Deliverable D2.7 Final Linked Media Layer and Evaluation

Tomáš Kliegr, Jan Bouchner, Barbora Červenková, Milan Dojchinovski, Jaroslav Kuchař, Ivo Lašek, Milan Šimůnek, Ondřej Zamazal, Václav Zeman / UEP
Raphaël Troncy, José Luis Redondo García, Giuseppe Rizzo, Benoit Huet, Maria Eskevich, Bahjat Safadi, Mathilde Sahuguet, Hoang An Le, Quoc Minh Bui / EURECOM
Jan Thomsen / CONDAT
Dorothea Tsatsou, Vasileios Mezaris/ CERTH
Adrian M.P. Brasoveanu, Lyndon J.B. Nixon/ MODUL
Lilia Perez Romero / CWI

31/03/2015

Work Package 2: Linking hypervideo to Web content

LinkedTV
Television Linked To The Web
Integrated Project (IP)
FP7-ICT-2011-7. Information and Communication Technologies
Grant Agreement Number 287911

Dissemination level	PU
Contractual date of delivery	31/03/2015
Actual date of delivery	31/03/2015
Deliverable number	D2.7
Deliverable name	Final Linked Media Layer and Evaluation
File	`D2.7_postcorrections.tex`
Nature	Report
Status & version	Released & v1.0
Number of pages	69
WP contributing to the deliverable	2
Task responsible	UEP
Other contributors	EURECOM, CONDAT, CERTH, MODUL, CWI
Author(s)	Tomáš Kliegr, Jan Bouchner, Barbora Červenková, Milan Dojchinovski, Jaroslav Kuchař, Ivo Lašek, Milan Šimůnek, Ondřej Zamazal, Václav Zeman / UEP Raphaël Troncy, José Luis Redondo García, Giuseppe Rizzo, Benoit Huet, Maria Eskevich, Bahjat Safadi, Mathilde Sahuguet, Hoang An Le, Quoc Minh Bui / EURECOM Jan Thomsen / CONDAT Dorothea Tsatsou, Vasileios Mezaris/ CERTH Adrian M.P. Brasoveanu, Lyndon J.B. Nixon/ MODUL Lilia Perez Romero / CWI
Reviewer	Stéphane Dupont / UMONS
EC Project Officer	Thomas Kuepper
Keywords	Web mining, Video Enrichment, Linked Media, Crawling, Information Retrieval, Entity Linking
Abstract (for dissemination)	This deliverable presents the evaluation of content annotation and content enrichment systems that are part of the final tool set developed within the LinkedTV consortium. The evaluations were performed on both the Linked News and Linked Culture trial content, as well as on other content annotated for this purpose. The evaluation spans three languages: German (Linked News), Dutch (Linked Culture) and English. Selected algorithms and tools were also subject to benchmarking in two international contests: MediaEval 2014 and TAC'14. Additionally, the Microposts 2015 NEEL Challenge is being organized with the support of LinkedTV.

History

Table 1: History of the document

Date	Version	Name	Comment
16/01/2015	v0.1	Kliegr, UEP	Deliverable structure
18/02/2015	v0.2.1	Kliegr, UEP	TAC'14 benchmark
19/02/2015	v0.2.2	Kliegr, UEP	LHD evaluation
20/02/2015	v0.2.3	Dojchinovski, UEP	THD salience evaluation
27/02/2015	v0.2.4	Dojchinovski, UEP	THD spotting & linking eval on SV content
27/02/2015	v0.2.5	Kliegr, UEP	Outlook subsection
02/03/2015	v0.3.1	Huet, EURECOM	MediaEval subsection
02/03/2015	v0.3.2	Rizzo, EURECOM	Microposts subsection
04/03/2015	v0.3.3	Redondo Garcia, EURECOM	News Entity Expansion Evaluation subsection
11/03/2015	v0.3.4	Troncy, EURECOM	General polishing over all sections
12/03/2015	v0.3.5	Dojchinovski, UEP	THD spotting & linking eval on RBB content
13/03/2015	v0.3.6	Zamazal, UEP	LHD evaluation on German and Dutch
13/03/2015	v0.3.7	Červenková, UEP	IRAPI evaluation and statistics
16/03/2015	v0.4	Kliegr, UEP	Content enrichment eval section
16/03/2015	v0.5	Thomsen, CONDAT	Final Linked media Layer section
16/03/2015	v0.6	Brasoveanu, MODUL	Recognyze subsection
16/03/2015	v0.7	Jones, MODUL	Linked News and Linked Culture trials
22/03/2015	v0.8	Troncy, EURECOM	Internal deliverable QA
25/03/2015	v0.8.1	Tsatsou, CERTH	Topic labelling subsection
25/03/2015	v0.8.2	Dojchinovski, UEP	Updated section 3.4
25/03/2015	v0.8.3	Brasoveanu, MODUL	QA comments worked in
26/03/2015	v0.9	Kliegr, UEP	QA comments worked in
26/03/2015	v0.91	Troncy, EURECOM	Final QA comments worked in
30/03/2015	v0.95	Tsatsou, CERTH	Final input and QA comments worked in
31/03/2015	v0.96	Kliegr, UEP	Gerbil results notice
31/03/2015	v1.0	Kliegr, UEP	Final editing changes
08/07/2015	v1.01	Kliegr, UEP	Update figures in section 3.4

0 Table of Content

1 Introduction

This deliverable presents the final linked media layer architecture and the evaluation of the content annotation and content enrichment systems developed within the project.

Content annotation is the key element in the semantic lifting of video content. Within this process, entities in the text are detected and linked to semantic knowledge bases such as DBpedia. To further help to represent the content, the salience of these entities is computed and additional entities, possibly relevant given the context, are proposed.

Content enrichment services deliver the business value to the editor, by suggesting links that might be interesting for the viewer given the current content and context. The integration with the Editor tool developed in WP1 ensures that the enrichment suggestions coming from multiple tools are syndicated in one place.

The evaluations were performed using the LinkedTV scenarios (Linked News and Linked Culture trial content) as well as other content annotated for this purpose. The evaluation spans three languages: German (Linked News), Dutch (Linked Culture) and English. The evaluations were performed in a realistic setting, where the content partner was providing the judgments about the relevancy of the content annotation/content enrichment output.

Selected algorithms and tools were also subject to benchmarking in two international contests: MediaEval 2014 and TAC'14. Additionally, the Microposts 2015 NEEL Challenge is being organized with the support of LinkedTV.

Chapter 2 briefly describes the final linked media layer architecture workflows, which encompass the WP2 services. It should be noted that more details on the architecture are presented in deliverables D5.5 *Final Integration Platform* and D5.6 *Final End-to-end Platform*.

Chapter 3 presents evaluations and benchmark for content annotation services. This chapter covers the following systems and resources: Recognyze, Entity expansion, Linked Hypernyms Dataset, THD and Topic Labelling service. The Recognyze and THD systems are evaluated on textual data associated with the LinkedTV trial videos: Recognyze on data from the RBB Linked News trial (German), and THD on both Linked News and Linked Culture (Dutch).

Chapter 4 covers the performance of the content enrichment services. The following systems have been evaluated: the IRAPI custom crawler/search engine, NewsEnricher and the TV2Lucene module for recommending related video chapters. The evaluation is performed using the logs of the editor activity from the Linked News and Linked Culture trials.

Chapter 5 reports on the participation of WP2 partners in the MediaEval'14 and TAC 2014 international competitions.

The deliverable is concluded with a summary and an outlook, which focuses on further plans for the individual services. The Annex I contains a table with a list of software tools used in the final WP2 processing workflow. The Annex II contains user instructions for the topic labelling evaluation presented in Subsection 3.6.

2 Final Linked Media Layer Architecture

This chapter describes briefly the final architecture of the Linked Media Layer in LinkedTV. For a more in-depth description of the architecture, see D5.5 Final Integration Platform and D5.6 Final End-to-end Platform. In some aspects, the following description updates the architecture based on experiences in LinkedTV application development and user trials. In the particular, we now distinguish between three main areas of the workflow processing:

1. The production workflow
2. The publishing workflow
3. The personalization and consumption workflow

2.1 Production workflow

The objective of the production workflow is to make the videos “LinkedTV-ready” and consists of the following sub-steps:

1. **ingestion** of the video itself and related metadata such as TV-Anytime metadata or subtitle files;
2. **analysis** of the video and audio tracks with all the various techniques as performed by WP1;
3. **serialization** of the results and metadata files into the common LinkedTV data model, which is an RDF-based description format making use of all kinds of existing ontologies such as the W3C Media Ontology and which provides annotated media fragments;
4. **annotation** using **named entity recognition** which provides information about basic entities detected in the video transcripts.

Figure 1: The LinkedTV Workflow

The outcome of the production workflow is RDF data, which is stored in a Virtuoso Triple Store within the LinkedTV Platform. For a 30 min rbb AKTUELL news show, approximately 50.000 triples are generated, with about 800 media fragments and about 3.000 annotations. This data is made accessible through the Linked Media Data Layer Interface.

2.2 Publishing workflow

The objective of the publishing workflow is to take the “raw” LinkedTV production data, evaluate it, correct it, filter out unwanted data, and most notably, enrich it by adding all kinds of related material to the various chapters or entities by making use of the rich set of the LinkedTV enrichment services as described in D2.6. Within the first version of the workflow process, enrichment was seen as part of the production workflow, but in the last stages of the project it turned out that this belongs more to the separate manual process of triggering and selecting those kinds of enrichments, which the editor wants to attach to the video’s side content. Concerning the data produced, only part of the data is stored back in the Platform repository, i.e. mainly that part which concerns data about the structure of the video, such as chapter titles, and start and end points of chapters. The publishing workflow is managed through the LinkedTV EditorTool and ends with actually publishing the video.

2.3 Personalization and consumption workflow

While both the production and the publishing workflow contribute to the Linked Media Layer, the personalization and consumption (or viewing) workflow uses the Linked Media Layer. The personalization and consumption (or viewing) workflow is the process of playing the video itself to a user/viewer, displaying the related content either on the same screen or a second screen depending on the respective scenario, adapting the related information to the viewer's profile, reacting to viewer events like pause, fast forward or switch channel, and building the user profile out of his or her implicit or explicit preferences. For a description of the final stage of this workflow, see D4.7 Evaluation and final results.

On the architectural side, the final Linked Media Layer Architecture that supports this workflow consists of three main sub-layers: 1) the Data/Video Layer, 2) the Integration Layer and 3) the Service Layer (Figure 2).

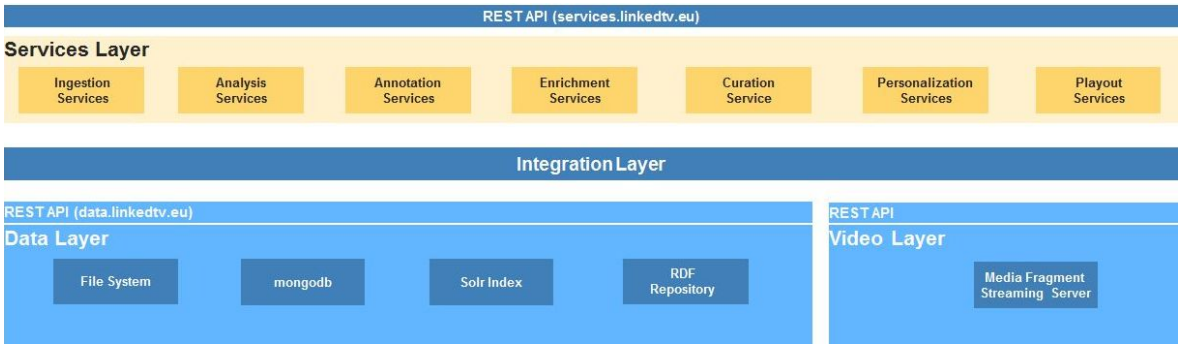

Figure 2: Linked Media Layer Architecture

The Data/Video Layer includes all persistent data, including metadata as well as the video resources themselves. Conceptually, the Data/Video Layer is a layer of Web resources and not necessarily limited to the LinkedTV Platform itself. The Data/Video Layer in LinkedTV provides a unified REST API under `http://data.linkedtv.eu` to all generated media fragments and their annotations. The Integration Layer, however, is a unique part of the LinkedTV platform, and provides the workflow orchestration which connects the different services and ensures the persistent and consistent storage of all data generated and aggregated throughout the LinkedTV workflow. The Service Layer includes all the different specific LinkedTV services for creation, ingestion, analysis, serialization, annotation, enrichment and personalization that create the richness of the Linked Media Layer. These services are distributed over the Web among all partners and can all be accessed and combined through genuine REST API interfaces. The Linked Media Layer also provides a unified access API under `http://services.linkedtv.eu`.

3 Content annotation evaluation

For content annotation, the following systems were subject to evaluation:

- **Recognyze** is a jointly developed system between several universities (Chur, WU, Modul) and the webLyzard company. Recognyze uses a set of SPARQL profiles, dictionaries, disambiguation algorithms and Linked Data dumps of well-known knowledge bases.
- **Entity Expansion** is a service that aims to return a ranked list of entities that fully describe a newscast and its context. This list of entities come from the transcript of the seed video as well as from relevant documents retrieved using the Google Custom Search Engine.
- **THD entity annotation tool** is an unsupervised entity discovery and classification system, which uses a purpose-built Linked Hypernyms Dataset to provide extended type coverage in addition to DBpedia and YAGO knowledge bases. For THD, we report the results of its named entity detection (spotting), linking and salience computation components. The type quality in the THD's **Linked Hypernyms Dataset** is evaluated separately.

3.1 Recognyze evaluation

Modul University has replaced STI International in the consortium. Recognyze is a Named Entity Resolution tool [55], jointly developed between several universities (Modul University, HTW Chur, Vienna University of Economics and Business) and the webLyzard company. Since Modul joined the consortium we decided to also evaluate Recognyze in this package even though it was not initially part of the plan.

Recognyze uses a set of SPARQL profiles, dictionaries, disambiguation algorithms and Linked Data dumps of well-known knowledge bases (DBpedia, Geonames, etc.) in order to perform disambiguation, ranking and linking of the named entities found in a text. It was initially developed for German and English NER tasks with a focus on the main entity types (Person, Location, Organisation), but it is currently extended to support multiple languages (including French, Russian, Arabic, Chinese, etc.) and additional entity types (Event, Product, Works of Art, etc.). In order to extend it for a new language, a user has to provide new knowledge base dumps and dictionaries for that language, and in some special cases, new disambiguation algorithms.

We have decided to evaluate Recognyze on RBB content (chapers extracted from live news), as due to the regionality of the content (news from Berlin-Brandenburg area focused on floods, highways, immigration, local derbys, local administration, or LGBT rights), the frequent use of shortened names for entities instead of the official names, and the language differences between the written German from newspapers or blogs and the German spoken in televion shows, the RBB content is much harder to disambiguate than the news media articles taken from newspapers or press agencies. Due to the way Recognyze was built, the insights obtained from such evaluations can be used to create much better lexicons or profiles, and these will be addressed in a future publication. We have not performed this evaluation in order to replace the established LinkedTV approach, but rather to examine if we can bring some improvements to it.

3.1.1 RBB annotated corpus

For evaluating our NER tool we have decided to create a larger corpus from subtitles extracted from news video chapters. The RBB corpus contains 80 documents that represent anonymized subtitles extracted from the RBB news show Abendschau (daily news show broadcasted between 19:30 and 20:00 CET).

We have extracted the subtitles from several hundreds of video fragments from the RBB index[1] and created a corpus by importing them into GATE. We have used two annotators that were asked to manually annotate the videos and provide the following information: surface forms, entity types, German DBpedia links for the entities wherever this was possible (as expected, not all entities were present in the German version of DBpedia). An expert was involved in assessing the quality of the agreement for typing and linking. All the documents contain at least one entity, but not necessarily any links, as there are at least several situations where even though the entities were quite clear for the annotators, they were not able to find any good links for them in the German version of DBpedia. The average duration of the clips was 132.45 seconds. The documents contained an average of 5.66 links available in the

[1] http://data.linkedtv.eu/solr/#/RBBindex

German version of DBpedia for an average of 8.85 entities per document. Table 2 and Table 3 present the inter-annotator agreements for types and links for the 80 documents that were manually annotated.

Table 2: Type agreements for current RBB ground truth

Document	Agreed	Total	Observed agreement	Cohen's Kappa	Pi's Kappa
Macro summary			0.9850	0.9766	0.9764
Micro summary	708	720	0.9833	0.9764	0.9764

Table 3: URI agreements for current RBB ground truth

Document	Agreed	Total	Observed agreement	Cohen's Kappa	Pi's Kappa
Macro summary			0.8926	0.8802	0.8746
Micro summary	453	518	0.8682	0.8666	0.8665

The corpus contains subtitles that cover a wide range of topics: sports, politics, entertainment, weather, disasters, immigration, healthcare, etc. Most of the news are focused on the area Berlin-Brandenburg. The disambiguation has been done in context. Therefore, a set of entities that might not have made any sense otherwise (people designated by a function and just family name, for example) were easier to disambiguate by the humans annotators since they knew that there are really good chances that the respective entity is related to the Belin-Brandenburg area (if you know it is a person and that it is the minister of Berlin land and a simple Google query does not help you find it, you are still allowed to look on the list of current people who are in office to find the right link, for example). The annotators used a simple ontology to annotate the corpus that contained the following types: Person, Organization, Location, Event, Product, Work, Miscellaneous (for all the other important objects that could not be classified in the previous types), and followed the conventions explained in annotation guideline provided to them before the tasks. The version of German DBpedia that was used for link collection is the one that is currently online[2], but the evaluations were done using the last available dump from DBpedia 2014[3].

3.1.2 Evaluation results

We have evaluated German profiles of Recognyze against the RBB ground truth. The results are presented in Table 4. Since the location profiles we use tend to be based on Geonames, we have converted DBpedia URIs that represent locations to Geonames URIs in post-processing. We first collected the link for the equivalent English entity and through a second SPARQL query collected the equivalent Geonames. For people and organizations, we have used the DBpedia URIs that were present in the gold standard.

The results for organization profiles are really bad if we consider the fact that Recognyze has performed much better in similar evaluations done on news media articles (see [55]). The difference, here, will be the fact that the corpus contains many local organizations for Berlin or Brandenburg, and Recognyze was not optimized for them. Due to the fact that in many of the documents, we can find lots of persons identified by a function and their family name (almost half of them) and that Recognyze was not yet optimized to find such persons, we find the results obtained for persons really good.

