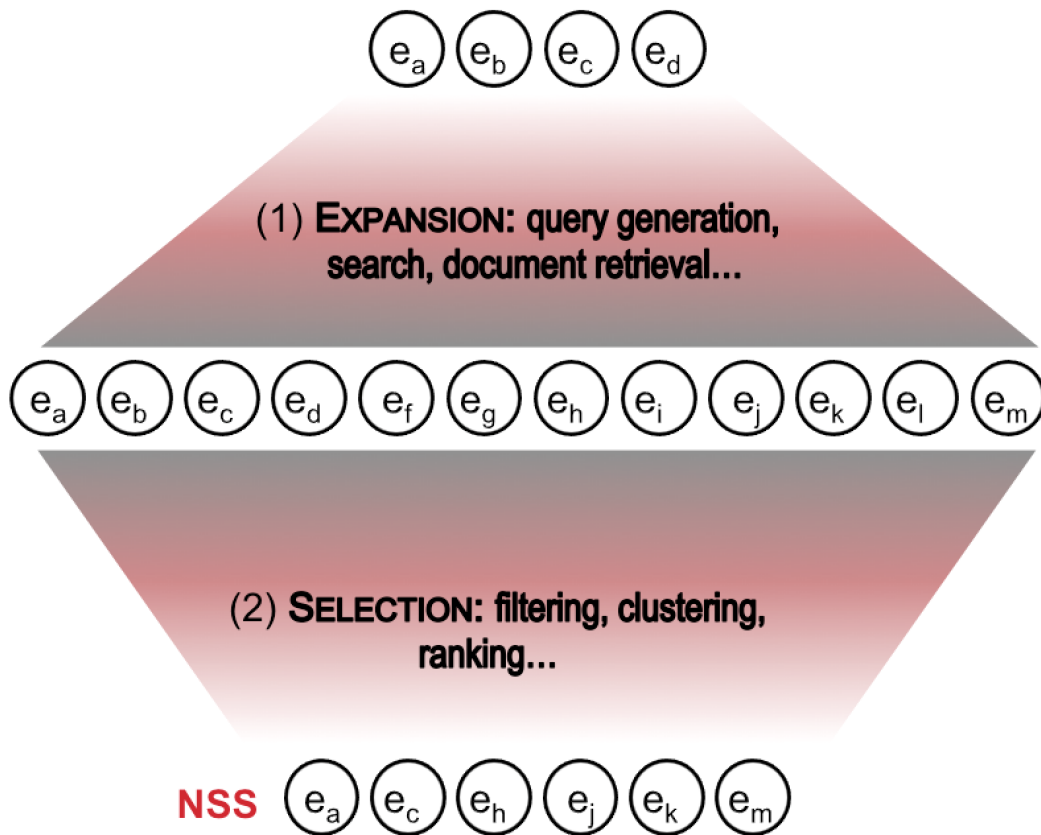
Figure 5.6: NSS generation approach: results obtained via Named Entity Expansion are filtered to build the list of entities being added to the final NSS

considered in the Gold Standard: the request of asylum made by Edward Snowden to the Russian government. In an airport in Moscow, he publicly express his desire to obtain political help while he can find a safe way to reach the Latin American countries that offer a safe harbor. The time period for which this particular event is relevant goes from 2013-07-06 to 2013-07-17.

### 5.5.1 Named Entity Extraction

In a first step, Name Entity Extraction techniques are applied over the video transcript using NERD. In Table 5.2 we show the list of entities directly extracted via this procedure, which brings a first approximation to the context of the news media item.

| Label | Relevance | Sentiment | Type | URI |
|---|---|---|---|---|
| Russia | 0.809216 | Mixed | Location | DBpedia:Russia |
| Edward Snowden | 0.717369 | Mixed | Person | DBpedia:/Edward_Snowden |
| South America | 0.56586 | Mixed | Location | DBpedia:South_America |
| president Putin | 0.459811 | positive | Person | DBpedia:Vladimir_Putin |
| president | 0.401138 | negative | JobTitle | DBpedia:President |
| Moscow | 0.352101 | Mixed | City | DBpedia:Moscow |
| CIA | 0.334887 | neutral | Organization | DBpedia:CIA |
| Bolivia | 0.324607 | neutral | Location | DBpedia:Bolivia |
| Obama | 0.321901 | negative | Person | DBpedia:Barack_Obama |
| human rights | 0.317419 | negative | Thing | DBpedia:Human_rights |
| law enforcement | 0.280055 | negative | Thing | DBpedia:Law |
| one month | 0.280055 | neutral | Quantity | DBpedia:Month |

Table 5.2: Raw result from Named Entity Extraction with NERD

| Label | Relevance | $F_{video}$ | $F_{Docu}$ | Type | URI |
|---|---|---|---|---|---|
| Russia | 1.0 | 7 | 264 | Location | DBpedia:Russia |
| Edward Snowden | 0.80479 | 2 | 227 | Person | http://dbpedia.org/resource/Edward_Snowden |
| US | 0.61643 | 5 | 160 | Location | DBpedia:United_States |
| Vladimir Putin | 0.39383 | 1 | 111 | Person | DBpedia:Vladimir_Putin |
| asylum | 0.32876 | 4 | 80 | Thing | DBpedia:Right_of_asylum |
| Barack Obama | 0.31506 | 1 | 88 | Person | DBpedia:Barack_Obama |
| Moscow, Russia | 0.30479 | 1 | 85 | Location | DBpedia:Moscow |
| Central Intelligence Agency | 0.19178 | 0 | 56 | Location | DBpedia:Central_Intelligence_Agency |
| Anatoly Kucherena | 0.147260 | 0 | 43 | Person | – |
| extradition | 0.116438 | 2 | 26 | Thing | DBpedia:Extradition |
| White House | 0.10616 | 0 | 31 | Location | DBpedia:White_House |
| Sheremetyevo | 0.0890 | 0 | 26 | Location | DBpedia:Sheremetyevo_International_Airport |
| WikiLeaks | 0.08219 | 0 | 24 | Organization | DBpedia:WikiLeaks |
| Washington's | 0.075342 | 0 | 22 | Location | DBpedia:Washington,_D.C. |

Table 5.3: Top entities obtained via Named Entity Expansion

## 5.5.2   Named Entity Expansion

The query generated for retrieving additional documents has the text "Edward Snowden asylum Russia", and is bounded to the period from 2013-07-06 to 2013-07-28. After analyzing the additional documents retrieved by the search engine, the instances are re-ranked according to their global frequency as shown in Table 5.3

The final result after the expansion operation is a bigger list of entities (more than 100). In some cases, the previously spotted entities (like *Extradition*) have been promoted in the hierarchy while others (like *Right_of_asylum*
) have been discovered in the new documents. For this use case, we show only a fixed number of top entities ($mainEntities = 15$).

## 5.5.3   Preliminar Evaluation

We first evaluate the set of entities obtained by NERD (see Table 5.2)against the entities available about the Snowden video in the ground truth. In terms of precision and recall, 8 entities out of the 12 originally spotted belong to the gold standard, which is contains a total of 28 entities. Entities like the lawyer of the case, *Anatoly Kucherena*, have been never mentioned in the subtitles and therefore not detected. Also, even the words such as *airport* and *Moscow* are present in the subtitles, we have not discovered the exact place the action is being taken, given that there are two airports in the city.

Next, we evaluate the set of *MainEntities* inside $E_{Expansion}$ depicted in Table 5.3. To be fair in the comparison with the results obtained via entity expansion on the same subtitles, we have only considered the two first 12 results on the table. Even

this is only the top of the list generated by expansion, it already shows how results are improved. The precision and recall indexes are better: there are 10 entities spotted versus the 8 we had for NERD, out of the 12 considered. Asylum, probably one of the more representative concepts behind this news item, has been correctly detected and disambiguated. *Anatoly Kucherena*, the lawyer involved in the defense of Edward *Snowden*, has been correctly proposed in the result, even if no disambiguation URL is provided. The name of the airport is now known: *Sheremetyevo* and it has been correctly disambiguated. Other important entities such as *extradition* are now present in the final set, helping to complete the context of the news.

The preliminary results indicate that we can successfully expand the initial set of recognized entities with more relevant concepts not detected by pure named entity recognition approaches. Analyzing the additional named entities occurring in related documents leads to a more accurate ranking of important concepts and it brings forward more relevant entities with additional information about the broader context of a news event.

However, taking a deeper look at the annotations available in the gold standard for the same news video (see Table 6.1 on Appendix6) and comparing them this time with the complete list of entities obtained by Entity Expansion, we identify some entities that are highly considered in the ground truth but have been placed at the very last positions of the expansion ranking. Two clear examples are *Glenn Greenwald*, a lawyer and journalist who published the first series of reports about the classified documents disclosed by Edward Snowden, and *Laura Poitras*, film director and producer that won the 2015 Academy Award for the best documentary for "Citizenfour", about Edward Snowden. By analyzing the original documents retrieved we have observed how, even those entities have been brought to the table, they appear in just a few related articles suggesting that a purely frequency based algorithm for judging entity relevance is not enough for promoting all the items considered in the gold standard. In order to promote such kind of entities we need to rely on alternative ranking methods incorporating other considerations that will be studied in Chapter 6.

## 5.6   News Expansion Service

In order to make the Named Entity Expansion algorithm publicly available and allow external parties to apply this logic over their video corpora, we have created a REST service that intends to recreate the context of a TV newscast through the identification of representative and relevant named entities. It relies on the idea introduced in Section 3.2.3 and used previously in this chapter, based on retrieving and analyzing additional documents from the Web where the same facts displayed on the video are also described. This service has been created under the scope of the LinkedTV

project and can be reached at: http://linkedtv.eurecom.fr/entitycontext/api/

This service makes use of the search capabilities offered by Google CSE [9]. This tool allows to create a customized search engine over a set of Web sites specified by their URLs. After the search engine is configured, it is possible to make search request through its API, which allows to specify the textual query to be executed and other different parameters like the number of results to be retrieved.

For the named entity extraction task, we use two different online services, in particular TextRazor and AlchemyAPI, already introduced in Section 3.2.1.3.

A live demo displaying the entities found by this service for an example newscast is available at http://linkedtv.project.cwi.nl/news/ and will be further described in Chapter 8.

### 5.6.1   Input Parameters

There are four main parameters that need to be provided to the service in order to perform the Named Entity Expansion logic:

1. The transcript of the video to be processed, as plain text.

2. The temporal window in which the action displayed in the video has taken place, via two parameters: the starting and ending date.

3. the ID of a Google Custom Search Engine. If no ID is provided, the journal based default CSE engine is used.

### 5.6.2   Implemented Features

The main reasons for the combined use of TextRazor and Semitags as NE extractors is the different behavior they offer: based on different experiments performed during their integration under NERD, the former is more conservative and prioritizes precision against recall, so it has been used for extracting the first round of entities from the subtitles (seeds). On the other hand, TextRazor proposes more candidates at the cost of being less precise, so it has been used annotating the second round of documents where the ranking algorithms try to get rid of the false positives.

The query generation phase described in Section 3.2.4.1 has been built on top of the annotated transcripts provided by Alchemy API through NERD framework, in order to align the 10 different NERD Core ontology classes to the corresponding *Five W's* concept of information gathering in journalistic reporting. For the collection phase, we have used a Google Custom Search Engine initialized with a list of well-know international newspapers which has reduced the scope of the search to semi-curated documents with a fair level of relatedness with the journalist domain.

---

[9] https://www.google.com/cse/

For more details about the particular newspapers considered, please visit https://www.google.com/cse/setup/basic?cx=014567755836058125714:alz73j11kbk. The annotation of documents retrieved after the collection phase has been performed using TextRazor with default parameters for entity recognition.

For clustering different instances of the same entities, we have applied Jaro-Winkler string metrics over the entity labels. In future versions we foresee to use more elaborated distances relying also on Wordnet and exploiting the entity types. Concerning the ranking phase explained in Subsection 3.2.4.4, we have implemented two different approaches further described below:

1. Named Expansion Ranking according to Representativity: The more frequent an entity is, the more "iconic" can be inside the story behind the newscast. Considering this frequency and the importance scores coming from the different extractors, we create a rank of entities that are representative of the newscast being played on TV.

2. Named Expansion Ranking according to Relevance (TF-IDF approach): an entity can be repeated many times and be representative for a story without being really attractive from a viewer point of view. For alleviating this problem, we divide the absolute entity frequency by the number of external documents where this entity was mentioned. We higher penalize extreme cases: entities appearing in a very low number of documents (probably errors or anomalies) and also those appearing in nearly the entire set of documents because they don't really add anything new.

### 5.6.3 REST API examples

There are two main REST API methods, each of them giving support to one of the Ranking operations presented in previous subsection:

```
curl -X POST --data-binary @snowden.txt "http://linkedtv.eurecom.fr/entitycontext
    /api/entities/relevant?startdate=20130703&enddate=20130716&cse=CSE_ID&limit
    =50" --header "Content-Type:application/json" -v
```

This call launches the Named Entity Expansion process using relevancy based ranking algorithm over the transcript file uploaded via the *–data-binary* field, with the following parameters:

1. startdate: day when the event began, in the format *YYYYMMdd*.

2. enddate: day when the event was finished. Clients can play with this dimension for being less/more strict in the relatedness of the retrieved documents.

3. limit: Number of entities to be displayed in the results.

4. cse: ID of the Google custom search engine to perform named entity expansion. If not provided, this service will use the one available by default.

```
curl -X POST --data-binary @snowden.txt "http://linkedtv.eurecom.fr/entitycontext
    /api/entities/representative?startdate=20130703&enddate=20130706&cse=CSE_ID&
    limit=10" --header "Content-Type:application/json" -v
```

This call launches the Named Entity Expansion process using representativeness-based ranking algorithm over the transcript file uploaded via the *–data-binary*, and uses the same parameters specified above.

### 5.6.4 Output

The results of this service are offered in the form of a JSON serialization containing an array of Named Entities serialized according to the NERD ontology. Some additional fields have been added for exposing the ranking scores that have internally generated the final scores, like the attributes "appearancesVideo", "appearancesDocuments", or "maxExtractorRelevance".

```
<pre>
 {
   "label":"Edward Snowden",
   "totalRelevance":0.9895833333333334,
   "maxExtractorRelevance":1.0,
   "appearancesVideo":4,
   "appearancesDocuments":1124,
   "startChar":10897,
   "endChar":10911,
   "nerdType":"http://nerd.eurecom.fr/ontology#Person",
   "extractorType":"Freebase:/award/award_winner,/influence/peer_relationship,/
       people/person",
   "uri":"http://en.wikipedia.org/wiki/Edward_Snowden"
 }
</pre>
```

In its current version, the implemented features take time to serve a response since it is invoking other external services that a certain latency when answering back.

## 5.7 Summary

In this chapter we have introduced the need of contextualizing news items in order to be properly consumed and interpreted, by both humans and machines. To make this context explicit, we have proposed a semantic knowledge structure called "News Semantic Snapshot" (NSS), defined as the set of Named Entities that are relevant for summarizing the big picture of the news story being reported. In order to be able to evaluate possible approaches that automatically generate this NSS, we have produced a Gold Standard described in Section 5.4. Preliminary implementation efforts suggest that Named Entity Expansion techniques relying on different information sources on

the Web are a good way to bring up the missing context entities, providing the original newscast with information that was not originally available in it. Ranking algorithms based on frequency have revealed to be effective to filter out the less important entities that are inevitably coming during the expansion process, and promote important concepts that therefore become part of the final NSS.

However, this implementation still does not consider integrating other different relevancy dimensions into the final ranking list, nor exploiting the semantic relationships between the spotted entities, so it fails at prioritizing certain entities that are barely mentioned along the retrieved related documents, but have been highly considered by users inside the ground truth and therefore should become part of the NSS. In future experiments we plan to further tuning the expansion's collection phase and study alternative ranking methods considering other entity relevancy aspects, in order to improve the scores obtained so far when evaluating the NSS generation approach against the Ground Truth.

# The Multidimensionality of the News Entity Relevance

## 6.1 Introduction

As introduced in previous Chapter, relying exclusively on the broadcasted news item is insufficient to fully grasp the context of the fact being reported. We have proved that machine driven approaches can alleviate the human difficulties for identifying and processing the huge amount of data available in the Web about news stories, but they struggle both in finding a good set of candidate documents, and filtering them. In this chapter we further develop the approach presented in Section 5.5 for generating the Semantic Snapshot of the considered Newscast (NSS) to provide viewers and experts of the domain comprehensive information to fully understand the news content. This extended approach takes as inputs the publication date and the newscast's title for gathering event-related documents on the Web that are automatically analyzed to detect the most relevant semantic annotations in them, bringing useful contextual information to the originally available knowledge. Named entities detected in the retrieved documents are merged with those found in the newscast subtitles for further disclosing hidden relevant concepts that were not explicitly mentioned in the original newscast. We leverage on different ranking algorithms based on entity frequency, popularity peak analysis, and domain experts' rules. The generated NSS has been benchmarked against the gold standard presented in Section 5.4. Results of the experiments show the robustness of the approach holding an average normalized discounted Cumulative Gain of 69.6%. This research has been published at [145].

The rest of the chapter is organized as follows: Section 6.2 highlights the limitations of the frequency based approach presented in Chapter 5 for generating NSS and introduces how this problem will be addressed, Section 6.3 introduces the new characteristics of the Entity Expansion method and formalizes the different parameters involved. Section 6.4 describes in depth the different ranking algorithms used for ordering the list of candidate entities generated in previous steps, bringing the desired multidimensionality. The experimental settings and results are reported in Section 6.5.

## 6.2 The Selection Problem: Towards a Multidimensional Entity Relevance

In this section we take a deeper look at the preliminary results obtained by applying the basic Named Entity Expansion approach over news items' subtitles as initially described in Section 5.5, in order to identify its weakness and justify the need of implementing multidimensional based ranking methods for promoting relevant entities when annotating news stories.

But before going into details we will illustrate the entity selection problem introduced in bottom half of Figure 5.6 through a diagram that helps to visualize how certain entities obtained via Named Entity Expansion are promoted for becoming part of the NSS.
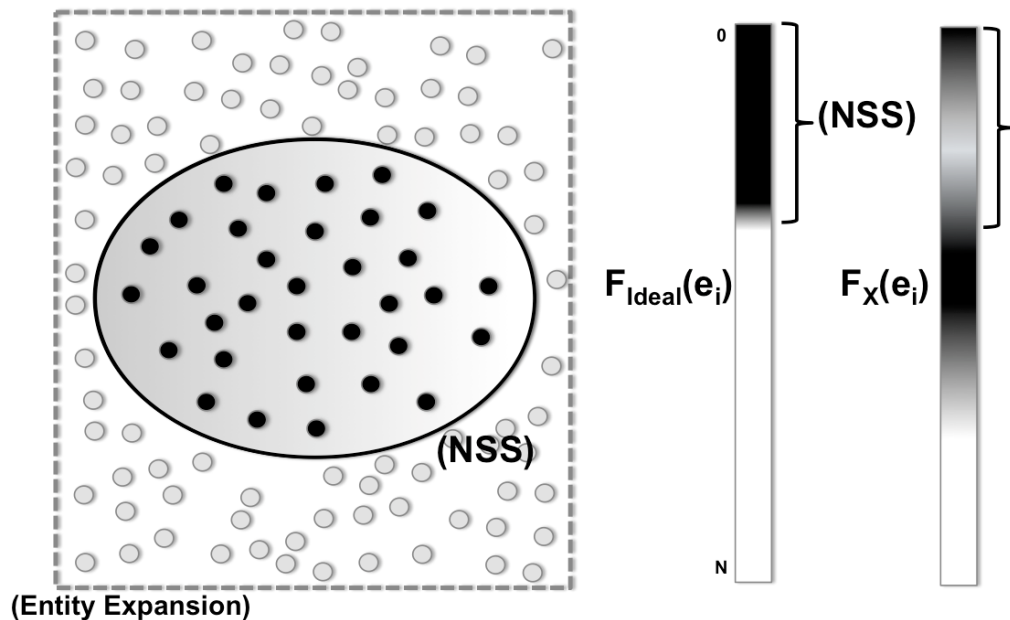


Figure 6.1: The Selection Problem: candidates from Expansion process need to be filtered to build the NSS of the news item

The squared region labeled as *EntityExpansion* in Figure 6.1 represents the set of entities (small elements inside the space) collected via the Expansion method. As introduced at the beginning of Section 5.5, this technique increases average recall to 0.91 from the 0.42 obtained using Named Entity Extraction, meaning that most of the entities in the gold standard are retrieved during the process. However together with the desired entities (positives, colored in black), many other concepts not relevant for the current news item are also collected (the negatives, colored in light grey). The objetive is to identify the firsts, in order to generate the News Semantic Snapshot of the analyzed video (circle containing entities in black).

In this situation we applied entity selection through ranking. In order to distinguish the positives from the negatives inside the entity expansion set, we aim to find a function $F_{ideal}(e_i)$ that perfectly ranks the entities in an unidimensional space $[0, N]$ so the positives are placed at the top positions. The NSS is therefore created by taking those $n$ first entities and discard the long tail of results placed below. Unfortunately, in real use cases ranking functions behave more as $F_x(e_i)$ does: it succeeds to promote some relevant entities to the top of the list, but many others are still placed in medium or lower positions (false negatives, colored in black-grey at the middle ranges), or get to be promoted when they have nothing to see with the current story (false positives, colored in white at the top positions).

Now in Figure 6.2 we extrapolate the selection problem to the implementation of NSS generation method described on Section 5.5, where the ranking function $F_x(e_i)$ is based on frequency measures.
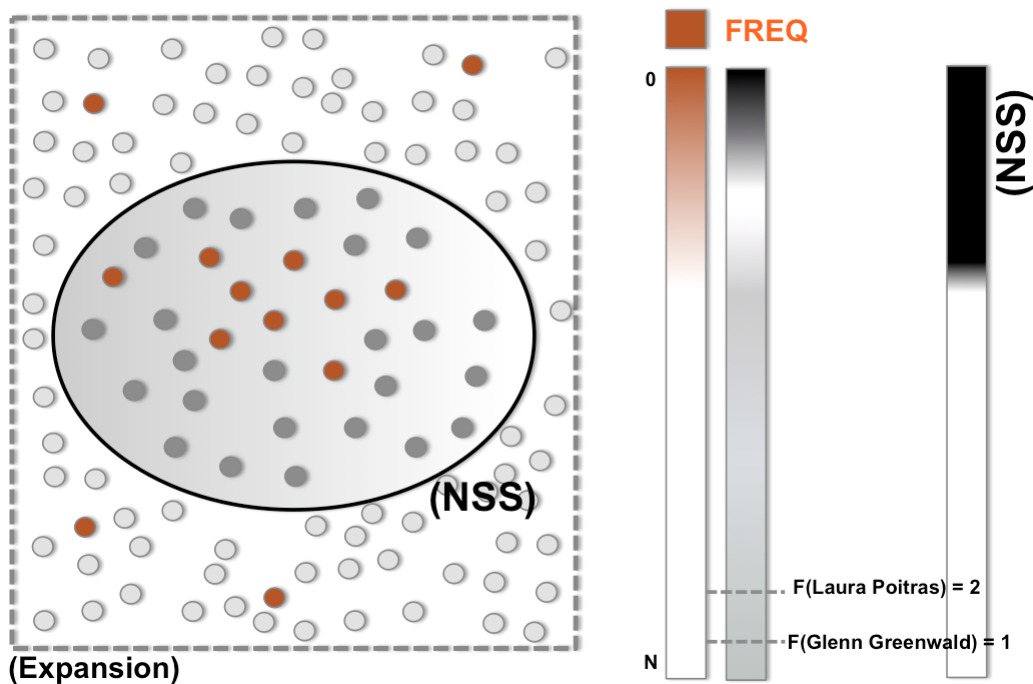


Figure 6.2: Frequency based function for selecting entities to be part of the NSS

As we can see, entities colored in orange are highly scored and therefore promoted to become part of the NSS by the frequency ranking function, labeled as $FREQ$ in the figure. Immediately we find that the ones falling inside the NSS region are the true positives, while the ones outside those boundaries are the false positives. Keeping in mind this function does not replicate the ideal ranking (see how the strong black colored area on the top is shorter than it should if all relevant entities would be placed there, and lower positions are not fully white meaning some false negatives

are present), it has succeeded in bringing to the NSS positions a fair amount of entities that were relevant for the news story.

Hence, the question arising after analyzing the frequency based ranking selection diagram is: how can we get to promote the rest of the entities inside the ideal NSS that have not been considered yet? Going deeper through some particular examples of entities already introduced in Section 5.5.3 like *Laura Poitras* and *Glenn Greenwald*, they are barely mentioned in the related documents ($F_x$(Poitras) $\approx$ $F_x$(Greenwald) $\lll 1$). Therefore they fall at the very end positions of the frequency ranking distribution as illustrated in the orange colored horizontal bar .Those entities will never be highly scored by this TF based functions. This has motivated us to study other different functions $F_x(e_i)$ that could select the entities inside the NSS round area that still remain on grey color.

In particular, in this chapter we will try to go over the results obtained by the former implementation of the NSS generation algorithm by making emphasis in two axes:

1. Further formalize the parameters involved in the collection phase, in order to fine tune the document retrieval process and therefore improve the entity space generated via named entity expansion (squared region where selection is applied over).

2. Implementing other ranking functions $F_x(e_i)$ working in different dimensions other than frequency, in order to reduce the number of false negatives obtained during the selection phase and be able to spot entities that remain relevant to the news story without being highly repeated along the related documents.

## 6.3   Revisited Entity Expansion Approach

In this section we further formalize the algorithm to automatically generate News Semantic Snapshots out of a particular newscast already introduced in Section 5.5 and 3.2.3, identifying the different parameters involved in each step of the sequential workflow. In a nutshell, the approach we use to generate Newscast Semantic Snapshot is composed of the following stages: query formulation, document retrieval, semantic annotation, annotation filtering, and annotation ranking. Fig. 6.3 depicts the whole expansion process.

**Query Formulation** Newscast broadcasters offer a certain amount of metadata about the items they publish, which is normally available together with the audiovisual content itself. In this work, we build the query $q = [h, t]$, where $h$ is the video heading, and $t$ is the publication date. Unlike the former NSS generation approach, we do not rely on subtitles during the collection phase, but they will be considered
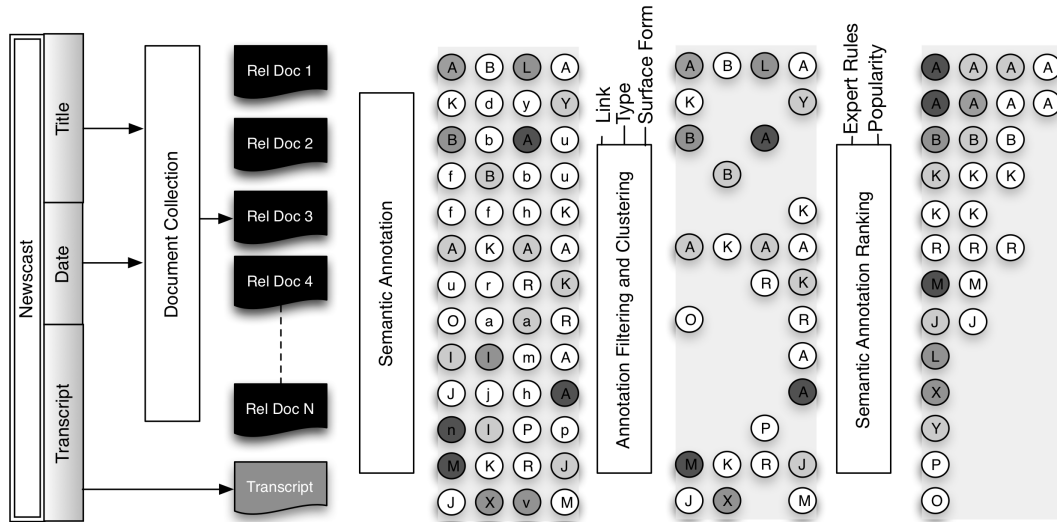
Figure 6.3: Schema of Named Entity Expansion Algorithm.

during the subsequent semantic annotation steps. The query is then used as input of the retrieval stage.

**Document Retrieval** The retrieval stage has the intent to collect event-related documents from the open Web. To some extents this process emulates what a viewer, who misses some details about the news he is watching, does: going to the Web, make a search, and get the most of the top ranked documents. Our programmatic approach emulates this human driven task by analyzing a much bigger set of related documents in a drastically smaller amount of time. The stage consists of retrieving documents that report on the same event discussed in the original video as result of the query $q$. It has a key role in the upcoming semantic annotation stage, since it selects a set of the documents $D$ over which the semantic annotation process is performed. The quality and adequacy of the collected documents sets a theoretical limit on how good the process is done.

**Semantic Annotation** In this stage we perform a named entity recognition analysis with the objective of reducing the cardinality of the textual content from the set $D$ of documents $\{d_1, ..., d_n, d_{n+1}\}$ where $d_{i=1,..,n}$ defines the $i_{th}$ retrieved document, while $d_{n+1}$ refers to the original newscast transcript. Since most of the retrieved documents are Web pages, HTML tags and other annotations are removed, keeping only the main textual information. The feature space is then reduced and each document $d_i$ is represented by a bag of entities $E_{d_i} = e_{1_{d_i}}, ..., e_{n_{d_i}}$, where each entity is defined as a triplet $(surface\_form, type, link)$. We perform a union of the obtained bags of named entities resulting in the bag of entities $E$ of the initial query $q$.

**Annotation Filtering and Clustering** The Document Retrieval stage expands the content niche of the newscast. At this stage we apply coarse-grained filtering

of the annotations $E$ obtained from the previous stage, applying a $f\left(E_{d_i}\right) \to E'_{d_i}$ where $\left|E'_{d_i}\right| < \left|E_{d_i}\right|$. The filtering strategy grounds on the findings we obtained in the creation of the gold standard. In fact, when watching a newscast viewers better capture Person-type entities, as well as Organization-type and Location-type entities. The rest of less-specific and wider enclosed entities are more vague to be displayed on a television user interface and potentially less relevant for complementing the seed content. Named entities are then clustered applying a centroid-based clustering operation. As cluster centroid we consider the entity with the most frequent disambiguation *link* that also have the most repeated *surface_form*. As distance metric for comparing the instances, we applied strict string similarity over the *link*, and in case of mismatch, the Jaro-Winkler string distance [210] over the *surface_form*. The output of this phase is a list of clusters containing different instances of the same entity.

**Semantic Annotation Ranking** The different related documents collected and annotated in previous steps algorithm contain the pieces of the puzzle that will allow to re-construct the big picture of the event. Unfortunately, those pieces come wrapped together with other less important information which is not related with the news event and need to be put aside. By ranking the entities using functions introduced in previous section, we get to identify important entities that are grouped at the top positions of the list, and we get rid of those which are not related with the considered event. The final NSS will be a subset of the top scored entities after the ranking algorithm. In a more formal way, the bag of named entities $E'_{d_i}$ is further processed to promote the named entities which are highly related to the underlined event. To accomplish such an objective, we propose different ranking strategies $F_x(e_i)$ that go beyond the TF approach presented in Section 5.5 to consider other aspect like entity appearance in documents, popularity peak analysis, and domain experts' rules in order to better sort the annotations and generate the Semantic Snapshot of the considered Newscast (NSS).

## 6.4 Multidimensional Ranking Strategy

The ranking of the entities coming from the entity expansion process is one of the most complex step to perform in order to recreate the NSS of a news item: while computers are good in processing much large amounts of information and it is theoretically possible to analyze hundreds of thousands of related documents, to decide which of the finally retrieved entities need be selected as relevant is much more challenging.

In essence, the task consists of developing functions that take the unordered entity list from the expansion process and rank them to promote those that are potentially relevant for the viewer. Following the reasons discussed in Section 6.2 that justified the need of exploring different relevancy dimensions apart from pure frequency meth-

ods, in this section we consider four different ranking methods that we will further describe below: (1) frequency-based as already reported in Section 5.5, (2) gaussian based frequency in documents, (3) popularity on the Web, (4) experts' judgements expressed in form of rules. The two first will be interchangeably applied over the raw set of entities extracted from the related documents, $E'_{d_i}$. (3) and (4) functions will be orthogonally plugged over the results obtained from the previous two in order to re-balance certain entities that would not be promoted by previous mechanisms.

### 6.4.1   Frequency-based Function

We first rank the entities according to their absolute frequency within the set of retrieved documents $D$. Let define the absolute frequency of the entity $e_i$ in the collection of documents $D$ as $f_a(e_i, D)$, we define the scoring function $S_F = \frac{f_a(e_i, D)}{|E|}$, where $|E|$ is the cardinality of all entities spotted across all documents. In Fig. 6.4 (a) we can observe how entities with lower absolute frequency are placed at the beginning of the distribution and discarded in the final ranking; instead those with high $S_F$ are in the right side of the plot, being then considered to be part of the NSS.

### 6.4.2   Gaussian-based Function

The $S_F$ scoring function privileges the entities which appear often. However from the perspective of a television viewer, this is not always the case: while it is true that entities appearing in just a few documents are probably irrelevant and not representative enough to be considered in the final results, entities spread over the whole set of related documents are not necessary the ones the viewers would need to know about. In fact, they often represent entities that have been so present in media before that have become fairly well-known to the viewer. This scoring function is therefore approximated by a Gaussian curve. By characterizing the entities in terms of their Bernoulli appearance rate across all documents $f_{doc}(e_i)$, and applying the Gaussian distribution over those values, we promote entities distributed around the mean $\mu = \frac{|D|}{2}$ , where $|D|$ is the cardinality of the retrieved documents (Fig. 6.4 (b)). In particular the scoring function has been formalized as: $S_G = 1 - \left| \frac{f_{doc}(e_i)}{|D|} - 1 \right|$.

### 6.4.3   Orthogonal Functions

#### 6.4.3.1   Popularity Function

We propose a weighting function based on a mechanism that detects variations in entity popularity values over a time window (commonly named as popularity peaks) around the date of the studied event. The functions proposed above exploit the frequency of the entities in documents as a factor to measure its importance. However the frequency based approaches fail to explain the phenomenon of certain found

(a) Frequency-based function

(b) Gaussian-based function
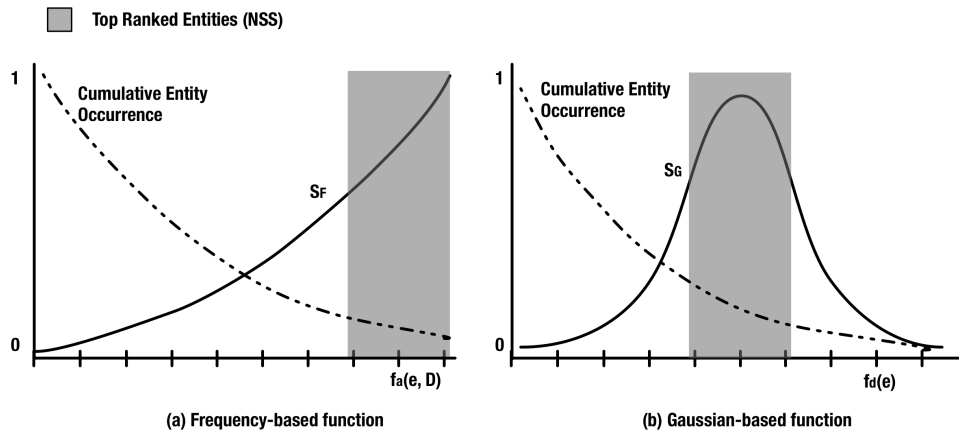
Figure 6.4: (a) depicts the Decay function of the entity occurrences in the corpus, and the $S_F$ which underlines the importance of an entity being used several times in the corpus. (b) represents the Gaussian-based function $S_G$, with the entities highly important over the mean.

entities which are barely mentioned in related documents but suddenly become interesting for viewers. These changes are sometimes unpredictable so the only solution is to rely on external sources that can provide indications about the entity popularity, like Google Trends[1] or Twitter.[2]
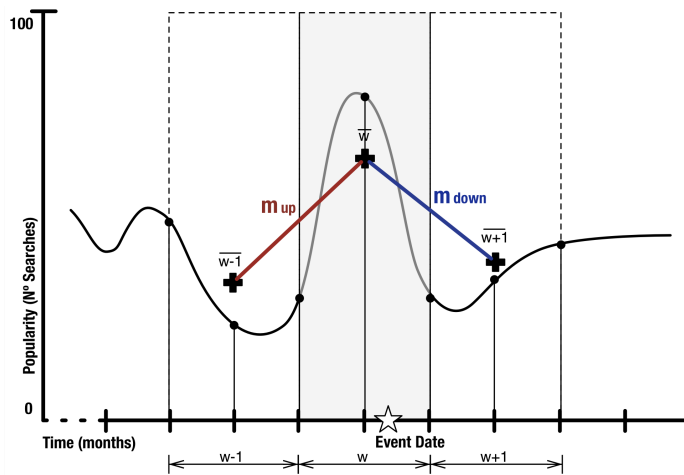


Figure 6.5: Popularity diagram of a considered event. On the x-axis is represented the time, and on the y-axis the magnitude of the popularity score. The star indicates when the event occurred. Given the discrete nature of the used platforms, the center of time window can be placed next to the day of the event.

The procedure for getting $P_{peak}(e_i)$ is depicted in Fig. 6.5. Using the label of an

---

[1] https://www.google.com/trends
[2] https://twitter.com

entity $e_i$, we obtain a list of pairs $[t, P]$ where $P \in [0, 100]$ is the popularity score of an entity at the instant of time $t$. Afterward we create three consecutive and equally long temporal windows around $t$, the first one $w_t$ containing the date itself, another one just immediately behind $w_{t-1}$ and a last one after the previous two $w_{t+1}$. In a next step we approximate the area inside the regions by calculating the average of the points contained in them, obtaining $\overline{w-1}$, $\overline{w}$ and $\overline{w+1}$. The slopes of the lines between $\overline{w-1}$ and $\overline{w}$, and $\overline{w}$ and $\overline{w+1}$ give the values $m_{up}$ and $m_{down}$ respectively, which are normalized and combined into a single score for measuring how significant the variation in volume of searches was for that studied entity label. When aggregating those two gradients, we scored $m_{up}$ higher in order to emphasize the irruption of a change, more than the later evolution of the search term.

By empirically studying the distribution of the popularity scores of the entities belonging to a newscast, we have observed that it follows a Gaussian curve. This fact will help us to better filter out popularity scores that do not trigger valid conclusions and therefore improving the merging of the ranking produced by the previous functions with the outcome from the popularity peaks detection algorithm.

### 6.4.3.2   Expert Rules Function

The knowledge of experts in the domain, like journalists or newscast editors can be materialized in the form of rules that correct the scoring output produced by former ranking strategies. The antecedent of those rules is composed by entity features such as type, number of documents where the entities appear, or the Web source from where documents have been extracted, while the precedent involves the recalculation of the scoring function according to the following equation: $S_{expert}(e) = S_{F-1}(e) * Op_{expert}$, being $Op_{expert}$ a factor which models the domain experts' opinions about the entities that match in the antecedent.

## 6.5   Experimental Settings and Evaluation of Multidimensional Approach

In this section we measure the effectiveness of our approach for building the NSS of a newscast against the gold standard presented in Section 5.4. We present the measures considered to carry out the study, we describe the experimental settings, and we conclude with the results.

### 6.5.1   Measures Considered

Inspired by similar studies in Web search engines, we have based our evaluation procedure in measures which try to find as many relevant documents as possible, while keeping the premise that the top ranked documents are the most important.

In order to summarize the effectiveness of a the different algorithm across the entire collection of queries considered in the gold standard, we have considered different averaging measures that are listed below:

- Mean precision/recall at rank N. It is probably the most used measure in information retrieval tasks. It is easy to understand and emphasize the top ranked documents. However it does not distinguish between differences in the rankings at positions 1 to p, which may be considered important for some tasks. For example, the two rankings in Figure 6.6 will be considered equally good when measured using precision at 10.

- Mean average precision at N. Also called $MAP$, it takes in consideration the order of the relevant items in the top N positions and is an appropriate measure for evaluating the task of finding as many relevant documents as possible, while still reflecting the intuition that the top ranked documents are the most important.

- Average Normalized Discounted Cumulative Gain $MNDCG$ at $N$. The Normalized Discounted Cumulative Gain is a popular measure for evaluating Web search and related applications [32]. It is based on the assumption that there are different levels of relevance for the documents obtained in results. In adition, the lower the ranked position of a relevant document the less useful it is for the user, since it is less likely to be examined.

As the relevant documents in our gold standard are scored in relevance for the user, we have mainly focused on the last measure since it can provide a more exhaustive judgment about the adequacy of the generated NSS. Concerning the evaluation point N, we have performed an empirical study over the whole set of queries and main ranking functions observing that from $N = 0$ $MNDCG$ decreasingly improves until it reaches a stable behavior from $N = 10$ on. Finally, we will not perform measures in terms of temporal efficiency. Even this kind of studies are easier to quantify, this falls outside the scope of this thesis.

### 6.5.2 Experimental Settings

#### 6.5.2.1 Document retrieval

We have relied on the Google Custom Search Engine (CSE) API service[3] by launching a query with the parameters specified by $q = [h, t]$.

Apart of the query itself, the CSE engine considers other parameters that need to be tuned up. First, due to quota restrictions the maximum number of retrieved

---

[3]https://www.google.com/cse/all
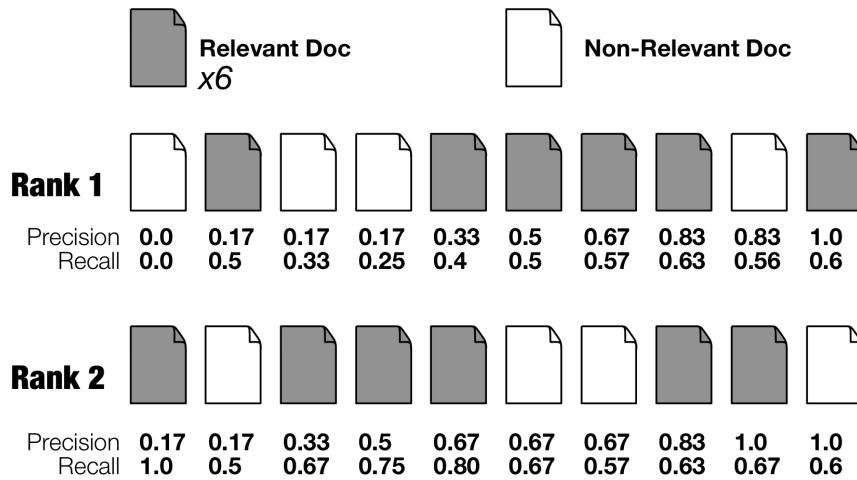
Figure 6.6: Inability of $P/R$ for considering the order of the relevant documents: rankings 1 and 2 share the same Precision and Recall at 10 .

document is set to 50. But in addition, we have also considered 3 different dimensions that influence the effectiveness in retrieving related documents:

1. Web sites to be crawled. Google allows to specify a list of web domains and subdomains where documents can be retrieved. This reduces the scope of the search task and, depending on the characteristics of the considered sources, influence the nature of the retrieved items: from big online newspapers to user generated content. At the same time, Google allows to prioritize searching over those whitelists while still considering the whole indexed Web. Based on this, in our study we considered five possible values for this parameter:

   **Google** : search over the whole set of Web pages indexed by Google.

   **L1** : A set of 10 internationals English speaking newspapers.[4]

   **L2** : A set of 3 international newspapers used in the gold standard creation.

   **L1+Google** : Prioritize content in L1 whitelist but still consider other sites.

   **L2+Google** : Prioritize content in L2 whitelist but still consider other sites.

2. Temporal dimension. This variable allows to filter those documents which are not temporarily close from the day where the newscast was published. Assuming that the news item is fresh enough, this date of publication will also be fairly close to the day the event took place. Taking $t$ as a reference and increasing the window in a certain amount of days $d$, we end up having $Time_{Window} = [t - d, t + d]$ The reason why we expand the original event period is because documents concerning a news event are not always published

---

[4] http://en.wikipedia.org/wiki/List_of_newspapers_in_the_world_by_circulation

during the time of the action is taking place but some hours or days after or before. The final $Time_{Window}$ could vary according to many factors such as the nature of the event itself (whether it is a brief appearance in a media, or part of a longer story with more repercussion) or the kind of documents the search engine is indexing (from very deep and elaborated documents that need time to be published, to short post quickly generated by users). In this study we have considered two possible values for it: 2 weeks and one week temporal windows.

3. In addition, Google CSE makes possible to filter result according to the Schema.org types; for our experiments we use the following settings: [NoFilter, keep sites containing entities of type *Person* or *Organization*]

That makes in total $5 * 2 * 2 = 20$ different parameter configurations that will be considered that we will study in the Section 6.5.3 in order to discover which configuration optimizes the expansion algorithm.

### 6.5.2.2   Semantic Annotation

We use [153] which applies machine learning classification of the entity type given a rich feature vector composed of a set of linguistic features, the output of a properly trained Conditional Random Fields classifier and the output of a set of off-the-shelf NER extractors supported by the NERD Framework.[5] We used it as an off-the-shelf entity extractor, using the offered classification model trained over the newswire content.

### 6.5.2.3   Annotation Filtering and Clustering

After some first trials it became evident that there were many non-pure named entities detected in the semantic annotation phase which are not well considered by viewers and experts. We have then applied three different filtering approaches:

**F1** : Filter annotations according to their NERD type[6]. In our case, we keep only Person, Location, and Organization.

**F2** : It consists in getting rid of entities with confidence score under first quarter of the distribution.

**F3** : Intuitively, people seem to be more attracted by proper names than general terms. Those names are normally capitalized. This filter keeps only named entities matching this rule.

---

[5]http://nerd.eurecom.fr
[6]http://nerd.eurecom.fr/ontology

By concatenating those filters, we obtain the following combinations: F1, F2, F3, F1_F2, F1_F3, F2_F3, F1_F2_F3 ). In order to reduce the number of runs, we did a first preselection of filters by setting the rest of steps of the approach to default values and averaging the scores obtained over the different queries. We ended up discovering that 3 of the filters (F1 and F3, and the combination F1_F3) were producing best results in the final MNDCG,

### 6.5.2.4   Semantic Annotation Ranking

For the current experiment we run both Frequency and Gaussian based functions, together with the orthogonal strategies based on popularity and expert rules. This makes a total of $2 * 2$ possible ranking configurations that will be considered and reported in result section. Regarding the particular details of the orthogonal functions, we have proceeded as follow:

**Popularity** We have relied on Google Trends,[7] which estimates how many times a search-term has been used in a given time-window. Since Google Trends gives results with a monthly temporal granularity, we have fixed the duration of such $w$ to 2 months in order to increase the representativity of the samples without compromising too much the validity of the selected values according with the time the event took place. With the aim of being selective enough and keep only those findings backed by strong evidence, we have filtered the entities with peak popularity value higher than $\mu + 2 * \sigma$ which approximately corresponds to a 2.5% of the distribution. Those entities will have their former scores combined with the popularity values via the following equation: $S_P(e) = R_{score}(e) + Pop_{peak}(e)^2$.

**Expert Rules** *i)* Entity type based rules: we have considered three rules to be applied over the three entity types considered in the gold standard. The different indexes per type have been deduced by relying on the average score per entity type computed in the survey $\overline{Sgt}_{entityType}$. Taking as input the average scores already reported in Section 5.4 and normalizing them on the interval $[1, -1]$, *organizations* have gotten a higher weight ($Op_{expert} = 0.95$), followed by *persons* ($Op_{expert} = 0.76$), and by *locations* ($Op_{expert} = 0.48$) that are badly considered and therefore lower ranked in general.
*ii)* Entity's documents based rules: each entity has to appear at least in two different sources in order to become a candidate. All entities whose document frequency $f_{doc}(e_i)$ is lower than 2 are automatically discarded ($Op_{expert} = 0$).

---

[7]https://www.google.com/trends

### 6.5.3 Results and Discussion

Given the different settings for each phase of the approach ($N_{runs_{Collection}} * Runs_{Filtering} * Runs_{Ranking}$), we have a total of $20*4*4 = 320$ different runs that have been launched and ranked according to $MNDCG_{10}$. In addition we have also executed two baseline approaches for comparing them with the better performing strategies in our approach. More details about them are shown below.

#### 6.5.3.1 Baselines

**Baseline 1: Former Entity Expansion Implementation**. As reported in the related work, a previous version of the News Entity Expansion algorithm was already published in [143] and reported in Section 5.5. The settings are: Google as source of documents, temporal window of 2 Weeks, no Schema.org selected, no filter strategy applied, and only frequency based ranked function with no orthogonal appliances. Results are reported in Table 6.1 under the run id *BS1*.

**Baseline 2: TFIDF-based Function**. To compare our absolute frequency and Gaussian based functions with other possible approaches already reported in the literature, we selected the well-known TF-IDF. It measures the importance an entity in a document over a corpus of documents $D$, penalizing those entities appearing more frequently. The function, in the context of the named entity annotation domain is as follows:

$$tf(e_i, d_j) = 0.5 + \frac{0.5 \times f_a(e_i, D)}{max\{f_a(e_i', D) : e_i' \in d_j\}}, idf(e_i, d_j) = log\frac{|D|}{\{d_j \in D : e_i \in d_j\}} \qquad (6.1)$$

We computed the average of the TF-IDF for each entity across all analyzed documents, resulting in aggregating the different $tf(e_i, d_j) \times idf(e_i, d_j)$ into a single function $tfidf^*(e_i, D)$ via the function $S_{TFIDF}(e) = \frac{\sum_{j=1}^{n} tf(e, d_j) \times idf(e)}{|D|}$. Results are reported in Table 6.1 under the run id *BS2*.

#### 6.5.3.2 Analysis of the Experiment's Results

In Table 6.1 we present the top 20 runs for our approach in generating NSS, together with some lower configurations at position 78 and following that are worth to be reported and the scores of the baseline strategies. We summarize the main findings of the experimental settings and evaluation as follows:

- Our best approach has obtained a $MNDCG_{10}$ score of 0.698 and a $MAP_{10}$ of 0.91, which are reasonably good in the document retrieval domain. The whole list of entities automatically obtained by executing this approach with the best-performing settings is available in Appendix 8.

| Run | Collection | | | Filtering | Functions | | | Result | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sources | $T_{Window}$ | Schema.org | | Freq | Pop | Exp | $MNDCG_{10}$ | $MAP_{10}$ | $MP_{10}$ | $MR_{10}$ |
| Ex0 | L1+Google | 2W | | F3 | Freq | | ✓ | 0.698 | 0.93 | 0.68 | 0.35 |
| Ex1 | L2+Google | 2W | | F3 | Freq | | ✓ | 0.695 | 0.93 | 0.68 | 0.35 |
| Ex2 | L1+Google | 2W | ✓ | F1+F3 | Freq | | ✓ | 0.689 | 0.93 | 0.62 | 0.31 |
| Ex3 | L1 | 2W | ✓ | F3 | Freq | | ✓ | 0.681 | 0.9 | 0.64 | 0.35 |
| Ex4 | L2+Google | 2W | | F1+F3 | Freq | | ✓ | 0.679 | 0.92 | 0.7 | 0.36 |
| Ex5 | L1+Google | 2W | ✓ | F1+F3 | Freq | | ✓ | 0.67 | 0.91 | 0.62 | 0.31 |
| Ex6 | L1 | 2W | ✓ | F3 | Freq | ✓ | ✓ | 0.668 | 0.86 | 0.6 | 0.32 |
| Ex7 | L2+Google | 2W | | F3 | Freq | ✓ | ✓ | 0.659 | 0.85 | 0.56 | 0.29 |
| Ex8 | Google | 2W | | F3 | Freq | | ✓ | 0.654 | 0.88 | 0.66 | 0.34 |
| Ex9 | L1 | 2W | | F3 | Freq | | ✓ | 0.654 | 0.88 | 0.66 | 0.35 |
| Ex10 | Google | 2W | ✓ | F1+F3 | Freq | | ✓ | 0.653 | 0.9 | 0.62 | 0.31 |
| Ex11 | Google | 2W | | F3 | Freq | ✓ | ✓ | 0.653 | 0.81 | 0.56 | 0.29 |
| Ex12 | L1+Google | 2W | ✓ | F1+F3 | Freq | | | 0.652 | 0.93 | 0.64 | 0.32 |
| Ex13 | L2 | 2W | ✓ | F3 | Freq | | ✓ | 0.651 | 0.89 | 0.64 | 0.34 |
| Ex14 | Google | 2W | | F1+F3 | Freq | | ✓ | 0.649 | 0.88 | 0.64 | 0.33 |
| Ex15 | L2+Google | 2W | | F1+F3 | Freq | | | 0.649 | 0.94 | 0.72 | 0.37 |
| Ex16 | L1+Google | 2W | | F3 | Freq | | | 0.649 | 0.9 | 0.68 | 0.35 |
| Ex17 | Google | 2W | | F1+F3 | Freq | | | 0.648 | 0.93 | 0.72 | 0.37 |
| Ex18 | L1 | 2W | | F1+F3 | Freq | | ✓ | 0.646 | 0.89 | 0.66 | 0.34 |
| Ex19 | L1+Google | 2W | | F1+F3 | Freq | | | 0.646 | 0.94 | 0.7 | 0.37 |
| Ex20 | L1+Google | 2W | | F1+F3 | Freq | | ✓ | 0.646 | 0.89 | 0.66 | 0.34 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Ex78 | Google | 2W | ✓ | F1+F3 | Gaussian | | ✓ | 0.552 | 0.66 | 0.66 | 0.34 |
| Ex80 | L2+Google | 2W | ✓ | F1+F3 | Gaussian | | ✓ | 0.55 | 0.69 | 0.7 | 0.36 |
| Ex82 | L1 | 2W | ✓ | F3 | Gaussian | | ✓ | 0.549 | 0.68 | 0.64 | 0.33 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| BS2 | Google | 2W | | | Freq | | | 0.473 | 0.53 | 0.42 | 0.22 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| BS1 | Google | 2W | | | TFIDF | | | 0.063 | 0.08 | 0.06 | 0.03 |

Table 6.1: Executed runs and their configuration settings, ranked by $MNDCG_{10}$

- Our approach performs much better than BS1 and by far better than BS2. The very low score of this last baseline is explained in the fact that traditional TF-IDF function is designed to measure the relevance of an item referred to the document that contains it and not to the whole collection. In addition, the absence of filters drop drastically the scores.

- Regarding the Document Retrieval step, we see that using whole set of sites indexed Google as source alone of together with other WhiteLists gives in general better results than restricting only to particular whitelist. The biggest $T_{Window}$ of 2 weeks performs better in all cases, while the use of Schema.org seems to be beneficial in a certain degree that we expect to increase in the future with the increasing adoption of this vocabulary by different Web sites.

- The best Filter strategy is F3, followed by the combination F1_F3. In conclusion, capitalization is a very powerful tool for making a first candidate list with those entities that a priori users consider more interesting.

- The absolute frequency function performs better than the Gaussian in all top cases.

- The Expert Rules based function improves the final NSS for almost every con-

figuration possible, *probing the importance of considering alternative relevancy dimensions.* The popularity function makes its first appearance in run number 6, suggesting it is bringing to the top important entities but in a lower degree than the Expert Rules dimension. However through a manual assessment we have discovered that this ranking method has a huge potential for the future when the ground truth will be extended, since it is bringing up relevant entities like for example *David Ellsberg*[8] for the query "Fugitive Edward Snowden applies for asylum in Russia". This person is barely mentioned in the collected documents, but his role in the story is representative since he published an editorial with high media impact in The Guardian praising the actions of Snowden in revealing top-secret surveillance programs of the NSA.

### 6.5.3.3 The Selection Problem when applying Multidimensionality

The scores in terms of $MNDCG_{10}$ obtained before confirm that a well tuned collection strategy and different ranking functions working together to identify relevant entities have considerably improved the NSS generation process. To understand why this is happening, we have extrapolated the selection problem diagram depicted in Figure 6.1 to the case of multidimensional ranking implemented in this chapter.
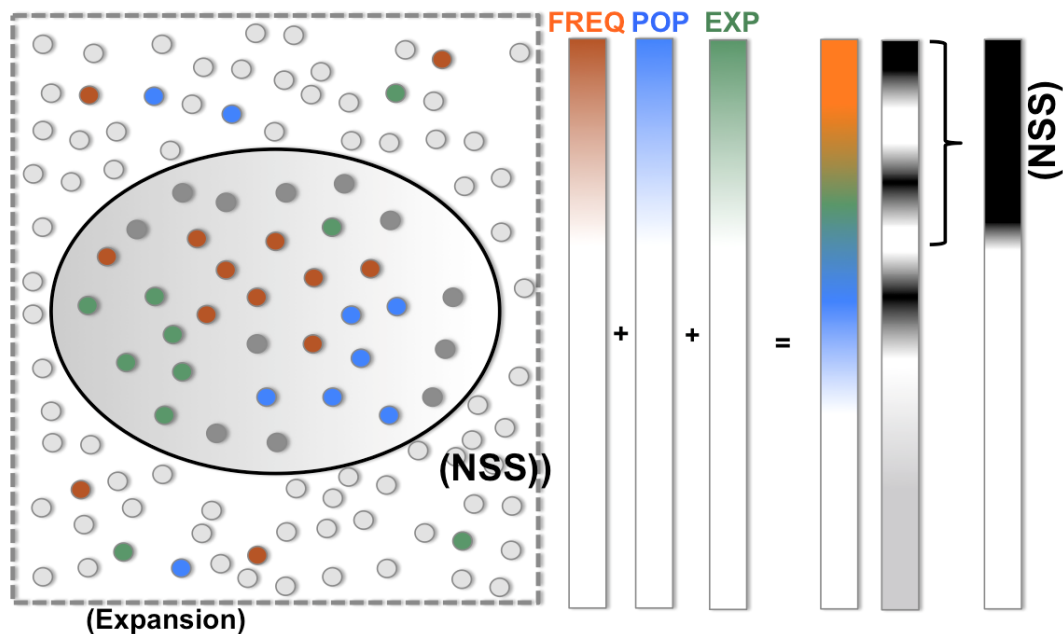


Figure 6.7: Multidimensional approach selecting entities to become part of the NSS

In Figure 6.7 we have illustrated the behavior of the following three ranking functions when working together to promote entities from the set of expansion documents:

---

[8]http://en.wikipedia.org/wiki/Daniel_Ellsberg

*FREQ* or frequency-based ranking already considered in Figure 6.2, *POP* or popularity based function, and *EXP* or expert's opinions-based ranking (colored in orange, blue and green respectively). By just looking at the number of selected entities falling inside the NSS region, we can intuitively verify that there is a higher number of true positives, and much less grey-colored entities are present suggesting a significant improve in recall. A similar conclusion can be inferred by looking at the projection of the entities over a single list: the top black colored entities that were promoted by the frequency function are now accompanied by others selected by the two remaining methods and therefore fall in upper regions of the list, which have more chances of becoming part of the top $n$ entities shaping up the Semantic Snapshot.

Coming back to the two examples previously introduced in Section 6.2, the entities *Laura Poitras* and *Glenn Greenwald* are not highy scored by the *FREQ* function, but they get promoted by the *POP* method to upper positions of the ranking that later conform the final NSS.

### 6.5.3.4   Comments on Google CSE Influence

Google CSE is a very flexible tool that allows to parametrize searches over Web documents in a very simple and effective way, hence it has been selected in our research. In order to probe that the good results of our solution are not funded upon a particular search engine, and the high quality ranking algorithms that Google internally uses are not boosting the scores of our approach, we have conducted a set of empirical experiments consisting in comparing the output of Google Custom Search service in default operation mode with the results of the Bing[9] engine for the same set of queries. The results have suggested that there is not significant difference in using one of the search tool versus the other when retrieving related documents in the named entity expansion phase. After injecting in both candidates one query per ground truth video generated out of their titles, we have looked at the mean overlap in the sets document obtained per search engine, when restricting the number of results per query to 50. The score obtained (0.47) proves that almost half of the documents returned by both search engines are essentially the same. As our approach does not rely on the order of the documents inside the set of 50 elements retrieved, we can assume that difference in the results obtained by using one search service or the other is minimum. Digging more into the details, we analyzed what happens for annotations of the video "Fugitive Edward Snowden applies for asylum in Russia". The recall of the set of entities coming from expansion method applied over Google (0.91) is fairly the same than the recall of the entity expansion set using Bing (0.89). In the case of entities like *Kucherena* for example, we found 132 mentions on the Google's 50 related documents while in Bing's 50 hits we obtained around 121,

---

[9]https://www.bing.com/

suggesting again there is not significant advantage in using one engine against the other. In future experiments we plan to reassure those initial findings by plugging other search solutions over our named entity expansion method for generating NSS.

## 6.6 Summary

In this paper we have further developed an approach for automatically generating Newscast Semantic Snapshots. By following a well specified and parametrized entity expansion process we retrieve additional event-related documents from the Web, in order to enlarge the niche of initial newscast content. The bag of retrieved documents, together with the newscast transcript, is analyzed with the objective of extracting named entities referring to people, organizations, and locations. By increasing the size of the document set, we have increased the completeness of the context and the representativeness of the list of entities, reinforcing relevant entities and finding new ones that are potentially interesting inside the context of that news item. The named entities have been then ranked according to different dimensions: pure frequency-based measures, the entity appearance in the sampled collection of documents, popularity of the entity on the Web, and experts' rules. We assessed the entire workflow against our gold standard presented in Section 5.4. The evaluation has showed the strength of a fine tuned collection phase together with a multidimensional based entity ranking, holding an $MNDCG_{10}$ score of 0.69, outperforming the two studied baselines. The whole list of entities automatically obtained by executing this approach with the best-performing settings is available in Appendix 8.

# The Concentric Nature of the News Semantic Snapshot

## 7.1 Introduction

As seen in previous chapters, the Web enables to have access to silo-ed information describing news articles, often offering a multitude of viewpoints that, once combined, can provide a broader picture of the story being reported on the news. A single presentation of a news item taken individually generally fails to illustrate the complexity of the event being reported. The Web has offered a data space where it is possible to find very diverse information such as citizen-based blogs, journalistic articles or social media posts, teaming up for generating a multi-rich ecosystem of complementary news content.

In Chapter 5 we proposed an approach that automatically extracts representative features of a news item, namely named entities, from textual content attached to a video item (subtitles) and from a set of documents from the Web collected using entity expansion techniques. Those techniques work over the open Web to produce a ranked list of entities in order to generate a so-called Newscast Semantic Snapshot (NSS), which complements the initial set of detected entities in subtitles with other item-related entities captured from Web documents.

In Chapter 6 we studied how entity expansion techniques succeed in collecting the important facts behinds a particular news item, but cannot uniquely rely on frequency-based functions to distinguish the relevant entities from the ones that are not. Instead, we need to consider the multiple dimensions ruling the importance of the candidate entities and build adequate functions to spot them. In particular, we have implemented a set of functions that rely on the absolute entity frequency of entities, appearances across the collected documents, expert rules, and global popularity of the entity.

However after performing an experimental and critical assessment of this method we have observed how the frequency functions and pure information retrieval techniques that were used during the NSS generation are complicated to combine together and neglect the intrinsic relationships that the entities hold. We have also observed that frequency-based rankings and their variants are appropriate for spotting essen-

tial or inherent entities, while other different functions working over relevance criteria such as interestingness, informativeness or popularity are needed for identifying the rest of relevant entities, therefore suggesting a duality on the news annotations that needs to be further exploited.

In this Chapter we recast the problem of generating a NSS by exploiting and harmonizing in a single model different semantic relationships established between the news' entities. Instead of tackling the problem from a pure list-based oriented model, where all the different news related phenomena are projected into a single dimension, we propose a concentric-based approach with two main layers called **Core** and **Crust**. This knowledge representation model better supports the complex and multi-dimensional relations established among the entities involved in a news item and allows to formalize the distinction between the representative entities, which better characterize the essence of the news item, and the relevant ones, that are potentially interesting because of different reasons that link them to the *Core*. This graph-based knowledge representation considers the multidimensional nature of those relationships, allowing us to focus on different desired features for the final NSS, like representativeness and compactness. The final ranking order is then delegated to the final applications that display the data and inevitably project the rich spectrum of relationships among entities describing an event into a single and easier to consume dimension. We compare our approach with a baseline by analyzing the compactness of the generated summary against the gold standard presented in Section 5.4. Results of the experiments show that our approach converges faster to the ideal compact news snapshot with an improvement of 36.9% over the baseline. This research has been published at [58].

The remainder of this chapter is organized as follows: Section 7.2 presents a critical assessment of other previous NSS generation approaches and lists the key motivations of the work carried out in this Chapter. We later state our hypothesis in Section 7.3. In Section 7.4, we describe our approach implementing the concentric based NSS generation. We propose an evaluation for our experiments in the Section 7.5. Finally, we summarize our main findings and outline some future work in Section 7.6.

## 7.2   Follow up of Multidimensional News Semantic Snapshot Generation

This section summarizes the different research efforts made for critically assessing and extending the experiments described in [145].

### 7.2.1   Improving the NSS Generation Baseline

In this section, we try to extend the strategy described in Chapter 6 by further exploiting additional relevance indicators that could have been missing during our previous study, with the objective of improving the Average Normalized Discounted Cumulative Gain ($MNDCG$) at $N$. This measure [32] considers different levels of relevance and gives more priority to items ranked at top positions since they are more likely to be examined by a user. Some changes that brought some improvement over the original approach are:

- Exploit Google relevance: Documents obtained from Google Custom Search Engine (CSE)[1] come ordered, so the ones on the top are potentially more relevant, and therefore related, to the studied news item. Assuming that entities spotted within those higher ranked documents are more important than the ones found in less interesting documents, we can weight them differently when summing up scores in frequency functions (see Section 4.1 in [145]). This adaptation enables to gain 1.8% in $MNDCG$ over the best configuration from the original approach.

- Promote subtitle entities: Entities detected in subtitles can be better considered since they are explicitly mentioned in the video speech and therefore, are more likely to be relevant to what is being reported. By analyzing different ratios for weighting subtitle entities versus related document's ones, the combination (1:4) brought the best outcome, obtaining a percentage increase of 2.5% of $MNDCG$.

Some attempts that did not improve or even slightly reduced $MNDCG$ are:

- Exploit Named Entity Extractor's confidence: Similarly to the Google relevance, confidence scores produced by the entity annotators [153] can be used to differently ranked entities when accumulating them on frequency. This new dimension brought a percentage decrease in $MNDCG$ of 0.2%.

- Interpret popularity dimension: Candidate entities proposed by the popularity function (see Section 6.4.3.1) need to be combined together with the outcome of the frequency measures to provide a single ranked list of entities. Scores coming from both dimensions were simply summed. In order to go a step further, we have created a function $F : R_{Pop} \rightarrow R_{Freq}$ which linearly transforms scores produced by the popularity function into values inside the range of the frequency functions $R_{Freq}$ before performing the addition. Unfortunately, this lead to a percentage decrease of 1.4% in $MNDCG$.

---

[1]https://www.google.com/cse/all

- Perform the clustering before entity filtering: By filtering entities first, we get rid of many noisy annotations and partially correct results that can contaminate the generation of the NSS. It is possible to proceed the other way around: run the clustering operation over the whole set of annotations and then filter out clusters according to their entity centroids. Intuitively, the clustering phase can benefit from partially correct annotations since they could still balance clusters by becoming part of some of them. However, the results following this modified workflow are less performing than the former approach because the filtering stage becomes too aggressive by removing some important entity clusters that have low representative centroids (we observed a percentage decrease of 0.60% in $MNDCG$).

According to the experiments conducted, the situation did not bring a significant improvement in the original scores. A deeper study of the results reveals that prioritizing certain ranking dimensions quickly brings valid results but also discards relevant entities that were selected before. This phenomena can is better illustrated on Figure 7.1. We can observe how shuffling the priority of the results coming from different ranking functions do not significantly change the number of relevant entities available in the spectrum for the final result. Fine tuning of certain functions can bring small improvements by squeezing the top entities per function in a shorter region of the spectrum, but the final $MNDCG$ is too dependent on the first position of the ranking so final scores do not notably change. In this situation and to be in line with the philosophy of a multidimensional ranking, we could think in adding more ranking dimensions. But this solution become to complex and barely scalable: how many additional dimensions would be need? how would we combine them with the ones already considered? The difficulty in answering those question motivates to look for alternative approaches.

In addition, the complete workflow is too dependent on frequency functions and pure information retrieval techniques thus neglecting the semantic relationships present in the ideal NSS of a news event. Frequency driven rankings and their hybrid approaches are successfully spotting essential or inherent entities that must be inside the final result, but the rest of dimensions for considering other entities that are relevant in the NSS creation for particular reasons are difficult to identify, time consuming to implement, and very challenging to combine together to produce the final ranking.

### 7.2.2 Thinking Outside the Box

Given the limitations found in the approach described in previous Chapter 6, we try to tackle the problem from a different angle and reconsider the conditions that a NSS should match in order to be properly consumed by other users and applications.

After studying that state-of-the-art ranking algorithms have reached a ceiling in
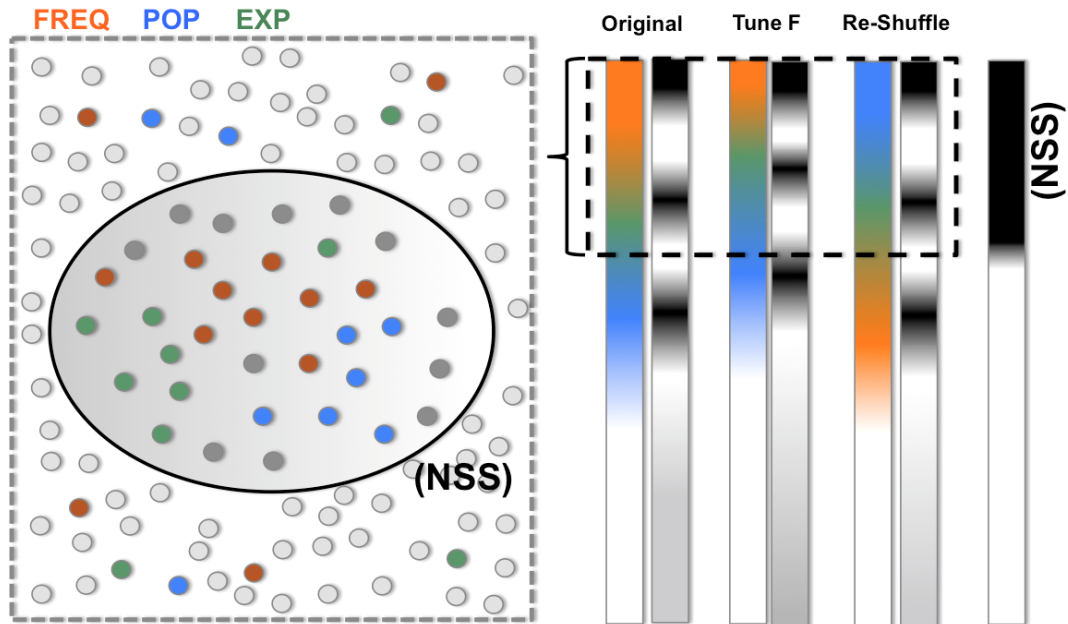
Figure 7.1: Illustrating the lack of significant improvement when fine-tuning a multidimensional NSS generation approach

performance in terms of $MNDCG$, the first question to be answered is if the data retrieved via the collection phase can still offer some rooms for improvement. By looking at further positions ($10 < n < 50$) in the ranking generated by the best run at 10 in Section 6.5.3 ($Best_{MNDCG_{10}} = \{$L1+Google, 2Weeks, NoSchema, F3, Freq, ExpertRules$\}$ we can plot the score evolution when NSS gets bigger in size. The curve in Figure 7.2 suggests that cumulative gain keeps increasing at bigger $n$ values even if the gradient is not pronounced.

Cumulative gain is a measure that weights more the matches in the top ranks. In order to make this analysis more ranking agnostic in Figure 7.3, we analyzed the recall $R$ when the size of the NSS goes till the position 50. The slope is now steeper and clearly reflects that more relevant entities are still found at lower positions in the ranking and can potentially be moved to the top.

To check how good $Best_{MNDCG_{10}}$ performs compared to other configurations in terms of Recall, the complete set of configurations has been re-run in order to see which one is working the best. The top 10 configurations, labeled from $Ex1$ to $Ex10$ are shown in Table 7.1.