While the location results look bad, the real results are much better (at least 2-2.5 times better, therefore suggesting a real F1 measure between 0.36 and 0.40). The main issue is that we used a Geonames profile, and converting German DBpedia links to Geonames via English DBpedia is a buggy process that currently looses more than half of the possible links. We have manually checked the results and Recognyze does indeed find more real Geonames results that correspond to the German DBpedia entities than simple matching algorithm used in the evaluation. The matching algorithm had two SPARQL queries: i) one to get the English DBpedia link that corresponds to the German DBpedia entity because there is no `owl:sameAs` for Geonames in most of the German DBpedia entries; ii) the second query just took the corresponding Geonames link from the English DBpedia page. We will try to create a better algorithm for getting the links between various versions of DBpedia, Geonames and other knowledge bases in the near future. Since the Geonames profile was also optimized for populated

[2] `http://de.dbpedia.org`

[3] `http://data.dws.informatik.uni-mannheim.de/dbpedia/2014/de/`

places and administrative entities, it is also likely that by adding the other types of geographical entities (hydro, roads, etc.) the results will improve even more.

The current results are intermediary and can be improved in the future. It is likely that for the current corpus, all Recognyze results (or the results of other NER systems) can be improved if a) the links between the entities are considered, and if b) the various profiles used for disambiguation put an emphasis on location.

Table 4: Recognyze results on the RBB ground truth.

Entity Type	Knowledge Base	P	R	F1
Organization	German DBpedia 2014	0.16	0.05	0.07
Person	German DBpedia 2014	0.51	0.38	0.44
Location	Geonames 2014	0.11	0.37	0.18

3.1.3 Outlook

It was expected that Recognyze will not perform really well in this evaluation due to the regionality of the content, use of shortened names, and language differences between German used in news media and television. Based on the results we obtained, we have already started creating new lexicons that are much better for this type of content (a much improved lexicon was especially needed for the Organisations, as it can easily be seen from the bad results obtained for this type of entity). The issue of shortened names (particularly relevant for royalty people, banks, universities) is not always easy to fix. If the relevant triples for the entities do not feature an alternate name (dbpprop:alternativeNames) that includes the shortened form, another solution would be to parse the abstracts of the entities (as these abstracts often contain the shortened name). A fix can be implemented as a disambiguation algorithm binded to the relevant fields retrieved through the SPARQL query (dbpedia-owl:abstract, dbpprop: alternativeNames). Of course choosing the right binding handlers for a particular profile can be really hard, and future work on Recognyze will focus on building such profiles automatically based on the type of content and the entity features a user might be interested in. The problems faced when we adapted Recognyze for different types of content (news media, social media, television), together with some of the solutions we will continue to develop, will be discussed in a further publication.

As it can be seen from the previous paragraph, by creating the RBB corpus we have established a basis for improving the NER services for multimedia content. The plans for the RBB corpus include the extension of the corpus to contain more documents (probably around 150-200 compared to the current 80 documents), its integration into GERBIL [51] and comparison with other systems.

3.2 Entity Expansion Evaluation

In this section, we evaluate our algorithm called Newscast Named Entity Expansion that is semantically annotating news items in the LinkedNews scenario. This approach retrieves and analyzes additional documents from the Web where the same event is described in order to automatically generate semantic annotations that provide viewers and experts of the domain a additional information to fully understand the context of the news item. 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.

The approach takes as inputs the publication date, the transcripts and the newscast's title. It outputs a ranked list of entities called Newscast Semantic Snapshot (NSS), which includes the initial set of detected entities in the subtitle and other event-related entities extracted from the associated documents. We have evaluated this method against a gold standard generated by domain experts and assessed via a user survey for five different BBC newscasts. The results of the experiments show the robustness of our approach holding an Average Normalized Discounted Cumulative Gain of 66.6%.

3.2.1 Newscast Entity Expansion Approach

The approach we use to generate Newscast Semantic Snapshot is composed of the following steps: query formulation, document retrieval, semantic annotation, annotation filtering, and annotation ranking. Figure 3 depicts the workflow.

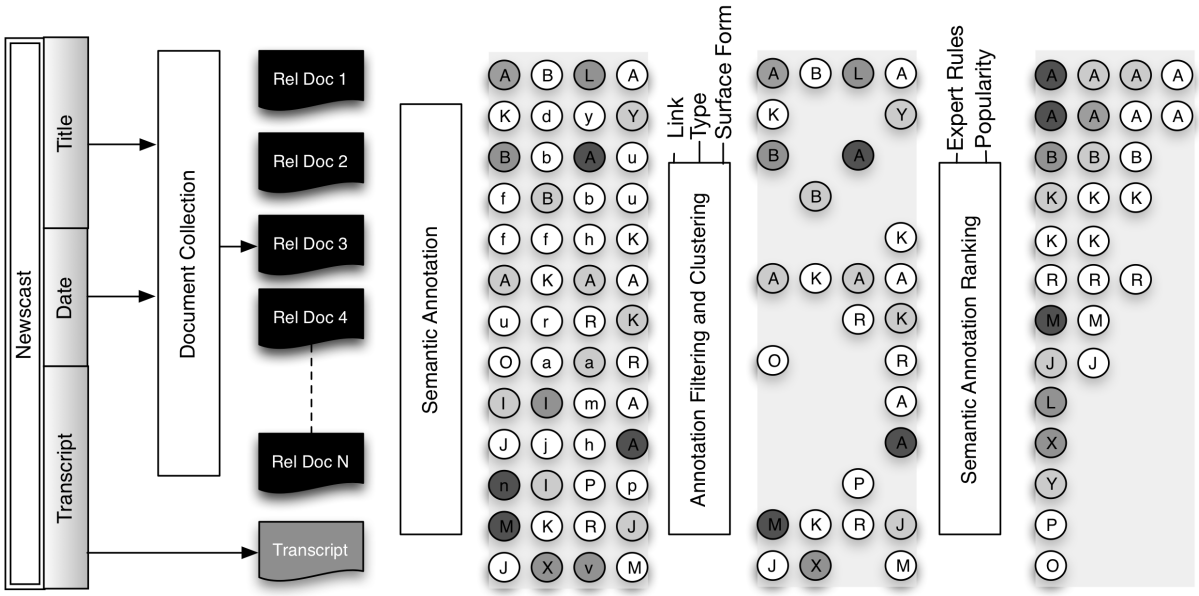

Figure 3: Workflow of the Named Entity Expansion algorithm

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 $q = [\mathrm{h}, \mathrm{t}]$, where *h* is the video headline, and *t* is the publication date. The query is then used as an 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, making a search, and looking at 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 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 $f\left(E_{d_i}\right) \rightarrow E'_{d_i}$ 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 other type of entities are generally more vague to be displayed on a second screen user interface and are 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 has 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 [56] 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 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

implemented a ranking strategy based on: entity appearance in documents, popularity peak analysis, and domain experts' rules. We finally generate Semantic Snapshot for the Newscast being analyzed (NSS).

3.2.2 Ranking Strategy

We have considered two different scoring functions for weighting the frequency of the entities. We then considered two orthogonal functions which exploit the entity popularity in the event time window, and the domain experts' rules.

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 Figure 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. Those with high S_F are instead on the right side of the plot, being then considered to be part of the NSS.

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. 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 $\mu = \frac{|D|}{2}$, being $|D|$ is the cardinality of the number of retrieved documents (Figure 4 (b)).

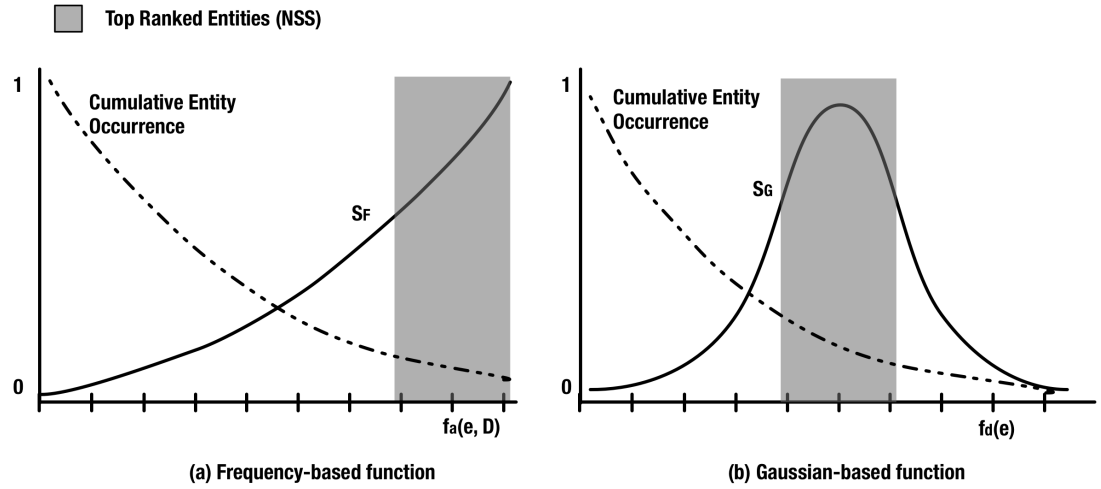

Figure 4: (a) depicts the Decay function of the entity occurrences in the corpus, and the S_F which underlines the importance of an entity being used several times in the corpus. (b) represents the Gaussian-based function S_G, with the entities highly important over the mean.

Popularity Function. Frequency-based approaches fail to capture the phenomenon when particular relevant entities are barely mentioned in the 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, such as Google Trends[4] or Twitter[5].

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 event. The procedure for getting $P_{peak}(e_i)$ is depicted in Figure 5. 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 a particular entity label.

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 helps us to better filter out popularity

[4] https://www.google.com/trends
[5] https://twitter.com

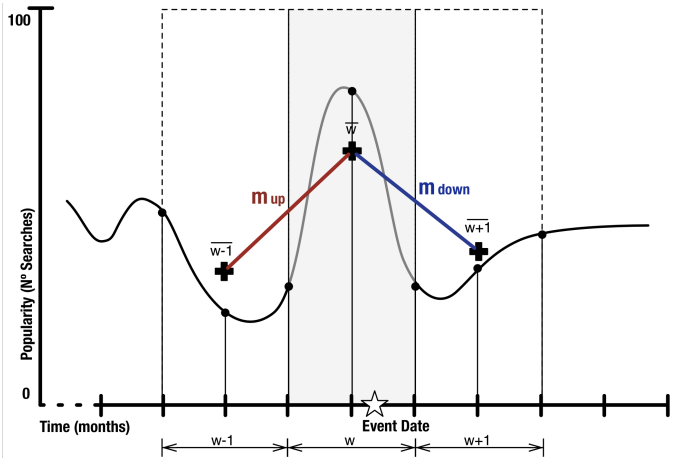

Figure 5: Popularity diagram for an event. On the x-axis, the time is represented, while the y-axis corresponds to the magnitude of the popularity score. The star indicates when the event occurred. Given the discrete nature of the platforms used, the center of the time window can be placed next to the day of the event.

Table 5: Breakdown entity figures per type and per newscast.

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

scores that do not trigger valid conclusions and therefore improve the merging of the ranking produced by the previous functions with the outcome from the popularity peaks detection algorithm.

Expert Rules Function. The knowledge of experts in the domain, like journalists or newscast editors, can be encoded in the form of rules that correct the scoring output produced by our 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 a factor which models the domain experts' opinions about the entities that match the antecedent.

3.2.3 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. 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. To the best of our knowledge there is no evaluation dataset suited to this context. The title of the newscasts and the breakdown figures per entity type are shown in Table 5, so we built our own Golden Set following the procedure described below. The dataset is freely available[6].

Golden Standard Generation Process. 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 topics: 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.

The annotation of those selected newscasts involved two human participants: an annotator and a journalist (expert of the domain). No system bias affected the annotation process, since each annotator

[6] `https://github.com/jluisred/NewsConceptExpansion/wiki/Golden-Standard-Creation`
[7] `http://www.bbc.com/news/video_and_audio/`

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 subtitles;

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 following sources were considered: The Guardian, New York Times, and Al Jazeera online (English). The results were filtered of one week time, where the median is represented by the day when the event took place.

The journalist, with more than 6 years of experience as a writer and editor for important American newspapers and web sites, acted as the expert of the domain. He was asked to watch the newscasts and to identify the entities that best serve the objective of showing interesting additional information to an end-user. He was completely free to suggest any named entity he wanted.

Afterwards, 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. The final step consisted in conducting a crowdsourcing survey with the objective to gather information about the degree of interestingness of the entities for each newscast. Based on [53], 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 one of the newscasts. 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”.

3.2.4 Experimental Settings

Document retrieval. We rely on the Google Custom Search Engine (CSE) API service[8] by launching a query with the parameters specified by $q = [\mathrm{h}, \mathrm{t}]$. Apart from the query itself, the CSE engine considers other parameters that need to be tuned. First, due to quota restrictions, the maximum number of retrieved document is set to 50. 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 sub-domains where documents can be retrieved. This reduces the scope of the search task and, depending on the characteristics of the sources considered, 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 white lists while still considering the entire indexed Web. Based on this, in our study, we considered five possible values for this parameter:

   Google : search over the complete set of Web pages indexed by Google.

   L1 : A set of 10 internationals English speaking newspapers[9].

   L2 : A set of 3 international newspapers used in the gold standard creation.

[8] https://www.google.com/cse/all

[9] http://en.wikipedia.org/wiki/List_of_newspapers_in_the_world_by_circulation

L1+Google : Prioritize content in Newspaper whitelist but still consider other sites.

L2+Google : Prioritize content in Ground Truth's whitelist but still consider other sites.

2. Temporal dimension. This variable allows to filter out those documents which are not temporarily close to the day when the newscast was published. Assuming that the news item is recent 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 $Time_{Window} = [\mathrm{t}-d, \mathrm{t}+d]$. The reason why we expand the original event period is because documents concerning a news event are not always published during the course of action but some hours or days after. 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: two weeks and one week temporal windows.

3. Schema.org type. Google CSE makes possible to filter results according to a Schema.org type. For our experiments, we use the following settings: [NoFilter, Person & Organization]

Semantic Annotation. We use [34], 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[10]. We used it as an off-the-shelf entity extractor, using the offered classification model trained over newswire content.

Annotation Filtering and Clustering. After initial trials, it became obvious that there are many 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[11]. In our case, we keep only Person, Location, and Organization.

F2 : Filter out entities which are extracted with a confidence score falling under the 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.

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 pre-selection 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 the best results in the final MNDCG score.

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 the result section. Regarding the particular details of the orthogonal functions, we have proceeded as follow:

Popularity. We rely on Google Trends[12] 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 representativeness of the samples without compromising too much the validity of the selected values according to the time when the event took place. With the aim of being selective enough and keeping 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

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

[11] http://nerd.eurecom.fr/ontology

[12] https://www.google.com/trends

deduced by relying on the average score per entity type computed in the survey $\overline{Sgt}_{entityType}$. Organizations have gotten a higher weight ($Op_{expert} = 0.95$), followed by Persons ($Op_{expert} = 0.74$), and by Locations ($Op_{expert} = 0.48$) that are less well-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 ($Op_{expert} = 0$).

3.2.5 Evaluation Results

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 best performing strategies in our approach. More details are presented below.

Measures. 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 proposed 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 will be the same 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 ones.

– Average Normalized Discounted Cumulative Gain $MNDCG$ at N. The Normalized Discounted Cumulative Gain is a popular measure for evaluating Web search and related applications [7]. It is based on the assumption that there are different levels of relevance for the documents obtained in results. According to this, 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 documents in our gold standard are scored in terms of relevance for the user, we have mainly focused on the last measure since it provides a general judgment about the adequacy of the NSS generated. 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.

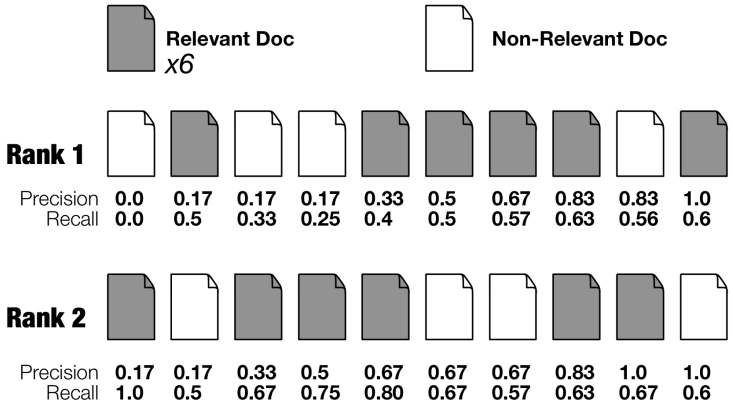

Figure 6: Inability of P/R for considering the order of the relevant documents: rankings 1 and 2 share the same Precision and Recall at 10.

Baselines.

1. Baseline 1: Former Entity Expansion Implementation. A previous version of the News Entity Expansion algorithm was already published in [33]. The settings were: Google as source of documents, temporal window of 2 weeks, no Schema.org selected, no filter strategy applied, and only frequency-based ranking function with no orthogonal appliances. The results are reported in the Table 6 under the run id *BS1*.

2. 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 of 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:

$$tf(e_i,d_j) = 0.5 + \frac{0.5 \times f_a(e_i,D)}{max\{f_a(e'_i,D):e'_i \in d_j\}}, idf(e_i,d_j) = log\frac{|D|}{\{d_j \in D:e_i \in d_j\}} \quad (1)$$

   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 the Table 6 under the run id *BS2*.

Launching the Experiments.

Table 6: Executed runs and their configuration settings, ranked by $MNDCG_{10}$

Run	Collection			Filtering	Functions			Result			
	Sources	T_{Window}	Schema.org		Freq	Pop	Exp	$MNDCG_{10}$	MAP_{10}	MP_{10}	MR_{10}
Ex0	Google	2W		F1+F3	Freq		✓	0.666	0.71	0.7	0.37
Ex1	Google	2W		F3	Freq		✓	0.661	0.72	0.68	0.36
Ex2	Google	2W		F3	Freq	✓	✓	0.658	0.64	0.6	0.32
Ex3	Google	2W		F3	Freq			0.641	0.72	0.74	0.39
Ex4	L1+Google	2W		F3	Freq		✓	0.636	0.71	0.72	0.37
Ex5	L2+Google	2W		F3	Freq		✓	0.636	0.72	0.7	0.36
Ex6	Google	2W		F1+F3	Freq			0.626	0.73	0.7	0.38
Ex7	L2+Google	2W		F3	Freq			0.626	0.72	0.72	0.37
Ex8	Google	2W		F1+F3	Freq	✓	✓	0.626	0.64	0.56	0.28
Ex9	L2+Google	2W		F1+F3	Freq		✓	0.624	0.71	0.7	0.37
Ex10	Google	2W		F1	Freq		✓	0.624	0.69	0.62	0.32
Ex11	L1+Google	2W		F3	Freq			0.623	0.7	0.72	0.37
Ex12	L2+Google	2W		F3	Freq		✓	0.623	0.68	0.66	0.35
Ex13	L2+Google	2W		F3	Freq	✓	✓	0.623	0.61	0.56	0.3
Ex14	L2+Google	2W		F3	Freq			0.62	0.69	0.74	0.4
Ex15	L1+Google	2W	✓	F1+F3	Freq		✓	0.617	0.69	0.66	0.34
Ex16	L2+Google	2W		F1	Freq		✓	0.616	0.68	0.62	0.32
Ex17	Google	2W	✓	F1+F3	Freq		✓	0.615	0.7	0.64	0.32
Ex18	L1	2W	✓	F3	Freq	✓	✓	0.614	0.65	0.6	0.32
Ex19	L1+Google	2W		F1+F3	Freq			0.613	0.72	0.72	0.38
Ex20	L1+Google	2W		F1+F3	Freq		✓	0.613	0.7	0.66	0.35
...	...	...	...	...	...	...	...	...	...	...	...
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

In Table 6, we present the top 20 runs for our approach in generating NSS, together with other 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 **0.662** and a MAP_{10} of 0.71, which are reasonably good in the document retrieval domain.