We first observe that $Best_{MNDCG_{10}} \neq Best_{R_{50}}$. In other words, at further positions in the ranking, other strategies bring up a higher amount of relevant entities, even if they were not that well performing in ranking at $n <= 10$. Additionally, the popularity dimension, which was getting down cumulative gain scores, seems to bring up relevant entities at higher $n$ values according to the configuration of the four best
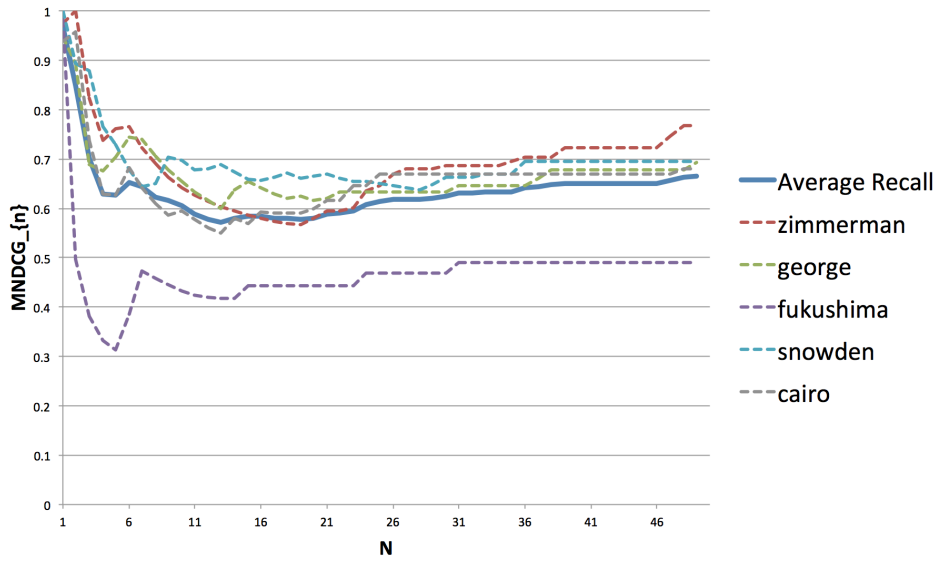
Figure 7.2: $MNDCG_{1-50}$ for run $Best_{MNDCG_{10}}$ in Section 6.5.3.
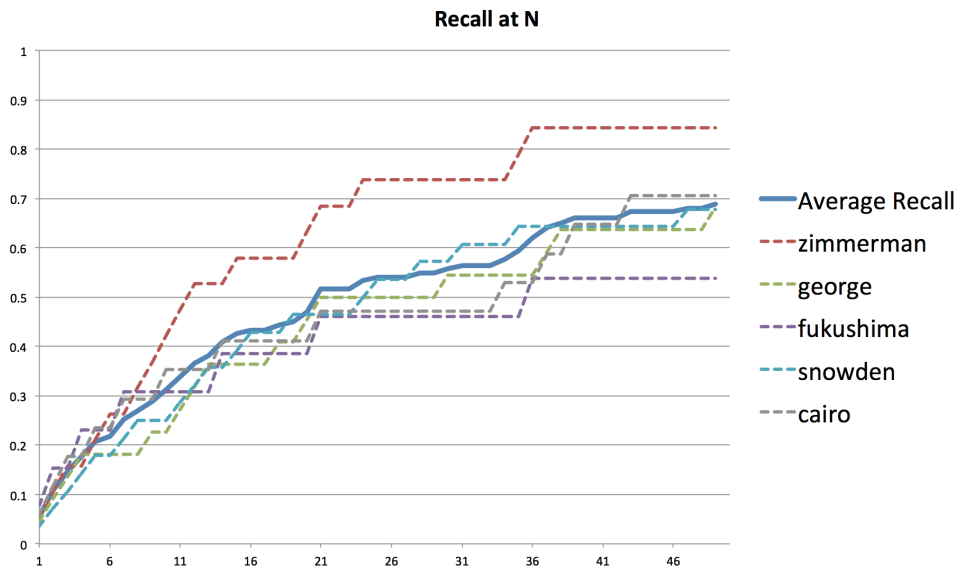


Figure 7.3: $Recall_{1-50}$ for $Best_{MNDCG_{10}}$ = in Section 6.5.3.

performing runs. In conclusion, there is room for improvement, but also a need for:

- Changing the knowledge acquisition method behind the approach. As already explained before, pure information retrieval techniques are not enough to explain the relevancy of an entity for a particular news item. Certain entities are important for summarizing the context of the news items, while others are informative for users who just want to discover something beyond the obvious facts. Dimensions like popularity or semantic relatedness cannot be projected

Table 7.1: Executed runs and their configuration settings, ranked by $R_{50}$.

| Run | Collection | | | Filter | Ranking | | | Result |
|-----|------------|------|------------|--------|---------|-----|-----|--------|
| | Sources | $T_W$ | Schema.org | | Freq | Pop | Exp | $R_{50}$ |
| Ex1 | L2+Google | 1W | | F3 | Gaussian | ✓ | ✓ | 0.7224 |
| Ex2 | L2+Google | 1W | | F3 | FreqGoogle | | ✓ | 0.7119 |
| Ex3 | L2+Google | 1W | | F3 | Freq | ✓ | ✓ | 0.7115 |
| Ex4 | L2+Google | 1W | | F3 | FreqGoogle | ✓ | ✓ | 0.707 |
| Ex5 | L2+Google | 1W | | F3 | Gaussian | ✓ | | 0.707 |
| Ex6 | L1+Google | 1W | | F3 | Gaussian | ✓ | ✓ | 0.7031 |
| Ex7 | L2+Google | 1W | | F3 | Gaussian | | | 0.6944 |
| Ex8 | L2+Google | 1W | | F3 | Freq | | | 0.693 |
| Ex9 | L1+Google | 2W | | F3 | Gaussian | ✓ | ✓ | 0.6919 |
| Ex10 | Google | 2W | | F3 | Gaussian | | ✓ | 0.6908 |

into a single ranking dimension. Instead, they need to be equally considered when promoting important entities so they can take part on the NSS.

- Changing the final objective. The task of bringing as many relevant entities as possible inside the NSS will be prioritized against being too precise in ranking. This is in line with the idea of generating a flexible and application-independent NSS, which intends to be comprehensive enough to contain as many entities as possible for better representing the context of a news item. Our outcome aims to produce a solid semantic representation that can properly feed different prototypes and tools with very broad information needs.

## 7.3   The Hypothesis: a concentric-based Model

This section formally describes the problem we are addressing and the main hypothesis we are formulating. In a nutshell, the semantic snapshot of a news item (NSS) can be modeled following a schema of concentric entity layers. This kind of representation helps to better reproduce the context of a news event and ease the task of identifying the different relevant entities for various dimensions.

In this model, we are considering two main entity layers that can annotate a news item: **Core** and **Crust**.

**Core:** It is composed of a small number of key entities which are essential to identify an event. Those entities have the highest degree of representativeness and can better summarize the main facts attached to the event. They are frequently mentioned in related documents and are therefore spottable via frequency-based functions. Sometimes, they are too obvious for the user, but they are the key elements for describing the facts. They are *semantically compact* in the sense that it exists numerous semantic relationships between each entity. Let $E_{d_i} = e_{1_{d_i}}, ..., e_{n_{d_i}}$ be the
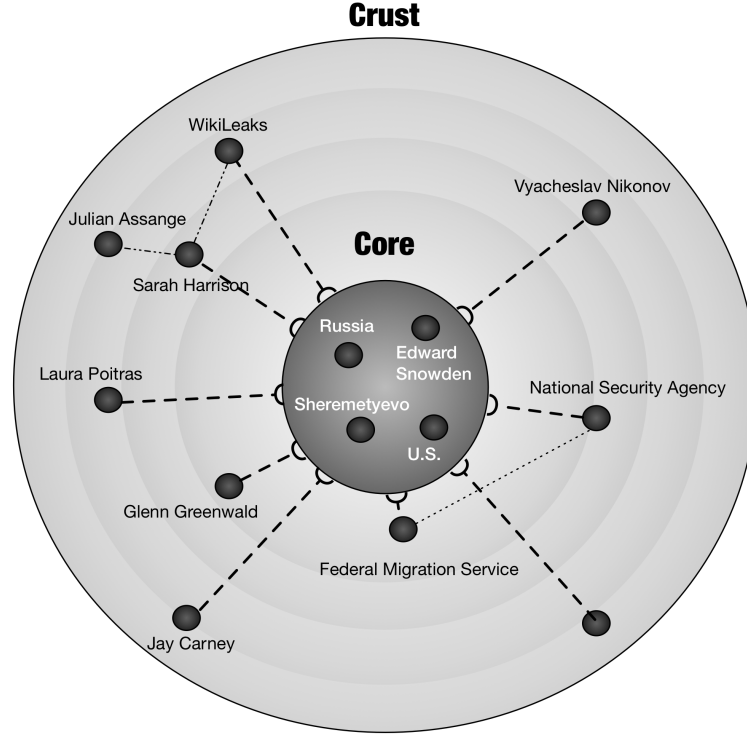
Figure 7.4: Concentricity of the news item *"Fugitive Edward Snowden applies for asylum in Russia"*

bag of entities belonging to the related document $d_i$, and $E_T$ the union of all $E_{d_i}$:

$$Core = G\left(e_1, e_2, ...e_c, e_i\right), e_i \in E_{d_i} \tag{7.1}$$

$$(\text{Frequency Prominence}) \forall e_i \in Core, f(e_i) > t \mid 0 \ll t < 1 \tag{7.1a}$$

$$(\text{Coherence}) \forall e_i, e_j \in Core, S(e_i, e_j) > s_{core} \mid 0 \ll s_{core} \leqslant 1 \tag{7.1b}$$

**Crust:** It is composed of a larger number of entities that describe particular details of a news items. Those entities are mentioned in some specific related documents, but they are not always spottable via frequency-based measures. They are not necessarily pairwise related (not semantically compact). Their relevancy is instead grounded on the existence of special relations (including popularity, serendipity, etc.) between those entities and the *Core*.

$$Crust = G\left(e_1, e_2, ...e_c\right), e_i \in E_{d_i} \tag{7.2}$$

$$(\text{Core Attached}) \forall e_i \in Crust, S_{(}e_i, Core) > s \mid 0 \leqslant s \ll 1 \tag{7.2a}$$

Those two layers can be aggregated into a single structure in order to build the so called **News Semantic Context (NSS)**. In our hypothesis, and differently than

in previously implemented methods, this structure will be a graph of entities since the relationships established among the elements inside the NSS are more important than their absolute final ranking, therefore providing a data structure that remains as flexible as possible, ready to be reused by very different news prototypes.

$$SSN_{concentric} = Core \oplus Crust \tag{7.3}$$

Given this new nature characterizing a NSS, maximizing cumulative gain is not a priority for this study. Therefore, we need to define a new and more ranking-agnostic **Objective Function**. One possibility is measuring the recall $R$. However, this metric does not consider different degrees of entity relevance which are available in the ground truth (see Section 5.4). In order to exploit this additional information, an extended recall index $R^*$ has been defined. This measure takes into account the different scores of the relevant entities and gives a more accurate idea about the coverage provided by a particular NSS:

$$R^*(\text{NSS}) = \frac{\sum Score_{GT}(e_{NSS_i})}{\sum Score_{GT}(e_{gt_i})} \tag{7.4}$$

Furthermore, this research aims to tackle the problem from a bigger perspective by studying a wider region of the entity annotations spectrum and not only focusing in a particular NSS at $n$. Let us define $Res_{Ap}$ as a list of $N$ entities produced by a certain approach $Ap$, we define a Semantic Snapshot $NSS_{Ap}$ as the $n$ first entities in $Res_{Ap}$.

$$NSS_{Ap} = \{e_0, e_1, ..., e_n\} \mid n < |Res_{Ap}| = N \tag{7.5}$$

We introduce the so-called compactness $Com$ of an entity set $Res$, given a certain function $f$ and a value $v$:

$$Com\,(R, f, v) = |min(NSS \in Res)| \mid f(NSS) \geq v \tag{7.6}$$

This measure helps to indicate if a particular set of entities is able to produce concise NSS while still keeping the goal of $f(NSS) > v$. In Figure 7.5 we show three different ranking functions $A$, $B$ and $C$. Considering $f(NSS)$ the Recall function, and being $v = 0.66$, we can observe how function $A$ is able to get to this value before (position $n = 27$) than the other two scoring methods (positions $n = 33$ and $n = 54$ respectively), therefore being able to produce a smaller and compact NSS (noted as $S_a$ in the figure) than $S_b$ and $S_c$. Our objective is to produce entity sets with minimum compactness $Com$. The function $R^*$ will be used to privilege coverage.

Having defined the different concepts and dimensions above, and being the results $Res_{Exp}$ and $Res_{Conc}$ sets of $N$ relevant entities about a particular news item generated via our former implementation and the concentric model approach respectively,
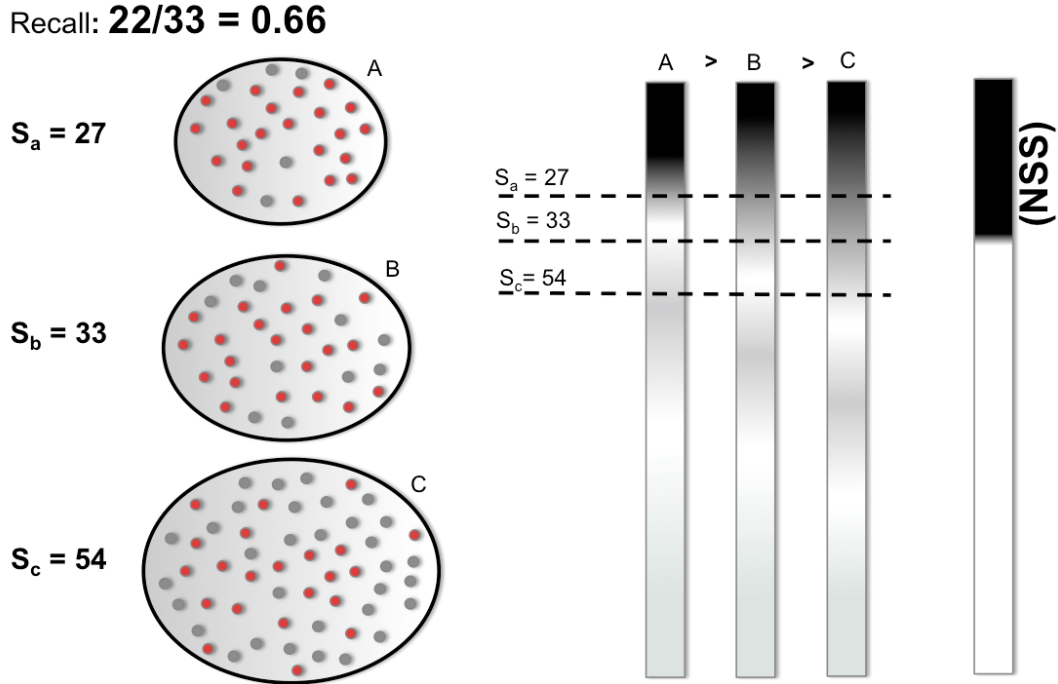
Figure 7.5: Example of compactness over different set of entities and distributions of true positives in their corresponding ranks

we formulate the following **Hypothesis**:

$$Com\left(Res_{Exp}, R^*, R^*(Res_{Exp})\right) > Com\left(Res_{Conc}, R^*, R^*(Res_{Exp})\right) \tag{7.7}$$

According to this hypothesis, the concentric model approach has to be able to produce more concise, cleaner, and potentially easier to consume New Semantic Snapshots than the related research efforts implementing an unidimensional ranking.

## 7.4 The Approach

This section presents our proposed approach for generating News Semantic Snapshots based in a concentric model. The workflow is composed of the following steps (Figure 7.6): after executing state-of-the art entity expansion and ranking strategies with the best configurations possible to bring to the top positions as many relevant semantic annotations as possible (see the grey part on the left side, labeled as (**1**), we build the concentric model of the news items in three different steps: generating the *Core*, the *Crust* and the final *NSS*, as depicted in the white part on the right side (**2**).

**Named Entity Expansion and Ranking.** The first step consists of executing the expansion approaches presented in previous Chapter 6 for generating a list of
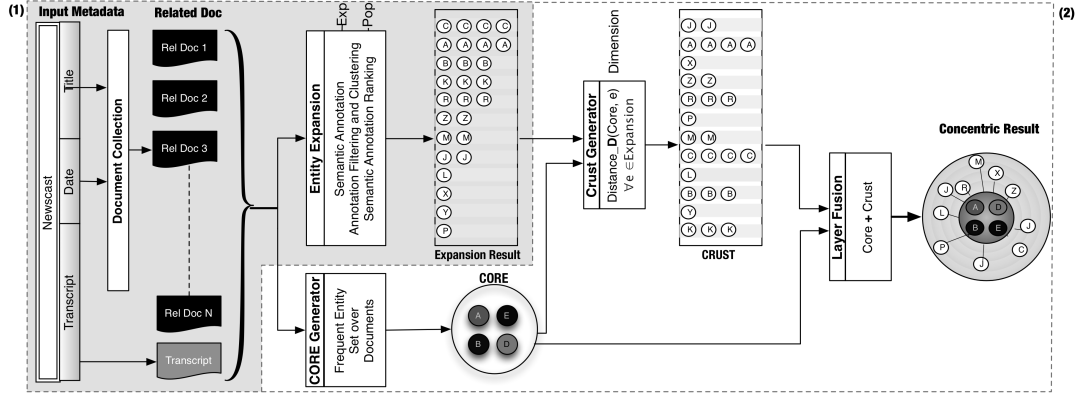
Figure 7.6: Concentric-based approach for generating News Semantic Snapshot using Named Entity Expansion

entities $Res_{Exp}$. The objective is to reduce the size of the spectrum of annotations to be considered while building the concentric model and, therefore, relaxing the complexity of having to work over the entire set of entities. Taking as input the metadata that news broadcasters offer about the items they publish, the query $q = [h, t]$ is built, where $h$ is the video headline and $t$ is the publication date. This query allows us to collect a set of event-related documents $D$ from the open Web over which the semantic annotation process is performed. After removing HTML tags and other markup annotations, the feature space is then reduced and each document $d_i$ is represented by a bag of entities $E_{d_i} = e_{1_{d_i}}, ..., e_{n_{d_i}}$, where each entity is defined as a triplet $(surface\_form, type, link)$. A filtering process prepares entities to be clustered applying a centroid-based algorithm based on strict string similarity over the $link$ and $surface\_form$. The output of this phase $E'_{d_i}$ is further processed to promote the named entities that are highly related to the underlined event, based on entity appearance in documents, popularity peak analysis and domain experts' rules in order to produce a ranked list of entities $Res_{Exp}$, which feed the concentric NSS generation approach.

**Core Generation.** The Core generation process works over the set of filtered entity annotations per document $E'_{d_i}$ in order to identify the entities with the higher level of representativeness for a particular event. As stated in our hypothesis, we exploit the frequency prominence principle expressed in Definition 7.1a to spot the candidates. In particular, the absolute frequency of an entity within the set of retrieved documents $D$, noted as $f_a(e_i, D)$, and the Bernoulli appearance rate across all documents $f_{doc}(e_i, D)$ will be considered according to the following formula:

$$f_{Core}(e, D) = f_{doc}(e_i, D) + \frac{f_a(e_i, D)}{f_{doc}(e_i, D)} \tag{7.8}$$

After ordering the entities according to $f_{Core}(e, D)$, top ranked entities start to be added in the *Core* until we found one which is not semantically connected to *all* the other ones already included. This way, we ensure the second condition for the *Core* generation expressed in Equation 7.1b, the semantic coherence.

In order to check if an entity $e_i$ is connected with other $e_j$, we identify existing paths between them in a particular Knowledge Base $KB$. The process of detecting those paths enables to identify other resources $r \in KB$ that can materialize such connections. Those intermediate resources are promoted via dimensions such as "popularity" and "rarity" that are essential components in the original PageRank algorithm [134]. The implementation makes use of the Jaccard coefficient to measure the dissimilarity and assign random walks based weights that are able to highly rank those rare resources, guaranteeing that paths between resources promote specific relations against more general ones [120]. Assuming there is a number $p$ of paths between the entities $e_i$ and $e_j$ ($path_{i,j}$) and being $|path_{i,j}|$ a path length as number of links among resources $r$, we define the similarity function $S_{KB}$ as:

$$S_{KB}(e_i, e_j) = \sum_{1}^{p} \frac{1}{|path_{i,j}|} \tag{7.9}$$

As stated in Equation 7.1b, two entities are considered well-connected if $S_{KB}(e_i, e_j) > s$. The whole logic for identifying the candidate entities that take part of the *Core* is further described in the pseudocode included below:

---

**Algorithm 1** Algorithm for generating the *Core* of a news item based on frequency measures and entity cohesiveness

---

1: **procedure** CoreGeneration($D, E$)      ▷ Set of documents from expansion $D$
   and their corresponding annotations $E$
2:     $Core \leftarrow \varnothing$
3:     $CoreCandidates \leftarrow \varnothing$
4:     **for** $e$ **in** $E$ **do**
5:         $CoreCandidates$.insertionSort(e, $f_{Core}(e, D)$)
6:     **end for**
7:     **for** $candidate$ **in** $CoreCandidates$ **do**
8:         **for** $c$ **in** $CoreCandidates$ **do**
9:             **if** $S_{KB}(candidate, c) \geq s$ **then**
10:                 **return** $Core$
11:             **end if**
12:         **end for**
13:         $Core$.add($candidate$)
14:     **end for**
15:     **return** $Core$
16: **end procedure**

---

In Figure 7.7 we can observe the *Core* for the "Fugitive Edward Snowden applies

for asylum in Russia" news item, with the more representative entities *Edward Snow-den*, *Russia*, *U.S.* and*Sheremetyevo* remaining semantically cohesive as noted by the red links between the entities.
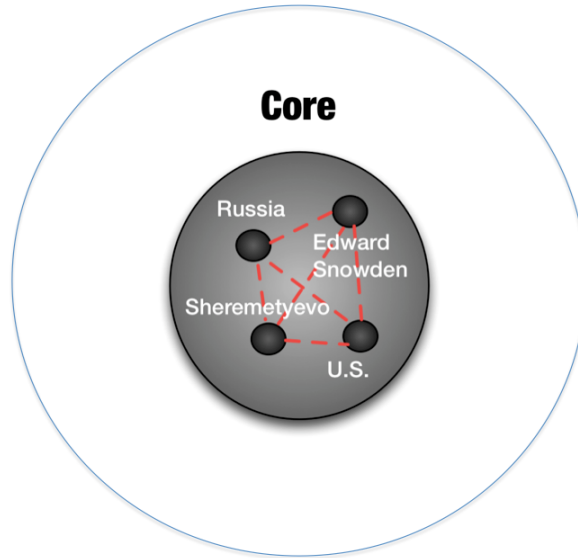


Figure 7.7: Entities inside the *Core*, spotted via frequency functions and semantically cohesive (red links between them)

**Crust generation.** Taking again as input the results $Res_{Exp}$, we use different similarity functions working in certain relevancy dimensions in order to detect which entities are next to the *Core* (as shown in Definition 7.2, $S_{(e_i,Core)}$). Therefore, the *Core* acts like the contextual anchor where *Crust* candidate entities are attached to (see Figure7.8).

In the current approach, two functions grounded on different principles have been considered:

- The semantic relationships between resources in knowledge bases, via the number and length of paths between an entity $e_i \in Crust$ and the entities in the *Core*. Based on the definition of $S_{KB}(e_i, e_j)$ in Equation 7.9, we define the similarity function $S_{KB}^*(e_i, Core)$ as the sum of the different similarities between $e_i$ and $ej \in Core$:

$$S_{KB}^*(e_i, Core) = \sum S_{KB}(e_i, e_j) \mid e_j \in Core \tag{7.10}$$

- The number of web documents talking simultaneously about a particular entity $e_i$ and the *Core*. This function, noted as $S_{Web}(e_i, Core)$, identifies documents in the Web talking about a candidate entity and the *Core* at the same time, while keeping in mind the original volume of documents containing them sepa-
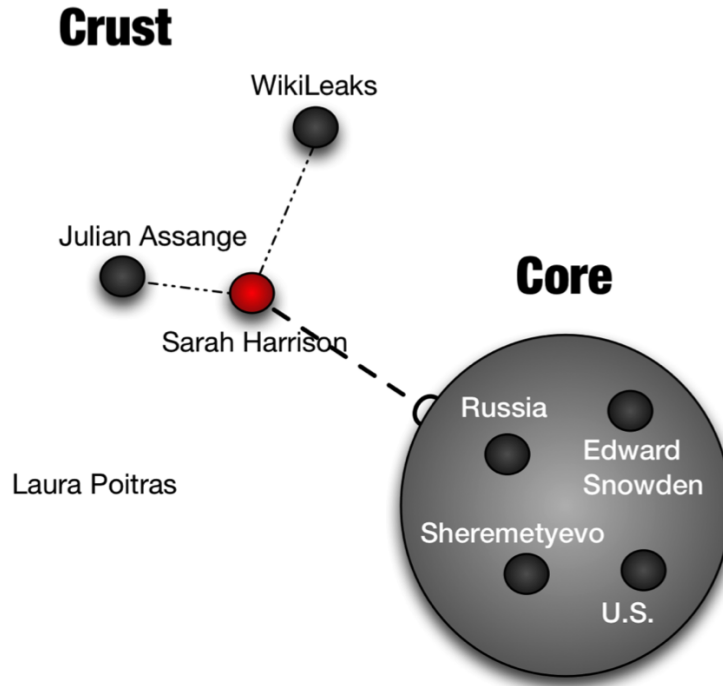
Figure 7.8: *Crust* entity *Sarah Harrison* is semantically attached to the *Core*

rately. Let $E$ be a set of entities and the function $hits_s(E)$, the number of web documents where all $e_i \in E$ are mentioned, $S_{Web}(e_i, Core)$ is:

1. Directly proportional to the square of the number of pages talking about the *Core* and the candidate entity at the same time $hits_s(e_i + Core)$.

2. Inversely proportional to the number of pages talking about the *Core* ($hits_s(Core)$).

3. Inversely proportional to the number of pages talking about the candidate entity alone ($hits_s(e)$). If the entity was already highly mentioned all over the Web, like in the example of very famous persons, the volume of documents mentioning that entity together with the *Core* has to be also big enough in order to be considered.

$$S_{Web}(e_i, Core) = \frac{hits_s(Core + e_i)^2}{hits_s(Core) * hits_s(e_i)} \tag{7.11}$$

The different similarity functions helps to populate the *Crust* with the first top $c$ entities spotted via each method.

**NSS Generation.** In a last step, the entities coming from the *Crust* are attached to the *Core* via the scores produced by the similarity functions described above in order to generate the final NSS of a news item. At this stage, the result is a graph of different event related entities. To evaluate this approach, different projection

functions $f : G \rightarrow L$ have been created, being $G$ a graph based structure and $L$ the resulting list.

## 7.5 Evaluation and Discussion

This section describes the experimental settings and the results of the concentric model approach against a gold standard. We re-used a dataset composed of 5 ranked list of named entities that each semantically annotate one video item. Named entities are extracted from the subtitles, video image, text contained in the video, articles related to the subject of the video and event suggested by a journalist expert. After building a candidate set of entities, this set was presented to 50 participants via an online survey that were asked to rate their level of interestingness. The methodology for building this dataset is thoroughly described at https://github.com/jluisred/NewsEntities, where links to the list of entities and scores per video are also available.

### 7.5.1 Experimental Settings

This section explains the configuration settings used during the execution of the experiments in order to produce the concentric model based news annotations $Res_{Conc}$. The first important parameter to be considered is the length of the entity spectrum that will be analyzed. In order to go beyond the $n = 10$ used in previous expansion approach (see Section 6.3) and to target the positions studied in the Section 7.2.2, the experiments have been configured to work over the first 50 entities coming from the previous expansion phase.

**Named Entity Expansion and Ranking.** In order to identify which entity expansion configuration can potentially serve as the best basis for generating NSS, we have studied the values of $R_{50}^*$ over the complete set of runs considered in Section 6.5.3. Table 7.2 shows the top 8 configurations:

| Run | Collection | | | Filter | Ranking | | | Result |
|-----|-----------|-----|-----------|--------|----------|-----|-----|---------|
| | Sources | $T_W$ | Schema.org | | Freq | Pop | Exp | $R_{50}^*$ |
| Ex1 | L2+Google | 1W | | F3 | Gaussian | ✓ | ✓ | 0.755 |
| Ex2 | L2+Google | 1W | | F3 | Freq | ✓ | ✓ | 0.7532 |
| Ex3 | L2+Google | 1W | | F3 | Gaussian | ✓ | | 0.7457 |
| Ex4 | Google | 2W | | F3 | Gaussian | | ✓ | 0.745 |
| Ex5 | L2+Google | 1W | | F3 | FreqGoogle | | ✓ | 0.7448 |
| Ex6 | L2+Google | 1W | | F3 | FreqGoogle | ✓ | ✓ | 0.7424 |
| Ex7 | L1+Google | 1W | | F3 | Gaussian | ✓ | ✓ | 0.7346 |
| Ex8 | L2+Google | 1W | | F3 | Gaussian | | | 0.7333 |

Table 7.2: Expansion runs ranked by $R_{50}^*$

We have selected the first two runs in Table 7.2 as candidates for feeding the concentric model approach:

- First run $Ex1$ uses the second whitelist and Google, F3 filtering, one week temporal window, Gaussian function, expert rules and popularity. This configuration is also the top result when ranking by $R_{50}$.

- Second run $Ex2$ uses the second whitelist and Google, F3 filtering, one week temporal window, absolute frequency ranking function, expert rules and popularity. This configuration is also third in $R_{50}$ and fourth in $MNDCG_{50}$. It is to be noted that restricting web sites to the ones embedding Schema.org markup lead to poor performance when one wants to maximize the recall of named entities.

**CORE Generation.** Entities have been ranked according to the frequency based Function 7.8. In order to perform the clustering operation that explores the frequency of the entities, we have considered two different strategies: $Core1$, based on Jaro-Winkler string distance [210] over the $surface\_form$, and $Core2$ based on exact string matching of the $link$.

As entities in the core usually express the most general, upper-level concepts that drive the story behind the news item, we will use DBpedia in order to discover relationships between the candidate entities via the similarity Function 7.9. In particular, we have used the optimized path-finding algorithm [34] implemented in the Everything is Connected Engine (EiCE). During the process of filtering only the n-top semantically well-connected entities, two entities have been considered as properly connected if there is at least a path of length 5 between them $S_{KB}(e_i, e_j) > \frac{1}{5} = 0.20 = t$

**CRUST generation.** For generating the Crust, the two Functions 7.10 and 7.11 have been considered with the following parameters:

- $S_{KB}^*(e_i, Core)$ has been configured to work over DBpedia in order to find connectivity between the entities and the $Core$. However, this general purpose knowledge base does not work well with the more fine-grained entities available in the $Crust$, even after relaxing the threshold $t$ in the function $S_{KB}(e_i, e_j)$ for also considering paths of length up to 10. We have empirically detected many missing relations among some entities that, according to the story being told in the news item, should be connected each other. For future work, we plan to rely in other news domain specific dataset potentially containing more links about the studied event.

- $S_{Web}(e_i, Core)$ has been configured to work over an instance of a Google Custom Search Engine where $hits_s(e)$ is the number of documents retrieved. In particular, we have set up the engine with no particular sites to crawl, no temporal filtering and the English language.

After discarding the first similarity function, there will be only one possible configuration for this phase.

**NSS Generation.** In order to project the final graph-based structure into a list that could be evaluated against the Ground Truth, two possible approaches have been taken into account:

- $Core + Crust$: entities in the $Core$ are placed at the top positions of the result list $Res_{Conc}$. Entities in the $Crust$ are added just after those.

- $CrustBased$: all entities in the $Crust$ are added to the list of results $Res_{Conc}$. We also calculate $S_{Web}(e_i, Core)$ for the entities in the $Core$ and we place them in the right position according to this similarity score.

The behavior of those two approaches is further clarified in Figure 7.9.
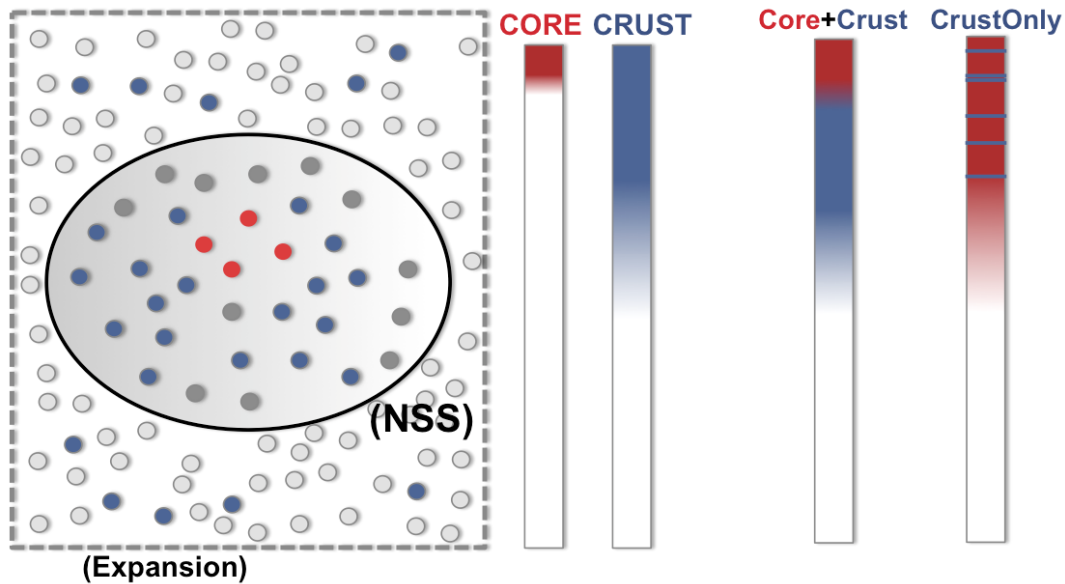


Figure 7.9: Spectrum of true positives in final ranking for each of the $Crust$-$Core$ fusion methods considered: $Core + Crust$ and $CrustBased$

### 7.5.2   Results

Initially, we have performed a specific evaluation of the $Core$, consisting in calculating the precision $P$ of the entities contained in this layer. Results are close to 95%, which means that the great majority of the $Core$ entities are in the ground truth, underling their importance as the driving force to build the NSS of a news item.

Afterward, we have executed the concentric model approach with the different configurations selected in the experimental settings (a total of $2*2*1*2 = 8$). We have also considered two baselines produced by traditional entity expansion techniques. They are identified by the run names **BAS01** and **BAS02** respectively. In addition, we have contemplated the existence of an ideal system able to generate the same

| Run | Expansion | | | | | $Com\,(R, f, v)$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Collection | | Core | Crust | Fusion | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | Avg |
| IdealGT | | | - | - | - | 16 | 11 | 22 | 27 | 19 | **19** |
| Cm4 | Ex2 | | CoreA | $S_{Google}$ | Core_Crust | 21 | 9 | 41 | 44 | 45 | **32** |
| Cm5 | Ex2 | | CoreA | $S_{Google}$ | CrustBased | 20 | 14 | 41 | 44 | 45 | **32.8** |
| Cm6 | Ex2 | | CoreB | $S_{Google}$ | Core_Crust | 27 | 10 | 43 | 44 | 42 | **33.2** |
| Cm0 | Ex1 | | CoreA | $S_{Google}$ | Core_Crust | 22 | 13 | 42 | 43 | 47 | **33.4** |
| Cm1 | Ex1 | | CoreA | $S_{Google}$ | CrustBased | 21 | 16 | 42 | 43 | 47 | **33.8** |
| Cm7 | Ex2 | | CoreB | $S_{Google}$ | CrustBased | 27 | 13 | 43 | 44 | 42 | **33.8** |
| Cm2 | Ex1 | | CoreB | $S_{Google}$ | Core_Crust | 28 | 13 | 43 | 43 | 44 | **34.2** |
| Cm3 | Ex1 | | CoreB | $S_{Google}$ | CrustBased | 28 | 16 | 43 | 43 | 44 | **34.8** |
| BAS01 | L2+AllGoogle, 1W F3 Gaussian + EXP + POP | | - | - | - | 41 | 45 | 34 | 41 | 37 | **39.6** |
| BAS02 | L2+AllGoogle, 1W F3 Freq + EXP + POP | | - | - | - | 24 | 39 | 49 | 48 | 39 | **39.8** |

Table 7.3: Compactness of concentric model results VS compactness of baselines and ideal ground-truth-based result set

perfect ranking available in the ground truth, in order to understand how good we could potentially get. Assuming $R^*(Ex1) = 0.755 \approx 0.753 = R^*(Ex2)$, in Table 7.3, we order the results in terms of compactness $Com\,(Cmn, R^*, R^*(Exn))$ for each of the 6 concentric model configurations, breaking down the scores by video ($v_1$, $v_2$, $v_3$, $v_4$, $v_5$), and showing the final average in the right column in bold.

We can first observe that the average compactness of the concentric model approach is already smaller than the original ones represented by the baselines $BAS01$ and $BAS02$ for all 5 videos. This proves our original hypothesis: $Com\,(BAS01|BAS02) > Com\,(Cm0\text{-}7)$. Additionally, if we compare the best concentric model run $Cm4$ ($Com = 32$) with the best baseline $BAS01$ ($Com = 39.6$), and having as ideal objective the compactness of $IdealGT$ ($Com = 19$), we can report a percentage decrease of 36.9% over the best baseline thus getting closer to the ideal smallest possible NSS. The whole list of entities automatically obtained by executing this approach with the best-performing settings is available in Appendix **??**.

In order to see in a more intuitive way how better is the evolution of $R^*$ values over the whole spectrum, we plot the scores from the concentric based run $Cm0$ against its baseline $BAS01$ (Figure 7.10).

Approximately from $n = 0$ to $n = 22$, we can see how dashed line $Res_{Conc}$ gets faster to higher values of $R^*$, which means it can potentially produce more representative NSS's from lower $n$ positions in the obtained results.