– Our approach performs much better than BS1 and BS2. The very low score of this last baseline is explained by the fact that traditional TF-IDF function is designed to measure the relevance of an item in the encompassing document and not with respect to a collection. In addition, the absence of filters drops drastically the score.

- Regarding the Document Retrieval step, we see that using Google as a single source or together with other white list gives better results than restricting only to particular white lists. The biggest T_{Window} of 2 weeks performs better in all cases, while the use of Schema.org does not bring anything back except when it is applied over the Gaussian function (see runs 78, 80, 82) where it turns to be an influential factor.
- The best Filtering 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 possible configuration.
- Popularity based function does not seem to improve significantly the results. However, a further manual study of the promoted entities has revealed that in fact, the method is bringing up relevant entities like for example *David Ellsberg*[13] for the query "Fugitive Edward Snowden applies for asylum in Russia". This entity is rarely mentioned in the collected documents, but *David Ellsberg*'s role in the newscast is quite 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.

3.2.6 Outlook and Future Work

In this section, we have presented an approach for automatically generating Newscast Semantic Snapshots. By following an entity expansion process that retrieves additional event-related documents from the Web, we have been able 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 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 a gold standard, which is also proposed in this section. The evaluation has showed the strength of this approach, holding an $MNDCG_{10}$ score of 0.666, outperforming the two studied baselines.

Future research includes tailoring the entity ranking functions to particular news categories: sport, politics, regional, international, opinion. We are investigating the role of entity relations in generating of the Newscast Semantic Snapshot: usually, entities are linked by tight relations extracted from a knowledge base, or simply from the documents collected, in order to generate a directed graph of entities instead of a list. We also plan to refine the ranking process, applying supervised techniques (Learning to Rank) that tailor the solution on particular domains.

3.3 Linked Hypernyms Dataset

The Linked Hypernyms Dataset (LHD) Core dataset[14], associates DBpedia entities (corresponding to Wikipedia articles) with a type which is obtained by parsing the first sentences of the respective Wikipedia article. This dataset is powering the THD system, which is evaluated in Subs. 3.4.

In this section, we report on the accuracy of the LHD Core dataset and compare it with the accuracy of types in DBpedia. To increase the coverage of LHD, there is an extended version of the dataset called LHD *Inferred*, which uses statistical and machine learning techniques to assign additional entities a type from the DBpedia Ontology. These algorithms are described in [24, 58], here we report on their accuracy.

3.3.1 Generating the gold standard

In order to evaluate the general quality of types in the LHD dataset and to compare it with the quality of types in DBpedia we generated a gold standard dataset using the crowdsourcing platform Crowdflower[15]. The CrowdFlower, similarly to the well-known *Amazon Mechanical Turk* (AMT), is an online

[13] http://en.wikipedia.org/wiki/Daniel_Ellsberg
[14] ner.vse.cz/datasets/linkedhypernyms
[15] http://www.crowdflower.com

Figure 7: The user interface of the CrowdFlower taxonomy annotation tool. The user can navigate through the taxonomy either by clicking on a concept, which shows its subtypes, or by fulltext search, which shows all concepts with substring match in the concept name along with the full path in the ontology.

labor market. The advantage of the use of a third-party operated service is the high credibility of the resulting annotations and easy repeatability of the experiment setup.

We asked the annotators to assign *the most specific category* (categories) from the presented *taxonomy of categories* for each Wikipedia article describing certain entity from the given list. The taxonomy used corresponds to the DBpedia 2014 ontology, which contains almost 700 DBpedia types. To collect the judgments, we used the advanced taxonomy annotation tool offered by the CrowdFlower platform, which enables the annotators to quickly browse through the taxonomy using fulltext queries issued against a taxonomy lookup service hosted at UEP. A screenshot of the tool is present at Figure 7.

The CrowdFlower platform allows a wide range of setting for controlling the quality of the work done by its workers. Our setup was as follows:

- Only workers residing in the following countries were eligible
  - English dataset: Australia, Canada, Denmark, Germany, Ireland, Netherlands, Sweden, United Kingdom and United States.
  - German dataset: Germany
  - Dutch dataset: the Netherlands
- The workers were Level 1 Contributors, which are described by the CrowdFlower service as accounting for 60% of monthly judgments and maintaining a high level of accuracy across a basket of jobs.
- Amount of 0.02 USD was paid for each annotated entity to a worker.
- The workers were given a quiz before starting a task with minimum of four test questions (entities to annotate). Only workers with accuracy of 30% or higher could continue in the task.
- To maintain high accuracy, additional test questions were asked as the workers were completing their job.
- A speed trap was put in place that eliminated workers who took less than 10 seconds to complete a task.

Each entity was typically annotated by three to four annotators. The CrowdFlower platform ensured that the annotations from workers who failed the test questions were replaced by untainted annotations.

The gold standard for given entity consists of all categories that were assigned by at least two annotators to the entity.

3.3.2 Evaluation Metrics

Performance is measured by four evaluation metrics based on accuracy. Evaluation metrics apply four variants of gold standard (*GS*).

– Acc_{exact} evaluation metric uses original gold standard GS_{exact}.

– $Acc_{dir\_supertypes}$ uses gold standard $GS_{dir\_supertypes}$ which extends GS_{exact} with all *direct* supertypes. $GS_{dir\_supertypes}$ is defined as follows:

$$\{x|x \in directSuperTypes(y), y \in GS_{exact}\} \cup GS_{exact} \quad (2)$$

– $Acc_{supertypes}$ uses gold standard $GS_{supertypes}$ extending GS_{exact} with *all* supertypes. $GS_{supertypes}$ is defined as follows:

$$\{x|x \in superTypes(y), y \in GS_{exact}\} \cup GS_{exact} \quad (3)$$

– $Acc_{[sub|super]types}$ uses gold standard $GS_{[sub|super]types}$ extending $GS_{supertypes}$ with all subtypes. $GS_{[sub|super]types}$ is defined as follows:

$$\{x|x \in subTypes(y), y \in GS_{exact}\} \cup GS_{supertypes} \quad (4)$$

Since there is a total order between the variants of the gold standard, $GS_{exact} \subseteq GS_{dir\_supertypes} \subseteq GS_{supertypes} \subseteq GS_{[sub|super]types}$, it follows that: $Acc_{exact} \leq Acc_{dir\_supertypes} \leq Acc_{supertypes} \leq Acc_{[sub|super]types}$.

Since the variants of the gold standard, except GS_{exact}, can have more than one type per entity, type assignment by a classifier is considered as correct if this type is included in a set of types associated with an entity in the gold standard.

3.3.3 Evaluation results

We performed evaluation on three languages: English, German and Dutch. For English the annotation was performed on 1165 randomly drawn articles from English Wikipedia. In total there were 1033 entities with assigned agreement category in the gold standard for English. Additionally, 22 entities were assigned to the 'not found' category, 47 entities the 'disambiguation page' category, and in 63 cases there was no agreement.

For German the annotation was performed on 300 randomly drawn articles from German Wikipedia. In total there were 248 entities with assigned agreement category in the gold standard for German. Additionally, 15 entities were assigned to the 'not found' category, 19 entities the 'disambiguation page' category, and in 18 cases there was no agreement.

For Dutch the annotation was performed on 239 randomly drawn articles from Dutch Wikipedia. In total there were 222 entities with assigned agreement category in the gold standard for Dutch. Additionally, 8 entities were assigned to the 'not found' category, 6 entities the 'disambiguation page' category, and in 3 cases there was no agreement. For the evaluation we used the most up-to-date version of the DBpedia Ontology (2014).

Figure 7 shows that the results for LHD Core (denoted as LHD 1.0 in [25]) and DBpedia are very similar for all three languages. There is just one larger difference in terms of Acc_{exact} for German, where LHD Core achieves improvement of 42% over DBpedia. By inspecting German DBpedia it seems that this can be explained by the fact that German DBpedia assigns mostly general concepts from the DBpedia ontology, while the more specific concepts are assigned from the German DBpedia ontology, which was not involved in the gold standard and evaluation.

In general, while LHD Core has exact types with higher accuracy than DBpedia (e.g. 19% improvement in Acc_{exact} for English), DBpedia has higher accuracy in the relaxed metrics but always only by a thin margin (e.g. of 3% in $Acc_{supertypes}$ for English). This result shows that DBpedia Ontology types assigned by the lexico-syntactic pattern based LHD Core extraction framework are of equal quality with types extracted from infoboxes.

However, it should be noted that LHD Core framework has much smaller coverage than DBpedia. In our evaluation datasets, it covered only 345 entities out of the 1033 for English (cf. 733 by DBpedia), 37 entities out of the 248 for German (cf. 165 by DBpedia) and 53 entities out of 222 for Dutch (cf. 180 by DBpedia) with DBpedia Ontology type.

For English, we tried to extend the number of entities with assigned DBpedia Ontology type using the following three algorithms: Statistical Type Inference (STI) algorithm [24], hierarchical Support Vector Machines (hSVM) classifier [58], which exploits the words in the short abstracts and links in the article categories to perform the classification, and a fusion of the two algorithms. The results depicted at Figure 8 show that STI has the highest accuracy from the type inference approaches in the Acc_{exact} metric, however, in the more relaxed metrics it is outperformed by hSVM. The types generated by the machine-learning algorithms are published as the LHD Inference dataset.

Table 7: DBpedia and LHD is evaluated on the English (1033 entities), German (248 entities) and Dutch (222 entities) gold standards. For hSVM we report result for the β run and for the hSVM-STI fusion from the "propβ-1" run described in [58].

Classifier	Acc_{exact}	$Acc_{dir\_supertypes}$	$Acc_{supertypes}$	$Acc_{[sub\|super]types}$
DBpedia (en) (on 733 entities)	.526	.690	.785	.878
LHD Core (en) (on 345 entities)	.655	.761	.787	.856
DBpedia (de) (on 165 entities)	.296	.521	.854	.915
LHD Core (de) (on 37 entities)	.702	.783	.783	.891
DBpedia (nl) (on 180 entities)	.638	.727	.877	.911
LHD Core (nl) (on 53 entities)	.655	.761	.787	.856

Table 8: Evaluation of STI and hSVM algorithms on the English dataset (1033 entities). For hSVM we report result for the β run and for the hSVM-STI fusion from the "propβ-1" run described in [58].

Classifier	Acc_{exact}	$Acc_{dir\_supertypes}$	$Acc_{supertypes}$	$Acc_{[sub\|super]types}$
hSVM	.275	.457	***.690***	.747
STI (on 324 entities)	***.438***	.462	.487	.666
STI/hSVM-fusion (699 STI types)	.429	***.566***	.666	.757

3.3.4 Outlook: automating LHD Inference dataset generation

Since the results by the LHD Inference dataset are promising, we carried out steps in order to automate the generation of the dataset to improve the odds of the lasting contribution to the community.

The core component of the LHD Inference dataset is a machine learning algorithm. In our earlier work we obtained promising results with the use of association-rule based classifier in a related text categorization task. We developed a new web service for association rule learning within our EasyMiner framework, which is based on the high performance *arules* library [4].

This web service can be used to build an association rule-based classifier that could possibly replace the third-party hSVM component in the Inference dataset generation workflow. The advantage of this approach is that there would be one in-house machine learning service that could meet the needs of multiple work packages as association rule learning is already used as a core WP4 component, however using a slower underlying implementation [46].

Table 9 provides a preliminary benchmark of our brCBA classifier [26], which builds classification models from association rules, with other commonly used classifiers, including linear SVM models. The experiments were performed on 28 datasets used to train the hSVM classifier in [58]. These datasets were derived from the short abstracts of DBpedia resources, which correspond to first sentences of

Table 9: Comparison of linear SVM, brCBA with other common classifiers

metric	Naive B.	SVM (linear)	SVM (poly)	Log Reg	brCBA
accuracy	0.74	**0.90**	0.87	0.83	0.80
run time	4s	3m 24s	10m 6s	2m 54s	11m 53s

Wikipedia articles. Ten percent of each dataset was used for testing, the rest for training (stratified selection). The feature set was pruned by removing features with less than 0.1 standard deviation in each dataset. No parameter tuning for any of the classifiers was performed, the default values from the RapidMiner[16] implementation of the respective reference classifier was used, i.e.:

- **SVM linear kernel**: C=0.0, $\varepsilon = 0.001$, shrinking applied.
- **SVM polynomial kernel**: degree 3, $\varepsilon = 0.001$, C=0.0, $\gamma = 0.0$, shrinking applied.
- **Logistic regression**: dot kernel used, convergence $\varepsilon = 0.001$, C=1.0, value scaling applied.
- **brCBA** minConfidence=0.01, minSupport=0.01, maxRules=2000, rule length from 1 to 5, R/arules used for association rule learning

The results depicted in Table 9 show that SVMs with linear kernels provide the best accuracy and at the same time have acceptable run time (aggregate result for training and testing phase). This result underpins the choice of linear SVM as the base classifier for the next generation of the LHD Inference dataset.

It should be noted that the runtime and accuracy of the result of the brCBA classifier is highly sensitive to the setting of the minConfidence parameter. Higher accuracy than reported in Table 9 by decreasing the threshold, however, at the cost of prolonged learning time. While the results of brCBA do not match those obtained with the best model, we hypothesize that additional improvement can be obtained by changes to the algorithm. Run time can be considerably decreased as well, as the used brCBA implementation did not undergo performance optimization.

3.4 THD Entity spotting and linking evaluation

In this evaluation we conducted several experiments to evaluate the performance of spotting and linking entities with the THD system[17], which was developed for content annotation in LinkedTV as described in the preceding deliverable D2.6.

The evaluation was performed on transcripts of LinkedNews and LinkedCulture videos, which were annotated by the Dutch Sound & Vision archive and the German RBB broadcaster. In the following we describe the annotation process, the evaluated entity spotting and linking methods, and we present the results. Additional evaluation of THD on English content was performed within the TAC 2014 Entity linking task, the results provide a benchmark against the state-of-the-art in entity detection and linking. A summary report is provided in Subs. 5.2.

3.4.1 Groundtruth Datasets

The subtitles for processing were selected and annotated by professionals from Sound&Vision and RBB. The professionals were asked to select sequences of videos lasting around 15 min in total. Next, we generated an annotation sheet containing the subtitles content. From the provided subtitles, we excluded words which indicate a specific sound or strange happening (e.g. laughing or audience gasps). The annotators were given strict annotation guidelines and their task was to:

- identify each entity occurrence in the given subtitle by providing its surface form,
- provide URL of Wikipedia page which describes the entity,
- provide information whether the entity refers to a named entity (proper nouns) or common entity (nouns with modifier), and
- provide information whether an entity is relevant or not for the video episode.

[16] `http://rapidminer.sourceforge.net`

[17] Available at `http://ner.vse.cz/thd` and `http://entityclassifier.eu`

Table 10: Size metrics for the Sound and Vision and RBB datasets.

Dataset	Num. of entities	Named entities	Common entities	Video length
Sound and Vision	289	42	247	14 min 55 sec
RBB	397	75	322	13 min 50 sec

3.4.2 Evaluated Approaches

3.4.2.1 Entity spotting Entityclassifier.eu supports two approaches for entity spotting, which were evaluated.

- **Pattern based (GRAM)** - a manually crafted lexico-syntactic patterns written as JAPE grammars. We have developed grammars that can be applied to detect common and named entities in Dutch, German and English texts. The grammars, can be used for detection of both, common and named entities.
- **Conditional Random Fields (CRF)** based - an entity spotting approach based on the state-of-the-art Conditional Random Fields (CRF) model. The model is trained on the CoNNL 2003 dataset. See Section 5.2.3.2 for more information.

3.4.2.2 Entity linking We evaluated following entity linking approaches.

- **Basic Lucene index (LB)** - uses a specialized Lucene index, which extends the Apache Lucene search API. It primarily ranks pages based on the number of backlinks and the Wikipedia articles' titles. It uses the detected entity name when performing the entity linking.
- **Lucene Skip Disambiguation (LSD)** - same as the previous, only as a correct link it considers the first non-disambiguation page.
- **Surface Form Index (SFI)** - this approach approach uses a custom entity candidate index. The candidate index contains all surface forms found in Wikipedia articles together with their candidates.
- **Surface Form Similarity (SFS)** - this approach first performs entity linking with the SFI and LSD. And then, the article with the most similar title to the entity surface form is considered as correct. For measuring similarity we opted for the widely used Jaro-Winkler string similarity measure.

3.4.3 Results

The evaluation was performed using GERBIL [51], a benchmarking framework for entity annotation and disambiguation tools. For the evaluation we run the *A2KB* experiment [51]. We report on thee metrics computed by the GERBIL framework: micro precision, micro recall and micro F-measure. The macro measures are not reported, since scores obtained on such a short documents (one document corresponds to one subtitle fragment) are not very meaningful, as many of the documents have no ground truth annotation, which results in increased macro score measures.

The results reported by the GERBIL framework in the micro recall measure (and consequently for the F1 measure) in Table 11-16 were updated in the final published version of the deliverable as foreseen in the initially submitted version.[18]

The details of the metrics can be found in the description of the underlying BAT framework [6]. Table 11 shows the results from the evaluation on the Dutch dataset with focus on the named entities only. The results show that the best F1 micro score 0.2737 was achieved by the approach which uses Conditional Random Field (CRF) model for entity spotting and a custom Surface Form Index (SFI) for entity linking.

Table 12 shows the results from the evaluation on the Dutch dataset which contains both, named and common entities. In this evaluation, the best micro 0.2912 F1 score was achieved by the Surface Form Similarity (SFS) based approach.

[18] A bug in the evaluation framework was discovered by the deliverable authors shortly before the deliverable submission due time and reported to the GERBIL community, which confirmed it and later fixed it.

Table 11: Evaluation results for Dutch. An experiment type A2KB - named entities only.

spotting/linking	Micro F1	Micro P	Micro R
GRAM/LB	0.2247	0.1887	0.2778
GRAM/LSD	0.2247	0.1887	0.2778
GRAM/SFS	0.2273	0.1923	0.2778
GRAM/SFI	0.3	0.2727	0.3333
CRF/LB	0.2376	0.1846	0.3333
CRF/LSD	0.2376	0.1846	0.3333
CRF/SFS	0.2376	0.1846	0.3333
CRF/SFI	**0.2737**	0.2203	0.3611

Table 12: Evaluation results for Dutch. An experiment type A2KB - named and common entities.

spotting/linking	Micro F1	Micro P	Micro R
GRAM/LB	0.2841	0.2396	0.3488
GRAM/LSD	0.2841	0.2396	0.3488
GRAM/SFS	**0.2912**	0.2476	0.3535

Table 13 shows the results from the evaluation on the Dutch dataset which contains common entities only. The results show that the best micro F1 score 0.2995 was achieved by the Surface Form Similarity (SFS) approach.

Table 13: Evaluation results for Dutch. An experiment type A2KB - common entities only.

spotting/linking	Micro F1	Micro P	Micro R
GRAM/LB	0.287	0.2423	0.352
GRAM/LSD	0.287	0.2423	0.352
GRAM/SFS	**0.2995**	0.2549	0.3631

Table 14 shows the results from the evaluation on the German dataset with focus on the named entities only. According to the results the method which uses the Conditional Random Fields (CRF) method for entity spotting and Surface Form Index (SFI) for linking achieved the best micro F1 score of 0.5047.

Table 14: Evaluation results for German. An experiment type A2KB - named entities only.

spotting/linking	Micro F1	Micro P	Micro R
GRAM/LB	0.3654	0.4872	0.2923
GRAM/LSD	0.4038	0.5385	0.3231
GRAM/SFS	0.3883	0.5263	0.3077
GRAM/SFI	0.396	0.5556	0.3077
CRF/LB	0.4545	0.5556	0.3846
CRF/LSD	0.4909	0.6	0.4154
CRF/SFS	0.4771	0.5909	0.4
CRF/SFI	**0.5047**	0.6429	0.4154

Table 15 shows the results from the evaluation on the German dataset which contains both, named and common entities. It can be observed that the best micro F1 score 0.4658 was achieved by the method which performs entity linking based on the Surface Form Similarity (SFS) method.

Table 16 shows the results from the evaluation on the Dutch dataset which contains common entities only. In this experiment, the best micro F1 score of 0.4495 has been achieved by the Surface Form Similarity (SFS) linking method.

Table 15: Evaluation results for German. An experiment type A2KB - named and common entities.

spotting/linking	Micro F1	Micro P	Micro R
GRAM/LB	0.4238	0.3903	0.4636
GRAM/LSD	0.3906	0.3597	0.4273
GRAM/SFS	**0.4658**	0.4315	0.5061

Table 16: Evaluation results for German. An experiment type A2KB - common entities only.

spotting/linking	Micro F1	Micro P	Micro R
GRAM/LB	0.411	0.3598	0.4792
GRAM/LSD	0.3625	0.3173	0.4226
GRAM/SFS	**0.4495**	0.3954	0.5208

3.5 THD Entity salience evaluation

The task of identification of *salient entities* aims at finding the set of entities that play an important role in the story described in the document. Figure 8 illustrates the methodology towards identification of salient entities. As shown in the figure, there are two sources for creating features for training. One source is the document and the set of features with a *local scope* derived from the information available within the document. The second source are knowledge graphs and a set of features with the *global scope* derived from information available outside the scope of the document.

In D1.4, we described the new experimental *entity salience* feature of THD which can provide salience estimation of the entity within the document. Since the algorithm works on the semantic level of entities (as opposed to words or terms which were focus of the WP1 processing) we further extended this algorithm with new features and learning algorithms, and incorporated it into the WP2 THD tool and API, which is available via `http://Entityclassifier.eu`.

In this deliverable, we provide an extended evaluation of the updated algorithm. We report on the results achieved on a large scale corpus containing over a million of annotations (New York Times corpus) in addition to the Reuters-128 corpus.

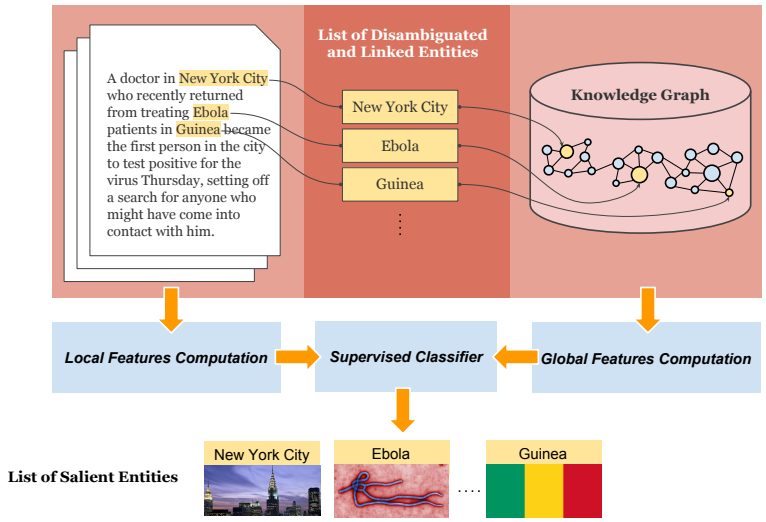

Figure 8: Schematic overview of the methodology for identification of salient entities.

3.5.1 Groundtruth Datasets

In this evaluation, we re-used the Reuters dataset which was described in D1.4. The Reuters-128 salience corpus is an extension of the entity linking corpus Reuters-128, part of the N3[19] datasets

[19] `http://aksw.org/Projects/N3nernednif`

Table 17: Size metrics for the Reuters-128 and New York Times entity salience corpora.

Corpus	Documents	Entity mentions	Unique entities	Entities linked with DBpedia	Salient entities	Not salient entities
Reuters-128	128	4,429	2,024	3,194	804 (18%)	3,625 (82%)
NYT (train)	100,834	1,990,355	173,462	1,990,355	255,572 (13%)	1,734,783 (87%)
NYT (eval)	9,706	184,457	42,251	184,457	24,438 (13%)	160,019 (87%)

collection [35]. The Reuters-128 dataset is an English corpus in the NLP Interchange Format (NIF) and it contains 128 economic news articles. The dataset contains information for 880 named entities with their position in the document (beginOffset, endOffset) and a URI of a DBpedia resource identifying the entity.

Since the dataset only provides information about named entities found in the corpus, we further extended the dataset with common entities. To this end, we used our Entityclassifier.eu NER tool to enrich the dataset with common entities. This resulted in additional 3551 common entities.

Furthermore, aiming to obtain a gold standard entity salience judgments we used a crowdsourcing tool to collect judgments from non-expert paid judges. For each named and common entity in the Reuters-128 dataset, we collected at least three judgments for each entity from annotators based in 15 different countries, including English-speaking countries, such as United Kingdom, Canada and United States. We also manually created a set of test questions, which helped us to determine contributor's trust score. Only judgments from contributors with trust score higher than 70% were considered as trusted judgments. If the trust score of a contributor falls bellow 70%, all his/her judgments were disregarded. In total we collected 18,058 judgments from which 14,528 we considered as "trusted" and 3,530 as "untrusted" judgments. The interannotator agreement, in cases where the annotators judgments differed, was determined by the crowdsourcing tool.

Additionally, we used also the New York Times dataset. The salience annotations in the NYT dataset have been automatically generated by aligning the entities in the abstract and the document and considering that every entity which occurs in the abstract is salient. The New York Times dataset consists of two partitions. A *training* partition which consists of about 90% of the data, and a testing partition consisting of the remaining 10%. The NYT dataset [18] provides only information about the begin and end index of the entities, the entity name, document ID and salience information. The annotations are shared without the underlying document's content. Thus, we have converted only the available information in the NIF format; without the documents' content.

Annotation statistics for both the crowdsourced Reuters-128 and the converted New York Times dataset are presented in Table 17.

3.5.2 Baseline methods

In the experiments we consider the following three baseline methods against which we compare our method.

- *Positional Baseline.* An entity is considered as salient only if the begin index of the first occurrence in the document is within the first 100 characters. This also corresponds to a typical sentence length, which in average is around 100 characters long. Since the entities of the Reuters-128 dataset are classified with three classes {most salient, less salient, not salient}, an entity is considered as "most salient" only if its first occurrence is within the first 100 characters in the document. The entities in the NYT dataset are classified with two salient classes {salient, not salient}, and an entity is considered as "salient" if it occurs withing the first 100 characters in the document.

- *Majority Vote Baseline.* This baseline method always predicts the majority class. For the Reuters-128 and the NYT datasets that is the "not salient" class.

- *Entity Frequency Baseline.* This baseline method is learning from the frequency of entity occurrence in a document. As a learning algorithm for this method we used the Random Forest decision tree learning algorithm.

3.5.3 Learning Algorithms

We experimented with various learning algorithms to find the most suitable one for the task of learning entity salience. In D1.4, we reported results for Support Vector Machines (SVM) with polynomial kernel,

Naive Bayes (NB) and k-Nearest Neighbor (k-NN) with Euclidean distance function and k=1. In this deliverable we extended the list of learning methods including a C4.5 decision tree classifier and a Random Forest tree classifier.

We also extended the list of features used for learning. The new features are listed in Table 18. This list of features were derived from information available in English DBpedia.

Table 18: Extended list of features.

Feature	Type	Description
`in-degree`	numeric	Number of incoming links exits for the given entity.
`out-degree`	numeric	Number of outgoing links exits for the given entity.
`num-triples`	numeric	Total number of triples describing the entity.
`num-props`	numeric	Number of properties describing the entity.
`object-props`	numeric	Number of object properties describing the entity.
`datatype-props`	numeric	Number of datatype properties describing the entity.

3.5.4 Results

For the evaluation we used the two created NIF entity salience corpora, the Reuters-128 and the NYT. For the Reuters-128 dataset we performed ten-fold cross validation by partitioning the dataset into ten equal partitions and performing ten cross-validations while training on nine partitions and one for validation. Since the NYT already has been split into training and testing partition, in our evaluation, we have used these partitions for training and testing.

The same set of learning algorithms has been evaluated on the two entity salience corpora: the Reuters-128 and the NYT corpus. The results from the experiment are presented in Table 19.

The evaluation measures we consider in our experiments are *precision*, *recall* and *F–measure*. We report weighted average across all classes, where the weight corresponds to the proportion of instances in that class (micro average).

The results show that the best performing algorithm for both datasets is the Random Forest decision tree-based classifier with F1 0.607 for the Reuters-128 and 0.898 for the NYT dataset. The second best performance has the C4.5 decision tree based classifier with 0.586 F1 for the Reuters-128 and 0.897 for the NYT dataset. The worst performance for the Reuters-128 is the NaiveBayes classifier with 0.391 F1, and the k-NN classifier with 0.858 for the NYT dataset. For comparison, the Random Forest compared to NaiveBayes shows improvement of nearly 55% for the Reuters-128, and 5% compared to the k-NN for the NYT dataset.

It can be concluded that the decision tree based classification algorithms are more suitable for learning entity salience than the instance-based learning algorithms (k-NN), probabilistic classifiers (NaiveBayes) or kernel-based classifiers (SVM). Since in this experiment the Random Forest algorithm achieves best results, in the following experiments that compare our results against the baseline approaches, we use the Random Forest classifier.

Table 19: Results for different learning algorithms. † - learning using SVM on the NYT corpus takes more than 24 hours.

ML algorithm	Reuters-128			New York Times		
	Precision	Recall	F1	Precision	Recall	F1
Naive Bayes	0.518	0.488	0.391	0.862	0.879	0.866
SVM†	0.534	0.504	0.416	/	/	/
k-NN	0.566	0.564	0.565	0.857	0.860	0.858
C4.5	0.586	0.586	0.586	0.897	0.906	0.897
Random Forest	**0.612**	**0.608**	**0.607**	**0.899**	**0.908**	**0.898**

Table 20: Evaluation results for different baseline methods for the class "*most salient*" for the Reuters-128 dataset and "*salient*" for NYT dataset.

Method	Reuters-128 (most salient class)			New York Times (salient class)		
	Precision	Recall	F1	Precision	Recall	F1
Positional baseline	0.518	0.488	0.391	0.620	0.262	0.369
Majority vote baseline	0.000	0.000	0.000	0.000	0.000	0.000
Entity frequency baseline	0.437	0.133	0.204	0.706	0.305	0.426
Our with Random forest	**0.693**	**0.516**	**0.592**	**0.611**	**0.629**	**0.620**

The scores reported in Table 19 are computed as "weighted average" for all the classes. Since we are interested in representing the aboutness of the Web documents in terms of salient entities, we report also the scores for the salient classes. Table 20 summarizes the results for the "most salient" class for the Reuters–128 dataset and the "salience" class for the NYT dataset.

The results show that our model outperforms all other considered baseline methods. For the Reuters–128 dataset our model based on the Random Forest algorithm achieves 0.592 F1, while the positional baseline 0.391 F1 and the entity frequency baseline 0.204 F1. Similarly, for the NYT dataset our method achieves 0.620, while the positional baseline is at 0.369 F1, and the entity frequency baseline at 0.426 F1. The *entity frequency baseline*, which is very close to a typical TF-IDF keyword extraction approach, is improved by our model by 290% for the Reuters–128 dataset and by 45% for the NYT dataset.

Since the Random Forest learning algorithm shows best performance for learning entity salience, we also use Random Forest for learning entity salience in Entityclassifier.eu.

3.5.5 Outlook: Entity-based clustering

As an outlook for further development of LinkedTV technologies, we investigated the utility of entity-annotated text for text clustering. While not incorporated into the WP2 pipeline, within the LinkedTV context the availability of clustering could serve the following purposes: i) the most discriminative features of each of the clusters can be used as "keyword" labels describing the documents in the cluster, ii) since the resulting cluster labels are LOD entities this will provide unambiguous semantics for the clusters. Additionally, we hypothesize that the entity-based text representation could also improve the quality of the clustering as opposed to the standard bag-of-words representation.

As the dataset, we used the Reuters-21578 text categorization collection. The Reuters-21578 collection contains 21,578 documents, which are assigned to 135 different categories (topics). Example topics are "earn" or "wheat". On average, one document belongs to 1.3 categories. For the experiments, we used only a subset consisting of 9,809 documents which are assigned to the ten most frequently populated categories (same list of categories was used e.g. in [2]).

Two versions of the dataset were prepared: the bag of words (BoW) and bag of entities (BoE). In the BoW version, the following preprocessing was performed: all terms were converted to lower case, numbers were removed, punctuation was removed, stop words were removed, whitespace was stripped and the documents were stemmed. The resulting words were used as features (attributes).

In the BoE version, the text was analyzed with THD and represented with the list of entities (DBpedia resources) and their types (DBpedia Ontology concepts) which THD returned. The features in both BoW and BoE documents were assigned a TF-IDF score in each document. Finally, for each version, ten datasets were created by selecting 10, 50, 100, 200, 500 features with highest TF scores.

For our experiments, we employed the bisection K-Means clustering algorithm implemented within the LISp-Miner data mining system (`lispminer.vse.cz`). The LISp-Miner system was adapted and extended to perform this analysis as described in the technical report [54], which contains additional details on algorithm settings and parameters. The results depicted in Table 21 show that the entity-based representation leads to consistently higher quality of clustering than the standard bag-of-words approach. The results also indicate that the BoE representation is more condensed, achieving peak cluster quality at only 200 term vector size. This demonstrates the utility of entity representation (and the THD tool) for the clustering task.

3.6 Topic labelling evaluation

The topic labelling service associates content with topics from the LUMO ontology (`http://data.linkedtv.eu/ontologies/lumo/`, [50]) Topics subhierarchy, as described in the previous deliverable D2.6. In

Table 21: Evaluation of clustering results using cluster quality metric defined in [54]

term vector length	**10**	**50**	**100**	**200**	**500**
Bag of entities (BoE)	0.43	0.47	0.52	**0.56**	0.54
Bag of words (BoW)	0.41	0.42	0.43	0.44	0.52

D2.6, the performance of the service for video chapters of LinkedTV content was evaluated, while in this experiment the focus is set on evaluating the service over enrichments of LinkedTV content. The goal of the evaluation was to measure the accuracy and coverage of the topics extracted for enrichments.

The dataset used was the set of enrichment articles identified during the user trials of WP6 (cf. D6.5). Therefore, it consisted of 46 articles related to all the chapters of five RBB news shows, for the Linked News scenario, and of 82 articles, related to all the chapters of three TKK shows, for the Linked Culture scenario.

3.6.1 Experiment setup

The topics detection was based on entities automatically extracted from the text of the articles. The steps to extract these entities are listed below.

1. The service received the HTML content of the web page for each web article via the trials' web-player. It is worth noticing that the player's embedded content proxy retrieved the contents of the page, stripped it of irrelevant text (e.g. menus, sidebars), and stored it locally.

2. The body text of the stripped HTML was run thought the THD Entity Extraction service, in order to extract DBPedia entities (resources) from the text. The salience score of extracted entities a) was employed in order to prune the less significant entities and b) for the remaining entities, it was conveyed across the next steps to take into account in the final step of the topic detection.

3. The THD Entity Classification service was employed in order to retrieve DBPedia types (from DBPedia Ontology and/or the Linked Hypernyms Dataset) for the extracted entities.

4. For the DBPedia ontology types extracted in the last step, a custom mechanism was built to filter out the more generic types per entity (e.g. "Agent" for every person that appears in the texts) and keep only the most specific and characteristic types for an entity (e.g. 'Politician' that gives an outlook of the topic of the text), where applicable.

5. These types were translated to LUMO classes, again where applicable, via the LUMO wrapper service and based on the LUMO mappings ontology (cf. D2.4 for more details). This step is necessary in order to bring the entities into the LUMO concept space, where relations between types of agents, objects, events and their respective topics can be retrieved.

The final input for the topic labelling service was the retrieved LUMO classes, along with a degree per class that represented the salience of the original entity that each class derived from. Essentially, the degree represented the relevance of the class to the article at hand.

Ultimately, the LiFR reasoner [49], the core of the topic labelling service as described in D2.4, was employed in order to retrieve the topics related with the input classes within the LUMO ontology for each article. The topics retrieved carried the salience degree of the classes they derived from, thus denoting the relevance of each topic to the article.

3.6.2 Evaluation setup

For the evaluation of the topics retrieved, one person was called to rate the results per scenario, assuming the hypothetical role of a content editor for a broadcaster. For the Linked News (i.e. German language content), a colleague from partner MODUL served as the editor, and for Linked Culture (i.e. Dutch language content), a colleague from partner S&V served as the editor.

Each evaluator was presented with an online form containing a) information about all articles and b) their respective topics, and were asked to evaluate the precision and recall of the topics per article based on their informed opinion.

In more detail, the information the evaluators were presented with per article was:

a. The **URL of the article**.

b. The **body text of the article**, as it was presented in the trials' webplayer, the specific day it was crawled to be presented. I.e. if it was a dynamic web page that is refreshed hourly/daily/weekly (like e.g. the home page of a sports news site) and the evaluators clicked on the link, they might see content other than the one they were displayed with in this form. This is because the webplayer visited the page at some point during trials' preparations, it parsed the HTML of that time point and kept it stored, while today the contents of the page may have changed. The topic labelling service detected topics on the stored contents. This is also the reason why the form presented the body text (instead of just letting the evaluators click on the URL), so they can assess the topics based on the actual (static) piece of text that they represent.

c. The collection of **topics** that describe the contents of the corresponding body text along with their respective **degrees**. The degrees were in the [0, 1] scale. So if there is a topics-degree pair "arts - 1.0", this means that the topic "arts" was found to be 100% relevant to the text. If, in the same article, there is a topics-degree pair "science - 0.3", this means that the topic "science" was also found in the text, but it has a medium-to-low relevance (30%).