## 7.6   Summary

The different applications and tools consuming information about news can benefit from the computation of a News Semantic Snapshot (NSS), which summarizes and explicitly describes the context of a news items. For generating such a data structure, we cannot rely exclusively on the metadata of that particular news item.
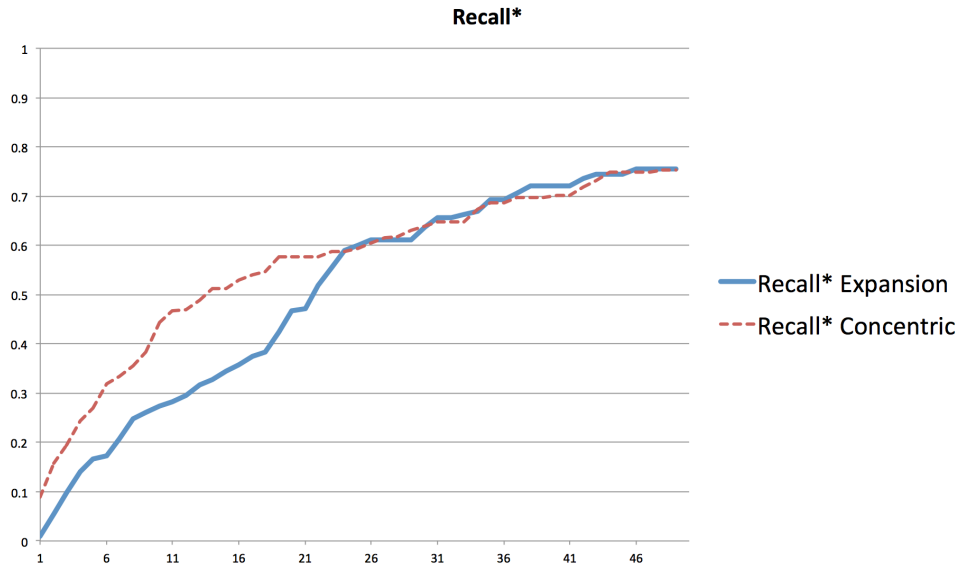
Figure 7.10: $R^*_{1-50}(Res_{Exp})$ vs. $Recall^*_{1-50}(Res_{Conc})$: the concentric model approach gets faster to higher values of $R^*$

Instead, there is a need of relying on the Web and to complement the available information via a process called named entity expansion. However, this step brings in numerous non-relevant entities that need to be discarded. One way of promoting news related entities is to rely on different functions considering aspects such as popularity, serendipity, or semantic proximity, but each dimension requires a dedicate method that is difficult to implement and integrate together with other ranking functions. For overcoming those difficulties and be able to exploit the entity semantic relations, we have proposed a concentric based model for generating the NSS. This model proposes two layers: the *Core*, composed of the most representative entities, which are well-connected between them and spottable via frequency measures, and the *Crust*, which sometimes includes unfrequent entities that are attached to the *Core* via particular similarity functions. In order to ensure the semantic compactness of the *Core*, we have looked at the existence of DBpedia paths between each entity pair. For establishing connections between the *Core* and the entities in the *Crust*, we observed that a general purpose knowledge base such as DBpedia is not necessarily ideal due to the fine grained nature and freshness of the entities in the *Crust*. However, other dimensions like the Web presence of both the *Core* and *Crust* entities have successfully highlighted relationships promoting relevant entities. The experiments in terms of $R^*$ over a set of results produced by the concentric model approach have revealed a significant improvement in the level of compactness of the new method compared with traditional expansion methods, which allows to produce more concise and at the same time representative NSS. The list of entities auto-

matically obtained by executing this approach with the best-performing settings is available in Appendix **??**.

Our future work includes: *i)* Unsupervised NSS Creation. The process of expanding the initial set of entities and build up the NSS has been initially conceived to be unsupervised. However we think that applying supervised techniques like *Learning to Rank* can improve the quality of the result obtained, leading to a better reconstruction of the context of the news in the video; *ii)* increasing the length of the spectrum of annotations used for feeding up the concentric model, up to the whole list of annotations (which it has been impossible due to quota restrictions in search services such as Google CSE). This is already being studied in some parallel investigations in the same line and will be reported in upcoming research papers; *iii)* being able to spot not only the degree of connectivity between the entities in the *Crust* and the *Core* but also the predicates that characterize those connections. This is equally being studied already through some further analysis of the related document collected in order to generate snippets that help to understand the causes behind those entity relationships. *iv)* studying the role of the model in tracking the evolution of news events over the time, where we expect that the *Core* remains more or less stable, and the entities in the *Crust* will vary reflecting the particular facts that together compose the entire event.

CHAPTER 8

# The NSS in the News Consumption Paradigm

## 8.1 Introduction

We live in a constantly evolving world where news stories and relevant facts are happening every moment. In this Chapter we will make emphasis on those stories to understand how they are offered to the viewers and consumed: millions of news articles, posts, and social media reactions are created, providing a multitude of viewpoints about what is happening around us. Many applications have tried to deal with this complexity from very different angles, targeting particular needs, reconstructing certain parts of the story, and exploiting certain visualization paradigms. In this chapter, we identify those challenges and study how an adequate news story representation can effectively support the different phases of the news consumption process. We analyze how an innovative data model such as the News Semantic Snapshot (NSS) presented in Chapter 5 and further developed in Chapter 6 and 7 can capture the entire context of a news story and make it available for machines to be exploited. This model can feed very different applications assisting the users before, during, and after the news story consumption. It formalizes a duality in the news annotations that distinguishes between *representative* entities and *relevant* entities, and considers different relevancy dimensions that are incorporated into the model in the form of concentric layers. Finally, we analyze the impact of this NSS on existing prototypes and how it can support advanced features that are expected to come in the near future. This research has been published in [57].

## 8.2 Motivation: Assisting Viewers in News Consumption

Even the a-priori conventional stories that we daily consume have some underlying facts that, during certain situations and for some particular users, become important and need to be unveiled. Those facts can be described in very different ways: publishers may aim to emphasize certain aspects of the story, target a specific audience, or respond to particular viewer's needs. The role of the consumers can also evolve

over the time, from a passive and less engaging behavior to a deep-into-the-details mode that requires deeper knowledge about the facts being reported. In this highly challenging ecosystem where stories are spread all over different pieces of information, interpreted by many different users and presented by various data sources, the existence of a model representing the entire context of the news item becomes highly relevant.

The construction of such advanced story representation has already been addressed during previous chapters in Part II. Under the hypothesis that a single video news item is often not enough to capture the complete story being reported and can be biased or even partially wrong, we have put in practice various information retrieval techniques for combining the original news content with additional data collected from other external sources. This process, called Named Entity Expansion, is able to produce a ranked list of named entities that complements the initial set of detected entities in video subtitles with other item-related entities captured from Web documents. The top $n$ items in this list build the conceptual structure called the Newscast Semantic Snapshot (NSS) of a news story. In Chapter 7, this NSS evolves from a plain ordered list of entities to a multi-layered concentric model, which is more appropriate for representing the duality between the most representative entities and the other ones that are relevant to the context of the news item due to diverse reasons such as interestingness, informativeness or popularity.

In this chapter, we analyze how this NSS can support the different requirements derived from the news consumption process. This structure needs to (1) be easy to exploit and flexible enough for giving an answer to different applications, (2) deal with the duality present in the news annotations, by differentiating between entities that better summarize a story and the ones that acquire relevancy as the story is further consumed, and (3) emphasize the relationships established between the different entities inside the context of the story, focusing more in the reasons for having such a connection and less in their absolute importance inside the story. In the last section of the chapter, we will analyze some prototypes that project the rich spectrum of relationships within entities into a simpler and human easier way to consume the story, in order to understand how they can benefit from such a news item representation.

## 8.3 The News Semantic Snapshot in the News Consumption Paradigm

The News Semantic Snapshot (NSS) is a graph structure that tries to represent the entire context of a news story, where nodes are named entities and edges represent relationships among them. According to the hypothesis stated in previous Chapter 7,

a NSS can be modeled following a schema of concentric entity layers.

The NSS aims to exploit and harmonize in a single conceptual model different semantic relationships established between the news' entities. This model makes explicit a duality in the entities via two main layers namely the **Core** and the **Crust**. The former is composed of a small number of key entities that are essential to identify a story. Those entities have the highest potential to better summarize the main facts behind the news story. They are frequently mentioned in related documents and therefore spottable via frequency-based functions. In Figure 7.4, the $Core$ is composed of the entities Russia, Snowden, U.S., and Sheremetyevo (the airport where the action is taking place), which are the seeds for a good understanding of the story. On the other hand, the $Crust$ is composed of the entities expressing the particular details around the news items. They are mentioned in some specific related documents, but they are not always spottable via frequency-based measures. Their relevancy is instead grounded on the existence of relations such as popularity, serendipity among those entities and the $Core$. In Figure 7.4, some entities such as Anatoli Kucherena (Edward Snowden's lawyer), are not so prominent at a first glance, but they definitely play a role in the story and can contribute to a better understanding of the facts. The semantic context of a news item can be therefore built by combining the $Core$ and the $Crust$ into a single data structure $NSS_{concentric} = Core \oplus Crust$. In order to go deeper in the formalization of the NSS, we focus on some other aspects of this conceptual model that are important in the news consumption scenario.

**Reconciling Relevancy Dimensions**. The concept of relevancy is extremely wide and complex. It depends on several variables that two different persons that ought to judge the relevancy of an entity would rarely agree on. However, the layer based representation used in the NSS better supports the complex and multi-dimensional relevancy relations established among the entities involved in a news item and allows to formalize the potential reasons that are linking them to the $Core$. The $Crust$ becomes then a place for hosting different relevancy dimensions [206], which bring diversity to story description: entities denoting opinions, informativeness, serendipity, popularity, interestingness, unexpectedness.

**Finding Predicates to Entity Relations**. In Section 7.3, we discussed the importance of discovering and explicitly establishing relations among the entities inside the NSS. We propose now a step further by considering, not only the unlabeled relations, but also explicit predicates characterizing the entity links. Finding such property names and formalizing those entity dependencies is still an open challenge. In Section 7.4, co-occurrences of entities in documents collected from the Web revealed how tight were the relations in the context of the story. A further analysis of those documents could help to provide additional information enabling to label the predicates.

As the NSS is a graph-based structure, this information can be straightforwardly

incorporated into the model. The predicates established between the elements inside can become labeled links thanks to the flexible nature of the model. Prototypes can exploit such kind of annotations in order to make the users aware of the reasons that make those entities relevant.

**Tracking Stories over Time**. In most of the cases the different facts shaping the story plot happen in a chronological order. This implies that the entities and predicates involved in such facts are especially relevant during a particular period of time within the time span of the story. Once more, the flexible graph-based NSS model can support edges annotated with temporal references in order to reflect when a particular entity plays an important role in the plot of the story or holds a specific relationship with others. Tracking the evolution of those relationships in time opens a room for timeline based summarization prototypes that highlight the milestones characterizing the story evolution.

## 8.4 The News Consumption Paradigm

Potentially, there are numerous ways of consuming a news story $St_{News}$. Each tool or application displaying information about a news item follows a different philosophy, targets a different audience, and presents the main facts from a different angle. In addition, the news consumers also evolve over time: at the very early stages of the consumption process, it happens that they do not even know what they want to watch; after glancing the variety of content available, he mades a selection and consumes the item. Many questions start to pop up during and after the item is consumed. As the viewer's knowledge about the matter grows, his information requirements get bigger as well. In last phase of the evolution some users satisfy this higher information demand by exploring related documents and more elaborated diagrams complementing and extending the original content. This news consumption evolving process has been illustrated in Figure 8.1, which has been inspired by the classic "evolution of man" diagram: the monkey on the left side represents the viewer who browses the content without a clear idea of what he/she wants to watch; in the following state, a more upright monkey already knows what to consume; third state represents the content visualization, where some questions about the depicted facts start to arise; a human figure is on fourth place representing the users that actively find an answers to the aforementioned questions with the use of dedicated applications, and on the very right; and in the case of the last man on the right, it represents the kind of user that have evolved so much during time than now require advanced ways of representing the different details about the process story.

Despite this variety of alternatives in this section, we propose a model for classifying those news consumption approaches. Having for reference the time, the user is actually consuming the news document $d_{news}$ describing the story $St_{News}$, we have
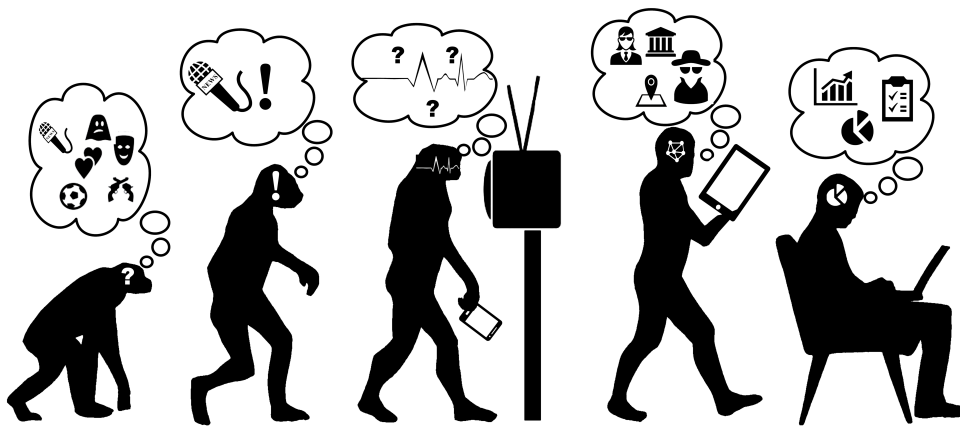
Figure 8.1: Evolution of a viewer consuming news: from content selection, to comprehension and further exploration

identified three main phases: the *before*, the *during*, and the *after*. In each of those phases, a user's behavior is different: there is an evolution in the understanding of the story and consequently in the information requirements of the applications presenting the story.

### 8.4.1　The before

Users, in this phase, have normally not consumed the main news item yet, so their understanding about the story is limited. Most of the times, they require a quick and easy way to interpret the main facts so they get to know in a glance what the news is talking about. In other cases, additional content is displayed in order to illustrate the news context. This type of recommendations look for content that is very similar to $d_{news}$, leaving diversity aside. A special application category under this phase includes the advanced summaries that aim to fully tell the story without having consumed the original content.

### 8.4.2　The during

It corresponds to the time the user is watching the main document illustrating the story, $d_{news}$. It normally implies a passive information activity where the users are pretty much focused in the task of consuming the document without engaging in any other actions. During this phase, the user's knowledge grows from the background information provided in the *before* phase to a most detailed understanding of the news. A good example of prototypes under this category is a second screen application aiming to illustrate what is being said on the news with minimal user interaction.

### 8.4.3 The after

The user became fully aware of the basics of the story and wants to go deeper into the details, switching to an active mode: browsing description of entities categorized into dimensions or ultimately jumping to other related stories. The level of interaction drastically increases since the user becomes more engaged, moved by the curiosity of discovering more details. The main document $d_{news}$ can be enriched with additional content, focusing on diversity, and detailing some specific facts of $St_{News}$. Other applications falling under the same consumption phase are advanced interactive summaries with browsing capabilities.

### 8.4.4 The NNS in the Consumption Process

We formulate, as hypothesis, that the NSS of a news item is a knowledge representation model that effectively captures the context of a story and can support different existing news applications. This graph layer-based structure helps to populate very diverse prototypes aiming to support users in interpreting the news. For the sake of illustrating our hypothesis and without claiming to provide an exhaustive plot based on quantitative data, Figure 8.2 shows how the duality between *Core* and *Crust* in the concentric model can better satisfy the evolution of user consumption needs across the different consumption phases, as a result of changes in aspects like viewers' knowledge about the story, user's engagement and content diversity.
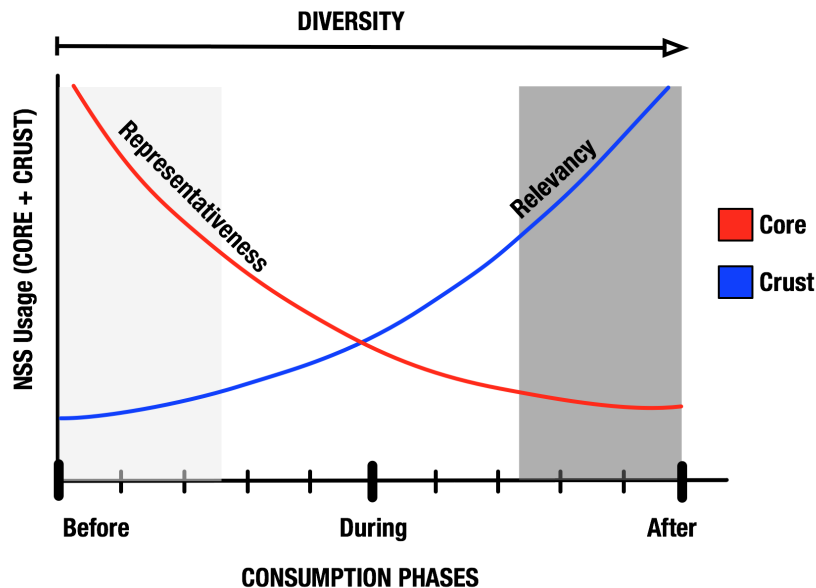


Figure 8.2: *Core* and *Crust* usage along the different consumption phases

As defined in the Formula 7.1 in Section 7.3, entities in the *Core* usually express

the most general, upper-level concepts that drive the story behind the news item. Even if those entities will be present in almost every stage of the news consumption, they have a stronger decisive role in the *before* phase (left side of Figure 8.2), decreasing in importance as $d_{news}$ is consumed and the *after* visualizations come into play.

The *during* is a special phase that requires both *Core* and *Crust* entities (middle part of Figure 8.2), specially when they are mentioned in the document $d_{news}$. In particular, *Core* entities start to be less demanded since they have often been consumed during the *before*, while *Crust* entities start bringing added value in revealing the non-obvious facts around the story plot.

In the last phase of the consumption process, the so-called *after*, users have already a fair understanding of the news story. The entities in the *Core*, which were highly present during the previous phases, become often too obvious and are therefore not so critical to be used. Instead, the entities in the *Crust* bring those particular details that users want to consume, in an attempt to move from a general understanding of the news item to explore specific story details (right side of Figure 8.2). Since the *Crust* considers different relevancy dimensions, applications can easily move along them and bring the diversity desired at those latest stages of the consumption process. In addition, for those applications displaying timeline based summaries, we propose an additional hypothesis stating that the *Core* remains stable in time and have less interest, while the *Crust* contains the entities that bring the stand-out information in particular periods of time and need to be displayed.

## 8.5    An Ecosystem of News Applications

In this section, we review existing applications and prototypes for consuming news. We classify them according to the different consumption phases identified in Section 8.4, and we analyze how they would benefit from a graph representation model like the News Semantic Snapshot in order to make a first qualitative evaluation of our hypothesis.

**Prototypes for Before Consuming the News Item.** In [146], we presented an approach for getting a quick overview of a video content enabling the user to decide if he is interested or not in the story. We automatically select some fragments within the video called Hotspots, which contain annotations with higher frequency scores. This notion of representativeness is clearly aligned with the definition of the *Core*. Entities inside this layer could be straightforwardly used for the Hotspots creation.

Something similar occurs in [6] where some small video fragments are hyperlinked to others based on some visual or topical similarity. Even if such a task can be tackled using multimodal analysis techniques, applications at early stages of the news

consumption focus mostly in finding content as similar as possible to the original. Therefore, entities in the *Core* are suitable for triggering searches on document indexes where that additional content can be found.

**Prototypes During the Consumption of the News Item.** In [144], we presented a second screen application implementing a slideshow that gives the user access to factual information about Person-type, Location-type and Organization-type entities that are related to a news story displayed in the main screen (see Figure 8.3). *Core*-like entities are still shown when mentioned in the video to illustrate the story as a whole, but there is an increasing use of other entities clarifying more specific facts in the video as they are displayed (*Crust*-like entities). Given the absence of a NSS in [144] to feed such prototype we relied on the first implementation of the Entity Expansion Algorithm as explained in Section 5.5 in order to recreate a very basic version of the *Crust*.
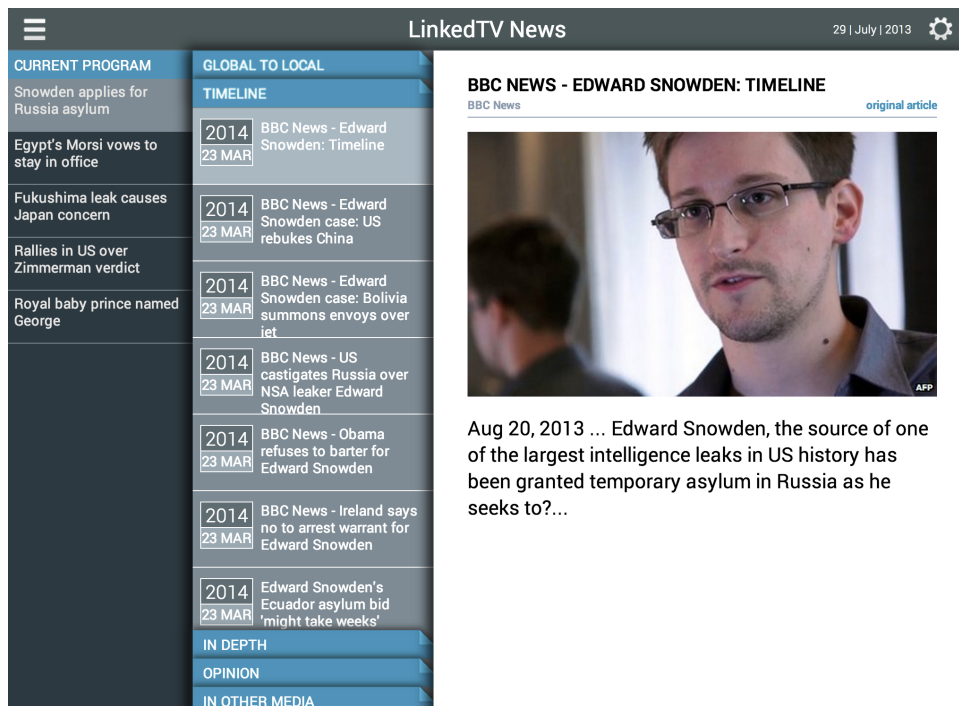


Figure 8.3: Context browsing during passive mode

Other example of a second screen application supporting the user during the news consumption is the Kinect[1] prototype described at [115]. This application aims to enrich the user experience when watching television by visualizing contextual information on a second screen device and controlling the video watched using a Kinect device (see Figure 8.4). The user can grab, at any time, a fragment from this video to obtain more information about it. Based on the results of performing

---

[1]https://dev.windows.com/en-us/kinect

named entity recognition on the subtitles of the video fragment, a first set of relevant entities are spotted according to their frequency in the transcripts. Those entities are used to gather information from the Linked Open Data cloud, discover what the *vox populi* says about this program, and generate media galleries that enrich the seed video fragments grabbed by the user. This application would largely take advantage of the additional set of non-explicitly-mentioned entities available in the NSS and its far more sophisticated ranking. A video of this prototype in action is available at http://youtu.be/4mSC685AG7k.
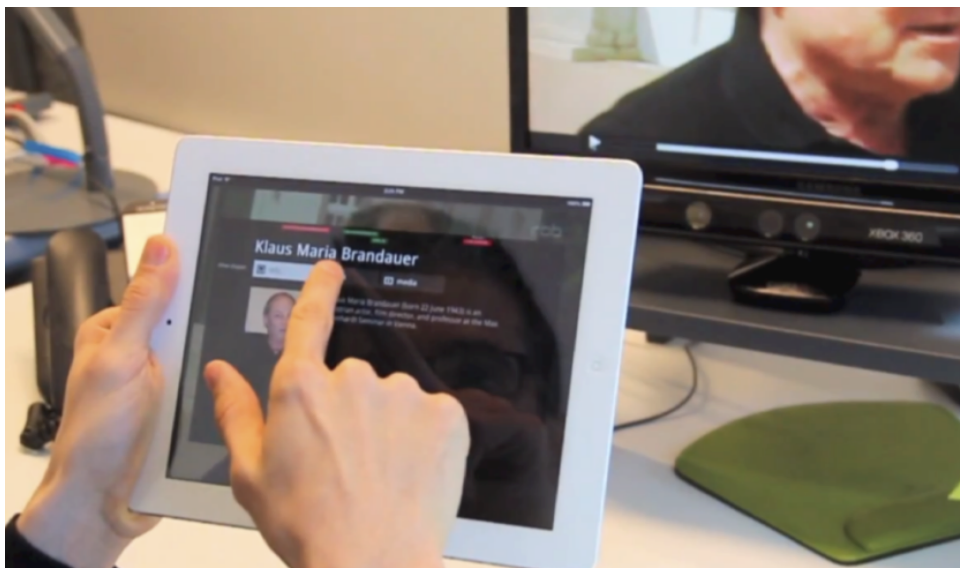


Figure 8.4: Consuming contextual news information by interacting through Kinect

**Prototypes for After Consuming the News Item.** We find an example of an application illustrating the story after consuming the news item in [144]. The active mode of the demo targeted the idea of a user who wants to further dig into the details of the story via some additional content that is proposed along different dimensions. Some of those additional content facets can easily match the layers envisioned in the *Crust* definition, like *Opinions from Experts* that aligns to *opinions*, or *In Other Sources* that aligns to *informativeness*, revealing the importance of the multi-layer philosophy inside the News Semantic Snapshot.

Under this category of prototypes we can also include applications offering advanced interactive visualizations for summarizing the entire context of the story. This is an extremely challenging task even when performed by experts in the domain. The representation model offered by the NSS considers relations between entities that can help to implement conceptual diagrams where entities are related to each other via particular connections like in Figure 8.5. In [116] and [148] we describe the prototype

MediaFinder[2] that creates such kind of visualization out of the information retrieved from social networks. By applying some of the annotation techniques reported in Part II, this application makes sense out of the different record streams of heterogeneous data about human's activities, feelings, emotions and conversations that can help to shape news stories in real-time. This is a extremely challenging task due to the heterogeneity of the data and its dynamics making often short-lived phenomena. The developed framework collects microposts shared on social platforms containing media items as a result of a user query, for example a trending event, and automatically creates different visual storyboards that reflect what other people have shared about this particular event. More precisely it leverages on: *(i)* visual features from media items for near-deduplication, and *(ii)* textual features from status updates to interpret, cluster, and visualize media items. A screencast showing these functionalities is available at: http://youtu.be/8iRiwz7cDYY. The agents displayed in this prototype and the links between them could be directly taken from a data structure like the News Semantic Snapshot, which already follows the same graph based philosophy where the relevant nodes related each others through different edges.
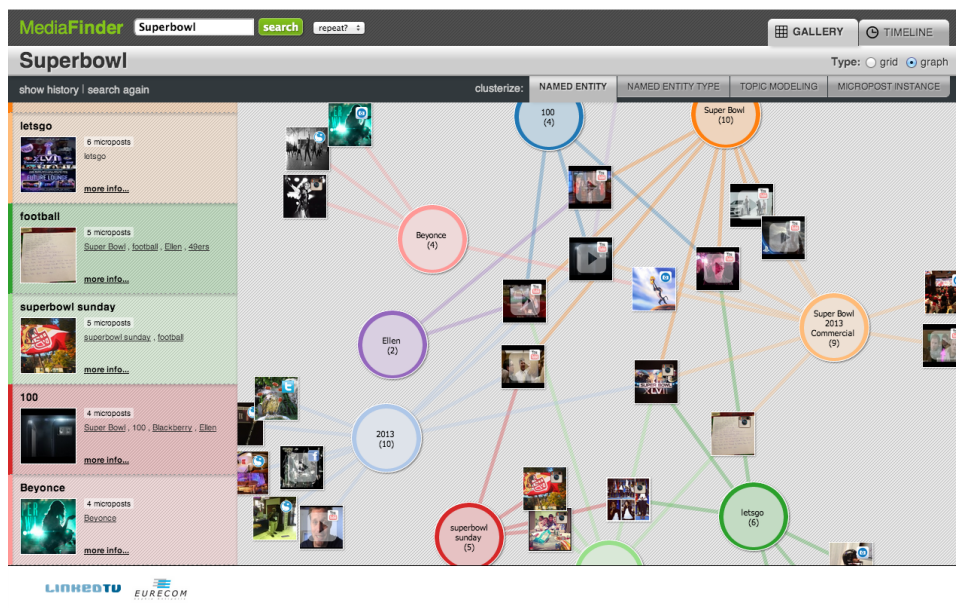


Figure 8.5: Advanced summarization prototypes

A third example in this category are the time-based representations that break down the story in relevant facts that are chronologically represented over the X axis. In [114], we analyzed the story of the Italian Elections 2013[3] during one week after the voting process. One of the main difficulties encountered during its implementation was to filter out those entities that were buzzing during the entire week so

---

[2]http://mediafinder.eurecom.fr/
[3]http://mediafinder.eurecom.fr/story/elezioni2013

they became obvious like *Italy*, and to promote instead those entities that peaked in relevance during certain moments of the week for particular reasons, like *Merkel* who had a meeting with the Italian president on the 1st of March (see Figure 8.6). This phenomena reinforces our hypothesis about the *Core* entities remaining stable in time and becoming useless for such timelines prototypes, while the ones in the *Crust* bringing the interesting facts that need to be shown.
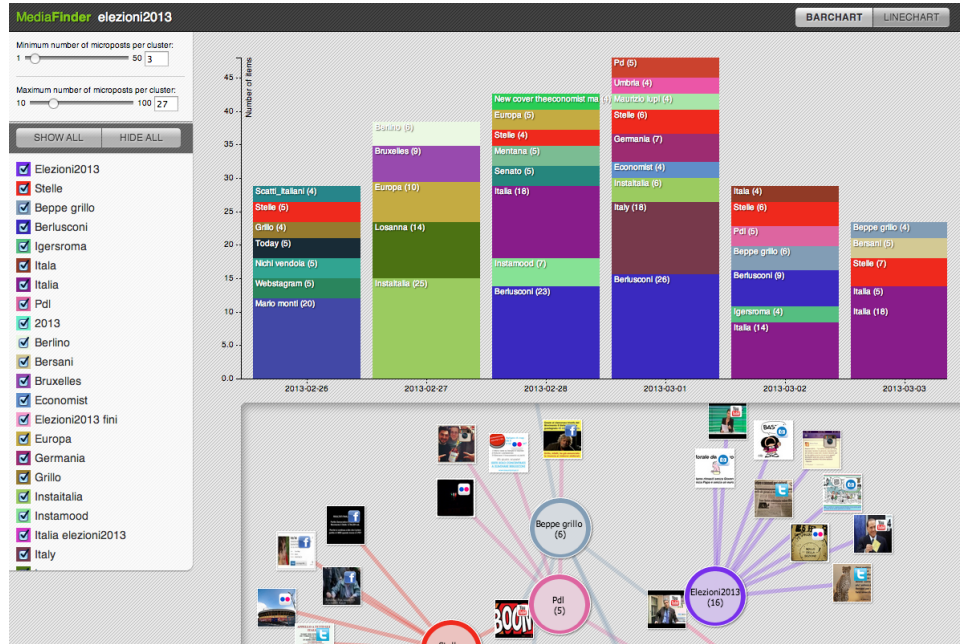


Figure 8.6: Temporal distribution of relevant concepts during Italian Elections 2013

**Other Prototypes.** In addition to the previous prototypes, there are also other applications targeting a bigger portion of the news consumption phases spectrum. A good example can be found in the online demo Hyperted[4] already introduced in Section 4.3.2.3, that offers an innovative way to consume TED talks, by supporting the user not only at the pre-consumption stage via Hotspots calculation, but also during the viewing through entity highlighting and in the period just after by linking to similar courses and other related TED chapters (see Figure 8.7).

## 8.6 Summary

The different applications consuming news can benefit from the existence of a graph-based model able to capture the entire context of the story. In this Chapter we have studied how to exploit in the News Semantic Snapshot model introduced in previous sections inside this Part, which grounds on the existence of semantic relations between

---

[4]http://linkedtv.eurecom.fr/Hyperted/
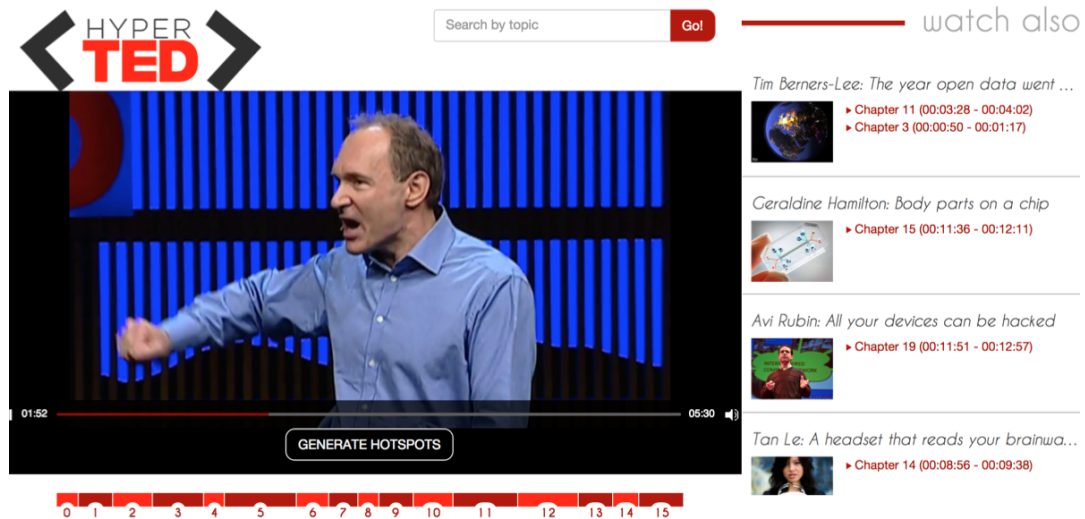
Figure 8.7: HyperTED prototype: consuming TED talks at the level of fragments

entities describing a news story. This model has a concentric nature considering two main layers: the *Core*, which includes the most representative and frequent entities, and the *Crust* which is composed of additional entities that become relevant because of certain relationships happening between them and the *Core*. In addition, we have identified three phases of the news consumption process: the before, the during, and the after. The NSS can support applications falling inside each of those categories. As support of the underlined hypothesis, we have reviewed some of our developed prototypes and existing applications today, identifying the challenges we faced when capturing the story information, and explaining how the use of NSS to feed their needs would have led to a mode adequate, straightforward implementation.

In future work we plan to experiment with other innovative techniques for automatically populating this concentric model and display it to the viewer in an intuitively and seamless way. We aim to use existing domain specific knowledge bases for further contextualizing the relationships present in the NSS, and exhaustively evaluate what are the main relevancy layers inside the *Crust* that should be considered to ensure the users get the best consuming experience possible.

# Conclusion of Part II

In this Part, we have tackled the problem of annotating news items in terms of the context of the story being presented in it. After justifying and motivating the need of such bigger picture in order to interpret complex news on different media channels, we have proposed a knowledge representation structure called the News Semantic Snapshot that formalizes this context in the form of relevant entities playing a role on the story. In order to evaluate different automatic approaches for generating such NSS, we have developed a Gold Standard of news items annotated considering different dimensions of the video content, from entities being mentioned in the video, to persons depicted on them or places written on the banners.

Our experiments against this Gold Standard have probed how particular configurations of the News Entity Expansion (see Section 3.2.3) algorithm developed during the research period of this Phd can bring the missing parts of the story context and help to recreate the NSS of international news. However, together with the relevant information the Entity Expansion phase also brings other less relevant items that need to be discarded. In Chapter 5 we analyze how frequency based ranking functions can effectively promote relevant entities. This logic has been published via the News Expansion REST API service[5]. Unfortunately those features alone are not able to explain the presence of certain entities in the Ground Truth which are barely mentioned on related documents retrieved during the expansion process, but have been considered important for the current news story. Hence, in Chapter 6 other functions working on different dimensions have been implemented and evaluated. The good scores obtained prove how they over-perform the initial approaches purely based on entity frequency, revealing the multidimensionality nature of the NSS.

But dealing with so many dimensions increases the complexity of the problem and leads to situations where further tuning the different scoring functions, or change the way they are projected in the final ranking does not bring any significant improvement. In Chapter 7 we propose a concentric model approach for generating NSS's that exploits a duality in the entity relevance detected during our experiments, between the highly representative, frequency spottable entities and other entities becoming relevant because very particular reasons. This method remains agnostic to the different relevancy dimensions that can be explaining the promotion of certain entities. Finally, in Chapter 8 we elaborate on the importance of the NSS for feeding different news applications assisting users in the task of consuming news stories. In particular, we analyze how a concentric structure can better support the information needs of prototypes working in different phases of the news consumption paradigm, the so

---

[5]http://linkedtv.eurecom.fr/entitycontext/api/

called: *before*, *during* and *after*.

**Extending the Ground Truth: Future Experiments**.

The high level of detail used for thoroughly annotating the context of the news items in our Ground Truth has compromised the size of the corpora considered. We have already made the first efforts in order to extend the corpora from 5 to 23 videos by adding 18 new BBC videos from last weeks of 2014 and beginning of 2015. In order to annotate them, we have used a more relaxed methodology, based on the manual assessment of 3 experts in the domain of the automatically generated annotations in 3 dimensions: 1) results of NSS generation algorithm presented in Chapter 6, 2) entities detected by NERD on video banners manually transcribed by humans, and 3) entities promoted by popularity based method described in Section 6.4.3.1. The gold standard annotations are already available online at the NewsEntities repository [6]. Those new videos are also divided in categories so it is possible to study how different configurations of the NSS generation process suit to different news genres.

Preliminary experiments have checked the performance of the multidimensional and the concentric-based approached over this bigger corpora, in terms of $MNDCG_{10}$. Score for best multidimensional NSS generation run is $MNDCG_{10} = 0.494$. This reveals the difficulty of annotating more diverse datasets. We want to overcome this issue by tailoring the expansion process to the different video categories. Thanks to an updated API quota plan on Google, we have also been able to apply the *Crust* generation function over the complete list of annotations retrieved from expansion. The concentric approach hasn't improve significantly the $MNDCG_{10}$ scores, which have fluctuated around the $0, 5$ for the best configuration settings.

However, and leaving aside the fact that $MNDCG_{10}$ is not the best metric for evaluating the NSS and a more adequate evaluation in terms of compactness will be performed soon, a deeper look at the results revealed very encouraging indications suggesting a brighter future: indeed, the concentric approach has been able to *identify various missing entities in the Gold Standard that were not initially considered because of human limitations* when exhaustively covering the whole semantic picture of the newscast. We then plan to use our approach as a means to suggest relevant entities in the process of the gold standard creation, as a much powerful alternative than the less diverse and time consuming human process followed during the first gold standard dataset creation (see Section 5.4). Results are planned to be published on a journal[7] paper in the upcoming weeks.

---

[6]https://github.com/jluisred/NewsEntities/tree/master/dataset2014
[7]http://tois.acm.org/

# Conclusions and Future Perspectives

In this last chapter we summarize the major achievements of this thesis and we outlook some future work that would be worth exploring for upcoming research efforts in this line.

## 9.1 Achievements

This thesis has studied how semantic technologies can be effectively applied to multimedia content in order to produce a new generation of video annotations that (1) turn multimedia documents into first citizens of the Web, and (2) give support to performing more advanced, human friendly operations over particular fragments inside the content, particularly in the news domain.

The accomplishments in Part I are the result of the efforts made on developing valid methodologies and applying semantic Web technologies over general multimedia content. Ontologies like **LinkedTV Ontology** reuses vocabularies in the domain and implement Linked Data principles in order to bring multimedia documents to the Web ecosystem at different levels of granularity, making them coexist with other information already existing on the Web. We have shown how this multimedia content and their corresponding fragments can be annotated following different semantic techniques like **Named Entity Extraction** or **Named Entity Expansion**. Semantic annotations open the door to a new set of possibilities: (1) they allow to refer to already existing resources in the Web of Entities automatically bringing additional knowledge and semantic relations between resources that algorithms can leverage on, and (2) they give support to advanced reasoning and inferencing techniques leveraging on the explicit semantic constraints available in the aforementioned ontologies. The text-based annotations can be combined with the result of visual-based approaches to empower multimodal algorithms that outperform the individual techniques. Through our participation in campaigns like **MediaEval** we have also probed how those semantic annotations can be used for interlinking content with other content in the same collection, with other multimedia documents in the Web from editorial blogs to fresh media, and with other textual documents further describing the original

facts. This whole expertise can be applied over entire video corporas like we have done in **HyperTED**, laying the foundations for a new generation of applications for consuming, recommending and browsing multimedia content at the fragment level.

The achievements obtained in Part II are derived from the application of the aforementioned techniques to the domain of the international news stories being daily offered in the media. We have highlighted the importance of properly reconstructing the context of those news items, materialized in the so-called **"News Semantic Snapshot" (NSS)**. This knowledge representation is composed of a set of Named Entities that bring in the big picture of the addressed story: persons involved in the background facts, locations that are somehow connected with the one where the video is taking place, etc. In order to reconstruct this semantic context, we have relied on external knowledge collected from the Web that complements the information described in the analyzed news items, coming mainly in two flavors: unstructured textual documents talking about the same news, and resources in general knowledge bases like DBpedia providing structured data. Our different experiments have revealed that we can fairly reproduce the NSS of different video items by comparing them with a **Gold Standard** of News Entities. In particular, **Named Entity Expansion** techniques have been shown to accomplish the task of bringing the entities absent in the original content. Applying ranking and filtering functions working in different relevancy dimensions identified in the domain, we successfully promote the entities that need to be part the NSS and discard the ones that do not. As the number of functions considered increase, the generation of NSS becomes more complex. In this regard we have proposed a **Concentric** approach for reproducing the context of news items that go beyond traditional information retrieval strategies in terms of **compactness** of the resulting NSS, while remaining agnostic to the possible relevancy dimensions involved in such selection. Finally, we have proved the necessity and adequacy of a concentric representation of the news story for supporting user oriented applications assisting the user in the different consumption phases: the **before**, **during**, and the **after**.

## 9.2   Future Perspectives

Apart of the assessing the achievements obtained by a particular research work, a good indicator to qualitatively measure the impact of those contributions is to study the number of alternative research directions that have been triggered. In this section we enumerate various research areas derived from the work accomplished in this thesis, which would be interesting to explore in future initiatives.

**Refine Ontology Models**. Nowadays data representation models in the Web tend to simplicity. Both publishers of data and experts developing applications are demanding more lightweight models that allow to bring the advantages of structured

semantic data without introducing extra complexity. For this reason and following what is happening already with other data in the Linked Open Cloud, multimedia annotations presented in Chapter 2 need to be refined to get maximum simplicity. For example, the use of very powerful vocabularies like The Open Annotation Data Model needs to be further analyzed in order to avoid the overuse of complex serialization patterns that theoretically expand consuming possibilities and data flexibility but become inefficient to be applied in real life scenarios.

**Study New Generation of Annotation Techniques**. The performance of annotation techniques such as Named Entity Extraction and others studied in Chapter 3 can be boosted if, instead of remaining domain-agnostic and do not adapt to the particular corpora being analyzed, they take into account the particular scenario the document is being analyzed for. In this regard, the efforts made in reproducing the context of news stories can be extrapolated to other scenarios in order to further tune the different annotation phases. Traditional techniques can still be important for triggering the generation of a background model that is later used as feedback for narrowing down the results to ultimately more context-aware and precise annotations. We can also foresee benefits in using a concentric approach for representing the contextual information in other tasks, such as entity disambiguation where entities in the *Core* can provide a first set of candidate resources to interlink to, while the *Crust* entities will perform a fine-grain selection of the more appropriate links to resources according to what is being explained in the original document.

**Extending the Ground Truth** The high level of detail used for thoroughly annotating the context of the news items in our Ground Truth has compromised the size of the corpora considered. Some efforts have been already made in order to extend the corpora from 5 to 23 videos using a more relax methodology, see [1]. Those new videos are also divided in categories so it is possible to study how different configurations of the NSS generation process suit to different news genres. In addition, in order to alleviate the efforts made by editors and target an even bigger set of videos to annotate, a good strategy would be to apply crowdsourcing to build the ground truth. This can significantly increase the number of users assessing NSS per video item, and therefore bring to the game a more realistic way of calculating the relevancy of the entities where more specific dimensions are considered (serendipity, interestingness, informativeness, etc) and different or even contradictory judgements from users can be taken into account [8].

**Alternative Ranking Techniques**. The approaches proposed in this thesis for generating the NSS of news items are fully unsupervised. If the Ground Truth corpora is extended, if would be also possible to experiment with Learn to Rank [23] techniques, in order to filter the entities collected during the Named Entity Expansion process. Such kind of research direction would also unveil details about which entity

---

[1]https://github.com/jluisred/NewsEntities/tree/master/dataset2014

features count the most when judging the relevancy of an entity along categories of news items: different entity attributes can be higher or lower weighted depending on the genre of the analyzed video. Along the same line, the knowledge structure generated by our concentric approach can be a good starting point for applying graph-based ranking algorithms on top, such as [93], which have already shown a good performance in similar information retrieval areas.

**Publish Generated Knowledge**. To be able to push the generated news annotations back to the open Web would be very beneficial for other agents that could use them for many other purposes, such as advanced search capabilities, reasoning over stories and facts happened in the past, or training themselves in better recognizing new events. It is therefore interesting to investigate how the generated knowledge can be correctly incorporated with the already existing information available in graphs like DBpedia. Different considerations concerning the vocabularies to apply or the need of human curators to ensure the quality of the generated data need to be taken into account. In addition, fresher information available in those sources would also mean more possibilities to rely on them for analyzing very recent news stories that are normally not covered in them, hence reducing the need of processing unstructured documents where the last minute information is normally available.

**Concentric Knowledge Representation in other IR tasks**. The concentric model establishes a capture and representation technique that can be used not only for annotating news stories but also for other information retrieval scenarios involving events and facts, such as educational and cultural resources, argumentation approaches, etc. Also it can be used to support certain Web search tasks, in particular concerning the so called exploratory engines [107] where users, instead of expecting to retrieve a specific answer for a particular question, feel motivated to learn, investigate or even be surprised by new information that somehow match their current information interests but in a much broader sense than traditional search tools. Those vague but always present concepts that drive and delimit the exploratory task could be represented as the *Core* of the knowledge being browsed, while the *Crust* contains those other entities that are related with the *Core* under certain reasons and particular similarity metrics and can interest the users.

**Offering Semantic Annotations to Consumers**. The news stories prototypes shown in Section 8 open room for innovative ways of interaction and information possibilities over the content being consumed, but are still far from being perfect. Many efforts are needed in order to determine what to display (just URL's to be followed, small excerpt of entities, more elaborated diagrams grouping them together) and when to display it (before the facts are displayed, during the entities are mentioned, only when the viewer explicitly demands them). Some applications would want to explicitly provide the reasons why those entities are relevant for the news without letting the users to figure it out. And all this, without being too intrusive.

# Part III

# Résumé de Thèse

# Introduction

Le contenu multimédia est l'un des moyens les plus largement et naturellement utilisés pour consommer de l'information. Le contenu vidéo est désormais roi sur le Web, avec des centaines de milliers d'heures de contenu déversés tous les jours sur des plate-formes sociales telles que YouTube. Mais, est-il possible de représenter ce contenu vidéo et ses métadonnées de telle une manière qu'il soit plus facilement trouvable et ré-utilisable non seulement comme vidéo entière mais également à différents niveaux de granularité ? Est-ce que l'information contenue dans ces vidéos est suffisamment contextualisée et formalisée pour être consommée tant par des humains que des machines?

Dans cette thèse, nous étudions comment tirer partie au mieux des standards disponibles, tels que la spécification Media Fragments URI et de techniques d'annotation avancées pour faire du contenu multimédia un objet de première classe sur le Web. Nous préconisons l'utilisation des technologies sémantiques comme un moyen de permettre aux machines d'effectuer automatiquement ces tâches. Nous présentons plusieurs approches basées sur la recherche d'information et la représentation des connaissances pour combler l'écart entre les caractéristiques visuelles de bas niveau pouvant être obtenues par des techniques d'analyse traditionnelles, et les concepts de plus haut niveau qui correspondent aux besoin des utilisateurs. Aligner ces concepts à des vocabulaires normalisés pour un domaine donné permet aux machines d'interpréter et de raisonner sur les informations disponibles dans ces documents afin d'offrir des services innovants pour la navigation, l'enrichissement et l'interconnection des Media Fragments, et, au final, d'améliorer la faon dont l'information multimédia est consommée.

L'absence de contexte qu'un unique document multimédia peut fournir est un frein pour bien comprendre l'histoire qui est rapportée. Les articles de presse inter-nationales sont un bon exemple de ce phénomène. Par conséquent, il est nécessaire de révéler tous les aspects de l'histoire qui, même s'ils ne sont pas explicitement présents dans le document de base, sont essentiels pour capturer la trame de fond. Pour faire face à ce problème, nous proposons un modèle conceptuel innovant appelé News Semantic Snapshot (NSS) qui est conu pour rendre explicite le contexte de l'événement d'une dépêche d'actualité. Suite à un processus appelé Named Entity Expansion, nous interrogeons le Web pour apporter d'autres points de vue sur ce qui se passe autour de nous, des milliers d'articles de presse et de messages où nous pour-

rions potentiellement trouver des détails de l'histoire. Nous avons également proposé une approche innovante concentrique afin de mieux répérer les entités contextuelles en misant sur la dualité entre le "Core", qui contient des entités représentatives fréquemment mentionnées dans les documents connexes, et celles qui détiennent des relations sémantiques particulières avec ce "Core" et faonnent des couches autour de lui: le "Crust".

## 1.1 Le Paysage du Contenu Multimédia Sémantique

L'information disponible sur le Web devient de plus en plus annoté sémantiquement en utilisant différentes techniques qui offrent de nouvelles méthodes d'exploiter automatiquement les connaissances disponibles et de raisonner sur ces données dune faon plus précise. Certaines initiatives qui mettent en valeur ce phénomène sont Google Knowledge Graph, fondé en partie sur Freebase [1], ou DBpedia [2] , qui sont devenus des vocabulaires largement utilisés comme cadre de référence pour la publication et consommation des données structurées dans de nombreux domaines différents: e-commerce, la gestion de l'entreprise, les affaires gouvernementales, etc.

Il existe déjà des outils qui transforment automatiquement les documents sur le Web (du texte, dans la plupart des cas), et les annotent selon des vocabulaires largement adoptés. Certains d'entre eux, comme [3] sont en mesure de permettre d'identifier des mots clé dans des documents textuels et de les relier à des concepts disponibles sur des bases de connaissance comme DBpedia. Ces concepts souvent aussi appelés entités, sont déjà devenus des citoyens du Web qui d'autres documents annotés font aussi référence. Ils sont conceptualisés dans un plus grand nuage d'entités et reliés à d'autres ressources, en jetant les bases d'une nouvelle génération d'algorithmes qui parviennent à raisonner sur ces annotations sémantiques pour répliquer la faon dont les humains effectuent ces mêmes opérations.

Cependant la situation est beaucoup moins mature en ce qui concerne les documents multimédias. Pour appliquer des techniques similaires à ces documents on aurait besoin dutiliser des méthodes de transcription automatique de l'audio et d'analyse visuelle des images qui introduisent un niveau de complexité supérieur au flux de travail. Dans le domaine des nouvelles et actualités sur le Web par exemple, il existe déjà différents prototypes qui tentent d'offrir des histoires d'une manière plus innovante[4]. Malheureusement pour l'instant, la plupart d'entre eux sont basés sur des annotateurs humaines ou des initiatives de crowdsourcing. Pour mettre ces solutions en conformité avec la quantité du nouveau contenu qui est quotidiennement

---

[1] http://freebase.com
[2] http://wiki.dbpedia.org/
[3] http://nerd.eurecom.fr/
[4] https://storify.com/

publié, nous avons besoin d'approches non supervisées qui peuvent se rapprocher de celles utilises par les humains.

Enfin, les documents multimédia sur le Web sont en constante évolution pour être mieux recherchés, consommés et partagés, soit dans leur ensemble ou selon différents niveaux de granularité temporelle et spatiale grâce à des normes comme les Media Fragments URIs [5]. De cette faon nous pouvons faire référence à un fragment en particulier appartenant à une vidéo ou même souligner une région spatiale que nous voulons mettre en valeur. Des opérations multimédia comme l'hyperlinking, la recommandation, ou la navigation des histoires publiées peuvent être basés sur des fragments vidéo encore plus réutilisables et plus adaptées au contexte pour améliorer la faon dont les médias sur le Web sont consommés.

## 1.2  Problématiques de Recherche

Nous présentons ci-dessous les principaux défis qui ont été abordés dans cette thèse afin d'améliorer l'annotation et consommation du contenu multimédia sur le Web.

### 1.2.1  Méthodologie Commune pour la Représentation de Contenu Multimédia

Nous voulons promouvoir des vocabulaires et des normes largement adoptés qui rendent possible la consommation transparente de l'information audiovisuelle sur le Web. En respectant les bonnes pratiques de publication des données Linked Data, nous permettons aux différents agents de faire face à la grande quantité de données disponibles, et d'alléger le traitement de l'information pertinente en temps opportun. Nous avons donc besoin des vocabulaires pour représenter des documents multimédias, mais aussi pour représenter leur essence et matérialiser la connexion avec les ressources.

### 1.2.2  Génération Automatique des Annotations Multimédia

être capable de consommer une vidéo et saisir correctement l'information qui nous est offerte, en réunissant les différents morceaux de l'histoire racontée, et même parcourir la vaste quantité de connaissances disponibles sur le Web pour trouver des informations connexes, rend le processus d'annotation extrêmement difficile. Nous avons besoin des algorithmes autonomes capables de produire des annotations similaires d'une faon plus rapide et avec moins d'intervention humaine.

---

[5]http://www.w3.org/TR/2012/REC-media-frags-20120925/

### 1.2.3 Exploitation des Annotations Sémantiques Multimédia

Identifier correctement les concepts et les entités d'une vidéo est une tâche importante, mais être capable de tirer parti de ces entits pour déclencher d'autres opérations telles que la recommandation et l'interconnexion des ressources est également crucial. Il faut se concentrer sur:

- **Filtrage et Classement des Annotations.** Pendant toute la durée de la vidéo, une énorme quantité de concepts peuvent être décrits ou mentionnés dans le signal audio. Nous avons besoin d'un processus de filtrage des éléments peu importants pour ne garder que ceux que nous pourrons éventuellement utiliser plus tard.

- **Compléter des Médias avec plus de Contenu.** Les annotations les plus importants qui résument le contenu peuvent être utilisés pour trouver des documents multimédias qui parlent de questions similaires.

- **Recréer le Contexte des Actualités sur le Web.** Une étape très importante dans l'exploitation des annotations sémantiques appartenant à des Actualités sur le Web est d'être capable de recréer le contexte de l'histoire racontée.

### 1.2.4 Présentation du Contenu Multimédia aux Utilisateurs

Il est extrêmement difficile d'identifier les besoins des utilisateurs et comment leur fournir des solutions, afin de dévoiler pleinement le potentiel des annotations sémantiques générées.

## 1.3 Contributions de la Thèse

Pour tenir compte des défis dans l'annotation sémantique des médias et l'interprétation de actualités décrites dans la section précédente, cette thèse apporte les contributions suivantes:

### 1.3.1 Contributions sur la Représentation des Données Multimédia

- Conception de l'ontologie LinkedTV. Dans cette ontologie, le contenu multimédia en général et les émissions télévisées en particulier, peuvent être annotées pas seulement au niveau de l'ensemble du programme, mais aussi avec différents degrés de granularité grâce à l'utilisation de la spécification Media Fragment URI 1.0.

- Développement du service TV2RDF [6] pour générer des annotations selon l'ontologie LinkedTV. Ce service prend en entrée des formats qui ne sont pas fondés sur le Web, tels que TVAnytime [7], et les convertit en instances du modèle de LinkedTV.

- Liaison des Annotations Multimédia avec des Ressources sur le Web. Nous avons étendu le modèle LinkedTV de pour sérialiser les relations établies entre les annotations et les ressources sur le Web qui servent à mieux les décrire. Nous avons créé des services pour relier des fragments vidéo entre eux et avec du matériel en ligne comme des messages sur les différents réseaux sociaux, d'autres documents textuels, etc.

### 1.3.2  Contributions sur l'Exploitation des Annotations Sémantiques

- Enrichissement du Contenu Multimédia. Complémenter des vidéos linéaires dune faon automatique, en offrant des informations connexes grâce à par exemple, les flux audio, des pages Web, ainsi que d'autres vidéos pertinentes.

- Classification vidéo. En prenant comme entrée la répartition temporelle des entités nommées détectées dans une vidéo et leur type, nous avons développé des méthodes très prometteuses pour les classer selon des catégories proposées.

- Récapituler et Souligner des Fragments Multimédia. Nous proposons un ensemble de fragments multimédia automatiquement annotés et appelés Hot Spots, qui tentent de mettre en évidence les principales idées contenues dans la vidéo, ainsi que d'indiquer à l'utilisateur des fragments d'intêret pour son visionnage ou pour les partager.

- L'algorithme d'extension d'entités nommées (Named Entity Expansion) utilise un sous-ensemble d'entités obtenues sur les transcriptions de la vidéo initiale, afin d'extraire et d'analyser des ressources supplémentaires sur le Web où d'autres entités décrivant davantage le contexte du document original.

- Création de Hyperliens Multimodales. Dans cette thèse, nous étudions comment combiner des techniques textuelles avec les résultats de l'analyse visuelle de la vidéo pour améliorer la faon dont nous établissons des liens entre des fragments multimédia qui sont hypothétiquement similaires.

### 1.3.3  Contributions sur l'Annotation des Actualités sur le Web

- Expansion des Actualités Internationales sur le Web. Nous avons effectué des études préliminaires qui montrent comment les actualités peuvent bénéficier

---

[6]http://linkedtv.eurecom.fr/tv2rdf/api/
[7]https://tech.ebu.ch/tvascope

d'un processus capable d'apporter d'autres annotations qui ne sont pas explicitement présentées dans la vidéo, mais sont utiles pour comprendre la trame de fond.

- News Semantic Snapshot. Nous proposons un nouveau modèle appelé News Semantic Snapshot (NSS), afin de capturer de une manière explicite et dans une seule structure les connaissances sur le contexte d'une actualité vidéo, qui peux finalement aider le public à comprendre ce qui est décrit dans les nouvelles.

- Référence pour le NSS. Nous avons produit un ensemble d'annotations de référence ou Gold Standard qui reproduisent le NSS idéal en suivant une méthodologie exhaustive[8].

- Explorer la Multidimensionnalité de la Pertinence des Entités dans les Actualités Videos. Nous avons sondé comme la pertinence des entités est un concept vaste qui doit être basée sur des considerations tels que la fréquence de l'entité, la popularité, les opinions des experts, etc.

- Prouver la nature concentrique du News Semantic Snapshot. Les entités appartenant au Semantic Snapshot peuvent être disposées en suivant une stratégie de représentation plus complexe qui transforme le sac d'entités en une sphère concentrique, où les entités du $Core$ indiquent les principaux aspects de l'histoire, alors que des autres ($Crust$) deviennent pertinentes grâce a des relations très particulières entre ces entités et celles du $Core$.

### 1.3.4 Contributions sur la Consommation des Actualités Web

- Étude sur les phases de Consommation. Nous proposons une classification des phases de consommation qui met en évidence la faon dont les besoins informatifs des téléspectateurs varient au cours du temps en termes de spécificité et de diversité, et comment le NSS peut répondre efficacement à ces besoins.

- Développement des Prototypes. Nous avons travaillé sur la mise en uvre de différentes applications comme des prototypes deuxième écran en montrant des expériences actives et passives pour les utilisateurs de télévision consommant des annotations sémantiques.

---

[8]https://github.com/jluisred/NewsEntities/

# Vers un Web Sémantique et Multimédia

Les documents textuels sont présents sur le Web depuis le début de son existence: articles encyclopédiques, blogs, journaux, etc. peuplent la plupart des connaissances disponibles en ligne. Ces documents sont liés entre eux et constamment annotés afin d'améliorer la faon dont ils sont recherchés et exploités par les humains et les machines. Malheureusement, les documents multimédias sont encore en publies d'une manière beaucoup plus primitive: sous forme d'unités indivisibles ou intégrées à l'intérieur des documents textuels, donc qu'ils ne sont pas facilement adressables ou annotées, ni connectés avec d'autres ressources. L'objectif de ce chapitre est de rendre les documents média citoyens de première classe du Web.

## 2.1 Une Ontologie Multimédia

Nous avons examiné l'état de l'art dans l'annotation multimédia dans des scénarios traditionnels, comme l'industrie télévisuelle, pour établir les exigences d'un modèle de connaissances capable de standardiser la publication des documents multimédia sur le Web selon différents niveaux de granularité et de incorporer le résultat de différentes techniques visuelles et sémantiques. Nous présentons donc l'ontologie LinkedTV, qui a l'intention de satisfaire à les exigences mentionnées ci-dessus et d'illustrer quelques exemples où ce modèle est utilisé pour représenter le contenu télévisuel. Cette ontologie fait usage de plusieurs vocabulaires largement utilisés, définit des nouveaux éléments (classes et propriétés) seulement lorsque nécessaire pour assurer la réutilisabilité. Cette ontologie a été développé dans le cadre du projet LinkedTV [1], d'où son nom. L'ontologie de LinkedTV est disponible à http://data.linkedtv.eu/ontologies/core/.

Les vocabulaires suivants ont été choisis comme base pour construire l'ontologie LinkedTV: (i) l'ontologie BBC Programmes pour représenter les métadonnées des programmes télévisés (séries, épisodes, chaînes, catégories, sous-titres), (ii) Ontologie for Media Resources, pour représenter les propriétés générales et les ressources

---

[1] http://www.linkedtv.eu/

*ma:MediaFragment* et *ma:MediaResource*, (iii) l'ontologie Ninsuna pour décrire explicitement les limites temporelles des fragments médias, (iv) l'ontologie Open Annotation pour relier les résultats d'analyse avec des fragments médias, (v) l'ontologie de NERD pour sérialiser les entités nommées, (vi) LSCOM pour représenter les concepts visuels, (vii) PROV-O pour représenter des informations de provenance, (viii) FOAF pour représenter les personnes reconnus dans les images et vidéos, et d'autres comme LODE, DBpedia ontologie, et WordNet 3.0. Pour plus d'informations sur la faon dont ils sont reliés entre eux, veuillez vérifier la figure 2.1.
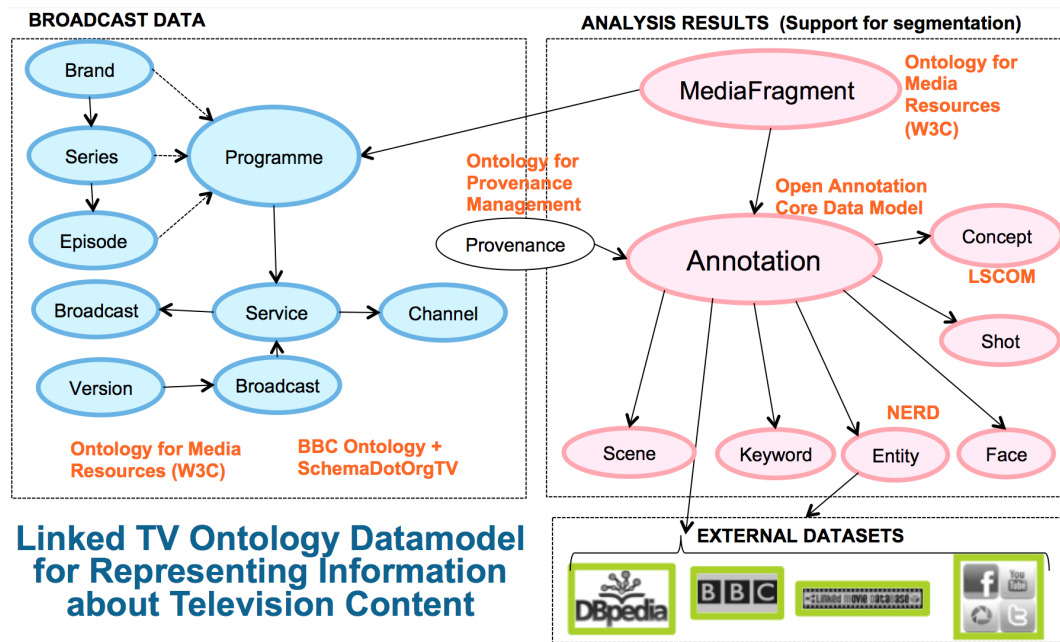


Figure 2.1: Le Modèle Ontologique LinkedTV

Afin de mieux mettre en valeur les possibilités du modèle, dans cette thèse nous analysons comment représenter un épisode du programme de télévision hollandais Tussen Kunst & Kitsch (Antiques Roadshow), qui est offert par le radiodiffuseur public AVRO [2], selon le modèle du téléviseur lié. En particulier, nous avons choisi une émission du 8 Décembre 2010[3]. Ce scénario a été conu dans le cadre du project LinkedTV, décrit à [97]. L'objectif général est d'étudier comment les téléspectateurs du programme Antiques Roadshow peuvent avoir leurs besoins d'information satisfaites même s'il s'agite d'un contenu complexe, avec des mentions fréquentes à des faits historiques et culturels, en leur fournissant des données provenant de sources externes telles que Europeana [4], les collections des musées, et différents ensembles de données encyclopédiques.

---

[2]http://web.avrotros.nl/tussenkunstenkitsch/
[3]http://web.avrotros.nl/tussenkunstenkitsch/player/8237850/
[4]http://www.europeana.eu/portal/

The legacy metadata for this program comes in the form of a spreadsheet. The automatic multimedia analysis results have been serialized in an EXMaRALDA[5] file and the subtitles of the show are available in SRT format[6]. All this metadata is converted in RDF using the LinkedTV ontology as depicted in the Figure 2.2.

Les métadonnées de ce programme se présentent sous la forme d'une feuille de calcul. Les résultats d'analyse automatique multimédia ont été sérialisé dans un fichier EXMaRALDA [7] , et les sous-titres de la série sont disponibles en format SRT [8]. Toutes ces métadonnées sont convertis en RDF en utilisant l'ontologie LinkedTV tel que c'est représenté dans la figure 2.2.

Afin de faciliter et de promouvoir l'adoption de ce modèle et toutes les principes du Web sémantique que cette ontologie promeut, nous avons développé l'outil de conversion *TV2RDF*[9]. Cet outil est offert sous la forme d'un Web API REST service qui sérialise les informations sur un contenu multimédia en RDF selon l'ontologie LinkedTV.

## 2.2    Génération des Annotations Vidéo

Les techniques d'analyse visuelle pour annoter des vidéos o ne sont pas nouvelles, mais malheureusement elles n'ont pas été utilisées pour travailler sur l'écosystème du Web. Par conséquent on ne profite pas des certaines caractéristiques souhaitables que ce scénario peut offrir, comme une meilleure interopérabilité grâce à des normes et vocabulaires communs, et la réutilisation de la vaste quantité de connaissances supplémentaires déjà disponible sur le Web. Dans cette thèse, nous introduisons des techniques du Web sémantique qui abordent ce processus d'annotation d'une manière plus automatique et autonome, en se fondant non seulement dans l'information qui peut être trouvé dans le contenu multimédia mais aussi d'autres connaissances disponibles sur le Web.

### 2.2.1    Annotations Textuelles

Nous décrivons des différentes techniques pour générer des annotations sémantiques provenant du texte. Même si le contenu se présente sous la forme de fichiers vidéo et audio, une partie considérable de l'information à l'intérieur de ces documents peut être collectée et exprimée sous forme écrite pour que les outils travaillant sur du texte soit aussi appliqué sur eux. Nous considérons deux dimensions vidéo: les

---

[5]http://www.exmaralda.org/en/tool/exmaralda/

[6]https://en.wikipedia.org/wiki/SubRip

[7]http://www.exmaralda.org/en/tool/exmaralda/

[8]https://en.wikipedia.org/wiki/SubRip

[9]http://linkedtv.eurecom.fr/tv2rdf/api/
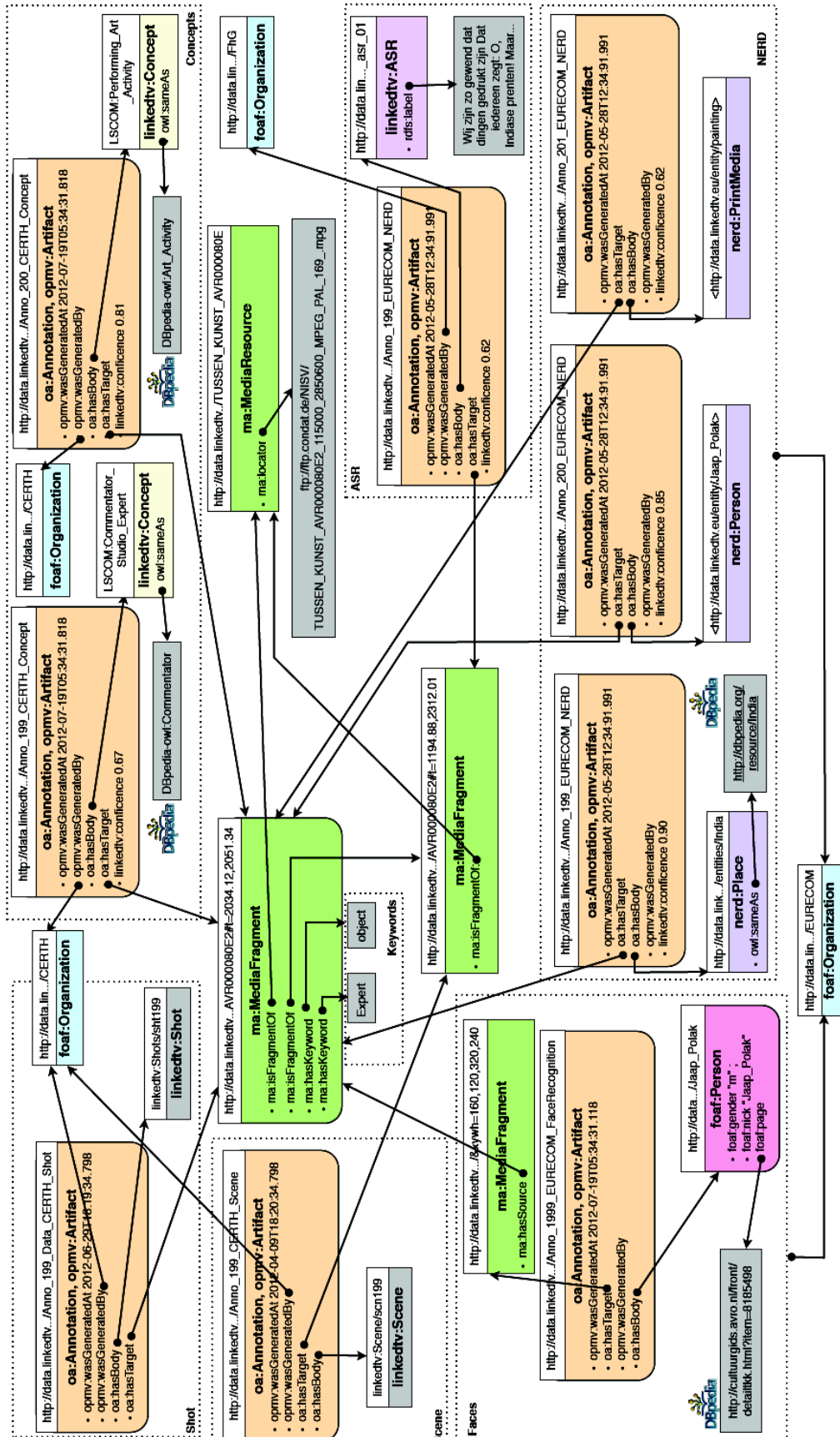
Figure 2.2: Instances impliquées dans le Sound & Vision Scénario

transcriptions audio, et le texte reconnu dans la vidéo par des techniques d'analyse visuelle.

### 2.2.1.1 Le Logiciel NERD

L'effort de recherche le plus remarquable réalisé au cours de cette période de doctorat sur la reconnaissance des entités nommées (NER) est l'exploitation et la personnalisation du logiciel NERD[10]. Ce logiciel est branché sur le dessus de divers extracteurs NER. Une partie de l'API a été examiné dans la version étendue de cette thèse. Son architecture suit les principes REST et comprend un HTML front-end pour les humains et une API pour les machines, en utilisant le standard JSON. Les deux interfaces sont alimentés par le moteur NERD Core REST. NERD a déjà été utilisé pour évaluer la qualité des résultats d'extraction provenant des différents extracteurs intégrés [149]. Dans [150] les auteurs ont offert des statistiques sur les mesures de précision pour chaque outil a fin de mettre en évidence les points forts et les points faibles de chaque solution.

### 2.2.1.2 Expansion des Entités Nommées

La deuxième technique à souligner appliquée sur l'information textuelle est l'algorithme d'extension d'entités nommées. Cet algorithme a été mis en uvre dans le cadre de cette thèse et a été appliqué sur différentes expériences présentées dans le chapitre 3.

D'un point de vue de l'exploitation, les techniques de NER ne résolvent pas certains problèmes qui sont présents lors qu'on essaie d'annoter une unité d'information: le contexte autour d'une histoire racontée est normalement plus large que ce qui est dit explicitement dans un document, de sorte que certaines informations importantes peuvent manquer. Dans cette thèse, nous présentons une approche qui étend l'ensemble des annotations initialement récupéré, a fin de soulager le manque de ressources textuelles qui détaillent l'événement en profondeur et qui permettent une application plus adéquate des techniques d'extraction sémantique. Nous étendons la description initiale d'un événement via des recherches sur Google et un regroupement d'entités afin de découvrir des ressources pertinentes qui sont ensuite filtrées en fonction du contexte. La logique se compose principalement de (1) la construction d'une requête approprié dans l'ensemble d'entités original, (2) la récupération de documents supplémentaires sur le même événement, et (3) l'analyse des documents trouvés pour fournir un ensemble plus complet d'entités, comme cela est illustré sur la figure 2.3.

Le résultat final de l'opération est une liste d'entités en même temps que leur score de pertinence. Ce mécanisme de classement est principalement basé sur la fréquence des entités dans les transcriptions et les documents connexes, Cette logique

---

[10]http://nerd.eurecom.fr/
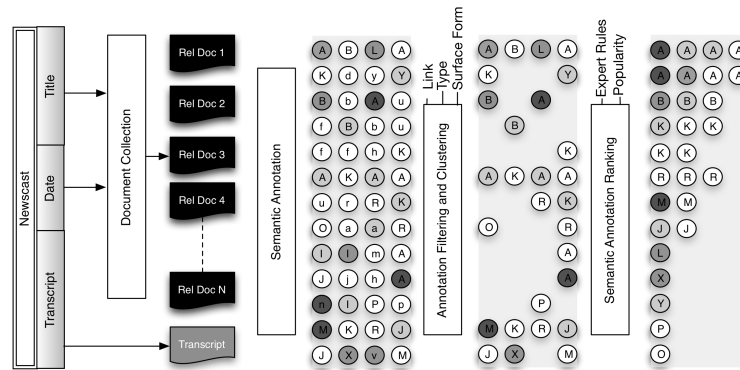
Figure 2.3: Schéma de l'Algorithme de Expansion des Entités Nommées.

est en outre étendu et amélioré au chapitre 3 avec heuristiques plus sophistiqués afin de promouvoir d'autres entités qui sont également très pertinentes pour l'histoire racontée, mais sont à peine mentionnés sur les documents connexes.

### 2.2.2 Annotations Visuelles: une Approche Multimodale

Nous avons également examiné des différentes techniques visuelles qui ont été étudiés dans la littérature depuis bien avant la philosophie Linked Data est apparue dans le scénario multimédia. Dans cette thèse, nous avons travaillé sur l'adéquation des annotations sémantiques multimédia à un nouveau scénario Web où elles peuvent être utilisés d'une faon universelle et exploitées à grande échelle. Cette expressivité sémantique plus avancée peut renforcer les résultats des différentes opérations déjà offertes sur le Web. Ci-dessous nous listons les techniques qui ont été examinés dans différentes expériences au cours de cette thèse:

- Détection Visuelle des Concepts. En particulier, nous avons suivi l'approche présentée dans [165], en utilisant un sous-ensemble de 10 détecteurs base.