A snapshot of the form can be seen in Figure 9.

ARTICLE URL	SOURCE HTML	TOPICS	TOPICS DEGREE	TOPICS PRECISION (0-5)	TOPICS RECALL (0-5)
http://www.tagess	Bauprojekt City West : Investor Hines startet Abriss der Schmuddelecke am Zoo Bild vergrößernVom Aufschwung überrollt. Hinter dem Segway-Fahrer wird an der Joachimsthaler Straße die als Beate-Uhse-Erotikmuseum, Aschinger- und Leineweber-Haus bekannte Passage abgerissen. - Foto: Doris Spiekermann-Klaas Am Bahnhof Zoo geht die Neugestaltung der Berliner City West weiter, die Passage in der Joachimsthaler Straße weicht dem geplanten Geschäftshaus des US-Investors Hines. Seit Dienstag ist es unübersehbar: Neben dem Bahnhof Zoo hat der Abriss der als Schmuddelecke verrufenen Passage in der Joachimsthaler Straße begonnen, die vor allem durch das Beate-Uhse-Erotikmuseum bekannt war. Nun sind Bauwagen im Einsatz, je eine Fahrspur in der Kant- und der Joachimsthaler Straße wurde für den Verkehr gesperrt. Bis Ende 2017 baut der US-Investor Hines für etwa 130 Millionen Euro ein neues sechsstöckiges Geschäftshaus mit Läden, Lokalen und Büros. Jetzt sind alle Mieter raus Die letzten alten Geschäfte schlossen am Sonnabend, darunter	architecture art sports politics arts_culture_entertainment music	1.0 1.0 0.45 0.45 0.35 0.35		

Figure 9: Topics Labelling evaluation form

Subsequently, the evaluators were asked to give their expert opinion on the precision and recall of topics presented by providing a ranking score in a scale of 0-5 for all topics retrieved per article, for each measurement (precision, recall), taking also into account the topics' degrees. In order to interpret the scores and the criteria for ranking, the reader should refer to the instructions towards the evaluators, available in the Annex, section 8.

It is worth noticing that, regarding recall, we did not want to restrain the evaluators to take into account the existing LUMO Topics subhierarchy, but rather provide their unbiased opinion. This way, while co-estimating recall of enrichment entities in the analysis of the results, we would be able to identify the coverage that LUMO offers over topics.

3.6.3 Results and outlook

The average precision, average recall and f-measure scores is displayed in Table 22. Based on these scores, it can be concluded that the overall performance of the topic labelling service was mediocre, while being better in the Linked News case than in the Linked Culture case. However, a closer inspection of the rest of the statistics of the two datasets, seen in Table 23, in correspondence with the results populations' histograms in Figure 10, reveals a significantly better performance of the service over the RBB dataset than over the TKK dataset.

Table 22: Precision, Recall, F-measure results for the topic labelling normalized to [0;1] interval.

	RBB	**TKK**
Avg Precision	0.617391304	0.402469136
Avg Recall	0.608695652	0.479012346
F-measure	0.613012643	0.437417436

Examination of the *standard deviation*, for both precision and recall, in Table 23 shows that although the scores for precision and recall in both datasets have a notably high deviation among them, the RBB data points deviate considerably more from their (mediocre) mean than the TKK ones, which appear

Table 23: Descriptive statistics of the topic labelling evaluation ranking scores (0-5).

	RBB		**TKK**	
	Precision	*Recall*	*Precision*	*Recall*
Median ranking score	3.5	3	2	3
Mode ranking score	4	5	2	3
StD	1.395645	1.645958	0.873124	0.957588
Variance	1.947826	2.709179	0.762346	0.916975

closer to their (low) mean. *Variance* also shows that the RBB data points are significantly spread out among the population for each measurement, especially so in the case of the recall score population. The lower variance of TKK data on the other hand, signifies a higher convergence of the distribution of the rating scores.

Both observations, combined with the fact that the *mode* (i.e. most frequently occurring score) for the RBB dataset is the high scores of 4 and 5 for precision and recall respectively, while for the TKK dataset the modes correspond to the lower scores of 2 and 3 respectively, hint that the majority of the population for the RBB data is oriented towards the higher scores ($\geq$ 3) for both precision and recall, while the population of the TKK data is oriented mostly around the 2-3 score range.

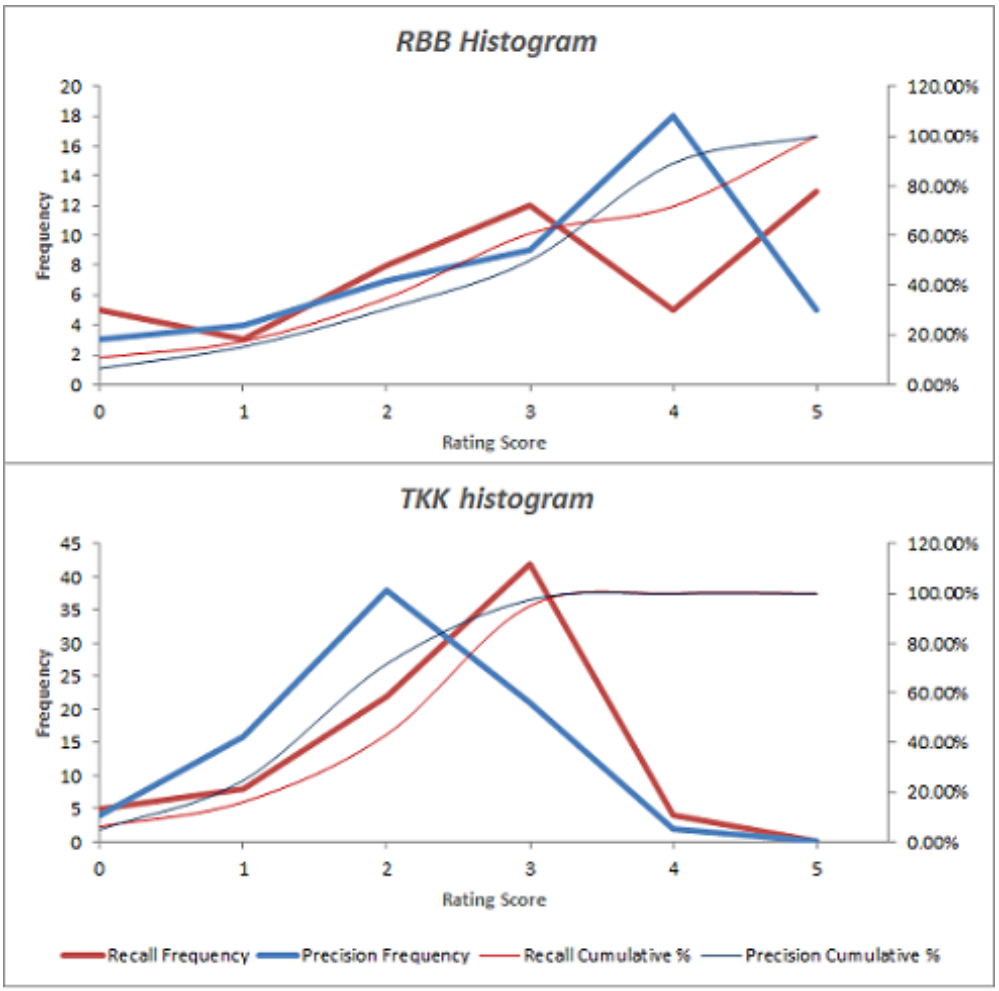

Figure 10: The distribution of data for the two datasets. Top: for the Linked News (RBB) content. Bottom: for the Linked Culture (TKK) content.

Indeed, this is confirmed by looking at the distribution of the data in the histograms of Figure 10. The frequency of appearance of each score in the precision population, for the RBB data, peaks at score: 4, with the majority of the scores gathered around scores 3-5. The latter is again confirmed by the cumulative percentage of precision, which shows that 50% of the scores are in fact accumulated under

scores 3-5. For recall, scores mostly peak at score:3 and score:5, while the cumulative percentage shows that 60% of the scores are between 3-5.

On the other hand, as indicated, the distribution of data for the TKK scenario show that for precision, the more frequent scores were along the lines of 1-3, peaking at 2. Their cumulative percentage reveals that by score:3, the data accumulation has reached a little less that 100%, i.e. almost no scores of 4 and 5 were recorded. Similarly, but relatively better, for recall, the peak is at score: 3, while the majority of the population is distributed around scores 1-4. Recall's cumulative percentage shows that 95% of the population is accumulated by score: 3.

From the above statistical analysis, it can be concluded that the performance of topic labelling for the RBB dataset was good, scoring mostly ≥ 3, with a lot of scores on the 4-5 rating. While for the TKK dataset, the performance was mediocre to poor, averaging at a 2-3 rating.

3.6.3.1 Observations While looking closely to some of the least successful articles in the RBB dataset, in collaboration with the evaluator, we noticed a dependency between the quality of topics detected and the quality of entities and types recognized by THD. This dependency was already observed and quantitatively evaluated in a previous experiment (cf deliverable D2.6, ch. 5.3: topics detection evaluation in video chapters). Another observation regarded the dependency of the service's performance and the domain coverage of the vocabularies used by the THD entity extraction.

For example, in an article about the war in Syria, words in the text like "Krieg" (war) and "Kampf" (battle) should have pointed towards the sub-topic "Armed_conflict". However, looking at the entities extracted from THD for this article, neither of these words were detected[20] On the other hand, entities that were correctly detected (e.g. `http://de.dbpedia.org/resource/Hisbollah`) failed to provide any meaningful information to the service, due to the fact that the German DBPedia did not provide any meaningful types for it other than "organisation", while it could also have been characterized as a "Political_Party", like its English DBPedia counterpart. These misses gave rise to unrelated topics, e.g., in this example, "Gastronomy", because of a statement of one of the officials in the article saying that things calmed down so that "they can now make some tea". "Tea" was recognized as "Beverage", which connected to culinary topics. Such outlier entities, and subsequently derived topics, had a low salience, however in lack of more relevant information, at times they were the only ones that were presented for this article.

Feedback from the evaluator also revealed that in general some topics, like "art_culture_entertainment" seem to be over-detected, but not necessarily wrongfully. This will be further scrutinized, but preliminary indications show that this is the result of a combination of facts: a) LUMO is more detailed in the arts subdomain, while is maintained more generic towards other more specific sub-domains, due to the LinkedTV scenario. b) A tendency of the entity extraction service to recognize more frequently named entities instead of common words within the German DBPedia was observed. This can also be tracked back to deficiencies of the German DBPedia, i.e. a focus in more efficiently modelling named entities over common words. Those named entities extracted often pointed to movies, plays, TV shows, musicians/bands and music albums that are entitled with the underlying common word. E.g. for a case where the word "Desaster" (disaster) was detected by THD, the German DBPedia resource extracted, pointed (via the " same-as" property) to an English DBPedia resource (`http://dbpedia.org/resource/Desaster`) of a German music band of that name. This resource and the type "Band" were the THD input for the underlying string "Desaster", misleading to the topic "Music".

For the TKK dataset, again the dependency between the quality of topics and the accuracy of extracted entities was prominent. Common entities (words) like "Material" or "Plate" (as an artifact) were not detected, supposedly pertaining to the coverage deficiencies of Dutch DBPedia. More significantly though, and as observed throughout the research phase of LinkedTV, the DBPedia types cannot cover efficiently arts or art related semantics, or any of the more specific sub-domains' semantics, since it focuses on covering generic domain semantics. This is why mappings to YAGO[23] (a more vast and detailed taxonomy) were opted, but for the setup of these trials[21] only DBPedia was used. The lack of art-related semantics was heavily observed as the cause of the topic detection's poor performance over the TKK dataset.

The generality of DBPedia was not conveyed only to missed types, but also to out-of-context semantics in extracted entities, which mislead the topic detection process. For instance, in an article about

[20] It is worth noticing that while THD has an option to restrict entity extraction to named entities only or common entities (words) only, we verify that the experiment uses the combined strategy, where both named entities and common words are detected.

[21] The trials followed the setup of the WP6 trials in relation to personalization, in which the omission of YAGO-extracted entities was deemed necessary to reduce the very large volume of information retrieved by THD per article, which caused delays in real-time data trafficking.

diamonds, the entity "Diamond" is, righteously so, typed in DBPedia in its more general sense as a mineral/chemical substance. Therefore it pointed to the scientific context of the item, and gave rise to topics such as "Science", "Science_Technology". These two observations converge to the general conclusion that generic vocabularies are mostly not sufficient enough to describe content of more specific domains. Therefore entity extraction, and services such as topic detection that depend on it, would benefit from more specific vocabularies in order to meaningfully annotate content of more specific domains.

3.6.3.2 LUMO Topics coverage analysis Regarding the TKK dataset, it was also observed that LUMO-arts (the arts-specific sub-ontology of LUMO, developed for the Linked Culture scenario, cf deliverable D4.6) still lacks important semantics to fully cover the arts sub-domain. This was expected, as LUMO-arts is still at its first version and requires expansion. However, the extend as to how much it should be expanded can be better assessed after examining the percentage of error that remains after a more conclusive vocabulary (like YAGO or an arts-specific vocabulary) is used as the basis of entity extraction.

Regarding the RBB dataset, while evaluating with respect to recall, the corresponding evaluator took note of the topics that, according to his opinion, should have been detected per article. Notably, these were not exclusively topics from within LUMO, but freely selected topics that had semantic relevance to the text. This feedback provided a chance to evaluate the coverage of LUMO in the aspect of topics for the general networked media super-domain.

The evaluator noted a total of 40 different topics and subtopics for the entire dataset. Of course, a topic could appear more than once in the dataset for different articles, but this analysis examines whether a given unique topic exists in fact within LUMO. The analysis of this set provided the following insights:

1. Two of the concepts noted (children, gender) are in fact classes of LUMO, but not under the Topics subhierarchy, nor are they deemed semantically relevant to topics, therefore the decision is that they will not be classified as such in following versions of LUMO.

2. Four of the concepts noted (events, event:funeral, radio (as media), work) are in fact classes of LUMO, but not under the Topics subhierarchy, and a synergy with topics is considered for following versions of LUMO. However, the semantic relevance of some of these concepts to the Topics subhierarchy is under evaluation (e.g. an event is not a topic semantically, but a general concept "Events_(topic)" under Topics might be opted and the parent concept "Event", that will remain under the appropriate subhierarchy, can be related to it via the "hasTopic" object property).

3. Out of the remaining 34 concepts that were noted by the evaluator and semantically consist of topics, 30 are covered within LUMO. This gives LUMO a coverage of 88.23% based on the examined dataset.

In such a generic domain as networked media, that concerns a vast and diverse plurality of subjects/concepts, and also given the design principles of LUMO[22] which trade off full domain coverage in order to maintain the ontology lightweight but also aim to keep it descriptive enough, a domain coverage percentage of 80-90% is the goal. In conclusion, given the results of this study and the engineering principles of LUMO, the coverage goal of LUMO for topics in the networked media domain is achieved.

3.6.3.3 Outlook The use and evolution of the Topic Labelling service is going to be pursued beyond LinkedTV, both research- and exploitation-wise, in the scope of semantic annotation and categorization of networked media content.

In the outlook, the following evaluation strategies and improvements will be followed:

– The correlation between detected entities' (from third-party software) accuracy and the performance of the topic labelling service is going to be further scrutinized.

– The use of YAGO is going to be investigated with regards to its efficacy to better describe types of entities, especially for more specific sub-domains of networked media, such as arts-related content.

– Given the results of these studies, mappings to further vocabularies, that model more specific sub-domains of the networked media super-domain, will be considered.

[22] cf deliverable D4.2 for more details on the principles in designing and engineering LUMO

– The LUMO topics, along with corresponding axiomatic relations between topics and other concepts and with corresponding mappings in the LUMO mappings ontology, will be adjusted in order to address the observations made in 3.6.3.2.

– An expansion of the topics (plus relations and mappings) in the LUMO-arts sub-ontology will be considered in order to more efficiently cover the arts sub-domain. Out of the scope of LinkedTV's Linked Culture scenario, the extent of this expansion will be restricted to a minimum, with an outlook to generally model arts-related content for broadcasters.

4 Content enrichment evaluation

For content enrichment, the following sources were evaluated:

- **IRAPI** is a purpose-built crawler and search engine, which is based on Apache Nutch and Solr frameworks. The system also features an on demand focused crawler with custom wrappers for selected websites.
- **NewsEnricher** is a component which identifies related documents from other media web sites. The results are classified into dimensions that follow a user-centered design study performed by WP3.
- **TV2Lucene** is a service that retrieves related chapters from the same collection of program that have been previously processed by the LinkedTV platform.

The evaluation was performed based on data generated by WP6 within the final LinkedTV trials. WP1 supported the evaluation by integrating a logging facility into the Editor tool, which was used by the editors in RBB and Sound&Vision content partners to curate the final trial content. This section presents the evaluation of a specific curation activity performed in the Editor tool (Figure 11) in addition of links to enrichment content.

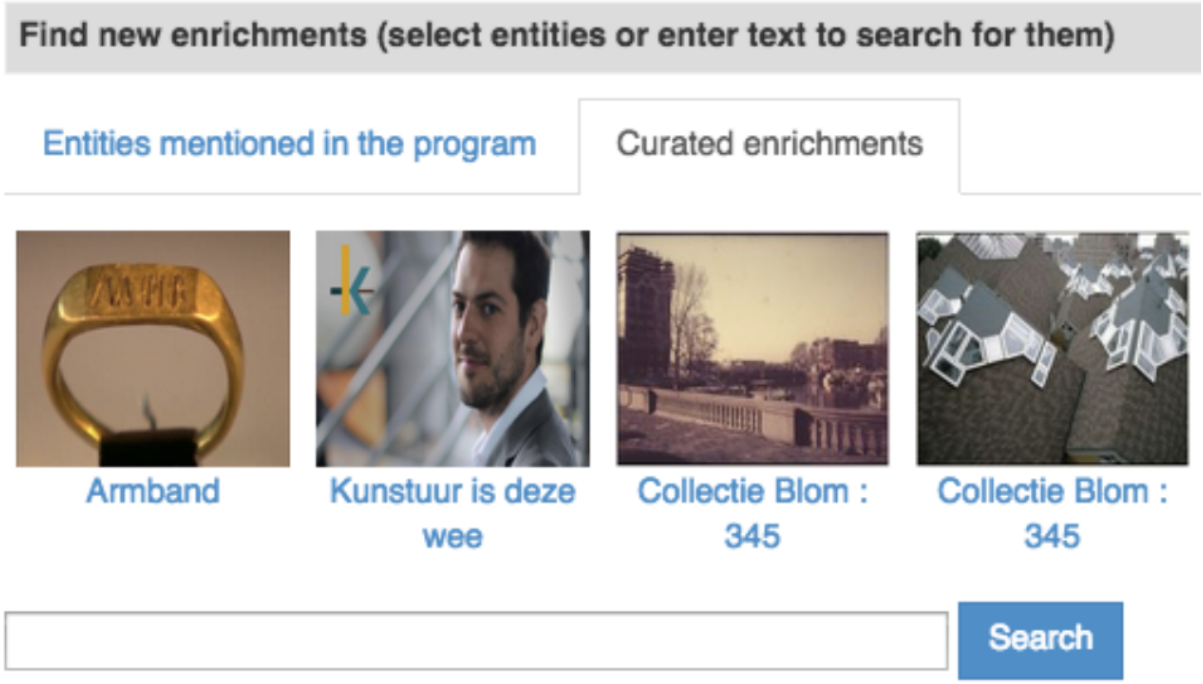

Figure 11: Editor Tool interface for issuing queries to obtain enrichment content

The workflow of the editor performing the enrichment activity is as follows:

1. Formulate a search query either by selecting entity(ies) from the automatically detected ones or by inputting a free text query;
2. Obtain a list of results from the individual WP2 enrichment services;
3. Save relevant enrichments.