- Détection des Changements de Plan. La segmentation temporelle des vidéos en plans est effectuée en utilisant l'algorithme proposé dans [194]. La précision de ces techniques sur de certain matériel LinkedTV a atteint une précision de 98,5% de [5].

- Détection des Scènes. Nous avons également examiné la segmentation en scènes, basé sur l'algorithme présenté dans [166].

- Reconnaissance Optique de Caractères (OCR). A partir du texte trouvé dans des photos et images ou vidéos, nous utilisons l'algorithme présenté à [179] pour extraire les caractères correspondants.

- Détection et Suivi de Visage. Pour cette tâche, nous utilisons la version C++ OpenCV du détecteur de visages "Viola et Jones" amélioré par Lienhart et Maydt [102]. La détection de visages est combinée avec un détecteur de couleur de peau [122] pour filtrer les détections qui ne sont pas des vraies visages.

- Reconnaissance Vocale Automatique. Nous comptons sur l'approche décrite à [178].

- Re-détection Rapide de Objects. Plus de détails sur notre approche pour re-détecter des objets peuvent être trouvés dans [4].

- Identification Localisée de Personnes. Dans [177], les auteurs soulignent le défi d'obtenir une base de données d'identification de la personne dans un contexte localisé.

- De l'Indices visuels à des Concepts Détectés. L'utilité de cette technique a été d'abord sondé par les bons résultats à MediaEval 2013 où ce processus a été décrite plus en détail.

## 2.3    Exploitation des Fragments Médias Annotés

Avoir le contenu annoté conformément aux vocabulaires largement utilisés rend plus facile de trouver d'autres ressources pertinentes à partir du Web, en permettant une nouvelle série de possibilités et un raisonnement plus humain et précis sur les connaissances.

### 2.3.1    Enrichissement des Fragments Médias

Les fragments médias annotés avec différentes techniques précédemment décrites peuvent être associés à d'autres contenus multimédia qui illustrent encore plus ce qui ils racontent. La nature du contenu étant attaché à les fragments peut être diverse: des publications sur les réseaux sociaux, billets de blogs, des articles dans des journaux, etc.

#### 2.3.1.1    Enrichissement avec des Réseaux sociaux

Dans cette thèse, nous avons développé une approche pour récupérer des éléments multimédias pertinentes aux événements racontées dans la video, qui sont publiées par les utilisateurs sur des plusieurs réseaux sociaux. Ce travail a été publié dans [116], et sa logique est disponible sous la forme d'un service REST appelé **MediaCollector**, qui est composé de différents extracteurs sur des réseaux comme Twitter, Facebook ou Instagram. Prenant comme entrée un terme de recherche, *par ex.,* "Obama" ce service effectue une recherche parallèle dans tous les réseaux sociaux. MediaCollector a

été initialement développé par Thomas Steiner à https://github.com/tomayac/media-server et révisé deux fois pour des fins d'expérimentation du projet LinkedTV et cette thèse à https://github.com/vuknje/media-server (afin de permettre des enquêtes temporelles dans l'ensemble des plates-formes de médias sociaux) et à https://github.com/MathildeS/media-server.

### 2.3.1.2   Enrichissement de Contenu Télévisuel avec TVEnricher

Le contenu télévisée structurée selon l'ontologie LinkedTV comme on explique dans la section 2.1, est enrichi avec d'autres ressources média extraites de plates-formes externes, soit en provenance d'une liste prédéfinie des sites Web ou des réseaux sociaux. Cette logique a été publiée comme service Web sous le nom *TVEnricher*[11], développé dans le cadre de cette thèse. TVEnricher détecte des ancres appropriées dans les ressources médias pour initier l'enrichissement du contenu, en fonction de la fréquence des entités nommées détectés dans les fragments. Les entités nommées deviennent les termes de la requête qui est issue dans Media Collector[12].

### 2.3.1.3   Enrichissement de Fragments Média avec Actualités Web

Cette approche d'enrichissement est plus ciblée à un type particulier de contenu (dans ce cas, les actualités) et il est organisé autour de différentes dimensions d'enrichissement identifiées par des experts dans le domaine (par exemple, trouver les mêmes faits dans d'autres sources journalistiques, des articles d'opinion, et d'autres catégories provenant d'études d'usagers [140]). Nous nous assurons un enrichissement plus personnalisée par deux mécanismes: (1) en utilisant des types particuliers d'entités nommées dans la vidéo pour déclencher des recherches pertinentes, (2) en effectuant des recherches sur des listes blanches prédéfinies de sources Web couvrant les dimensions d'enrichissement spécifiées par des études d'usagers.

L'entrée du service d'enrichissement est une liste des entités générées par les annotations du vidéo original, une date indiquant la période de temps où les faits étaient pertinents, et l'une des dimensions précédemment définis par les experts. La sortie est un ensemble de documents Web classés par pertinence et prêts à être exploités par les différentes applications qui aident les téléspectateurs à consommer le contenu. Afin d'indexer et d'exposer les fonctionnalités de requête nous nous sommes appuyés sur les outils de recherche fournis par Google Custom Search Engine via leur[13] (CSE ). Le processus d'enrichissement Avec des nouvelles élément est décrit dans le document [144]. La logique a-été publiée sur le service REST TVNewsEnricher disponible à: http://linkedtv.eurecom.fr/newsenricher/api/.

---

[11]http://linkedtv.eurecom.fr/tvenricher/api/
[12]http://linkedtv.eurecom.fr/api/mediacollector/
[13]https://cse.google.com/cse/

### 2.3.2   Raffinage et Promotion des Fragments Média

Nous présentons un ensemble de techniques qui utilisent les informations attachées aux fragments de médias annotées afin de renforcer ou de re-organiser leurs limites temporelles ou identifier des certaines parties de la vidéo qui semblent plus important et doivent donc être prioritaires ou mis en évidence dans la consommation du contenu.

#### 2.3.2.1   Utilisation d'Annotations pour redéfinir les Contours des Fragments

Les plans visuels détectés via différentes techniques d'analyse sont parfois trop artificielles et dénué de sens d'un point de vue de l'utilisateur pour être considéré comme l'unité à utiliser sur les successives étapes du flux de travail. Dans [6] nous présentons notre approche pour, à partir de les fragments visuels détectés dans le matériel vidéo et en se fondant sur certains techniques d'annotation des textes, identifier les limites des segments temporels correspondant à des sujets importants dans le contenu.

#### 2.3.2.2   Promoting Key Fragments: Hotspots

Dans cette section, nous présentons une approche qui tire parti des techniques d'analyse visuelle et les connaissances présentes sur le Web pour identifier des fragments pertinents (appelés Hot Spots) à l'intérieur des vidéos éducatifs en ligne, afin d'obtenir un aperu de ce qui est dit dans la video, et promouvoir la consommation des clips média à un niveau de granularité plus fin. Notre approche effectue une première segmentation en combinant les caractéristiques visuelles et les unités sémantiques dans les transcriptions (paragraphes). Les fragments vidéo résultants sont sémantiquement annotés par extraction des entités et la détection des sujets. En identifiant des chapitres consécutifs qui parlent de sujets et entités similaires, nous fusionnons les segments initiaux en unités plus grandes et sémantiquement indépendantes. Finalement nous les classons, filtrons les candidats le plus bas marqués, et proposons un résumé qui illustre les fragment dans une interface graphique Web. L'algorithme a été appliqué sur un ensemble de videos éducatives TED. Une démo en ligne en montrant la solution proposée est disponible à http://linkedtv.eurecom.fr/Hyperted/, et un document expliquant la démarche mise en uvre est disponible à [146].

### 2.3.3   Classification Vidéo en Utilisant des Annotations Média

Nous avons conu une expérience empirique qui tient compte le nombre d'entités nommées extraites des sous-titres vidéo, leurs types et leurs références temporelles pour classer 805 Dailymotion vidéos en 9 catégories de contenu. Avec ces expériences, nous avons abordé deux questions de recherche: *(i)* est-ce qu'il y a une corrélation entre la répartition temporelle des entités nommées et leurs types NERD, et les

catégories qu'ils appartiennent? *(ii)* quel algorithme de machine learning peut mieux trouver ces corrélations afin de prédire ces catégories? Même les scores obtenus sont loin d'être idéals et inférieurs à ce qu'est déjà publié dans la littérature, il est possible de déduire que le nombre d'entités nommées et leur type est un indicateur qui doit être pris en compte pour améliorer les algorithmes de classification vidéo, en supposant qu'il ya un nombre suffisant d'entités nommées détectés pour chaque type NERD. Le travail a été publié à [99].

## 2.4 Hyperliens Multimodales entre Fragments Médias

La gestion du contenu multimédia et des opérations avancées sont vitals pour une consommation approprié. Comme on peut être dépassés par l'énorme quantité d'informations disponibles étant créés chaque jour, les fonctionnalités de recherche des contenus pertinents sont cruciaux. Cette navigation avancée d'une vidéo sur l'autre à différents niveaux de granularité est similaire à l'activité de navigation des utilisateurs qui suivent des hyperliens HTML classiques pour se déplacer d'un document textuel à l'autre. Dans la version étendue de ce manuscrit, nous décrivons nos efforts en hyperlinking une collection fermée d'éléments vidéo au niveau des fragments, en utilisant à la fois dans les techniques sémantiques appliquées sur des informations textuelles et les annotations sur la base visuels attachés à des morceaux de vidéo, dans ce que nous appelons une approche de hyperliens multimodales. Cette multimodalité a amélioré les résultats par rapport à d'autres algorithmes qui tiennent en compte seulement un ensemble réduit des annotation média possibles.

Grâce à la participation à Mediaeval [41] et les successives campagnes, nous avons exploré la complémentarité entre les caractéristiques textuelles et visuelles: le texte a toujours été en mesure de donner des résultats intéressants, mais il manque de la spécificité de l'information visuelle, tandis que les caractéristiques visuelles exploitent cette partie plus intuitive mais elles ne sont pas suffisamment descriptives par eux-mêmes. Nous soutenons que c'est possible d'améliorer des opérations de récupération en combinant l'information textuelle et visuelle pour créer une requête multimodale comme déjà proposé auparavant dans la littérature [171]. L'originalité de notre travail réside dans le fait que nous étions prêtes à commencer d'une requête mixte (textuelle et visuelle). Par conséquent, nous essayons de surmonter le fossé sémantique en associant automatiquement le texte disponible à des concepts sémantiques. En ce qui concerne la tâche 2013 à Mediaeval, mesurée en MRR dans un intervalle temporel de 10 secondes, notre plus performante configuration a été choisi comme la meilleure parmi les candidats. Idem avec les expériences de hyperlinking où notre technique, en utilisant la segmentation en scènes, a obtenu le meilleur résultat, de ce fait révélant le grand potentiel de notre technique multimodal. Des résultats pour la campagne 2014 et 2015 ont également été signalés dans la version complète de cette thèse.

# Annotation Sémantique Avancée des Actualités

Parfois, les documents sont précédés ou suivis par d'autres extraits de l'information qui sont nécessaires pour bien comprendre le message original. Ces pièces autour influencent le sens du document initial et affinent ses effets, est ce que nous appelons le *contexte*. L'importance de ce concept a été bien noté par des travaux de recherche dans d'autres domaines, comme l'annotation multimédia et des technologies sémantiques [203] ou des techniques de recherche médicales [35] par exemple.

In general, the more complex an information unit is, the bigger the context we need to successfully consume it. In this regard, international news are deeply complicated by nature. News items in textual and audiovisual form come full of connotations, codes and conventions, related topics, implicit paradigms, unwritten referents, and very different ideologies [71]. In this Section we define a semantic structure called "Newscast Semantic Snapshot" (NSS) in order to explicitly represent the context of a particular news item. This knowledge abstraction intends to summarize into a single unit the necessary information for making sense of the news story being told, providing humans and machines with details that were not explicitly included in the original news document.

En général, plus complexe est l'unité d'information, plus grand est le contexte que nous avons besoin pour la consommer. à cet égard, les nouvelles internationales sont profondément compliquées par nature. Actualités en format texte ou audiovisuelles sont pleines de connotations, des codes et des conventions, des sujets liés, des paradigmes implicites, des référents non écrites, et influencées par très différentes idéologies [71]. Dans cette section, nous définissons une structure sémantique appelée " News Semantic Snapshot " (NSS) afin de représenter explicitement le contexte d'une actualité particulière. Cette abstraction de la connaissance a l'intention de résumer dans une seule unité les informations nécessaires pour donner un sens à l'histoire de nouvelles étant dit, fournir les humains et les machines avec des détails qui ne sont pas explicitement inclus dans le documents originales.

## 3.1 Le News Semantic Snapshot

Vu qu'un contexte qui porte l'ensemble des faits abordés dans des nouvelles est nécessaire, dans cette thèse nous définissons ce que nous avons appelé le "News Semantic Snapshot", une structure de connaissance sémantique ce qui représente le contexte d'une actualité internationale. La Figure 3.1 montre comment le $NSS$ est composé d'un ensemble d'entités sémantiques $e_i$ qui sont pertinentes pour les faits abordés et jouent un rôle dans l'histoire racontée.
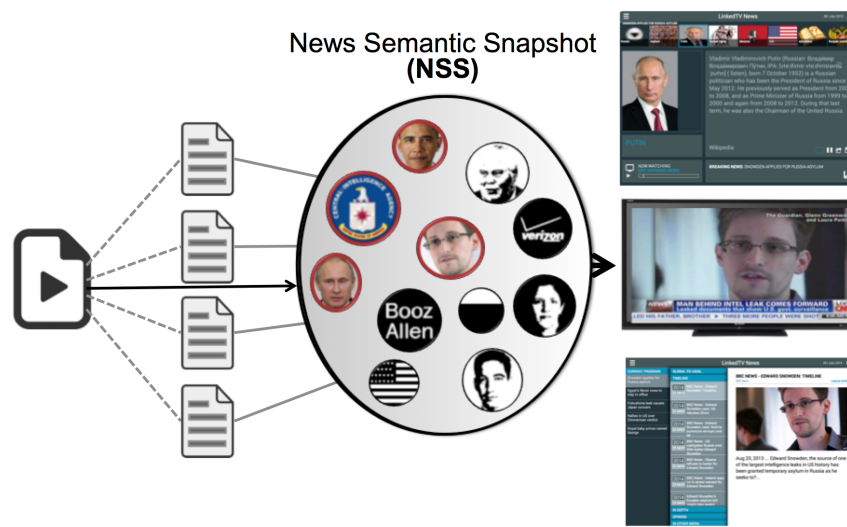


Figure 3.1: The News Semantic Snapshot (NSS): un ensemble d'entités pertinentes décrivant l'ensemble de l'histoire racontée, qui ne peut être généré à partir d'un seul témoignage

The NSS has as primary objective: to be able to serve machines and humans in advanced operations over the news items. Viewers can better consume the story being told via applications leveraging on data contained in the NSS, while machines can exploit certain entities to trigger enrichments, propose recommendations, or link to other related documents that the users would like to explore.

Le NSS a comme objectif principal d'être en mesure de servir les machines et les humains dans des opérations avancées sur les nouvelles. Les téléspectateurs peuvent mieux consommer l'histoire racontée en utilisant des applications tirant parti des données contenues dans le NSS, alors que les machines peuvent exploiter certaines entités pour déclencher des enrichissements, proposer des recommandations, ou un lien vers d'autres documents connexes que les utilisateurs peuvent être intéressés à explorer.

### 3.1.1    Norme de Référence pour l'Evaluation des NSS

Au meilleur de notre connaissance, il n'y a pas des ensembles de données sur des actualités qui aient été annotées avec des entités appartenant à leur contexte, mais seulement celles qui sont explicitement mentionnées dans eux. Pour cette raison, nous avons créé notre propre ensemble de données, suivant la méthodologie décrite à https://github.com/jluisred/NewsEntities, où l'ensemble de données est disponible gratuitement. En particulier, nous avons choisi au hasard 5 actualités Web du site BBC One Minute World News[1]. Chaque nouvelle dure de 1 à 3 minutes. La sélection a porté sur un large éventail de sujets: la politique, les conflits armés, les événements environnementaux, les litiges juridiques et de nouvelles sociales. Les sous-titres des vidéos ne etaient pas disponibles; par conséquent, un membre de l'équipe les a transcrit manuellement. En regardant les différentes dimensions, nous avons obtenu des entités qui ont été marqués en fonction du nombre de participants qui ont noté une entité et le score que chaque participant a donné à cette entité .

### 3.1.2    Une Première Approche pour la Génération du NSS

Afin de recréer le contexte d'une actualité, nous avons appliqué l'algorithme d'extension d'entités nommées décrit dans la section 2.2.1.2. Cette technique d'annotation d'entités ne s'appuie que sur les métadonnées de la vidéo, mais aussi sur les analyses d'autres documents Web pertinents qui sont trouvées a partir d'entités reconnues dans les sous-titres. Notre hypothèse initiale indique que ce mécanisme est en mesure de trouver un ensemble d'entités qui, malgré quelqu'unes ne sont pas explicitement mentionnés dans la vidéo d'origine, elles sont importants pour bien comprendre le contexte de l'histoire décrite. Comme le processus introduit du bruit sur la forme d'entités non pertinentes étant amenées par la méthode d'expansion, nous effectuons une sélection sur la liste des candidats initiale. L'ensemble de ce procédure est divisé en deux étapes et représenté sur la figure 3.2.

Au meilleur de notre connaissance, il n'y a pas des ensembles de données d'actualités annotées avec des entités appartenant à leur contexte, et donc pas toujours mentionnés d'une manière explicite dans les documents originaux. Pour cette raison, nous avons créé notre propre ensemble de données, suivant la méthodologie décrite à https://github.com/jluisred/NewsEntities, où l'ensemble de données est disponible gratuitement. En particulier, nous avons choisi au hasard 5 bulletins de nouvelles du "BBC One Minute News" site[2]. Chaque bulletin de nouvelles dure de 1 à 3 minutes. La sélection a porté sur un large éventail de sujets: la politique, les conflits armés, les événements environnementaux, les litiges juridiques et les nouvelles de dimension sociale. Comme les sous-titres des vidéos ne sont pas disponibles; un membre

---

[1] http://www.bbc.com/news/video_and_audio/
[2] http://www.bbc.com/news/video_and_audio/

de l'équipe les a transcrit manuellement. En regardant les différentes dimensions, nous avons obtenu une liste d'entités qui ont été choisis en fonction du nombre de participants qui ont noté chaque entité et le score qu'ils ont leur donné.
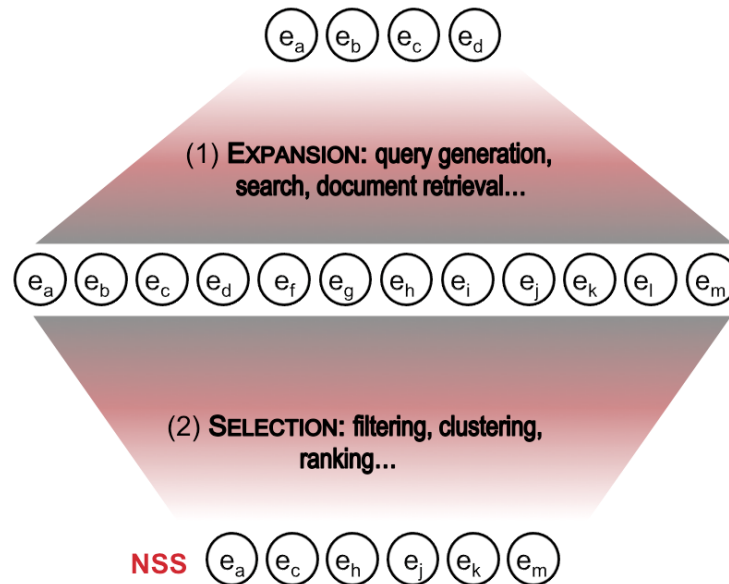


Figure 3.2: Approche de génération de NSS's: les résultats obtenus par extension d'entités nommées sont filtrés pour construire la liste des entités finalement présentes dans le NSS

Les résultats en termes de rappel, lorsque l'on compare avec l'ensemble de référence sont très prometteurs et montrent que le rappel moyen des entités obtenues directement par l'extraction d'entités nommées dans les sous-titres des vidéos $0, 42$ tandis que le rappel moyen en utilisant l'algorithme d'extension d'entités nommées va jusqu'à $0, 91$, ce qui prouve l'avantage de compter sur les documents connexes trouvés pendant l'étape d'extension.

### 3.1.3  Le service d'Extension des Actualités

Afin de rendre disponible l'algorithme d'extension d'entités nommées et permettre a des agents externes d'appliquer cette logique sur leur corpus, nous avons créé un service REST qui a l'intention de recréer le contexte d'une actualité international grâce a un processus avancé d'identification des entités nommées représentatives et pertinentes. Il est basé sur la même idée de récupérer et analyser des documents supplémentaires à partir du Web où les mêmes faits affichés sur la vidéo sont également décrits. Ce service a été créé dans le cadre du projet de LinkedTV et est disponible à http://linkedtv.eurecom.fr/entitycontext/api/.

## 3.2   La Multidimensionnalité de la Pertinence des Entités dans les Actualités

Les approches de classement fondées sur la fréquence des entités dans le documents originaux ont des limitations qui doivent être pris en compte. Prenons l'exemple d'un article de début de Juillet 2013, où Edward Snowden sollicite l'asile en Russie à l'aéroport de Moscou. Nous avons découvert comment certaines des entités provenant de la phase d'expansion, comme *Laura Poitras* ou *Glenn Greenwald* sont à peine mentionnés dans les documents connexes, et donc ne sont pas sélectionnés par des fonctions travaillant avec la fréquence des entités nommées, mais par contre elles sont vraiment importantes pour comprendre l'événement en cours. Afin de remédier ce problème, dans cette section nous montrons comment formaliser davantage les paramètres impliqués dans la phase de collection, afin d'affiner le processus de récupération de documents, et nous mettons en uvre d'autres fonctions de classement $F_x(e_i)$ travaillant dans des dimensions différentes autres que la fréquence, afin de pouvoir repérer les entités qui restent pertinentes à l'histoire sans être hautement répétée sur le texte des documents connexes. Les résultats des expériences montrent la robustesse de l'approche normalisée tenant un gain net cumulé moyen de 69,6%. Cette recherche a été publiée à [145].

### 3.2.1   Approche d'Extension d'Entités Révisée

Nous nous sommes appuyés sur l'API de Google Custom Search Engine (CSE)[3] en lanant une requête avec les paramètres spécifiés par $q = [\text{h}, \text{t}]$. En ce qui concerne les sites Web pour être analysées, Google permet de spécifier une liste de domaines Web et sous-domaines où les documents peuvent être récupérés. Sur cette base, dans notre étude, nous avons examiné cinq valeurs possibles pour ce paramètre: [**Google**]: la recherche se fait sur l'ensemble des pages Web indexées par Google, [**L1**] composée par l'ensemble des 10 journaux internationaux[4], [**L2**] contenant un ensemble de 3 journaux internationaux utilisés dans la création de la norme de référence, et [**L1 + Google**] et [**L2 + Google**] ou on donne la priorité au contenu dans la liste blanche, mais on considère encore d'autres sites Web qui Google indexe.

Comme l'intervalle temporel pour filtrer les documents, nous avons considéré une et deux semaines. Google CSE permet également de filtrer les résultats selon les annotations schema.org, donc nous avons également étudié l'influence de cette fonction. Enfin, nous nous sommes débarrassés de plusieurs entités nommées détectés dans la phase d'annotation en appliquant des fonctions de filtrage: filtre [F1] pour ne garder que les types NERD Person, Location, et Organization, filtre [F2] pour se

---

[3] https://www.google.com/cse/all
[4] http://en.wikipedia.org/wiki/List_of_newspapers_in_the_world_by_circulation

débarrasser des entités avec un score de confiance sous le premier quartier de la distribution, et le filtre [F3] pour garder les entités avec la première lettre en majuscule afin de considérer seulement les noms propres.

### 3.2.2 Stratégie de Classement Multidimensional

En ce qui concerne les différentes stratégies de classement, nous avons envisagé quatre méthodes différentes qui sont décrites ci-dessous.

**Fonction Basé sur la Fréquence**. Nous classons d'abord les entités en fonction de leur fréquence absolue dans l'ensemble des documents récupérés $D$.

**Fonction Gaussienne**. Le fonction $S_F$ privilèges les entités qui apparaissent le plus souvent. Cependant du point de vue d'un spectateur de télévision, ce n'est pas toujours le cas: les entités largement mentionnées sur ensemble de documents connexes sont parfois évidentes et assez bien connus pour le spectateur. Cette nouvelle fonction de score est donc approchée par une courbe de Gauss. Nous priorisons donc les entités réparties autour de la moyenne $\mu = \frac{|D|}{2}$, où $|D|$ est la cardinalité des documents récupérés.

**Fonction Basé sur la Popularité**. Nous vous proposons une fonction pondérée basée sur un mécanisme qui détecte les variations de popularité de l'entité sur une intervalle de temps (communément appelé "pics" de popularité) autour de la date de l'événement étudié. Si une entité a souvent été recherchée au cours de cette période de temps, nous pouvons déduire qu'elle est populaire et donc une bonne candidate pour faire partie du NSS. Nous avons compté sur Google Trends,[5] qui estime combien de fois un terme de recherche a été utilisé dans une fenêtre de temps donnée.

**Fonction Basé sur des Règles d'Experts**. La connaissance des experts dans le domaine, comme des journalistes ou des éditeurs télévisés peut être matérialisée sous la forme de règles qui corrigent le résultat offert par les stratégies de classement susmentionnées. Par exemple, en prenant comme entrée les scores moyens déjà mentionnés à la section 3.1.1 et après les normaliser sur l'intervalle $[1, -1]$, les *organizations* ont obtenu un poids plus élevé($Op_{expert} = 0.95$), suivie des*persons* ($Op_{expert} = 0.76$), et des *locations* ($Op_{expert} = 0.48$).

### 3.2.3 Résultats Expérimentaux

Si on tiens en compte les différents paramètres pour chaque phase de l'approche $N_{runs_{Collection}} * Runs_{Filtrage} * Runs_{Classement}$), nous avons un total de $20 * 4 * 4 = 320$ configurations différentes qui ont été lancé et classés en fonction de le score $MNDCG_{10}$ obtenu comme résultat. Nous les avons comparés avec deux lignes de base: **Baseline 1** ou l'ancienne algorithme d'extension des entités [143], et **Baseline**

---

[5]https://www.google.com/trends

**2** ou classement en fonction du score TF-IDF. Les résultats sont présentés dans le tableau 3.1.

| Run | Collection | | | Filtrage | Fonctions | | | Résultat | | | |
|-----|------------|--------------|------------|----------|------|-----|-----|--------------|-----------|----------|----------|
| | Sources | $T_{Intervalle}$ | Schema.org | | Freq | Pop | Exp | $MNDCG_{10}$ | $MAP_{10}$ | $MP_{10}$ | $MR_{10}$ |
| Ex0 | L1+Google | 2W | | F3 | Freq | | ✓ | 0.698 | 0.93 | 0.68 | 0.35 |
| Ex1 | L2+Google | 2W | | F3 | Freq | | ✓ | 0.695 | 0.93 | 0.68 | 0.35 |
| Ex2 | L1+Google | 2W | ✓ | F1+F3 | Freq | | ✓ | 0.689 | 0.93 | 0.62 | 0.31 |
| Ex3 | L1 | 2W | ✓ | F3 | Freq | | ✓ | 0.681 | 0.9 | 0.64 | 0.35 |
| Ex4 | L2+Google | 2W | | F1+F3 | Freq | | ✓ | 0.679 | 0.92 | 0.7 | 0.36 |
| Ex5 | L1+Google | 2W | ✓ | F1+F3 | Freq | | ✓ | 0.67 | 0.91 | 0.62 | 0.31 |
| Ex6 | L1 | 2W | ✓ | F3 | Freq | ✓ | ✓ | 0.668 | 0.86 | 0.6 | 0.32 |
| Ex7 | L2+Google | 2W | | F3 | Freq | ✓ | ✓ | 0.659 | 0.85 | 0.56 | 0.29 |
| Ex8 | Google | 2W | | F3 | Freq | | ✓ | 0.654 | 0.88 | 0.66 | 0.34 |
| Ex9 | L1 | 2W | | F3 | Freq | | ✓ | 0.654 | 0.88 | 0.66 | 0.35 |
| Ex10 | Google | 2W | ✓ | F1+F3 | Freq | | ✓ | 0.653 | 0.9 | 0.62 | 0.31 |
| Ex11 | Google | 2W | | F3 | Freq | ✓ | ✓ | 0.653 | 0.81 | 0.56 | 0.29 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Ex78 | Google | 2W | ✓ | F1+F3 | Gaussian | | ✓ | 0.552 | 0.66 | 0.66 | 0.34 |
| Ex80 | L2+Google | 2W | ✓ | F1+F3 | Gaussian | | ✓ | 0.55 | 0.69 | 0.7 | 0.36 |
| Ex82 | L1 | 2W | ✓ | F3 | Gaussian | | ✓ | 0.549 | 0.68 | 0.64 | 0.33 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| BS2 | Google | 2W | | | Freq | | | 0.473 | 0.53 | 0.42 | 0.22 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| BS1 | Google | 2W | | | TFIDF | | | 0.063 | 0.08 | 0.06 | 0.03 |

Table 3.1: Configurations exécutées et leurs paramètres, selon $MNDCG_{10}$

Notre meilleure configuration a obtenu un $MNDCG_{10}$ de 0,698 et un $MAP_{10}$ de 0,91, qui sont raisonnablement bien dans le domaine de recherche de documents. Le flux de travail conu est plus performant que BS1 et de loin mieux que BS2. Le plus grande intervalle $T_{Intervalle}$ de 2 semaines donne de meilleurs résultats dans tous les cas, alors que l'utilisation de Schema.org semble être bénéfique dans une certaine mesure que nous attendons à augmenter bientôt avec une plus large adoption de ce vocabulaire par les différents sites Web. La meilleure stratégie de filtrage est F3. La fonction basé sur la fréquence absolue est plus performante que la gaussien dans tous les cas, et la fonction basé sur des règles d'experts améliore le score final pour presque toutes les configurations possibles, en soulignant l'importance de tenir compte des dimensions de pertinence alternatives. La fonction de la popularité fait sa première apparition dans la position 6, mais une évaluation manuelle a suggéré que cette méthode de classement a un énorme potentiel pour l'avenir, car elle apporte des entités pertinentes qui ont été initialement pas pris en compte dans la norme de référence en raison des limitations humaines dans la méthodologie appliquée.

## 3.3 La Nature Concentrique du News Semantic Snapshot

Après avoir effectué une évaluation expérimentale du paradigme multidimensionnel pour générer le NSS, nous avons observé la faon dont les différentes fonctions sont compliqués à combiner ensemble et continuent a négliger les relations sémantiques intrinsèques que les entités tiennent. Par conséquent, dans cette section, nous refondons le problème de la génération du NSS en exploitant et en harmonisant dans

un modèle unique les différentes relations sémantiques établies entre les entités appartenant des actualités internationales. Nous proposons une approche concentrique à base de deux couches appelées **Core** et **Crust**. Ce modèle de représentation des connaissances prend en charge les relations complexes établies entre les entités impliquées dans les nouvelles et permet de formaliser la distinction entre le entités **représentatifs**, qui vont mieux pour caractériser l'essence même de l'histoire, et les **pertinents** qui sont intéressants pour différentes raisons qui les relient d'une faon ou d'une autre à le *Core*. Les résultats de nos expériences montrent que notre approche converge plus rapidement vers un NSS plus compact et idéal avec une amélioration de 36,9% par rapport à la ligne de base. Cette recherche a été publiée à [58].

### 3.3.1 Vers un Modèle Concentrique pour la Generation du NSS

Pour résumer, le News Semantic Snapshot d'une actualité (NSS) peut être modélisé en suivant un schéma de couches d'entités concentriques. Ce type de représentation permet de mieux reproduire le contexte de l'événement raconté et de faciliter la tâche d'identifier les entités pertinentes.
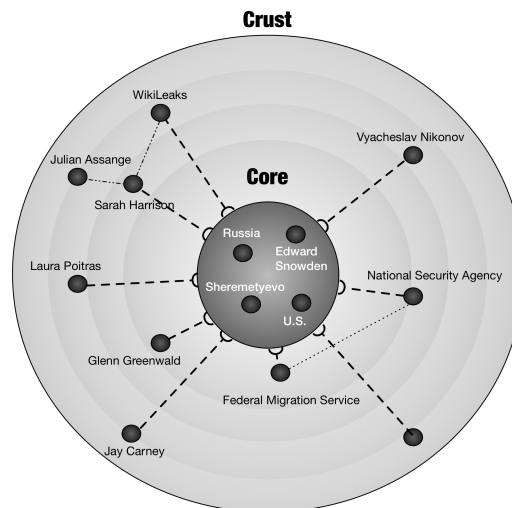


Figure 3.3: NSS for *"Le fugitive Edward Snowden demande l'asile en Russie"*

**Core:** Il est composé d'un petit nombre d'entités essentielles et tres représentatives qui peuvent mieux résumer l'événement raconté. Ils sont très fréquemment mentionnés dans les documents connexes et sont donc trouvables grâce à des fonctions basées sur la fréquence. Soit $E_{d_i} = e_{1_{d_i}}, ..., e_{n_{d_i}}$ le sac des entités appartenant au

Document connexe $d_i$, et $E_T$ l'union de tous les $E_{d_i}$:

$$Core = G\left(e_1, e_2, ...e_c, e_i\right), e_i \in E_{d_i} \tag{3.1}$$

$$(\text{Prominence en Fréquence})\forall e_i \in Core, f(e_i) > t \mid 0 \ll t < 1 \tag{3.1a}$$

$$(\text{Cohérence})\forall e_i, e_j \in Core, S(e_i, e_j) > s_{core} \mid 0 \ll s_{core} \leqslant 1 \tag{3.1b}$$

**Crust:** Il est composé des entités qui apportent les détails particuliers d'une actualité. Leur pertinence est plutôt fondée sur l'existence de relations spéciales (popularité, sérendipité, etc.) entre ces entités et le $Core$.

$$Crust = G\left(e_1, e_2, ...e_c\right), e_i \in E_{d_i} \tag{3.2}$$

$$(\text{Attaché au Core})\forall e_i \in Crust, S_{(}e_i, Core) > s \mid 0 \leqslant s \ll 1 \tag{3.2a}$$

Dans l'approche actuelle, nous avons examiné le nombre de documents Web qui parlent simultanément sur une entité particulière $e_i$ et le $Core$. Cette fonction, notée comme $S_{Web}(e_i, Core)$, identifie les documents Web qui parlent d'une entité candidate et el $Core$ au même temps, tout en tenant compte le volume initial de documents qui l'entité et le $Core$ avaient séparément. Soit $E$ un ensemble d'entités et le fonction $hits_s(E)$ le nombre de documents Web où toutes les entités $e_i \in E$ sont mentionnés, $S_{Web}(e_i, Core)$ est:

$$S_{Web}(e_i, Core) = \frac{hits_s(Core + e_i)^2}{hits_s(Core) * hits_s(e_i)} \tag{3.3}$$

$S_{Web}(e_i, Core)$ a été configuré pour fonctionner sur une instance du moteur de recherche personnalisé Google (CSE), où$hits_s(e)$ correspond avec le nombre de documents récupérés. Finalement, ces deux couches peuvent être regroupées en une structure unique afin de construire le **News Semantic Snapshot (NSS)**.

## 3.3.2 Résultats expérimentaux: un NSS Plus Compact

Après l'exécution des stratégies avancées pour agrandir l'ensemble d'entités disponible et de les classer pour situer dans les premières positions les annotations sémantiques les plus pertinentes, nous construisons le modèle concentrique des actualités sur trois différentes étapes: générer le $Core$, le $Crust$ et le $NSS$. Plus de détails sur cette approche et les différents expérimentations sont rapportés dans la version étendue de ce manuscrit.

Afin de mesurer la pertinence de notre solution, nous définissons le concept de **compacité** $Com$ d'un ensemble d'entités $Res$, donné une certaine fonction $f$ et une valeur $v$:

$$Com\left(R, f, v\right) = |min(NSS \in Res)| \mid f(NSS) \geq v \tag{3.4}$$

Cette mesure permet d'indiquer si un ensemble d'entités peut produire un NSS concise, tout en gardant l'objectif de $f(NSS) > v$. L'hypothèse que nous testons à travers cette expérience indique que l'approche du modèle concentrique produit des NSS plus concises, et potentiellement plus faciles à consommer que d'autres approches combinant différentes dimensions de classement.