The log of the editor activity containing the queries, the set of retrieved results and the set of saved enrichments was made available to WP2 via the *showlogs* editor tool REST interface. There are multiple enrichment tools provided by WP2. All of the tools were evaluated using the following two metrics:

- Average number of retrieved enrichments per query,
- Average number of saved enrichments.

Additional metrics are reported for the IRAPI service, which provides three search facets (video, image, webpage). The details on the IRAPI evaluation setup is covered in the subsection 4.1. The evaluation statistics are given for the Linked News and Linked Culture use cases in subsections 4.2 and 4.3 respectively. The subsection 4.4 gives an overview of the IRAPI index statistics and recent enhancements of the service.

4.1 IRAPI Evaluation methodology

This subsection describes how was the export from the Editor tool logging facility used to evaluate the IRAPI service.

IRAPI contains separate indexes for four media types: video, image, audio (podcast) and webpage. The editors issued queries against three of these facets: video, image and webpage. For the individual facets different crawling and data extraction strategies are in place, therefore they need to be also evaluated independently. However, the *showlogs* export of the Editor Tool did not make such detailed comparison directly possible. In order to perform the evaluation with a breakdown for each media type, two problems had to be overcome.

For some queries, one showlogs entry contained multiple queries against multiple media types, and as a result it was not possible to directly assign given proposed enrichment or saved enrichment to one media type. In these cases, we issued the queries again, and used their result to match the individual enrichment URLs in the log with the media type. If the given URL was on the result list of multiple queries (say webpage and video), it was counted for both types. In case the enrichment URL was not present in any of the result lists we weren't able to recognize original source media type of enrichments[23]. We excluded those URLs for which the media type could not be detected.

The second problem was caused by the fact that the logging system contained many near duplicate entries: the same query URLs, the same list of returned enrichments, but a different set of saved enrichments (typically one). Essentially, the entire record in the log was repeated for each saved enrichment. These cases were handled as follows: the duplicate entries were merged into one. E.g. if there were five log entries, with the exactly same set of urls in the retrieved enrichments list and each with one (different) entry in saved enrichments list, we merged these five log entries into one record containing five saved enrichment items.

Finally, it should be noted that from the analysis we excluded enrichment queries to one video[24] that was used for the editor tool tests.

4.2 LinkedNews Enrichment Experiment Setup & Evaluation

The LinkedNews trials are designed to be as ambitious as possible in terms of scope, requiring the active participation of all research, development, and exploitation work packages. They feature participants in the role of editors curating a nightly news show and producing content for other participants in the role of end-user using the LinkedTV player to observe the content. These evaluations were coordinated by Rundfunk Berlin-Brandenburg (RBB) and took place primarily in the users' homes during the evenings. As these evaluations employ the highest possible level of fidelity to proposed scenarios, they provide important qualitative data into the entire LinkedTV experience. The television content that was used for the evaluations were five episodes of the RBB nightly news broadcasts:

- rbb AKTUELL 01/03/2015 @ 21h45 (15m30)
- rbb AKTUELL 02/03/2015 @ 21h45 (30m30)
- rbb AKTUELL 03/03/2015 @ 21h45 (30m06)
- rbb AKTUELL 04/03/2015 @ 21h45 (29m45)
- rbb AKTUELL 05/03/2015 @ 21h45 (30m02)

Curated content was separated into three different section: "About", "Background", and "Related RBB Videos". For the first three days, content was curated by a single editor (an actual member of RBB production staff), with a LinkedTV member present for technical support. The last two days, content was curated by another member of RBB production staff, with the same technical support. Curation

[23] The typical cause was that the URL was removed from the index in the meantime due to the HTTP status codes 404 (Not found), 301 (Moved permanently) or 410 (Gone).

[24] ID d57dd4c2-4ca6-4ec4-8db3-f0731730f8a3

efforts took place the morning after the broadcast, with the content being consumed in the evening by participants in the role of end-users. More details on the trial setup are present in the deliverable D6.5.

Table 24: Average number of proposed enrichments and saved enrichments by an editor for the LinkedNews scenario

	RBB	
	AllEnrichments	SavedEnrichments
tv2lucene	10.9	1.0
newsEnricher	17.3333333333	1.33333333333

The results of the evaluation are present in Tables 24 and 25. For all enrichment services, except IRAPI image retrieval, the list of hits contained on average at least one document, which the editor saved. NewsEnricher provided, on average, the longest lists of enrichments.

4.3 LinkedCulture Enrichment Experiment Setup & Evaluation

The television content that is used for the LinkedCulture evaluations consists in three episodes of the Dutch television show Tussen Kunst & Kitch:

- Koninklijke Porceleyne Fles, Delft (45m13s, originally broadcast on 29/10/2008)
- Museum Martena Franeker, Franeker (49m09s, originally broadcast on 14/11/2007)
- Mu.ZEE, Oostende (44m59s, originally broadcast on 28/12/2011)

Curated content was separated into four different sections: “About”, “Background”, “Related Artworks”, and “Related Chapters”. In addition, to these three episodes, 80 additional program chapters from this collection were added so that the participants may be able to properly explore the “Related Chapters” dimension. Of these 30 “Related chapter” entities, 20 of them used IRAPI enrichments. All curation content was generated by an internal LinkedTV editorial teams comprised of three members from WP6: one principal editor and two others who double-checked content. Curation effort took place over a period of several weeks late 2014 and early 2015. More details on the trial setup are present in the deliverable D6.5.

The results of the evaluation are presented in Tables 26 and 27. Solr-based approach retrieved on average more enrichments candidates for SV users in comparison with RBB ones (see Tables 26 and 24). This can be due to the nature of the content in these videos .

For all enrichment services, the list of hits contained on average at least one document, which the editor saved. It should be noted that video search was not used in this trial.

4.4 IRAPI Index Statistics and New Developments

The evaluation of the IRAPI component presented in subsections 4.3 and 4.2 is based on the content inserted into the index by the IRAPI crawler module, which periodically visits domains designated by the content partners to download and index webpages, videos, images and podcasts. There is also an asynchronous on demand crawling module (Focused crawler), which is triggered by video queries obtained through the IRAPI query interface.

The quality of IRAPI responses as perceived by the editors is to a large extent determined by the size and comprehensiveness of the index. The following statistics listed below illustrate index size as of 16/03/2015:

Table 25: IRAPI evaluation per media type for the LinkedNews scenario, † indicates result on the subset of queries with at least one hit

	RBB			
	Queries	Queries with no hits	AllEnrichments	SavedEnrichments
webpage	30	5	13.6	1.73 (2.08†)
image	27	14	7.33 (8.25†)	0.67 (1.38†)
video	27	13	4.96	1 (1.92†)

Table 26: Average number of proposed enrichments and saved enrichments by an editor for the LinkedCulture scenario

SV		
	AllEnrichments	SavedEnrichments
tv2lucene	14.2574257426	1.23762376238

Table 27: IRAPI evaluation per media type for the LinkedCulture scenario, video is missing because of no query in this scenario to the video index

SV				
	Queries	Queries with no hits	AllEnrichments	SavedEnrichments
webpage	23	0	14.91	1.65
image	11	0	8.55	1.45

- Webpage: 811471
- Video: 78664
- Image: 1225316
- Audio: 32993

A brief overview of the index statistics is also given as a pie chart within the IRAPI Dashboard interface (Figure 12).

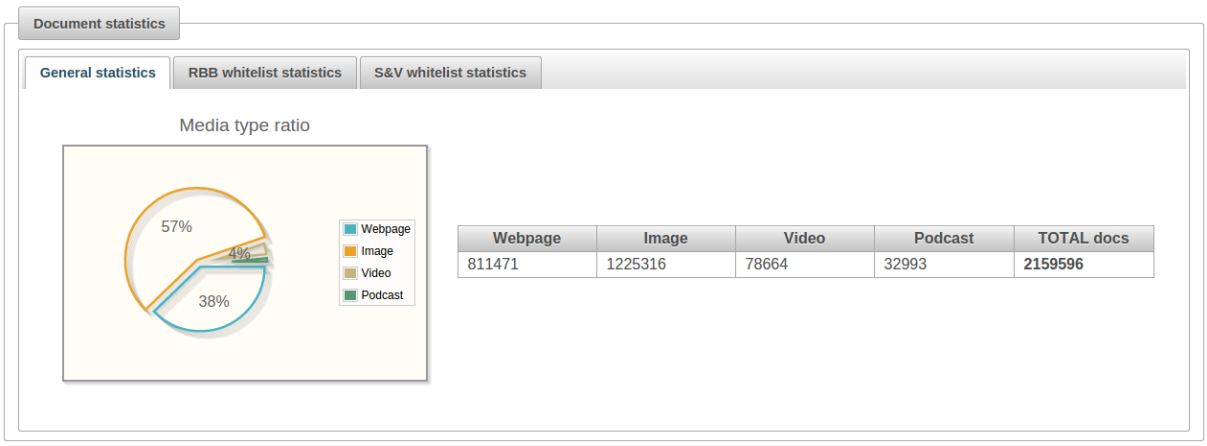

Figure 12: Dashboard - document statistics

There are 6 indexing cycles (generate-fetch-parse-index) per day. The average daily increment of all media type recorded in the period from 20/11/2014 to 26/02/2015 is as follows:

- Webpage: 1455
- Video: 323
- Image: 2340
- Audio: 119

The index increment statistics are also available in the IRAPI Dashboard interface (Figure 13).

Before the trials were performed, two significant updates in IRAPI were performed, reflecting on the feedback from the evaluation present in D2.6: new query execution strategy and on demand cleanup of media urls.

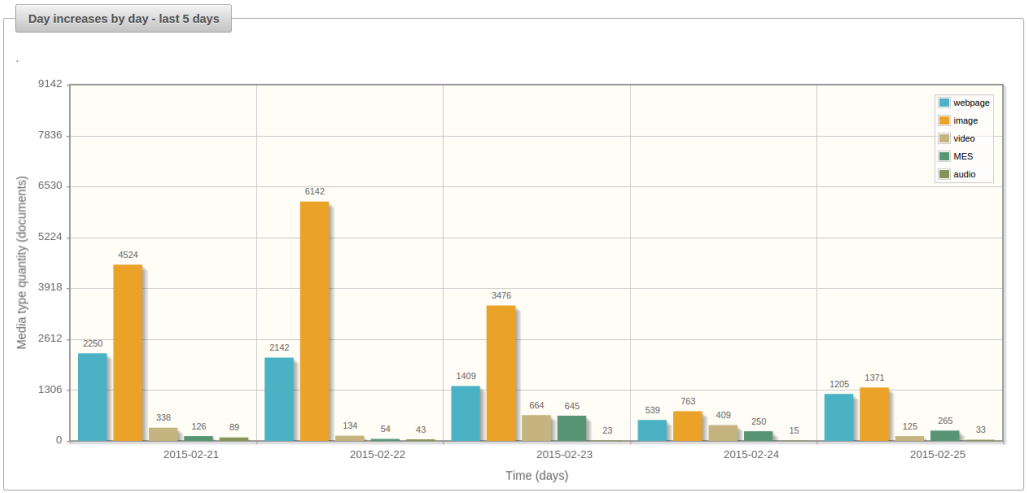

Figure 13: IRAPI Dashboard - IRAPI index increase by day - 5 days

Enhanced Query Execution workflow is shown in Figure 14. Depending on the form of the input query, the right strategy is selected - *simple term query*, *simple phrase query* or *multiquery*. First search is executed on the most reliable fields (webpage title or media title). Next, if the number of hits is smaller than the desired number, the supplemental fields (description, meta_description, source_webpage_title, picture_alt, ...) are used. Multiple queries are issued against these fields. Finally, the results are merged and the relevance score is normalized. Only documents with relevance higher than specified by the minRelevance parameter are returned.

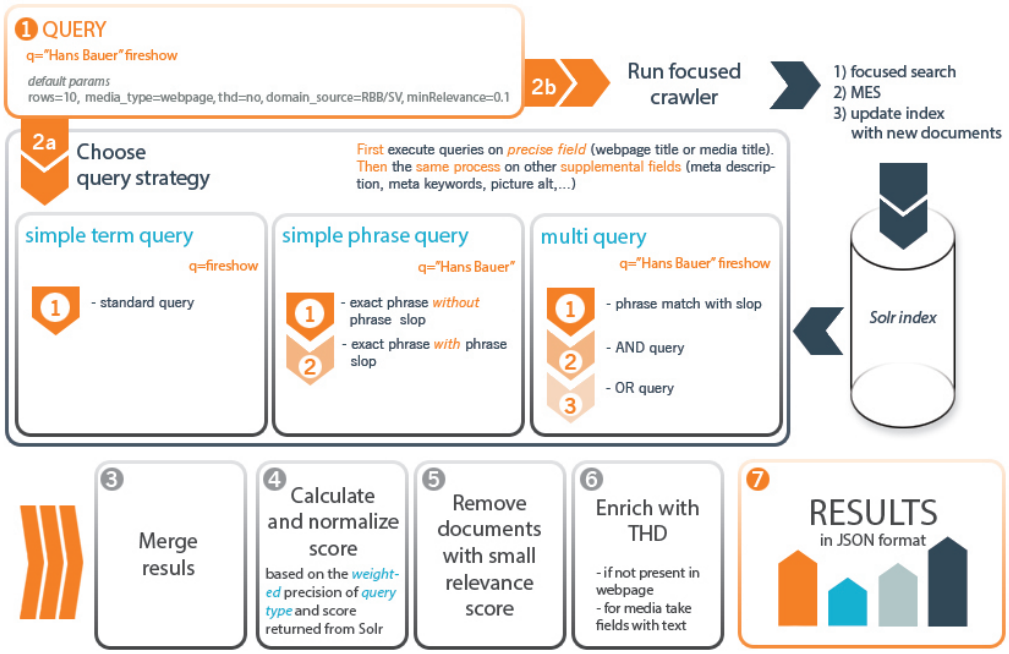

Figure 14: IRAPI workflow

Automated index cleanup was implemented. Index cleanup is an important search engine component. The automated index cleanup process removes all URLs that are not reachable (mostly 404) or are somehow invalid (unkonwn hostname, maformed url, etc.). One complete cleanup was scheduled to run once per week.

IRAPI newly checks the HTTP response code for all retrieved results, where N is a given parameter (default is 10), and marks the invalid urls for cleanup, removing them also from the search result.

Example log entry created by this process follows.

Listing 1: Strucutre of row from log and example

```
ACTION | media id |media type |REASON
DELETE|de.mdr:http/mp4dyn/video1.mp4|video|status code is 404
TO_CHECK|MES:http://avro.nl/WO_AVRO_013233/|webpage|java.net.ConnectException: Connection timed out
```

The entries in this log are processed once per day.

The process of checking all retrieved URLs prior they are returned in response to the query slightly increases query latency. However, this is not an issue in the LinkedTV workflow, where the editor identifies enrichment content “off-line”.

4.5 Outlook

After the end of the project, the gist of the IRAPI system will be made available under a free license in a public repository. Outside the complete LinkedTV workflow, the prospectively most useful reusable component is the Nutch plugin for crawling videos embedded on webpages in multiple formats, as well as podcasts and images. While these media wrapper were built for a set of specific websites indicated by the content partners, most of them will be directly, or with minor modification, applicable to other websites as well.

Another viable direction is the on-demand focused crawler module described in D2.6 section 3.2, which wraps on-site video search facilities of selected high-priority websites. The crawling is triggered immediately after IRAPI receives a query, which ensures that the index is supplied with most recent documents. It can be expected that if this on-demand focused crawling is extended to additional media types and websites, the ratio of hits perceived as relevant by the editors would improve.

5 Benchmarking activities

5.1 LinkedTV @ MediaEval 2014

LinkedTV partners (CERTH, UEP, FhG and EURECOM) achieved first class performance with a joint submission to the Search and Hyperlinking task at the MediaEval 2013 benchmarking campaign. This work was elaborated with an ICMR'14 publication based on the 2013 dataset, and further developed algorithms achieved high results in the 2014 edition of the Search and Hyperlinking task.

The MediaEval Search and Hyperlinking task is an evaluation activity that allows comparison of various approaches to media search and enrichment. As the latter focus fits at the heart of the ongoing EU FP7 project LinkedTV, we concentrated our efforts on the Hyperlinking sub-task. Obviously, not all LinkedTV components have been tested, but most of the techniques from the multimedia analysis chain (such as shot and scene detection, OCR, visual concept detection, named entity recognition, keywords extraction, semantic distance using knowledge bases and WordNet, multimodal indexing and fusion, etc.) were fused together when designing our 2013-2014 framework. The set up of the task enabled us to address the hyperlinking problem in a real big data setting scenario: as participants, we had to provide enrichments to media fragments of seed video content, and those enrichments were pooled across all submissions and later judged by real users according to their relevance.

The Search sub-task set up released within the 2013 campaign allowed us to investigate further and to improve our results when studying the intention gap which arises from the difficulty for the retrieval system to interpret accurately the user's query. We investigated a novel automatic approach to map the visual cues provided by the user in the form of a textual description or a query to visual concepts detectors. The proposed method makes extensive use of WordNet similarity measures to identify relevant visual concept detectors for the query at hand. Experimental results, conducted on the MediaEval 2013 Search sub-task, show that mapping text-based queries to visual concepts is not a straightforward task. Manually selecting relevant concepts requires impractical human intervention and does not necessarily lead to perfect results. The proposed strategy, which automatically maps visual cues from the query to the system visual concepts based on WordNet similarity, improves significantly the performance of the video search system (up to 41% MRR and 40% mGAP).