Nous avons examiné deux lignes de base correspondant aux techniques traditionnelles d'expansion des entités. Ils sont identifiés par le nom **BAS01** et **BAS02** respectivement. En outre, nous avons envisagé l'existence d'un système idéal capable de générer le même classement parfait disponible dans la norme de référence, afin de représenter l'algorithme ideal. Dans le tableau 3.2, nous avons rangé les résultats en termes de compacité $Com\,(Cmn, R^*, R^*(Exn))$ pour chacun des 6 configurations possibles, en divisant les scores par vidéo ($v_1$, $v_2$, $v_3$, $v_4$, $v_5$).

| Run | Expansion | | | | | $Com\,(R, f, v)$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Collection | Core | Crust | Fusion | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | Avg |
| IdealGT | | | - | - | - | 16 | 11 | 22 | 27 | 19 | **19** |
| Cm4 | Ex2 | | CoreA | $S_{Google}$ | Core_Crust | 21 | 9 | 41 | 44 | 45 | **32** |
| Cm5 | Ex2 | | CoreA | $S_{Google}$ | CrustBased | 20 | 14 | 41 | 44 | 45 | **32.8** |
| Cm6 | Ex2 | | CoreB | $S_{Google}$ | Core_Crust | 27 | 10 | 43 | 44 | 42 | **33.2** |
| Cm0 | Ex1 | | CoreA | $S_{Google}$ | Core_Crust | 22 | 13 | 42 | 43 | 47 | **33.4** |
| Cm1 | Ex1 | | CoreA | $S_{Google}$ | CrustBased | 21 | 16 | 42 | 43 | 47 | **33.8** |
| Cm7 | Ex2 | | CoreB | $S_{Google}$ | CrustBased | 27 | 13 | 43 | 44 | 42 | **33.8** |
| Cm2 | Ex1 | | CoreB | $S_{Google}$ | Core_Crust | 28 | 13 | 43 | 43 | 44 | **34.2** |
| Cm3 | Ex1 | | CoreB | $S_{Google}$ | CrustBased | 28 | 16 | 43 | 43 | 44 | **34.8** |
| BAS01 | L2+AllGoogle, 1W F3 Gaussian + EXP + POP | | - | - | - | 41 | 45 | 34 | 41 | 37 | **39.6** |
| BAS02 | L2+AllGoogle, 1W F3 Freq + EXP + POP | | - | - | - | 24 | 39 | 49 | 48 | 39 | **39.8** |

Table 3.2: Compacité du modèle concentrique versus compacité des lignes de base et de la norme de référence

Nous observons que la compacité moyenne de l'approche concentrique est déjà plus réduite que celles correspondant aux lignes de base $BAS01$ et $BAS02$ pour les 5 vidéos, ce qui prouve notre hypothèse de départ: $Com\,(BAS0X) > Com\,(\mathrm{CmY})$. De plus, si l'on compare le meilleur modèle concentrique exécuté $Cm4$ ($Com = 32$) avec la meilleure ligne de base $BAS01$ ($Com = 39,6$), et en ayant comme objectif idéal la compacité de $IdealGT$ ($Com = 19$), nous pouvons signaler une diminution en pourcentage de 36,9% par rapport à la meilleure ligne de base, se rapprochant ainsi à le NSS le plus compact possible.

## 3.4 Le NSS dans le Nouveau Paradigme de Consommation des Actualités

In a final research line we study how an adequate news story representation can effectively support the different phases of the news consumption process. We analyze how an innovative data model such as the News Semantic Snapshot (NSS) can capture the entire context of a news story and make it available for machines to be exploited.

This model can feed very different applications assisting the users before, during, and after the news story consumption. It formalizes a duality in the news annotations that distinguishes between *representative* entities and *relevant* entities, and considers different relevancy dimensions that are incorporated into the model in the form of concentric layers. Finally, we analyze the impact of this NSS on existing prototypes and how it can support advanced features that are expected to come in the near future. This research has been published in [57].

Dans une dernière ligne de recherche, nous étudions comment une représentation adéquate de l'histoire racontée peut soutenir efficacement les différentes phases du processus de consommation de nouvelles. Nous analysons comment un modèle de données innovant tel que le News Semantic Snapshot (NSS) peut capturer tout le contexte d'une actualité et de le rendre disponible pour les machines qui peuvent l'utiliser. Ce modèle peut alimenter des applications très différentes qui aident les utilisateurs avant, pendant, et après la consommation des nouvelles. Il formalise une dualité dans les annotations, qui distingue entre les entités *représentatifs* et les *pertinentes*, et considère les différentes dimensions de pertinence qui sont incorporées dans le modèle comme des couches concentriques. Finalement nous analysons l'impact de cette NSS sur des prototypes existants et comment il peut offrir des fonctionnalités avancées qui sont prévues dans un proche avenir. Cette recherche a été publiée dans [57].
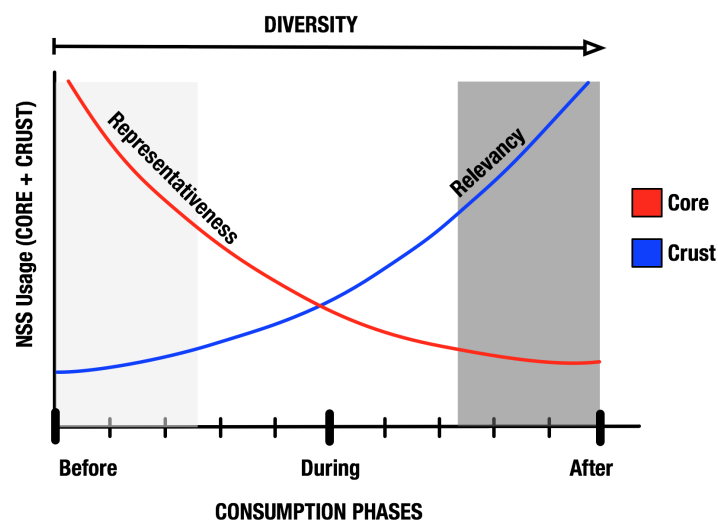


Figure 3.4: Utilisation du *Core* et *Crust* au cours des différentes phases de consommation

Figure 3.4 shows our hypothesis on how the duality between *Core* and *Crust* in the concentric model can better satisfy the evolution of user consumption needs across the different consumption phases, as a result of changes in aspects like viewers'

knowledge about the story, user's engagement, and content diversity.

La figure 3.4 montre notre hypothèse sur comment la dualité entre le *Core* et le *Crust* dans le modèle concentrique peut mieux satisfaire l'évolution des besoins de consommation de l'utilisateur à travers les différentes phases de la consommation, à la suite de changements dans des aspects comme la connaissance des téléspectateurs sur l'histoire, l'engagement de l'utilisateur ou la diversité des contenus.

# Achievements

Cette thèse a étudié comment les technologies sémantiques peuvent être efficacement appliquées au contenu multimédia afin de produire une nouvelle génération d'annotations vidéo qui (1) accordent aux documents Web l'importance qu'ils méritent, et (2) donnent un appui à l'exécution d'opérations plus avancées et humaines sur des fragments à l'intérieur du contenu multimédia, en particulier dans le domaine des actualités Web.

Les réalisations de la chapitre 2 sont le résultat des efforts déployés sur le développement de méthodologies valides et l'application des technologies sémantiques du Web sur le contenu multimédia. Ontologies comme **LinkedTV Ontology** réutilisent des vocabulaires dans le domaine et mettent en uvre les principes de Linked Data afin de rendre disponible le contenu multimédia sur le Web selon différents niveaux de granularité, ce qui les rend coexistants avec d'autres informations déjà présentes. Nous avons montré comment ce contenu multimédia et leurs fragments peuvent être annotées suivant les différentes techniques sémantiques comme **Named Entity Extraction** ou **Named Entity Expansion**. Les annotations sémantiques ouvrent la porte à une nouvelle série de possibilités: (1) elles permettent de se référer à des ressources déjà existantes sur le Web pour apporter automatiquement des connaissances supplémentaires, et (2) elles donnent un soutien à des techniques avancées de raisonnement et inférence qui tirent parti des contraintes sémantiques disponibles dans les ontologies susmentionnés, en jetant les bases d'une nouvelle génération d'applications pour la consommation, la recommandation et l'exploration du contenu multimédia au niveau du fragment.

Les réalisations obtenues dans la chapitre 3 sont dérivées de l'application des techniques susmentionnés au domaine des actualités internationales qui sont quotidiennement offertes dans les médias. Nous avons mis en évidence l'importance de reconstruire le contexte de ces nouvelles, matérialisée dans le **"Nouvelles Semantic Snapshot" (NSS)**. Nos différentes expériences ont révélé que nous pouvons reproduire le NSS de différents éléments vidéo, en nous les avons évaluées par rapport au **Gold Standard** d'entités dans des actualités Web. En particulier, la technique **News Entity Expansion** a été utilisée pour récupérer les entités absentes dans le contenu original. Le classement et les fonctions de filtrage qui considèrent différentes dimensions de pertinence dans le domaine nous aident a identifier les entités qui doivent faire partie du NSS et ignorer celles qui ne le font pas. Comme le nombre de

fonctions considérées peut augmenter avec les actualités plus complexes, la génération du NSS devient plus compliquée. à cet égard, nous avons proposé une approche **concentrique** pour reproduire le contexte des nouvelles qui vont au-delà des stratégies traditionnelles de recherche d'information en termes de **compacité** du NSS résultant, tout en restant agnostique aux dimensions de pertinence impliquées dans la phase de sélection. Enfin, nous avons prouvé la nécessité d'une représentation concentrique de l'histoire de nouvelles pour supporter des applications qui aident l'utilisateur dans les différentes phases de consommation: le **avant**, **pendant**, et **après**.

## 4.1 Perspectives

Outre l'évaluation des résultats obtenus par un travail de recherche particulier, un bon indicateur pour mesurer qualitativement l'impact d'une enquête est d'étudier le nombre de directions de recherche qui en ont découlées. Dans cette section, nous énumérons les différents domaines de recherche issues du travail accompli dans cette thèse, qu'il serait intéressant d'approfondir dans de futures initiatives

**Affiner les Modèles Ontologiques**. De nos jours les modèles de représentation de données sur le Web ont tendance à la simplicité. Les éditeurs de données et les développeurs d'applications exigent des modèles plus simples et légers qui permettent d'apporter les avantages de données sémantiques structurées sans introduire de complexité supplémentaire. Pour cette raison, et à la suite ce qui se passe déjà avec d'autres vocabulaires dans le Linked Open Cloud, les annotations multimédias présentés dans la section 2.1 ont besoin d'être raffinées pour obtenir un maximum de simplicité.

**Etude d'une Nouvelle Génération de Techniques de Annotation**. La performance des techniques d'annotation telles que l'extraction de entités nommées étudiées dans le section 2.2 peut être améliorée si, au lieu d'être agnostiques au domaine, elle prend en compte le scénario particulier du document analysé. à cet égard, les efforts déployés pour reproduire le contexte d'actualités peuvent être extrapolés à d'autres scénarios afin de régler plus précisément les différentes phases d'annotation. Nous pouvons également prévoir les avantages d'une application de la philosophie concentrique pour représenter l'information contextuelle dans d'autres tâches telles que la désambigusation sémantique, où les entités dans le *Core* peuvent fournir un premier ensemble de ressources candidates, tandis que les *Crust* apporte une sélection à grain fin des liens plus appropriés aux ressources en fonction de ce qui est expliqué dans le document original.

**Extension de la Norme de Référence NSS** Le niveau de détail utilisé pour annoter soigneusement le contexte des actualités qui font partie de la norme de référence a compromis la taille du corpus considéré. Des efforts ont déjà été fait pour agrandir

le corpus de 5 à 23 vidéos en utilisant une méthodologie plus relâchée, voir [1]. Une stratégie plausible serait d'appliquer le crowdsourcing pour construire la norme de référence. Cela peut augmenter de manière significative le nombre d'utilisateurs évaluant le NSS's des vidéos, et donc calculer d'une manière plus réaliste la pertinence des entités où les dimensions plus influentes [8].

**Techniques de Classement Alternatifs**. Les approches proposées dans cette thèse pour générer la NSS de news sont entièrement non dirigées. Si la norme de référence est agrandie, il serait également possible d'expérimenter avec des méthodes de Learning to Rank [23], afin de filtrer les entités collectées au cours du processus d'extension d'entités nommées. Dans la même ligne, la structure de connaissances générées par notre approche concentrique peut être un bon point de départ pour l'application des algorithmes de classement de graphes, comme [93], qui ont déjà fait preuve d'une bonne performance dans des tâches de recherche d'information similaires.

**Publier les Connaissances Générées**. Être en mesure de publier les annotations générées vers le Web serait adéquat pour les agents intéressés à sa consommation, pour améliorer ses capacités de recherche avancées, raisonner sur des histoires passés pour comprendre le présent, ou pour se former à mieux interpréter les nouveaux événements qui arrivent. Il est donc intéressant d'étudier comment les connaissances générées peuvent être correctement intégrées avec les informations déjà existantes dans les graphes comme DBpedia.

**Représentation Concentrique des Connaissances dans d'autres Tâches de Recherche d'Information**. Le modèle concentrique établit une technique de représentation qui peut être utilisée non seulement pour annoter de nouvelles histoires mais aussi pour d'autres scénarios de recherche d'information en décrivant des événements et histoires, comme les ressources éducatives et culturelles, etc. En outre, ce modèle peut aussi mieux mettre à exécution certains enquêtes sur le Web, en particulier concernant les moteurs exploratoires de recherche [107]. Ces concepts vagues mais nécessaires qui conduisent et délimitent la tâche d'exploration pourraient être représentés comme *Core* de la connaissance étant analysé, tandis que le *Crust* contient d'autres entités qui sont liées avec le *Core* en raison de certaines causes et considerations et peuvent intéresser les utilisateurs.

**Offrir des Annotations Sémantiques aux Consommateurs**. Les prototypes de consommation d'actualités Web offrent de nouvelles possibilités d'interaction et de compréhension du contenu multimédia, mais ils sont encore loin d'être parfait. Beaucoup d'efforts sont nécessaires afin de déterminer les éléments à afficher et quand les afficher, sans être trop intrusif.

---

[1]https://github.com/jluisred/NewsEntities/tree/master/dataset2014

# Bibliography

[1] E. Alfonseca and S. Manandhar. An unsupervised method for General Named Entity Recognition and Automated Concept Discovery. In *Proceedings of the First International Conference on General WordNet*, Mysore, India, 2002.

[2] James F. Allen and George Ferguson. Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4:531–579, 1994.

[3] Robin Aly, Maria Eskevich, Roeland Ordelman, and Gareth J. F. Jones. Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. *CoRR*, abs/1312.1913, 2013.

[4] E. Apostolidis, V. Mezaris, and I. Kompatsiaris. Fast object re-detection and localization in video for spatio-temporal fragment creation. In *Proc. MMIX Workshop at IEEE Int. Conf. on Multimedia and Expo (ICME)*, San Jose, CA, USA, July 2013.

[5] Evlampios Apostolidis and Vasileios Mezaris. Fast shot segmentation combining global and local visual descriptors. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6583–6587. IEEE, 2014.

[6] Evlampios Apostolidis, Vasileios Mezaris, Mathilde Sahuguet, Benoit Huet, Barbora Cervenková, Daniel Stein, Stefan Eickeler, José Luis Redondo Garcia, Raphaël Troncy, and Lukas Pikora. Automatic fine-grained hyperlinking of videos within a closed collection using scene segmentation. In *22$^{nd}$ ACM International Conference on Multimedia (ACMMM)*, Orlando, USA, 2014.

[7] Richard Arndt, Raphaël Troncy, Steffen Staab, Lynda Hardman, and Miroslav Vacura. COMM: Designing a Well-Founded Multimedia Ontology for the Web. In *6$^{th}$ International Semantic Web Conference (ISWC)*, 2007.

[8] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

[9] Masayuki Asahara and Yuji Matsumoto. Japanese Named Entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 8–15, Stroudsburg, PA, USA, 2003.

[10] Roberto Basili, Marco Cammisa, and Emanuale Donati. RitroveRAI: A Web Application for Semantic Indexing and Hyperlinking of Multimedia News. In *The Semantic Web ISWC 2005*, Lecture Notes in Computer Science, pages 97–111. Springer Berlin / Heidelberg, 2005.

[11] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008.

[12] Rachid Benmokhtar and Benoit Huet. An ontology-based evidential framework for video indexing using high-level multimodal fusion. *Multimedia Tools and Applications*, pages 1–27, 2011.

[13] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web: Scientific American. *Scientific American*, 2001.

[14] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201, Stroudsburg, PA, USA, 1997.

[15] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[16] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-lee. Linked data on the web (ldow2008. In *In WWW*, pages 1265–1266, 2008.

[17] 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, 2009.

[18] Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. NYU: Description of the MENE Named Entity System as Used in MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7*, 1998.

[19] Darin Brezeale and Diane J Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(3):416–430, 2008.

[20] Sergey Brin. Extracting Patterns and Relations from the World Wide Web. In *The World Wide Web and Databases*, volume 1590 of *Lecture Notes in Computer Science*, pages 172–183. Springer Berlin / Heidelberg, 1999.

[21] Razvan Bunescu, Marius Pasca, and Marius Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, Italy*, pages 9–16, 2006.

[22] Michael J. Cafarella, Doug Downey, Stephen Soderland, and Oren Etzioni. Knowitnow: Fast, scalable information extraction from the web. In *The Human Language Technology Conference (HLT-EMNLP-05*, pages 563–570, 2005.

[23] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 129–136, 2007.

[24] S-K Chang and Arding Hsu. Image information systems: where do we go from here? *IEEE transactions on Knowledge and Data Engineering*, 4(5):431–442, 1992.

[25] Shih-Fu Chang, R Manmatha, and T-S Chua. Combining text and audio-visual features in video indexing. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 5, pages v–1005. IEEE, 2005.

[26] Shu Chen, Maria Eskevich, Gareth J. F. Jones, and Noel E O'Connor. An investigation into feature effectiveness for multimedia hyperlinking. In *MMM14, 20th International Conference on MultiMedia Modeling*, pages 251–262, Dublin, Ireland, January 2014.

[27] Shruti Chhabra. Entity-centric Summarization: Generating Text Summaries for Graph Snippets. In *$23^{rd}$ ACM International Conference on World Wide Web (WWW)*, pages 33–38, Seoul, Korea, 2014.

[28] Philipp Cimiano and Johanna Völker. Text2Onto: a framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th international conference on Natural Language Processing and Information Systems*, pages 227–238, Berlin, Heidelberg, 2005.

[29] William W. Cohen. Automatically extracting features for concept learning from the web. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 159–166, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[30] William W. Cohen and Wei Fan. Learning page-independent heuristics for extracting data from web pages. In *In AAAI Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[31] Cédric Courtois and Evelien D'heer. Second Screen Applications and Tablet Users: Constellation, Awareness, Experience, and Interest. In *10<sup>th</sup> European Conference on Interactive Tv and Video (EuroITV)*, pages 153–156, Berlin, Germany, 2012.

[32] W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.

[33] Tom De Nies, Wesley De Neve, Erik Mannens, and Rik Van de Walle. Ghent university-iminds at mediaeval 2013: An unsupervised named entity-based similarity measure for search and hyperlinking. In *MediaEval*, 2013.

[34] Laurens De Vocht, Sam Coppens, Ruben Verborgh, Miel Vander Sande, Erik Mannens, and Rik Van de Walle. Discovering Meaningful Connections between Resources in the Web of Data. In *6<sup>th</sup> International Workshop on Linked Data on the Web (LDOW)*, 2013.

[35] Duy Dinh and Lynda Tamine. Combining global and local semantic contexts for improving biomedical information retrieval. In *Advances in Information Retrieval*, volume 6611, pages 375–386. Springer Berlin Heidelberg, 2011.

[36] Milan Dojchinovski and Tom Kliegr. Entityclassifier.eu: Real-Time Classification of Entities in Text with Wikipedia. In *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 654–658. Springer Berlin Heidelberg, 2013.

[37] Asmaa El Hannani and Thomas Hain. Automatic optimization of speech decoder parameters. *Signal Processing Letters, IEEE*, 17(1):95–98, 2010.

[38] Hariklia Eleftherohorinou, Vasiliki Zervaki, Anastasios Gounaris, Vasileios Papastathis, Yiannis Kompatsiaris, and Paola Hobson. Towards a common multimedia ontology framework. aceMedia Report, http://www.acemedia.org/aceMedia/files/multimedia_ontology/cfr/MM-Ontologies-Reqs-v1.3.pdf, Apr. 2006.

[39] D. W. Embley, C. Tao, and D. W. Liddle. Automatically extracting ontologically specified data from html tables of unknown structure. In *ER 2002*, pages 322–337, London, UK, 2002.

[40] M. Eskevich, R. Aly, D. Racca, R. Ordelman, S. Chen, and G.J.F. Jones. The Search and Hyperlinking Task at MediaEval 2014. In *MediaEval 2014 Workshop*, Barcelona, Spain, October 18-19 2014.

[41] M. Eskevich, J Gareth J.F., S. Chen, R. Aly, and R. Ordelman. The Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[42] M. Eskevich, G.J.F. Jones, C. Wartena, M. Larson, R. Aly, T. Verschoor, and R. Ordelman. Comparing retrieval effectiveness of alternative content segmentation methods for Internet video search. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pages 1–6, 2012.

[43] Maria Eskevich and Benoit Huet. EURECOM @ SAVA2015: Visual features for multimedia search. In *MEDIAEVAL 2015, Multimedia Benchmark Workshop*, 09 2015.

[44] Maria Eskevich, Walid Magdy, and Gareth J. F. Jones. New metrics for meaningful evaluation of informally structured speech retrieval. In *Proceedings of the 34th European conference on Advances in Information Retrieval*, pages 170–181, Berlin, Heidelberg, 2012. Springer-Verlag.

[45] Slim Essid, Marine Campedel, Gaël Richard, Tomas Piatrik, Rachid Benmokhtar, and Benoit Huet. *Machine learning techniques for multimedia analysis*. John Wiley & Sons, Ltd, 2011.

[46] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165(1):91 – 134, 2005.

[47] Richard Evans. A framework for named entity recognition in the open domain. In *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003*, 2003.

[48] Manaal Faruqui and Sebastian Padó. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany, 2010.

[49] Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[50] Norberto Fernandez, Jesus Arias Fisteus, Luis Sanchez, and Gonzalo Lopez. IdentityRank: Named entity disambiguation in the news domain. *Expert Systems with Applications*, 39(10):9207–9221, 2012.

[51] Roy T. Fielding and Richard N. Taylor. Principled design of the modern web architecture. *ACM Transaction Interneternet Technology*, 2:115–150, May 2002.

[52] Katja Filippova and Keith B Hall. Improved video categorization from text metadata and user comments. In *34$^{th}$ International ACM SIGIR Conference on Research and development in Information Retrieval*, pages 835–842, 2011.

[53] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for ComputationalLinguistics*, pages 363–370, Stroudsburg, PA, USA, 2005.

[54] M. Fuller, E. Tsagkias, E. Newman, J. Besser, M. Larson, G.J.F. Jones, and M. de Rijke. Using term clouds to represent segment-level semantic content of podcasts. In *2nd SIGIR Workshop on Searching Spontaneous Conversational Speech (SSCS 2008)*, Singapore, 07/2008 2008.

[55] Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. Structured and extended named entity evaluation in automatic speech transcriptions. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 518–526, Chiang Mai, Thailand, November 2011.

[56] Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. Sweetening Ontologies with DOLCE. In *13$^{th}$ International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 166–181, 2002.

[57] José Luis Redondo García, Giuseppe Rizzo, and Raphaël Troncy. Capturing news stories once, retelling a thousand ways. In *8$^{th}$ international Conference on Knowledge Capture (KCAP)*, 2015.

[58] José Luis Redondo García, Giuseppe Rizzo, and Raphaël Troncy. The Concentric Nature of News Semantic Snapshots. In *8$^{th}$ international Conference on Knowledge Capture (KCAP)*, 2015.

[59] Roberto Garcia and Oscar Celma. Semantic Integration and Retrieval of Multimedia Metadata. In *5$^{th}$ International Workshop on Knowledge Markup and Semantic Annotation*, pages 69–80, 2005.

[60] Roberto Garcia and Rosa Gil. Facilitating Business Interoperability from the Semantic Web. In *10$^{th}$ International Conference on Business Information Systems (BIS)*, pages 220–232, 2007.

[61] Roberto Garcia, Rosa Gil, and Jaime Delgado. A Web Ontologies Framework for Digital Rights Management. *Journal of Artificial Intelligence and Law*, 15:137–154, 2007.

[62] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1):89–108, 2002.

[63] Alfio Gliozzo, Claudio Giuliano, and Carlo Strapparava. Domain kernels for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 403–410, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[64] Ralph Grishman and Beth Sundheim. Message Understanding Conference-6: a brief history. In *16$^{th}$ International Conference on Computational linguistics (COLING'96)*, pages 466–471, Copenhagen, Denmark, 1996.

[65] Camille Guinaudeau, Guillaume Gravier, and Pascale Sbillot. Improving asr-based topic segmentation of tv programs with confidence measures and semantic relations. In *INTERSPEECH*, pages 1365–1368, 2010.

[66] Astrid Gynnild. Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets. *Journalism*, 15(6):713–730, 2014.

[67] Amirhossein Habibian, Koen E.A. van de Sande, and Cees G.M. Snoek. Recommendations for video event recognition using concept vocabularies. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, pages 89–96, Dallas, Texas, USA, April 2013.

[68] Thomas Hain, Asmaa El Hannani, Stuart N Wrigley, and Vincent Wan. Automatic speech recognition for scientific purposes-webasr. In *Interspeech*, pages 504–507, 2008.

[69] Abdelkader Hamadi, Georges Quénot, and Philippe Mulhem. Conceptual feedback for semantic multimedia indexing. In *CBMI13, the 11th International Workshop on Content-Based Multimedia Indexing*, pages 53–58, Veszprèm, Hungary, June 2013.

[70] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Proceedings of 4th Alvey Vision Conference*, pages 147–151, 1988.

[71] John Hartley. *Understanding news*. Routledge, 2013.

[72] A. Hauptmann, Rong Yan, Wei-Hao Lin, M. Christel, and Howard Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *Multimedia, IEEE Transactions on*, 9(5):958–966, 2007.

[73] M. Hausenblas, S. Boll, T. Brger, O. Celma, C. Halaschek-Wiener, E. Mannens, and R. Troncy. Multimedia vocabularies on the semantic web. W3c incubator group report, W3C Multimedia Semantics Incubator Group, 2007.

[74] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, pages 539–545, Stroudsburg, PA, USA, 1992.

[75] Marti A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, 1997.

[76] Sebastian Hellmann, Jens Lehmann, and Sören Auer. Towards an Ontology for Representing Strings. In *Proceedings of the EKAW 2012*, Lecture Notes in Artificial Intelligence (LNAI). Springer, 2012.

[77] Monika Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Query-free News Search. In *$12^{th}$ International Conference on World Wide Web (WWW)*, pages 1–10, Budapest, Hungary, 2003.

[78] Laura Hollink, Suzanne Little, and Jane Hunter. Evaluating the Application of Semantic Inferencing Rules to Image Annotation. In *$3^{rd}$ International Conference on Knowledge Capture (K-CAP)*, 2005.

[79] Chunneng Huang, Tianjun Fu, and Hsinchun Chen. Text-based video content classification for online video-sharing sites. *Journal of the American Society for Information Science and Technology*, 61(5):891–906, 2010.

[80] Jane Hunter. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *$1^{st}$ International Semantic Web Working Symposium (ISWC)*, pages 261–281, 2001.

[81] Jane Hunter. Combining the CIDOC/CRM and MPEG-7 to Describe Multimedia in Museums. In *Museums on the Web (MW)*, 2002.

[82] Jane Hunter. Enhancing the semantic interoperability of multimedia through a core ontology. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):49–58, 2003.

[83] Jane Hunter and Suzanne Little. A Framework to Enable the Semantic Inferencing and Querying of Multimedia Content. *International Journal of Web Engineering and Technology – Special Issue on the Semantic Web*, 2(2/3):264–286, 2005.

[84] Antoine Isaac and Raphaël Troncy. Designing and Using an Audio-Visual Description Core Ontology. In *Workshop on Core Ontologies in Ontology Engineering*, 2004.

[85] Juraj Kacur and Jan Korosi. An accuracy optimization of a dialog asr system utilizing evolutionary strategies. In *Proc. Image and Signal Processing and Analysis*, pages 180–184. IEEE, 2007.

[86] Saurabh S. Kataria, Krishnan S. Kumar, Rajeev R. Rastogi, Prithviraj Sen, and Srinivasan H. Sengamedu. Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1037–1045, New York, NY, USA, 2011.

[87] Polyxeni Katsiouli, Vassileios Tsetsos, and Stathes Hadjiefthymiades. Semantic video classification based on subtitles and domain terminologies. In *Workshop on Knowledge Acquisition from Multimedia Content (SAMT'07)*, 2007.

[88] Tomas Kliegr, Krishna Chandramouli, Jan Nemrava, Vojtech Svatek, and Ebroul Izquierdo. Combining image captions and visual analysis for image concept classification. In *Proceedings of the 9th International Workshop on Multimedia DataMining: held in conjunction with the ACM SIGKDD 2008*, pages 8–17, New York, NY, USA, 2008.

[89] Tomáš Kliegr. Entity classification by bag of Wikipedia articles. In *Proceedings of the 3rd workshop on Ph.D. students in information and knowledge management*, pages 67–74, New York, NY, USA, 2010.

[90] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466, New York, NY, USA, 2009.

[91] Carl Lagoze and Jane Hunter. The ABC Ontology and Model (v3.0). *Journal of Digital Information*, 2(2), 2001.

[92] Lori Lamel and Jean-Luc Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing (LNCS 5221)*, pages 4–15. Springer, 2008.

[93] Sangkeun Lee, Sang-il Song, Minsuk Kahng, Dongjoo Lee, and Sang-goo Lee. Random walk based entity ranking on graph for multidimensional recommendation. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, pages 93–100. ACM, 2011.

[94] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA, 1986.

[95] Sheldon Levine. How People Currently Share Pictures On Twitter, 2011. http://blog.sysomos.com/2011/06/02/.

[96] Mieke Leyssen, Lynda Hardman, and Jacco van Ossenbruggen. Specification of presentation interfaces for the three scenarios. Technical report, Television Linked To The Web (LinkedTV), 03 2012.

[97] Mieke Leyssen, Lynda Hardman, Jacco van Ossenbruggen, Lotte Baltussen, and Nico de Abreu. Specification of functionality requirements satisfying user information needs. Technical report, Television Linked To The Web (LinkedTV), 03 2012.

[98] Li-Jia Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *IEEE 11$^{th}$ International Conference on Computer Vision (ICCV'07)*, pages 1–8, 2007.

[99] Yunjia Li, Giuseppe Rizzo, José Luis Redondo Garcia, and Raphaël Troncy. Enriching media fragments with named entities for video classification. In *1$^{st}$ Worldwide Web Workshop on Linked Media (LIME)*, Rio de Janeiro, Brazil, 2013.

[100] Yunjia Li, Giuseppe Rizzo, Raphaël Troncy, Mike Wald, and Gary Wills. Creating enriched YouTube media fragments with NERD using timed-text. In *11$^{th}$ International Semantic Web Conference, Demo Session*, 2012.

[101] Yunjia Li, Giuseppe Rizzo, Raphaël Troncy, Mike Wald, and Gary Wills. Creating enriched YouTube media fragments With NERD using timed-text. In *11$^{th}$ International Semantic Web Conference (ISWC)*, 2012.

[102] R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference*, volume 1, pages I–900 – I–903 vol.1, 2002.

[103] Dekang Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[104] Berenike Litz, Hagen Langer, and Rainer Malaka. Sequential Supervised Learning for Hypernym Discovery from Wikipedia. In *Knowledge Discovery, Knowlege Engineering and Knowledge Management*, volume 128, pages 68–80, Berlin Heidelberg, 2011.

[105] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–110, November 2004.

[106] Brian Mak and Tom Ko. Automatic estimation of decoding parameters using large-margin iterative linear programming. In *Proc. Interspeech*, pages 1219–1222, 2009.

[107] Nicolas Marie, Fabien Gandon, Alain Giboin, and Emilie Palagi. Exploratory search on topics through different perspectives with DBpedia. In *SEMANTICS*, Leipzig, Germany, September 2014.

[108] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, pages 188–191, Stroudsburg, PA, USA, 2003.

[109] Olena Medelyan, Ian H. Witten, and David Milne. Topic Indexing with Wikipedia, 2008.

[110] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8, New York, NY, USA, 2011.

[111] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In $7^{th}$ *International Conference on Semantic Systems (I-Semantics)*, 2011.

[112] Rada Mihalcea. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 411–418, Stroudsburg, PA, USA, 2005.

[113] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA, 2007.

[114] Vuk Milicic, JoséLuis Redondo García, Giuseppe Rizzo, and Raphaël Troncy. Tracking and Analyzing the 2013 Italian Election. In $10^{th}$ *Extended Semantic Web Conference (ESWC), Demo Track*, pages 258–262, 2013.

[115] Vuk Milicic, Giuseppe Rizzo, José Luis Redondo García, and Raphaël Troncy. Grab your favorite video fragment: Interact with a Kinect and discover enriched hypervideo. In $11^{th}$ *European Interactive TV Conference, (EUROITV)*, Como, Italy, 2013.

[116] Vuk Milicic, Giuseppe Rizzo, José Luis Redondo García, Raphaël Troncy, and Thomas Steiner. Live topic generation from event streams. In *22$^{nd}$ International World Wide Web Conference, Demos Track (WWW)*, Rio de Janeiro, Brazil, 2013.

[117] George A. Miller. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41, November 1995.

[118] David Milne and Ian H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008.

[119] David Milne and Ian H. Witten. Learning to link with Wikipedia. In *17$^{th}$ ACM International Conference on Information and Knowledge Management (CIKM'08)*, pages 509–518, Napa Valley, California, USA, 2008.

[120] Joshua L Moore, Florian Steinke, and Volker Tresp. A novel metric for information retrieval in semantic networks. In *3$^{rd}$ International Workshop on Inductive Reasoning and Machine Learning for the Semantic Web (IRMLeS)*, pages 65–79, 2011.

[121] MPEG-7. Multimedia Content Description Interface. Standard No. ISO/IEC 15938, 2001.

[122] J. See N. A. Abdul Rahim, C. W. Kit. RGB-H-CbCr Skin Colour Model for Human Face Detection. In *MMU International Symposium on Information & Communications Technologies (M2USIC)*, Petaling Jaya, Malaysia, 2006.

[123] Frank Nack, Jacco van Ossenbruggen, and Lynda Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part II). *IEEE Multimedia*, 12(1), 2005.

[124] Frank Nack, Jacco van Ossenbruggen, and Lynda Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part II). *IEEE Multimedia*, 12(1), 2005.

[125] David Nadeau, Peter Turney, and Stan Matwin. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. In *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 266–277. Springer Berlin / Heidelberg, 2006.

[126] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[127] Roberto Navigli and Paola Velardi. Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1075–1086, 2005.

[128] Hwee Tou Ng and Hian Beng Lee. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 40–47, Stroudsburg, PA, USA, 1996.

[129] Usman Niaz, Bernard Merialdo, Claudiu Tanase, Maria Eskevich, and Benoit Huet. EURECOM at TrecVid 2015: Semantic indexing and video hyperlinking tasks. In *TRECVID 2015, 19th International Workshop on Video Retrieval Evaluation, 16-18 November 2015, Gaithersburg, MD, USA*, 10 2015.

[130] R. Nock and F. Nielsen. Statistical Region Merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26:1452–1458, November 2004.

[131] Jacco van Ossenbruggen, Frank Nack, and Lynda Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). *IEEE Multimedia*, 11(4), 2004.

[132] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F. Smeaton, and Georges Quénot. TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2012*, 2012.

[133] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F. Smeaton, and Georges Quenot. TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2012*, 2012.

[134] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab, 1999.

[135] Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, Alpa Jain, and Alpa Jain. Organizing and Searching the World Wide Web of Facts - Step One: The One-Million Fact Extraction Challenge. In *AAAI 2006*, 2006.

[136] Adam Pease, Ian Niles, and John Li. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, 2002.

[137] Ted Pedersen. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8, Stroudsburg, PA, USA, 2001.

[138] Fernando Pereira, Anthony Vetro, and Thomas Sikora. Multimedia retrieval and delivery: Essential metadata challenges and standards. *Proceedings of the IEEE*, 96:721–744, April 2008.

[139] Lilia Perez Romero, Rene Ahn, and Lynda Hardman. Linkedtv news: designing a second screen companion for web-enriched news broadcasts. Technical report, Technical Report, Eindhoven University of Technology, 2013.

[140] Lilia Perez Romero, Michiel Hildebrand, José Luis Redondo García, and Lynda Hardman. Linkedtv news: A dual mode second screen companion for web-enriched news broadcasts. In *ACM International Conference on Interactive Experiences for Television and Online Video (TVX)*, Newcastle, United Kingdom, 2014.

[141] Till Quack, Ullrich Mönich, Lars Thiele, and B. S. Manjunath. Cortina: a system for large-scale, content-based web image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 508–511, New York, NY, USA, 2004. ACM.

[142] José Luis Redondo-García, Vicente Boton-Fernandez, and Adolfo Lozano-Tello. Linked data methodologies for managing information about television content. *International Journal of Interactive Multimedia and Artificial Intelligence (IJI-MAI)*, 1(6):342–351, 09 2012.

[143] José Luis Redondo García, Laurens De Vocht, Raphaël Troncy, Erik Mannens, and Rik Van de Walle. Describing and Contextualizing Events in TV News Shows. In *2$^{nd}$ International Workshop on Social News on the Web (SNOW)*, pages 759–764, Seoul, Korea, 2014.

[144] José Luis Redondo García, Michiel Hildebrand, Lilia Perez Romero, and Raphaël Troncy. Augmenting TV Newscasts via Entity Expansion. In *11$^{th}$ Extended Semantic Web Conference (ESWC), Demo Track*, pages 472–476, 2014.

[145] José Luis Redondo García, Giuseppe Rizzo, Lilia Perez Romero, Michiel Hildebrand, and Raphaël Troncy. Generating the Semantic Snapshot of Newscasts using Entity Expansion. In *15$^{th}$ International Conference on Web Engineering (ICWE)*, 2015.