For the 2014 edition of the Search and Hyperlinking task, our submissions aimed at evaluating 2 key dimensions: temporal granularity and visual properties of the video segments. The temporal granularity of target video segments is defined by grouping text sentences, or consecutive automatically detected shots, considering the temporal coherence, the visual similarity and the lexical cohesion among them. Visual properties are combined with text search results using multimodal fusion for re-ranking. Two alternative methods are proposed to identify which visual concepts are relevant to each query: using WordNet similarity or Google Image analysis. For Hyperlinking, relevant visual concepts are identified by analysing 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.

5.1.1 Motivation behind the Search and Hyperlinking challenge

Since the last decade, more and more multimedia documents are being published and consumed in many forms, mainly over the Internet. In particular, videos constitute an increasingly popular mean to convey information, due to the ease of both capturing and sharing them: it has become very common to record a video on a mobile phone or a tablet and to upload it to a social sharing platform such as YouTube. Hence, searching for relevant content is a crucial issue, as one may be overwhelmed by the amount of available information. Media fragments enrichment further improves user experience, as the systems provide the users with more relevant content about the topic. This further navigation from one video to another is similar to the browsing activity when the users follow the hyperlinkins to move from one textual document to another on the Internet. In order to carry out this browsing behaviour through the visual archives, the network of hyperlinks has to be created beforehand and/or adjusted on the fly according to each user interests.

5.1.2 Task definition at MediaEval evaluation campaign

5.1.2.1 Search sub-task 2013-2014: from known-item queries with visual cues to purely textual queries within the ad-hoc search framework

In 2013, the task focused on the search of a known video segment in a archive collection using a query provided by a user in the form of text [12]. This framework is based on an assumption that writing text is the most straightforward mean for a user to formulate a query: the user doesn't need any input image (for which (s)he would need to perform a preliminary image search or need drawing skills). In this situation, a query is constituted of two parts: the first part gives information for a text search while the other part provides cues on visual information in the relevant segments using words. We give two examples of such queries below:

Query 1:

- *Text query:* Medieval history of why castles were first built
- *Visual cues:* Castle

Query 2:

- *Text query:* Best players of all time; Embarrassing England performances; Wake up call for English football; Wembley massacre;
- *Visual cues:* Poor camera quality; heavy looking football; unusual goal celebrations; unusual crowd reactions; dark; grey; overcast; black and white;

For the text-based search, the state-of-the-art methods perform sufficiently well. However, the visual cues are not straightforwardly understandable by a computer, since some queries are not so easy to interpret.

As these visual cues can be any text words, it is a challenging task to have a visual model for every word of the text query. Thus, a basic candidate solution is to have a set of models for predefined visual concepts (the maximum it covers, the better it is), and to map each word to its closest concept in the list. Then, the models of the mapped concepts will be used as visual content models for each query.

Ideally, this mapping process should be done manually to avoid any intent gap between the query and the mapped concepts. However, this is a very time consuming process, which may be subject to personal interpretation and therefore error prone. Strong of these facts, this process should be automated, even knowing that it will provide some noise in the mapping. Our framework uses a predefined mapping between keywords from the visual cues and the visual concepts automatically computed using WordNet distances. Each mapping is characterized by a confidence score, derived from the WordNet distances, indicating how related a keyword and a visual concept are [38].

Instead, we want to study how to perform a joint query combining text and visual concepts for video segment search. Using visual concepts relies on the accuracy of concept detectors, which can vary from one concept to the other. Hence, the query used should be carefully designed and take into account the confidence in different modules: concept mapping, concept detectors; It should also balance the part given to text and visual concepts in the search.

In 2014, the framework of the Search sub-task has been changed in favour 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 [11]. At the same time the visual cues were no longer available for these new queries, thus our investigation into the solutions for intention gap problem are confined within 2013 data set experiments, and 2014 data is used for testing implementations of LinkedTV components combination.

5.1.2.2 Hyperlinking sub-task

Overall goal of the hyperlinking task is to generate a network of links between video segments within an archive. At the MediaEval benchmark, the hyperlinking sub-task requires the participants to generate a ranked list of video segments within the same collection that provide information about these initially given list of video fragments that are otherwise named as anchors. The anchors are defined by their start and end times within the video by the users, and these users gave a textual description of potentially relevant target video segments. This textual description is not available for experiments, and is used only at evaluation stage. This approach is taken to better imitate the real case scenario when all the collection has to be indexed with anchors and targets before the actual user accesses it with a concrete information need or interest. In these experiments we test the ability of our system to generate the lists of video targets, given only the visual content of the anchor and corresponding transcript of the audio channel.

5.1.2.3 Evaluation metrics

For the known-item version of the Search sub-task there are 3 following metrics that address different aspects of results ranking:

- the Mean Reciprocal Rank (MRR) assesses the ranks of the relevant segment returned for the queries. It averages the multiplicative inverse of the ranks of the correct answers (within a given time windows, here 60s).
- 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 term of both precision of the retrieved segments and the length of the segments that should be watched before reaching the relevant content [13]. 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.

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[25]. 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 [1]

5.1.3 Motivation behind LinkedTV experiments within Search and Hyperlinking

5.1.3.1 Use of visual content to bridge the semantic gap in search

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 accompagnied with textual metadata. Several research works attempt to include visual information based on input images and/or on relevance feedback [43, 45, 32, 47].

The work of Hauptmann *et al.* [22] 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 [19], 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 concept than the accuracy of the detectors.

Hamadi *et al.* [21] 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 [5] 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) performs 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

[25] www.mturk.com

improves the performance. In this paper, we endeavor to estimate how visual concepts can improve a search, depending on the way they are used.

Another aspect of our framework is the automatic linking of a textual query to visual concepts through a semantic mapping. Several works achieve this step by exploiting ontologies. In [44], the authors developed an OWL ontology of concept detectors that they have aligned with WordNet [15]. They question whether semantically enriching detectors helps in multimedia retrieval tasks. Similarly, an ontology based on LSCOM taxonomy [30] has been developped[26], and has been aligned with ontologies such as DBpedia[27].

We focus on the use of visual information to improve content retrieval in a video collection. Indeed, videos are visually very rich and it is not straightforward 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 [42]. Indeed, there is a gap between the low-level features extracted from an image, and the high-level semantics that can be understood from it.

We propose and evaluate a video search framework using visual information, in the form of visual concepts, for video retrieval. We report how much improvement this information can provide to the search, and how we can tune this system to get better results. Indeed, we want to explore cross modality between textual and visual features: we know text is 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 retrieval can be achieved by combining textual and visual information to create an enriched query. Hence, we aim at designing a system that is able to query not only the textual features, but also the visual ones. The originality of our work lies in the fact that we start from a text query to perform the visual search: we attempt to overcome the semantic gap by automatically mapping input text to semantic concepts.

This work proposes to evaluate the use of high-level semantic concepts in complementing text for video retrieval. Text and visual concepts' scores are calculated separately and we apply a late fusion function to combine the results. We investigate the following two questions: i) to which extent can visual concepts add information when retrieving videos? ii) How can we cope with the confidence in visual concept detection? We answer to the above questions by first, studying an effective approach for combining visual and textual information, then, investigating how reliable visual concept detectors should be to achieve better improved performance, of multi-modal search on video database.

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 is more complex than last year [11]. We propose two alternatives to this problem. On one hand, WordNet similarity is employed to map visual concepts with query terms [37]. On the other hand, the query terms are used to perform a Google Image search. Visual concept detection (using 151 concepts from the TRECVID SIN task [31]) is performed on the first 100 returned images and concepts obtaining the highest average score are selected.

5.1.3.2 Temporal granularity to improve video segmentation for search and hyperlinking

As it is harder to browse through the video search results than in case of textual result lists, it is of importance to find relevant segments that start close enough to the beginning of the relevant content. Therefore we investigate into three temporal granularities of the content, i.e. the segmentation methods. The first, termed *Text-Segment*, consists in grouping together sentences (up to 40) from the text sources. We also propose to segment videos into scenes which consist of semantically correlated adjacent shots. Two strategies are employed to create scene level temporal segments. Visually similar adjacent shots are merged together to create *Visual-scenes* [40], while *Topic-scenes* are built by jointly considering the aforementioned results of visual scene segmentation and text-based topical cohesion (exploiting text extracted from ASR transcripts or subtitles).

5.1.4 LinkedTV Framework: Search sub-task 2013

Our proposed framework operates on any provided video collection with associated subtitles (or automatic speech recognition). First, we need to pre-process the video collection in order to extract and index features (i.e. text, concepts, scenes), which are needed by our work. Text search is straightforward

[26] http://vocab.linkeddata.es/lscom/

[27] http://www.eurecom.fr/~atemezin/def/lscom/lscom-mappings.ttl

with a search engine such as Lucene/Solr[28]. Nevertheless, it is different for a search based on visual features: incorporating visual information in the search task requires to design a complex framework that maps queries to a vocabulary of concepts and that is able to rank the videos segments accordingly. Figure 15 illustrates this framework.

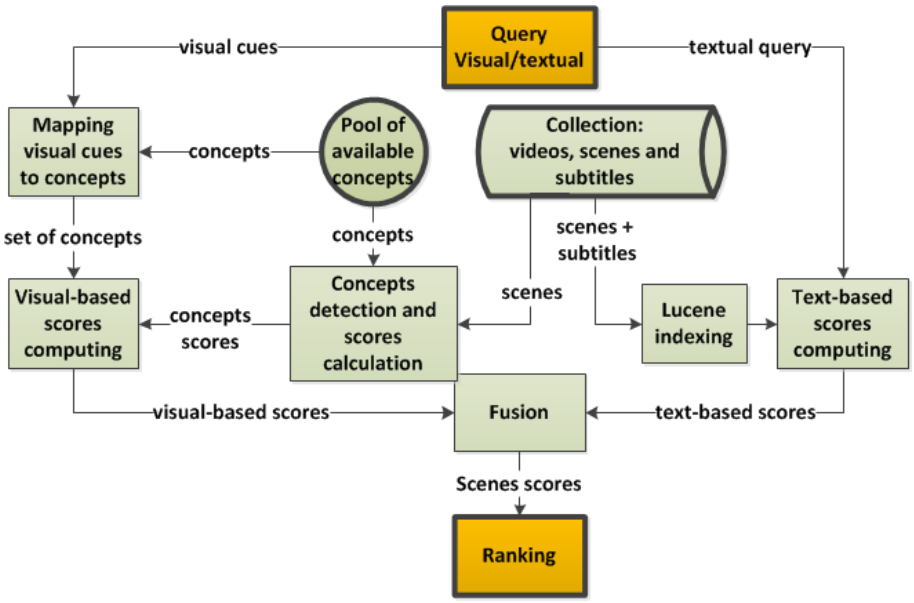

Figure 15: Our multimodal video retrieval framework

5.1.4.1 Pre-processing

We search for segments inside a video collection given a text query. For the experiments on 2013 dataset the videos are pre-segmented into *scenes* and we extract textual and visual features (visual concepts) in order to give grounds to the search.

Scenes segmentation

As a video by itself is too long to present to the user, and it may not be relevant as a whole, we want to retrieve meaningful segments of video. Shots are too short segments, hence we define scenes as combinations of adjacent shots, that have temporal and visual consistency. We use the work proposed in [41]. This algorithm, based on an extension of the Scene Transition Graph (STG) [57], groups video shots by taking into account visual similarity (using HSV histogram comparisons) between temporally adjacent shots represented by keyframes.

Concepts detection

For visual concept detection, we follow the approach presented in [39], which is based on the state-of-art for content-based multimedia indexing (CBMI). CBMI systems consists of two main phases: the modeling and indexing phases. In the modeling, the system should be extract different low-level features form a training set (the labeled set) to build different descriptors based on the content, such as Color-histograms, SIFT [28], Opponent-SIFT [52], bag-of-visual-words, etc. Then, for each concept a classifier (e.g. SVM) should be trained on each type of these descriptors to obtain a classification model. This model will be used to assign scores for new unlabeled samples (e.g. video-shots) as containing an instance of the learned concept.

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 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.

We directly use scores on the key-frame level, which are computed using a set of pre-calculated classification models. These models were trained on 151 predefined concepts from the complete list of concepts provided by TRECVid 2012 Semantic Indexing task (SIN) [31].

[28] http://lucene.apache.org/solr/

5.1.4.2 Text-based scores computation: T

We have used the search platform Lucene/Solr for indexing textual features. We temporally aligned text from the subtitles to the scenes, performed base processing (converting to lower-case, stop-words removal, etc) and indexed each scene in Lucene/Solr together with its corresponding text.

Then, we compute the text-based scores by using Lucene's default text search based on TF-IDF representation and cosine similarity.

5.1.4.3 Visual-based scores computation: V

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:

$$v'_{ij} = l + \frac{(u-l) \times (v_{ij} - min_j)}{max_j - min_j} \tag{5}$$

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.

Valid detection rate: w

Concept scores are not normalized against each other: it is not possible to compare them, or to define a threshold that provides a boolean result (whether the concept is present or not present). Nevertheless, in order to have an insight on their performance, we manually created a valid detection score by examining the top 100 images for each concept and counting the number of true positive. The percentage of true positives found will be designated by *valid detection rate* w in the remaining of this document.

Mapping text-based visual cues to visual concepts

In the visual cues description, the user provides a textual description of what are the visual characteristics of the video segment (s)he is looking for. As we propose to enhance the text-based search using visual concepts, we need a mapping between the text-based query and the concepts that should be found in the video, among the set of concepts that were computed.

For this mapping, we use the work reported in [38]. Keywords are extracted from the "visual cues" using the Alchemy API[29], and then each of those keywords is mapped with concepts for which a detector is available. This was done by computing a semantic distance between the keyword and the names of the concepts, based on Wordnet synsets [27]. Hence, each keyword was aligned to several concepts with a confidence score: this score gives a clue on the proximity between the keyword and the concept.

In this work, we will study the impact of the *confidence score* β on the set of concepts C^q associated to each query q, through its text-based visual cues. We plan to compute the performance of the system with different thresholds θ that will automatically define 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 θ, the selected concepts C'^q are those having $\beta \geq \theta$.

An example of concept mapping is given in table 28, where, the term *Castle* was mapped to five concepts (form the predefined set of concepts) with different associated confidence scores β-values.

Computing visual scores regarding each query

For each query q, we compute the visual score v_i^q associated to every scene i as the following:

$$v_i^q = \sum_{c \in C'^q} w_c \times v_i^c, \tag{6}$$

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

Table 28: Concepts mapped to the visual query from example "Castle", with their associated confidence score β

Concept	β
Windows	0.4533
Plant	0.4582
Court	0.5115
Church	0.6123
Building	0.701

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 θ is the main objective of this paper and this will be 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).

Fusion between text-based and visual-based scores
Scores of the scenes (T) based on the text feature are computed for each query. Independently, we compute scores (V) based on visual attributes and apply late fusion between both in order to obtain the final ranking of items. After these scores are calculated, the score of each scene is updated according to its t_i and v_i scores. Many alternative fusion methods are applicable to such situation [14, 3]. Here, we chose a simple weighting fusion function as follows:

$$s_i = t_i^{\alpha} + v_i^{1-\alpha} \tag{7}$$

where α is a parameter in a range of [0,1] that controls the "strength" of the fusion method. There are two critical values of α: $\alpha = 0$ 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. However, this parameter has to be tuned by cross-validation within a development set or different subsets.

5.1.5 LinkedTV Framework: Search sub-task 2014

5.1.5.1 Text-based methods

In this approach, relevant text and 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 [36], (*I*) LIMSI [17], (*S*) NST/Sheffield [20])).

5.1.5.2 Multimodal Fusion method

Motivated by [37], 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*) segmentation is evaluated.

5.1.6 LinkedTV Framework: Hyperlinking sub-task 2014

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 [8] (*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.

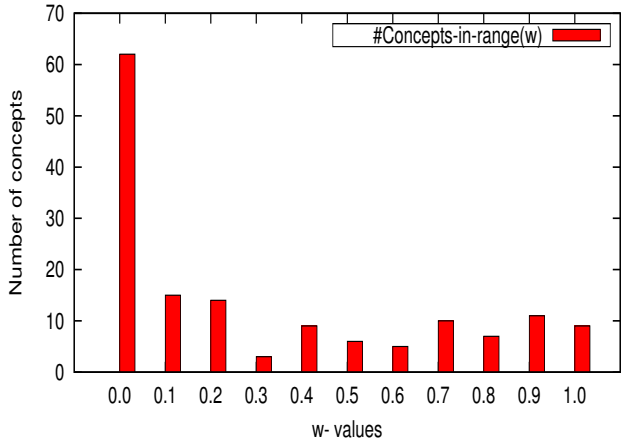

Figure 16: The predictor confidence scores of the visual concepts (w), for simplicity we show scores grouped in ten ranges.

5.1.7 Experiments

5.1.7.1 Dataset

We conducted our work 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 respectively. This represents the television content of all sort: 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 development 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 seeked by the user, described by the name of the video, the beginning 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.

5.1.7.2 Combining textual and visual information for effective multimedia search

Visual scores

To produce the visual scores we used the approach presented in [39], using a sub-set of ten different low-level descriptors calculated on key-frames. Each detector was used to train a linear SVM on 151 semantic concepts of TRECVid 2012 SIN task, these results in ten SVM-models for each concept. The same descriptors were computed on the considered dataset (i.e. Mediaeval 2013) and the models for each concept were used to predict the presence of the concepts at each key-frame of our dataset. A simple late fusion approach was applied on the ten scores for each key-frame and results in one score for each concept per key-frame. These scores are then normalized by the min-max function. We have no information about the quality of the models trained on TRECVid 2012, since only the scores on the key-frames were provided to us. Thus, we have computed manually the performance of the models on the first ranked 100 key-frames for each visual concept, which have the maximum predicted scores for each concept. We have used these scores as the valid detection rate of each concept, denoted as w.

Figure 16 shows, the histogram of w values. As this histogram shows, there are many concepts whose confidence score is equal to zero: $w = 0$. This means that these concepts will be ignored when calculating the visual scores according to function 6.

In this paper, we also compare the performance of the system using these confidence detector scores w and the case when having the same confidence for each detector, i.e. when $w = 1$.

Query mapping

Table 29 reports the minimum (Min), maximum (Max) and mean ($Mean$) number of concepts per query with different thresholds θ on the mapping confidence β. It also shows the number of queries that have

Table 29: Number of concepts associated to queries .