[146] José Luis Redondo Garcia, Mariela Sabatino, Pasquale Lisena, and Raphaël Troncy. Finding and Sharing Hot Spots in Web Videos. In *13<sup>th</sup> International Semantic Web Conference (ISWC), Demo Track*, 2014.

[147] José Luis Redondo García and Raphaël Troncy. Television meets the web: a multimedia hypervideo experience. In *12<sup>th</sup> International Semantic Web Conference, Doctoral Consortium Track (ISWC)*, Sydney, Australia, 2013.

[148] Giuseppe Rizzo, Thomas Steiner, Raphaël Troncy, Ruben Verborgh, and José Luis Redondo García. What Fresh Media Are You Looking for? Retrieving Media Items from Multiple Social Networks. In *1<sup>st</sup> International Workshop on Socially-aware Multimedia (SAM)*, pages 15–20, Nara, Japan, 2012.

[149] Giuseppe Rizzo and Raphaël Troncy. NERD: A Framework for Evaluating Named Entity Recognition Tools in the Web of Data. In *10<sup>th</sup> International Semantic Web Conference (ISWC'11), Demo Session*, pages 1–4, Bonn, Germany, 2011.

[150] Giuseppe Rizzo and Raphaël Troncy. NERD: Evaluating Named Entity Recognition Tools in the Web of Data. In *Workshop on Web Scale Knowledge Extraction (WEKEX'11)*, pages 1–16, Bonn, Germany, 2011.

[151] Giuseppe Rizzo and Raphaël Troncy. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *13<sup>th</sup> Conference of the European Chapter of the Association for computational Linguistics (EACL'12)*, 2012.

[152] Giuseppe Rizzo, Raphaël Troncy, Sebastian Hellmann, and Martin Bruemmer. NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud. In *5<sup>th</sup> International Workshop on Linked Data on the Web (LDOW'12)*, Lyon, France, 2012.

[153] Giuseppe Rizzo, Marieke van Erp, and Raphaël Troncy. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In *9<sup>th</sup> International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014.

[154] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[155] Stephen E. Robertson and Karen Sparck Jones. Document retrieval systems. *Journal of the American Society for Information Science*, pages 143–160, 1988.

[156] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estèv. LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *Proceedings of IWSLT 2011*, San Francisco, USA, 2011.

[157] Anthony Rousseau, Paul Deléglise, and Yannick Estève. Enhancing the tedlium corpus with selected data for language modeling and more ted talks. In *LREC 2014*, 2014.

[158] Bahjat Safadi, Mathilde Sahuguet, and Benoit Huet. When textual and visual information join forces for multimedia retrieval. In *ICMR 2014, ACM International Conference on Multimedia Retrieval, April 1-4, 2014, Glasgow, Scotland*, Glasgow, United Kindom, 04 2014.

[159] Mathilde Sahuguet, Benoit Huet, Barbora Cervenkova, Evlampios Apostolidis, Vasileios Mezaris, Daniel Stein, Stefan Eickeler, José Luis Redondo García, Raphaël Troncy, and Lukas Pikora. Linkedtv at mediaeval 2013 search and hyperlinking task. In *Multimedia Benchmark Workshop, (MEDIAEVAL)*, Barcelona, Spain, 2013.

[160] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, 1975.

[161] Leo Sauermann, Richard Cyganiak, and Max Völkel. Cool uris for the semantic web. Technical report, DFKI GmbH, Deutsches Forschungszentrum f ür Künstliche Intelligenz GmbH, 2007.

[162] Satoshi Sekine. NYU: Description of the Japanese NE system used for MET-2. In *Proceedings of Message Understanding Conference*, 1998.

[163] Satoshi Sekine, Chikashi Nobata, and Chikashi Nobata. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In *LREC*, 2004.

[164] Yusuke Shinyama and Satoshi Sekine. Named entity discovery using comparable news articles. In *Proceedings of the 20th international conference on Computational Linguistics*, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[165] Panagiotis Sidiropoulos, Vasileios Mezaris, and Ioannis Kompatsiaris. Enhancing Video concept detection with the use of tomographs. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2013.

[166] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso. Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21:1163–1177, 2011.

[167] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[168] John R. Smith and Shih-Fu Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia*, pages 87–98, New York, NY, USA, 1996.

[169] R. Smith. An Overview of the Tesseract OCR Engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, pages 629–633, Washington, DC, USA, 2007. IEEE Computer Society.

[170] C. G M Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *Multimedia, IEEE Transactions on*, 9(5):975–986, 2007.

[171] Cees GM Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications*, 25(1):5–35, 2005.

[172] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, pages 1297–1304, Cambridge, MA, 2005.

[173] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808, 2006.

[174] James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:3, March 1992.

[175] Valentin I. Spitkovsky and Angel X. Chang. A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of LREC 2012*, 2012.

[176] David McG. Squire, Wolfgang Müller, Henning Müller, and Jilali Raki. Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In *Pattern Recognition Letters*, pages 143–149. Elsevier, 1999.

[177] D. Stein, E. Apostolidis, V. Mezaris, N. de Abreu Pereira, J. Müller, M. Sahuguet, B. Huet, and I. Lašek. Enrichment of News Show Videos with

Multimodal Semi-Automatic Analysis. In *Proc. NEM-Summit 2012*, pages 1–6, Istanbul, Turkey, October 2012.

[178] D. Stein, J. Schwenninger, and M. Stadtschnitzer. Improved speed and quality for automatic speech recognition using simultaneous perturbation stochastic approximation. In *Proc. Interspeech*, pages 1–4, Lyon, France, 2013.

[179] Daniel Stein, Stefan Eickeler, Rolf Bardeli, Evlampios Apostolidis, Vasileios Mezaris, and Meinard Müller. Think Before You Link – Meeting Content Constraints when Linking Television to the Web. In *Proc. NEM Summit*, Nantes, France, October 2013.

[180] Daniel Stein, Alp Öktem, Evlampios Apostolidis, Vasileos Mezaris, José Luis Redondo García, Raphaël Troncy, Mathilde Sahuguet, and Benoît Huet. From raw data to semantically enriched hyperlinking: Recent advances in the linkedtv analysis workflow. In *NEM Summit 2013, Networked & Electronic Media*, Nantes, France, 2013.

[181] Thomas Steiner. A Meteoroid on Steroids: Ranking Media Items Stemming from Multiple Social Networks. In *$22^{nd}$ World Wide Web Conference (WWW'13), Demo Session*, Rio de Janeiro, Brazil, 2013.

[182] Thomas Steiner, Ruben Verborgh, Joaquim Gabarro Vallés, and Rik Van de Walle. Adding meaning to social network microposts via multiple named entity disambiguation apis and tracking their data provenance. *International Journal of Computer Information Systems and Industrial Management*, 5:69–78, 2013.

[183] Nicola Stokes, Joe Carthy, and Alan F. Smeaton. Select: A lexical cohesion based news story segmentation system, 2004.

[184] Partha Pratim Talukdar, Thorsten Brants, Mark Liberman, and Fernando Pereira. A context pattern induction method for named entity extraction. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 141–148, Stroudsburg, PA, USA, 2006.

[185] Franck Thollard and Georges Quénot. Content-based re-ranking of text-based image search results. In *ECIR13,35th European Conference on IR Research*, pages 618–629, Moscow, Russia, March 2013.

[186] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[187] Mai-Vu Tran, Tien-Tung Nguyen, Thanh-Son Nguyen, and Hoang-Quynh Le. Automatic named entity set expansion using semantic rules and wrappers for unary relations. In *Asian Language Processing (IALP), 2010 International Conference on*, pages 170–173, Dec 2010.

[188] Nam Khanh Tran, Andrea Ceroni, Nattiya Kanhabua, and Claudia Niederée. Back to the Past: Supporting Interpretations of Forgotten Stories by Time-aware Re-Contextualization. In *8th ACM International Conference on Web Search and Data Mining*, pages 339–348, Shanghai, China, 2015.

[189] Nam Khanh Tran, Andrea Ceroni, Nattiya Kanhabua, and Claudia Niederée. Time-travel Translator: Automatically Contextualizing News Articles. In *24th ACM International World Wide Web Conference (WWW)*, pages 247–250, Florence, Italy, 2015.

[190] Raphaël Troncy. Integrating Structure and Semantics into Audio-visual Documents. In *2nd International Semantic Web Conference (ISWC'03)*, pages 566–581, Sanibel Island, Florida, USA, 2003.

[191] Raphaël Troncy, Werner Bailer, Michael Hausenblas, Philip Hofmair, and Rudolf Schlatte. Enabling Multimedia Metadata Interoperability by Defining Formal Semantics of MPEG-7 Profiles. In *1st International Conference on Semantics And digital Media Technology (SAMT'06)*, pages 41–55, Athens, Greece, 2006.

[192] Raphaël Troncy and Jean Carrive. A Reduced Yet Extensible Audio-Visual Description Language: How to Escape From the MPEG-7 Bottleneck. In *4th ACM Symposium on Document Engineering (DocEng'04)*, Milwaukee, Wisconsin, USA, 2004.

[193] Raphaël Troncy, Vuk Milicic, Giuseppe Rizzo, and José Luis Redondo Garcia. Mediafinder: Collect, enrich and visualize media memes shared by the crowd. In *nd International Workshop on Real-Time Analysis and Mining of Social Streams (RAMSS'13)*, Rio de Janeiro, Brazil, 2013.

[194] E. Tsamoura, V. Mezaris, and I. Kompatsiaris. Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In *Proc. Image Processing*, pages 45 –48, oct. 2008.

[195] E. Tsamoura, V. Mezaris, and I. Kompatsiaris. Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In *Proc. Image Processing*, pages 45 –48, 2008.

[196] S. Tschpel and D. Schneider. A lightweight keyword and tag-cloud retrieval algorithm for automatic speech recognition transcripts. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association ISCA (INTERSPEECH)*, 2010.

[197] Chrisa Tsinaraki and Stavros Christodoulakis. Interoperability of XML Schema Applications with OWL Domain Knowledge and Semantic Web Tools. In $6^{rd}$ *International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE)*, 2007.

[198] Chrisa Tsinaraki, Panagiotis Polydoros, and Stavros Christodoulakis. Interoperability support for Ontology-based Video Retrieval Applications. In $3^{rd}$ *International Conference on Image and Video Retrieval (CIVR)*, pages 582–591, 2004.

[199] Chrisa Tsinaraki, Panagiotis Polydoros, and Stavros Christodoulakis. Interoperability support between MPEG-7/21 and OWL in DS-MIRF. *In Transactions on Knowledge and Data Engineering (TKDE), Special Issue on the Semantic Web Era*, 19(2):219–232, 2007.

[200] Koen van de Sande, Theo Gevers, and Cees Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1582–1596, September 2010.

[201] Jan C. van Gemert, Cor J. Veenman, Arnold W. M. Smeulders, and Jan-Mark Geusebroek. Visual Word Ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1271–1283, 2010.

[202] Jacco van Ossenbruggen, Frank Nack, and Lynda Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). *IEEE Multimedia*, 11(4), 2004.

[203] Ruben Verborgh, Davy Van Deursen, Erik Mannens, Chris Poppe, and Rik Van de Walle. Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform. *Multimedia Tools and Applications*, 61(1):105–129, 2012.

[204] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511 – I–518 vol.1, 2001.

[205] Vadim von Brzeski, Utku Irmak, and Reiner Kraft. Leveraging context in user-centric entity detection systems. In *Proceedings of the Sixteenth ACM*

*Conference on Conference on Information and Knowledge Management*, pages 691–700, New York, NY, USA, 2007.

[206] Tadej Štajner, Bart Thomee, Ana-Maria Popescu, Marco Pennacchiotti, and Alejandro Jaimes. Automatic Selection of Social Media Responses to News. In $19^{th}$ *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 50–58, 2013.

[207] Richard C. Wang and William W. Cohen. Language-independent set expansion of named entities using the web. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 342–350, Washington, DC, USA, 2007.

[208] Richard C. Wang and William W. Cohen. Iterative set expansion of named entities using the web. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 1091–1096, Washington, DC, USA, 2008.

[209] G.C. Whitworth. *Indian English: an examination of the errors of idiom made by Indians in writing English*. Bahri Publications, 1982.

[210] William E Winkler. Overview of record linkage and current research directions. In *Bureau of the Census*, 2006.

[211] Ian H. Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with Java implementations. In *The Morgan Kaufmann series in data management systems*. Morgan Kaufmann, 1999.

[212] Garrett Wolf, Hemal Khatri, Bhaumik Chokshi, Jianchun Fan, Yi Chen, and Subbarao Kambhampati. Query processing over incomplete autonomous databases. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, pages 651–662, 2007.

[213] Minerva Yeung, Boon-Lock Yeo, and Bede Liu. Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, 71:94–109, 1998.

[214] Hongliang Yu, Dongdong Zheng, Ben Y. Zhao, and Weimin Zheng. Understanding user behavior in large-scale video-on-demand systems. In *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006*, EuroSys '06, pages 333–344, New York, NY, USA, 2006.

[215] John R Zhang, Yang Song, and Thomas Leung. Improving video classification via youtube video co-watch data. In *Workshop on Social and behavioural networked media access*, 2011.

# Accessing Media Enrichments in RDF: SPARQL Queries over LinkedTV Ontology Datasets

In this appendix we show how to query both the annotations and the enrichments that have been serialized in RDF and stored in a triplestore following the LinkedTV ontology model. In the following, we provide 10 representative queries for accessing the data available inside the LinkedTV RDF graph. All the examples are applied over the media resource *http://data.linkedtv.eu/media/8a8187f2-3fc8-cb54-0140-7dd151100003* but are easy to generalize to any other media resource. For a better understanding of the different statements inside the SPARQL queries, please check Figure 2.14 and Figure 4.6 where those attributes and properties are graphically displayed.

### Query 1: list of `linkedtv:Shot` or `linkedtv:Chapter` for a media resource

```
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
PREFIX rdfs:   <http://www.w3.org/2000/01/rdf-schema#>
PREFIX oa: <http://www.w3.org/ns/oa#>

SELECT ?mediaFragment
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
    ?mediaFragment ma:isFragmentOf
        <http://data.linkedtv.eu/media/8a8187f2-3fc8-cb54-0140-7dd151100003> .
  ?mediaFragment a ma:MediaFragment .
  ?annotation a oa:Annotation .
    ?annotation oa:hasBody ?shot .
    ?annotation oa:hasTarget ?mediaFragment .
    ?shot a linkedtv:Shot .
}
```

Listing 5.1: SPARQL query for Retrieving Shots and Chapters

### Query 2: list of keywords attached to a media fragment

```
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX rdfs:   <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
```

```
SELECT ?keywordtext
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
    <http://data.linkedtv.eu/media/b82fb032−d95e−11e2−951c−f8bdfd0abfbd#t=1463.28,1480.48>
        ma:hasKeyword ?keyword .
  ?keyword a linkedtv:Keyword .
  ?keyword rdfs:label ?keywordtext
}
```

Listing 5.2: SPARQL query for keywords

## Query 3: list of LSCOM concepts attached to a media fragment

```
    PREFIX linkedtv: <http://data.linkedtv.eu/ontology/>
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX ma: <http://www.w3.org/ns/ma−ont#>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf−schema#>
PREFIX oa: <http://www.w3.org/ns/oa#>

SELECT  ?keyURL
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
    ?annotation oa:hasTarget
      <http://data.linkedtv.eu/media/b82fb032−d95e−11e2−951c−f8bdfd0abfbd#t
          =1463.28,1480.48>  .
    ?annotation oa:hasBody ?concept.
  ?concept a linkedtv:Concept .
  ?concept owl:sameAs ?keyURL .
}
```

Listing 5.3: SPARQL query for LSCOM concepts

## Query 4: list of entities spotted inside the transcript of a media resource, showing their label and disambiguation URI

```
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX nsa: <http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna#>
PREFIX ma: <http://www.w3.org/ns/ma−ont#>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf−schema#>
PREFIX dc:      <http://purl.org/dc/elements/1.1/>
PREFIX owl:     <http://www.w3.org/2002/07/owl#>

SELECT ?Entity, ?label, ?source, ?disURL
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
    ?MediaFragment ma:isFragmentOf
     <http://data.linkedtv.eu/media/b82fb032−d95e−11e2−951c−f8bdfd0abfbd> .
    ?MediaFragment a ma:MediaFragment .
    ?MediaFragment linkedtv:hasSubtitle ?subtitle .
    ?AnnotationEntity oa:hasTarget ?MediaFragment.
    ?AnnotationEntity oa:hasBody ?Entity.
    ?Entity a linkedtv:Entity .
    OPTIONAL { ?Entity rdfs:label ?label .  }
    OPTIONAL { ?Entity dc:source ?source . }
    ?Entity owl:sameAs ?disURL
}
```

Listing 5.4: SPARQL query for entities

## Query 5: list of entities of type `nerd:Person` attached to a media resource

```
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX nsa: <http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc:       <http://purl.org/dc/elements/1.1/>
PREFIX owl:      <http://www.w3.org/2002/07/owl#>
PREFIX nerd: <http://nerd.eurecom.fr/ontology#>

SELECT ?Entity, ?label, ?source, ?disURL
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
    ?MediaFragment ma:isFragmentOf
     <http://data.linkedtv.eu/media/b82fb032-d95e-11e2-951c-f8bdfd0abfbd> .
    ?MediaFragment a ma:MediaFragment .
    ?MediaFragment linkedtv:hasSubtitle ?subtitle .
    ?AnnotationEntity oa:hasTarget ?MediaFragment.
    ?AnnotationEntity oa:hasBody ?Entity.
    ?Entity a linkedtv:Entity .
    OPTIONAL { ?Entity rdfs:label ?label .  }
    OPTIONAL { ?Entity dc:source ?source . }
    ?Entity owl:sameAs ?disURL .
    ?Entity  a nerd:Person.
}
```

Listing 5.5: SPARQL query for entities of type person

## Query 6: list of entities for a media resource which are temporally located inside a given period of time (a,b)

```
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX nsa: <http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc:       <http://purl.org/dc/elements/1.1/>
PREFIX owl:      <http://www.w3.org/2002/07/owl#>

SELECT ?Entity, ?label, ?source, ?disURL
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
    ?MediaFragment ma:isFragmentOf
     <http://data.linkedtv.eu/media/b82fb032-d95e-11e2-951c-f8bdfd0abfbd> .
    ?MediaFragment a ma:MediaFragment .
    ?MediaFragment linkedtv:hasSubtitle ?subtitle .
    ?AnnotationEntity oa:hasTarget ?MediaFragment.
    ?AnnotationEntity oa:hasBody ?Entity.
    ?Entity a linkedtv:Entity .
    OPTIONAL { ?Entity rdfs:label ?label .  }
    OPTIONAL { ?Entity dc:source ?source . }
    ?Entity owl:sameAs ?disURL .
    ?MediaFragment nsa:temporalStart ?start ;
     nsa:temporalEnd ?end .
    filter(?start > 60 ) .
    filter(?end < 120 )
}
```

Listing 5.6: SPARQL query for entities inside a time interval

## Query 7: list of the more frequent entities for a media resource

```
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
```

```
SELECT ?mr ?label ?url count(?url)
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
  ?mr a ma:MediaResource .
  ?mf ma:isFragmentOf ?mr .
  ?ann oa:hasTarget ?mf.
  ?ann oa:hasBody ?entity.
  ?entity a linkedtv:Entity .
  ?entity dc:source "textrazor" .
  OPTIONAL { ?entity rdfs:label ?label . }
  OPTIONAL { ?entity owl:sameAs ?url . }
}
GROUP BY ?mr ?label ?url
```

Listing 5.7: SPARQL query for more frequent entities

## Query 8: list of enrichments triggered by a particular entity E given its label

```
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX nsa: <http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
PREFIX rdfs:   <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc:     <http://purl.org/dc/elements/1.1/>
PREFIX owl:    <http://www.w3.org/2002/07/owl#>
PREFIX nerd: <http://nerd.eurecom.fr/ontology#>
PREFIX prov:   <http://www.w3.org/ns/prov#>

SELECT ?MediaResources ?URL
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
    ?MediaFragment ma:isFragmentOf
    <http://data.linkedtv.eu/media/b82fb032-d95e-11e2-951c-f8bdfd0abfbd>  .
    ?MediaFragment a ma:MediaFragment .
    ?MediaFragment linkedtv:hasSubtitle ?subtitle .
    ?AnnotationEntity oa:hasTarget ?MediaFragment .
    ?AnnotationEntity oa:hasBody ?Entity .
    ?Entity a linkedtv:Entity .
    ?Entity rdfs:label Berlin .
    ?AnnotationRelResource a oa:Annotation .
    ?AnnotationRelResource prov:wasDerivedFrom ?Entity .
    ?AnnotationRelResource oa:hasBody ?MediaResources .
    ?MediaResources ma:locator ?URL
}
```

Listing 5.8: SPARQL query for obtaining enrichments

## Query 9: list of media fragments (initially, with any granularity) that have at least one enrichment result attached

```
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX nsa: <http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
PREFIX rdfs:   <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc:     <http://purl.org/dc/elements/1.1/>
PREFIX owl:    <http://www.w3.org/2002/07/owl#>
PREFIX nerd: <http://nerd.eurecom.fr/ontology#>
PREFIX prov:   <http://www.w3.org/ns/prov#>

SELECT ?MediaFragment
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
    ?MediaFragment ma:isFragmentOf
    <http://data.linkedtv.eu/media/b82fb032-d95e-11e2-951c-f8bdfd0abfbd>  .
```

```
    ?MediaFragment a ma:MediaFragment .
    ?AnnotationRelResource a oa:Annotation .
    ?AnnotationRelResource oa:motivatedBy oa:linking .
    ?AnnotationRelResource oa:hasBody ?MediaResources .
    ?AnnotationRelResource oa:hasTarget ?MediaFragment .
    ?MediaResources a ma:MediaResource .
    ?MediaResources ma:locator ?URLs
}
```

Listing 5.9: SPARQL query for getting fragments with enrichment

## Query 10: list of enrichment media resources interlinked to a certain media fragment

```
PREFIX linkedtv: <http://data.linkedtv.eu/ontologies/core#>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX nsa: <http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc:      <http://purl.org/dc/elements/1.1/>
PREFIX owl:     <http://www.w3.org/2002/07/owl#>
PREFIX nerd: <http://nerd.eurecom.fr/ontology#>
PREFIX prov:    <http://www.w3.org/ns/prov#>


SELECT ?MediaResources ?URLs
FROM <http://data.linkedtv.eu/graph/linkedtv>
WHERE {
    ?AnnotationRelResource a oa:Annotation .
    ?AnnotationRelResource oa:motivatedBy oa:linking .
    ?AnnotationRelResource oa:hasBody ?MediaResources .
    ?AnnotationRelResource oa:hasTarget
     <http://data.linkedtv.eu/media/b82fb032-d95e-11e2-951c-f8bdfd0abfbd#t=80.12,128.96> .
    ?MediaResources a ma:MediaResource .
    ?MediaResources ma:locator ?URLs
}
```

Listing 5.10: SPARQL query for media resources attached to fragments

# News Entities Gold Standard

The gold standard is composed by a set of 5 short videos from the BBC One Minute World News Website[1]http://www.bbc.com/news/video_and_audio/ together with a list of Named Entities that intend to illustrate the context of the stories being told in them. Each named entity has an associated score normalized from $-1$ to $1$ in order to indicate its importance inside the whole set of video annotations. Those annotations are available in a structured format, and with further documentation, at https://github.com/jluisred/NewsEntities/.

---

[1]\unskip\penalty\@M\vrulewidth\z@height\z@depth\dpff

Table 6.1: Fugitive Edward Snowden applies for asylum in Russia

| Entity Label | Score |
|---|---|
| Edward Snowden | 0.5555556 |
| National Security Agency | 0.3777778 |
| Central Intelligence Agency | 0.0666667 |
| Glenn Greenwald | 0.0666667 |
| Foreign Intelligence Surveillance Court | 0 |
| Booz Allen Hamilton | -0.1333333 |
| Laura Poitras | -0.2666667 |
| Barton Gellman | -0.3111111 |
| Verizon Wireless | -0.3333333 |
| Angela Merkel | -0.3777778 |
| Anatoly Kucherena | -0.4 |
| Federal Migration Service | -0.4444444 |
| US State Department | -0.4444444 |
| Vladimir Putin | -0.4666667 |
| Sheremetyevo International Airport | -0.4666667 |
| Nicaragua | -0.4888889 |
| Evo Morales | -0.4888889 |
| Patrick Ventrell | -0.5111111 |
| Moscow | -0.5555556 |
| Bolivia | -0.5777778 |
| Kremlin | -0.6 |
| Russia | -0.6222222 |
| Venezuela | -0.6222222 |
| United States of America | -0.6666667 |
| Latin America | -0.7333333 |
| Barack Obama | -0.7555556 |
| Jay Carney | -0.7555556 |
| South America | -0.8 |

Table 6.2: Egypts Morsi Vows to Stay in Power

| Entity Label | Score |
|---|---|
| Muslim Brotherhood | 0.441176471 |
| Mohamed Morsi | 0.352941176 |
| Islamist Nour Party | 0.323529412 |
| Arab Spring | 0.264705882 |
| Egypt opposition alliance | 0.235294118 |
| Abdul Fattah Sisi | 0.088235294 |
| Hosni Mubarak | -0.029411765 |
| Adly Mansour | -0.176470588 |
| Tahrir Square | -0.235294118 |
| Mohamed ElBaradei | -0.235294118 |
| Supreme Constitutional Court | -0.294117647 |
| Egypt | -0.352941176 |
| Cairo | -0.470588235 |
| Cairo University | -0.470588235 |
| Nasr | -0.529411765 |
| Jeremy Bowen | -0.705882353 |
| Nisha Pilla | -0.882352941 |

Table 6.3: Fukushima leak causes Japan concern

| Entity Label | Score |
|---|---|
| Fukushima Daiichi | 0.5 |
| Tokyo Electric Power Company | 0.476190476 |
| Reactor building no 3 | 0.404761905 |
| Nuclear Regulation Authority | 0.142857143 |
| Pacific Ocean | -0.357142857 |
| Nuclear Regulatory Commission | -0.357142857 |
| Japan | -0.380952381 |
| Naomi Hirose | -0.476190476 |
| Tokyo | -0.5 |
| Shunichi Tanaka | -0.523809524 |
| Shinzo Abe | -0.547619048 |
| Rupert Wingfield Hayes | -0.785714286 |
| TV Tokyo | -0.880952381 |

Table 6.4: Rallies in US after Zimmerman Verdict

| Entity Label | Score |
|---|---|
| George Zimmerman | 0.444444444 |
| Trayvon Martin | 0.407407407 |
| National A. for the Adv. of Colored People (NAACP) | -0.0370370 |
| Rachel Jeantel | -0.185185185 |
| General Eric Holder | -0.222222222 |
| Rev. Lowman Oliver | -0.259259259 |
| Mark O?Mara | -0.259259259 |
| Sanford | -0.333333333 |
| Angela Corey | -0.37037037 |
| United States Department of Justice | -0.407407407 |
| Matt Gutman | -0.555555556 |
| Barack Obama | -0.592592593 |
| New York | -0.666666667 |
| Florida | -0.666666667 |
| Los Angeles | -0.703703704 |
| San Francisco | -0.740740741 |
| Times Square | -0.740740741 |
| Washington DC. | -0.814814815 |
| United States of America | -0.851851852 |

Table 6.5: Royal Baby Prince Named George

| Entity Label | Score |
|---|---|
| Kate Middleton Duchess of Cambridge | 0.357142857 |
| King George the VI Bertie | 0.357142857 |
| Prince George of Cambridge George Alexander Louis | 0.285714286 |
| Prince William Duke of Cambridge | 0.285714286 |
| King George V | 0.142857143 |
| Queen Elizabeth | 0.071428571 |
| Earl Louis Mountbatten | -0.142857143 |
| Prince Harry of Wales | -0.214285714 |
| Charles Prince of Wales | -0.285714286 |
| Queen Victoria | -0.285714286 |
| Balmoral Castle | -0.285714286 |
| Princess Margaret | -0.357142857 |
| Cambridge | -0.357142857 |
| Queen Mary of Teck | -0.357142857 |
| Kensington Palace | -0.428571429 |
| The Queen Mother Elizabeth Angela Marguerite Bowes-Lyon | -0.428571429 |
| Irish Republican Army | -0.428571429 |
| St Mary's Hospital | -0.5 |
| England Country | -0.5 |
| Great Britain | -0.5 |
| Suzannah Lipscomb | -0.571428571 |

# Results from Multidimensional NSS Generation Approach

In this appendix we present the results obtained after automatically annotating the videos proposed in the News Entities dataset[1], containing different BBC One Minute World News[2] news items. For obtaining such annotations we have launched the multidimensional NSS generation approach described in Chapter 6, with the best settings according to the experiments performed: using *L1 whitelist* and Google as sources for the expansion, 2 weeks temporal window, *F3* filtering, no *schema.org* filtering, and *Expert Rules* applied.

In particular, we display the top 15 named entities proposed, together with an associated score normalized from −1 to 1 in order to indicate its importance inside the whole set of video annotations according to the aforementioned algorithm.

Table 7.1: Fugitive Edward Snowden applies for asylum in Russia

| Entity Label | Score |
|---|---|
| Edward Snowden | 1.0 |
| Russia | 0.48216106 |
| Vladimir Putin | 0.407747197 |
| National Security Agency | 0.308868502 |
| United States | 0.296636086 |
| Moscow | 0.201834862 |
| President of the Russian Federation | 0.180428135 |
| Federal Migration Service | 0.171253823 |
| Latin America | 0.150866463 |
| Human Rights Watch | 0.143730887 |
| United States | 0.137614679 |
| WikiLeaks | 0.113149847 |
| Sheremetyevo Airport | 0.095820591 |
| Barack Obama | 0.081549439 |
| Hong Kong | 0.074413863 |
| Amnesty International | 0.067278287 |

[1] https://github.com/jluisred/NewsEntities/
[2] http://www.bbc.com/news/video_and_audio/

Table 7.2: Egypts Morsi Vows to Stay in Power

| Entity Label | Score |
|---|---|
| Mohamed Morsi | 1.0 |
| Muslim Brotherhood | 0.998929336 |
| Egypt | 0.592077088 |
| Islamist movement | 0.308351178 |
| Cairo | 0.284796574 |
| Hosni Mubarak | 0.274089936 |
| Egyptian President | 0.203426124 |
| Egyptians | 0.175588865 |
| Tahrir Square | 0.13490364 |
| Al-Jazeera | 0.118843683 |
| Barack Obama | 0.111349036 |
| United States | 0.104925054 |
| Sissi | 0.104925054 |
| Mohamed Badie | 0.096359743 |
| Freedom and Justice Party | 0.086723769 |

Table 7.3: Fukushima leak causes Japan concern

| Entity Label | Score |
|---|---|
| Tokyo Electric Power Company | 1.0 |
| Fukushima Daiichi nuclear disaster | 0.698765432 |
| Japan | 0.387654321 |
| Nuclear Regulatory Commission | 0.288888889 |
| San Onofre Nuclear Generating Station | 0.202469136 |
| Barack Obama | 0.192592593 |
| Pacific Ocean | 0.175308642 |
| BBC News | 0.162962963 |
| Japanese government | 0.138271605 |
| United States | 0.135802469 |
| Fukushima Daiichi nuclear power plant | 0.130864198 |
| California | 0.108641975 |
| 23-Jul | 0.10617284 |
| Nuclear Regulation Authority | 0.101234568 |
| Dear | 0.096296296 |

Table 7.4: Rallies in US after Zimmerman Verdict

| Entity Label | Score |
|---|---|
| Martin | 1.0 |
| George Zimmerman | 1.0 |
| U.S. Justice Department | 0.631578947 |
| African-American | 0.534412955 |
| Barack Obama | 0.412955466 |
| Florida | 0.317813765 |
| National Association for the Advancement of Colored People | 0.206477733 |
| United States of America. | 0.198380567 |
| New York City | 0.157894737 |
| CNN | 0.139676113 |
| CRS | 0.133603239 |
| SANFORD Fla | 0.127530364 |
| Los Angeles California | 0.109311741 |
| National Public Radio | 0.097165992 |
| Police | 0.0951417 |

Table 7.5: Royal Baby Prince Named George

| Entity Label | Score |
|---|---|
| George | 1.0 |
| HRH The Duke of Cambridge | 1.0 |
| The Prince of Wales | 0.648318043 |
| Duchess of Cambridge | 0.556574924 |
| Queen Elizabeth II | 0.519877676 |
| Kate | 0.4617737 |
| Lord Mountbatten | 0.324159021 |
| Alexander the Great | 0.29969419 |
| Kensington Palace | 0.266055046 |
| Alexander III of Macedon | 0.201834862 |
| His Royal Highness | 0.186544343 |
| Prince Albert | 0.183486239 |
| London | 0.162079511 |
| Queen Victoria | 0.159021407 |
| Diana Princess of Wales | 0.159021407 |

# Results from Concentric-based NSS Generation Approach

In this appendix we present the results obtained after automatically annotating the videos proposed in the News Entities dataset[1], containing different BBC One Minute World News[2] news items. For obtaining such annotations we have launched the concentric-based NSS generation approach as described in Chapter 7, with the best settings according to the experiments performed: using *Ex2* as set of 50 candidates from multidimensional second best multidimensional run according to $R_{50}^*$ for *Crust* generation, *CoreA* strategy to cluster entities and create the *Core*, $S_{Google}$ as similarity function between *Crust* candidates and the *Core*, and *Core_Crust* strategy for merging *Core* and *Crust*.

In particular, we display the top 15 named entities proposed, together with an associated score normalized from $-1$ to $1$ in order to indicate its importance inside the whole set of video annotations according to the aforementioned algorithm.

---

[1] https://github.com/jluisred/NewsEntities/
[2] http://www.bbc.com/news/video_and_audio/

Table 8.1: Fugitive Edward Snowden applies for asylum in Russia

| Entity Label | Score |
|---|---|
| Edward Snowden | 1.0 |
| United States | 0.9583 |
| Russia | 0.958 |
| Moscow | 0.895 |
| Barack Obama | 0.024321144 |
| The Guardian | 0.013002178 |
| National Security Agency | 0.003663278 |
| Vladimir Putin | 0.001598574 |
| Kremlin | 0.001014826 |
| Glenn Greenwald | 8.68E-04 |
| China | 7.90E-04 |
| White House | 6.79E-04 |
| WikiLeaks | 6.01E-04 |
| Russians | 5.14E-04 |
| Reuters | 4.67E-04 |

Table 8.2: Egypts Morsi Vows to Stay in Power

| Entity Label | Score |
|---|---|
| Mohamed Morsi | 1.0 |
| Muslim Brotherhood | 0.998929336 |
| Egypt | 0.592077088 |
| Islamist movement | 0.308351178 |
| Cairo | 0.284796574 |
| Hosni Mubarak | 0.274089936 |
| Egyptian President | 0.203426124 |
| Egyptians | 0.175588865 |
| Tahrir Square | 0.13490364 |
| Al-Jazeera | 0.118843683 |
| Barack Obama | 0.111349036 |
| United States | 0.104925054 |
| Sissi | 0.104925054 |
| Mohamed Badie | 0.096359743 |
| Freedom and Justice Party | 0.086723769 |

Table 8.3: Fukushima leak causes Japan concern

| Entity Label | Score |
| --- | --- |
| Fukushima nuclear disaster | 1.0 |
| Tokyo Electric Power Company | 0.902 |
| Japan | 0.6095 |
| Pacific Ocean | 0.609 |
| Fukushima Daiichi nuclear power plant | 0.616160076 |
| Nuclear Regulation Authority | 0.015505721 |
| San Onofre Nuclear Generating Station | 0.002797357 |
| Radioactive | 0.002165414 |
| Nuclear Regulatory Commission | 0.001744183 |
| Shinzo Abe | 7.43E-04 |
| Seawater | 5.30E-04 |
| Tokyo | 1.58E-04 |
| Officials | 1.35E-04 |
| United Nations | 1.31E-04 |
| National Security Agency | 1.31E-04 |

Table 8.4: Rallies in US after Zimmerman Verdict

| Entity Label | Score |
| --- | --- |
| Florida | 1.0 |
| United States of America. | 0.775 |
| U.S. Justice Department | 0.7 |
| Benjamin Jealous | 0.032807734 |
| Eric Holder | 0.016400893 |
| George Zimmerman | 0.010390413 |
| SANFORD Fla | 0.008996256 |
| Jesse Jackson | 0.007637556 |
| Michael Bloomberg | 0.006067228 |
| Al Sharpton | 0.005806011 |
| Stand Your Ground law | 0.005145025 |
| Rick Scott | 0.003617693 |
| Benjamin Crump | 0.003601372 |
| Don West | 0.002986261 |
| Angela Corey | 0.002508299 |

Table 8.5: Royal Baby Prince Named George

| Entity Label | Score |
|---|---|
| HRH The Duke of Cambridge | 1.0 |
| Duchess of Cambridge | 0.886 |
| Queen Elizabeth II | 0.681 |
| Prince Albert | 0.670 |
| Prince Philip Duke of Edinburgh | 0.015882624 |
| Edward VIII | 0.010753411 |
| House of Windsor | 0.003646611 |
| Princess Diana | 0.003197463 |
| Royal Family | 0.002985726 |
| Michael Middleton | 0.002033262 |
| Clarence House | 0.001667531 |
| The Prince of Wales | 0.001333869 |
| Prince Harry | 0.001060707 |
| Kensington Palace | 0.00102576 |
| William's | 8.91E-04 |