THR (θ)	Min	Max	Mean	$\#Q(\#c'^q > 0)$
0.0	5	45	20	50
0.1	5	45	19	50
0.2	5	41	18	50
0.3	2	37	15	50
0.4	0	25	11	49
0.5	0	19	7	49
0.6	0	19	5	48
0.7	0	12	3	44
0.8	0	6	1	29
0.9	0	2	1	21

Table 30: The optimal α- values with different concepts selection thresholds θ

	$\theta = 0.0$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.4$	$\theta = 0.5$	$\theta = 0.6$	$\theta = 0.7$	$\theta = 0.8$	$\theta = 0.9$
$w = Score(c)$	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.8	0.8
$w = 1.0$	0.9	0.9	0.9	0.9	0.9	0.5	0.5	0.7	0.7	0.7

at least one concept at each confidence level of θ ($\#Q(\#c'^q > 0)$). It is clear that when θ increases, the number of associated concepts decreases (see the Max and Mean values), and when $\theta > 0.7$ very few concepts will be included for each query. Furthermore, there are only 21 out of the 50 queries that have at least one concept with a strong confidence score (i.e. β) for the mapping (see $\#Q(\#c'^q > 0)$ with $\theta = 0.9$).

Optimizing the α parameter of the fusion function

MediaEval does not provide a relevant development set for the search task. However, we chose to tune the α parameter (equation 7) using the aforementioned initial results (the text-based and the concept-based scores) with different subsets of 20 queries. We have randomly chosen ten different subsets, each includes 20 queries out of the 50. As mentioned before, the α parameter controls the range, in which we expect the visual content to improve the text-based search. The optimal value for this parameter is likely to depend on the collection and the queries themselves. We run the evaluations with different values of α, including the two following cases: $\alpha = 1$ which is the baseline when using only text-based search, and $\alpha = 0$ that means only visual contents were used. The aim of the tuning is to get the values of α that enable to obtain the best performance of our system.

Table 30 reports the optimal values of α for each threshold θ using the (manually computed) visual predictor confidences $w = Score(c)$ and the case when all concept confidences are the same $w = 1$. These values were chosen after applying the majority vote on the ten selected subsets of different 20 queries each. As we can see, the values of α for $\theta < 0.5$ are close to 0.9 in both cases, which means the effectiveness of the visual scores is very small comparing to the text-based system. Furthermore, for $0.5 \geq \theta < 0.7$, the α values are different between both cases, they are between 0.5 and 0.7. When $\theta \geq 0.7$, the values are stable and the influence of the visual scores is coherent.

Evaluation on all 50 queries

The goal, of this experiment, is to study the influence of the visual concept mapping to text-based queries, that was done based on WordNet. 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 (θ). We have set the values of the α parameter as obtained by cross-validation (see Table 30), with the two confidence scoring ($w = Score(c)$ and $w = 1$).

Figure 17 shows the system performance (with MRR measure) when combining the visual content (selected using threshold θ) with the text-based search approach. The performance is shown with the two studied cases: when having a concepts validation rates $w = Score(c)$ and when $w = 1$. When $\theta = 0$, all mapped concepts (using the WordNet-based mapping) are selected, and as the θ value increases, the number of selected concepts decreases. In other words, the θ values perform as a noise remover in the concept mapping, and as it increases the number of mapped concepts decreases. Indeed, we want to study the impact of combining visual concepts with the text-based scores for query searching task.

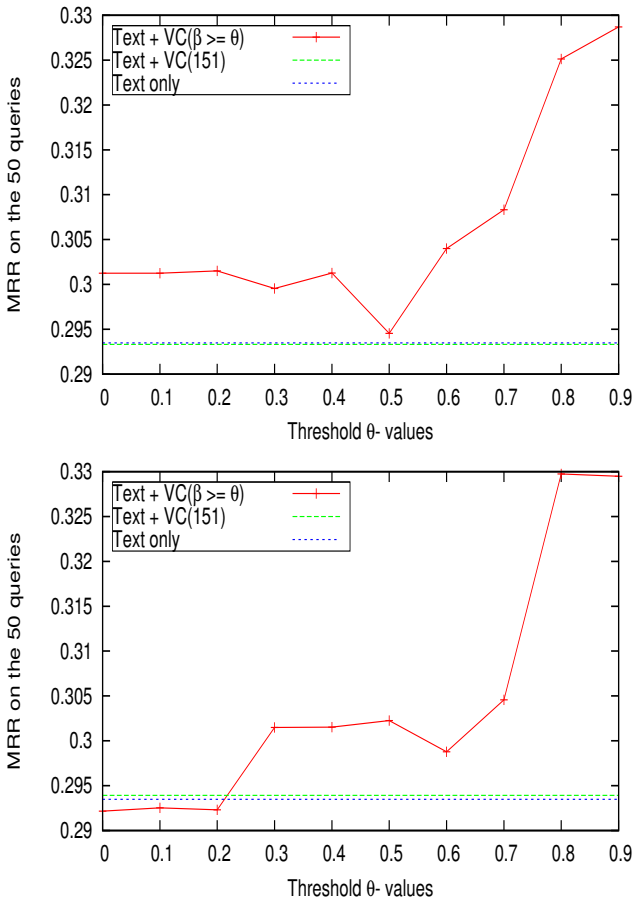

Figure 17: MRR values on the 50 queries with different θ-values using concepts validation rates: $w = Score(c)$ (a) and $w = 1$ (b).

The system performance with the evaluation of θ is compared to the two aforementioned baselines: i) using 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 in the two sub-figures, combining the visual scores of all concepts does not improve the text-based approach, while significant improvement can be achieved by combining only mapped concepts with $\theta \geq 0.3$ to each query. However, best performance is obtained when $\theta \geq 0.8$ and the gain comparing to the baseline approaches is about $11 - 12\%$ in both cases. The impact of the concept detector confidence (i.e. w) is not of that much importance, we believe that this may be due to the fact that many concepts have a valid detection rate $w = 0$. Thus, the use of $w = 1$ for each concept is a good choice for large values of θ. There is a strange bottom value with $\theta = 0.5$ using $w = Score(c)$ (the top figure in 17). We believe this is due to the noise in concept mapping, as well as the fact that many concepts were mapped with $w = 0$ as a valid detection rate. However, when θ increases this noise is removed. The same performance was observed with both mGAP and the MASP measures, but for simplicity we report only the results with the MRR measure.

This experiment considers all the MediaEval search task queries (i.e. 50 queries), whether the visual task can be mapped to visual concepts or not. We believe that the real improvement should be computed on only the 21 queries that contain at least one mapped visual concept when $\theta \geq 0.9$ (according to table 29). In the next section we will report the performance on the subset of these 21 queries only.

Evaluation on a subset of 21 queries

We have run the same evaluation as mentioned in the previous section but on only 21 appropriately selected queries. Each of these queries was mapped to at least one visual concept with high confidence mapping $\beta \geq 0.9$. This results on the 21 queries for which the visual information is important, and where the textual description maps to visual concept detectors with a high probability. Figure 18 shows the performance (in terms of MRR) of the 21 queries in function of the threshold θ, and again this is compared to the baselines on the same set of query.

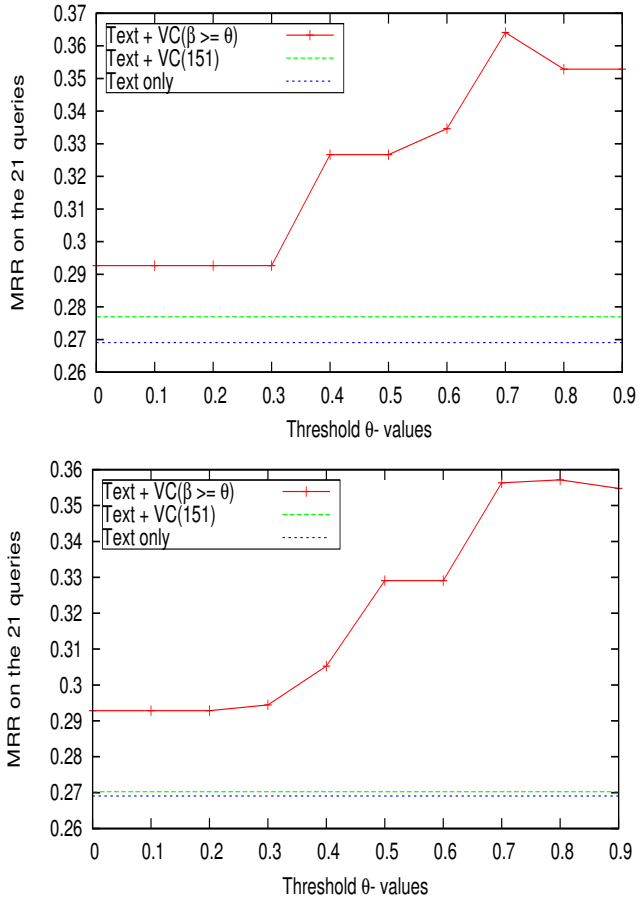

Figure 18: MRR values on only 21 queries that have minimum one concept with high confidence ($\beta \geq 0.9$) from WordNet, with different θ-values using concepts validation rates:$w = Score(c)$ (a) and using $w = 1$ (b)

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 with $\theta \geq 0.7$ in both cases, 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. Furthermore, using only concepts with high confidence values $\beta \geq 0.7$ leads to better performance with gain about $32 - 33\%$.

Conclusions and lessons learnt in 2013
While popular search engines retrieve documents on the basis of text information only, this investigation aimed at proposing and evaluating an approach to include high-level visual features in the search of video segments. A novel video search framework using visual information in order to enrich a text-based search for video retrieval has been presented. Starting from a textual query that includes some description of visual components of the searched segment, we performed a search on a large video collection of television broadcast material by fusing text-based and visual-based scores at the scenes level in order to compute the final ranking. We attempted to overcome the so-called problem of semantic gap by automatically mapping text from the query to semantic concepts, for which we have associated detectors.

Experimental results show that carefully selecting the visual concepts related to a query improves the performance of the search system. Moreover, with an appropriate concept mapping ($\beta \geq 0.7$) a significant improvement of about $32 - 33\%$ in MRR measure of the system's performance was achieved.

5.1.7.3 Search 2014

Table 31 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

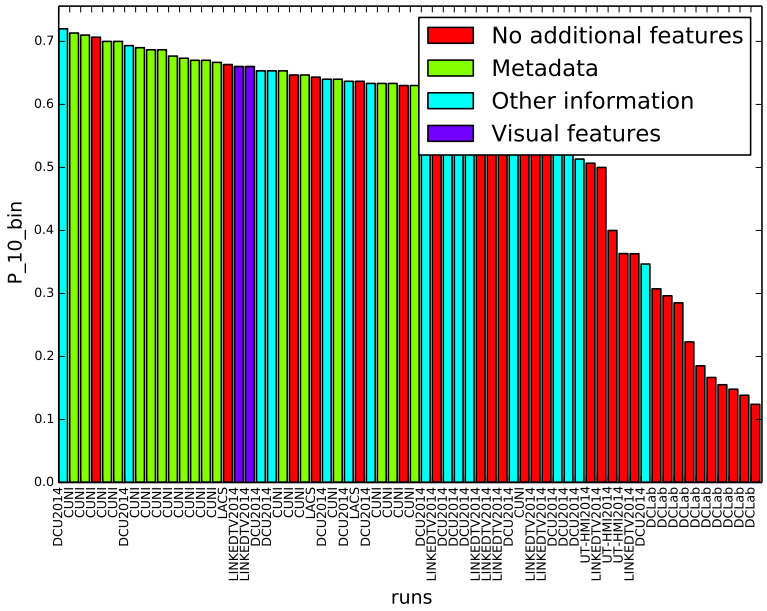

Figure 19: Mediaeval 2014 Search Performance (All Participant's Runs)

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.

Table 31: 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 19 shows in details the performance of all participants runs in terms of precision at rank 10 as evaluation score. Here the binned version of the metric is used, so that the difference in segmentation between different submissions is normalized across them by using bins of certain length as relevant units. 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. As the query set of 2014 did not have a description of the visual content, we were limited in options for use of the visual stream, at the same time the queries changed their nature becoming less specific and shorter, however the LinkedTV results in context of overall sub-task performance have shown to be competitive and og high standard.

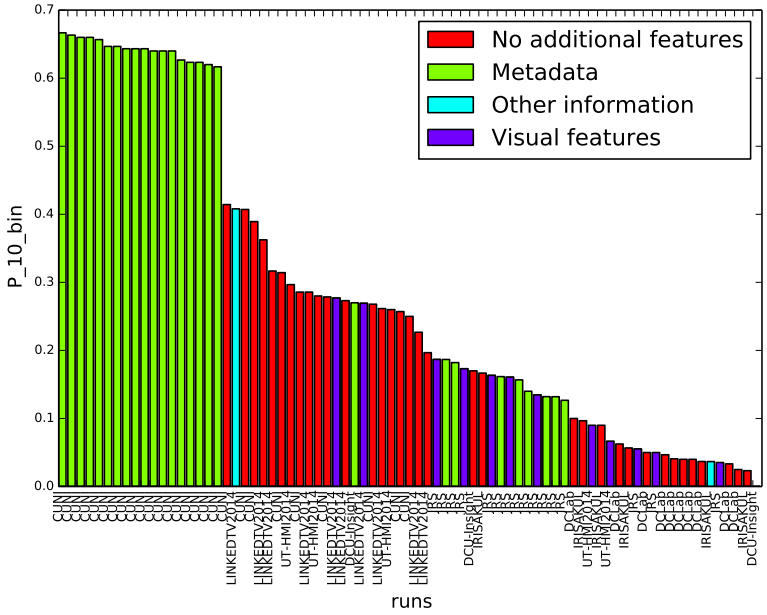

Figure 20: Mediaeval 2014 Hyperlinking Performance (All Participant's Runs)

5.1.7.4 Hyperlinking 2014 Table 32 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*). Multi-modal approaches did not produce the expected performance improvement. We believe 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.

Table 32: 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

Figure 20 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 feautres is still better than the other groups. As the anchors became shorter this year, the metadata proved to become important for the task performance.

Conclusions and lessons learnt in 2014

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. Overall, our S&H algorithms performance on this year's dataset have decreased compared to 2013, showing that task definition changes have made the task harder to solve.

5.1.7.5 Overall conclusion

LinkedTV participation in the MediaEval Search and Hyperlinking task and follow up work on the datasets

has shown that our approach to the task can significantly improve the quality of the results. 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 manage to achieve high score results. Unfortunately the changes in the 2014 task framework prevented us from direct development of methods that have proven to achieve significantly better results on 2013 data set when dealing with user queries featuring visual cues (somewhat detailing the user intention when performing the query). However, even in new conditions of an ad-hoc task, our techniques that vary size of segments depending on the structure of the scenes showed promissing results that can be further advanced when combined with the usage of metadata that has been shown of high importance for this task by other participants.

Our work was distiguished with a "Distinctive Mention Award" during the closing session of the MediaEval 2014 workshop. Such results indicate that LinkedTV is currently providing one of the leading technologies for Multimedia Search and Hyperlinking. We have also joined the team of researchers who proposed to organise the hyperlinking sub-task as part of the TRECVID evaluation campaign [30].The task has been accepted and already has 25 international registrations. This show high demand on the development of these techniques. At the same time, at the MediaEval venue, the search sub-task stays and will investigate further into the automatic anchor creation challenge. Representatives of our team help to shape this new task for the 2015 edition.

5.2 TAC'14

In this section we report on the results on the participation of the UEP team in the English Entity Discovery and Linking (EDL) track. The challenge was organized at the Text Analysis Conference 2014 (TAC) under the Knowledge Base Population (KBP) track. Twenty teams submitted a total of 74 runs to the TAC 2014 EDL track. The comparison with the team at position 10 depicted at Table 33 shows that our best submission #2 underperformed the average F1 by 17%.

Table 33: The scores for the teams at rank 1, 10,11 and the LinkedTV submission in the strong typed link match metric.

Team	Precision	Recall	F1
Rank 1	0.717	0.642	0.678
Rank 10	0.445	0.595	0.509
Rank 11	0.433	0.583	0.497
linkedtv	0.409	0.415	0.412

This section is abridged report, which will be published in full within the TAC 2014 proceedings as [10].

5.2.1 Task Description

This year, the Entity Linking task was extended also with full entity detection and classification. There were the following subtasks: spotting entities in a document corpora, linking those entities with their representation in a given reference knowledge base and classifying the entities with one of the following types: PER (Person), GPE (Geo-Political Entity) and ORG (Organization). If the entity does not belong to one of these classes, then it should not be added in the list of detected entities.

During the evaluation window each participation team was given a set of 138 documents to process. The documents contained XML markup; sometimes also not valid. The participants were asked to process not just the text content of the XML markup elements, but also the values of the XML attributes. Furthermore, the organizers asked the participants not to extract entities from quoted text (inside the quote element). The participants were also asked to adopt their systems for specific cases of entity detection, e.g. recognizing two entities, poonam and poonam8, from the content <POSTER> poonam <poonam8...@gmail.com> </POSTER>.

The systems had to output the results and provide *"mention query file"* containing the query id, document id, namestring of the mention, its begin and end offset. Additionally, the system is required to provide also *"link ID file"* providing information about the query id, reference knowledge base link (or NIL link), entity type (PER, ORG or GPE) and a confidence score – if available.

[30] http://www-nlpir.nist.gov/projects/tv2015/tv2015.html

5.2.2 Data Preparation

The reference knowledge base provides identification of the entities with custom identifiers (E.g., E0522900), while our systems identify the entities with DBpedia URIs. Therefore, it was necessary to perform mapping of these identifiers.

In the reference knowledge base, each entity, in addition to the custom identifier, is also identified with the name of the corresponding Wikipedia page. Since DBpedia derives the resource names from the corresponding Wikipedia pages' names, we could easily map a DBpedia resource URI to a Wikipedia name and, consequently, to its unique identifier in the reference knowledge base. For example, the person Sam Butler is identified with the DBpedia resource `http://dbpedia.org/resource/Sam_Butler`, which is in the reference knowledge base identified with the Wikipedia page name `Sam_Butler` and the custom identifier E0522900.

5.2.3 Entity Mention Detection

5.2.3.1 Pattern-Based Mention Detection This entity detection approach uses a manually crafted lexico-syntactic pattern which utilize Part-Of-Speech tags. The pattern was written as JAPE grammar: `NNP+`, where NNP is a proper noun tag. Before execution of the JAPE grammars, POS tags were assigned using the ANNIE POS Tagger in the GATE framework.

5.2.3.2 Stanford NER Mention Detection In this entity detection approach was used the Stanford Named Entity Recognizer [16], which is based on the state of the art Conditional Random Fields model [29]. We used models trained based on the CoNLL 2003 [48] dataset.

5.2.4 Entity Linking

5.2.4.1 Lucene-based Entity Linking The Lucene-based linking is a context independent method, which only uses the detected entity name when performing the entity linking. This approach links the entity with the most-frequent-sense entity found in the reference knowledge base. To this end, we used a specialized Lucene index, which extends the Apache Lucene search API. It primarily ranks pages based on the number of backlinks and the Wikipedia articles' titles. Note that this Lucene index is also used as by the Wikipedia Search API. This baseline approach corresponded to our best performing submission from TAC 2013 [9].

5.2.4.2 Lucene-based Entity Linking enhanced with Context To choose the most relevant entity candidate, this approach combines the most frequent sense approach described in Subs. 5.2.4.1 with the context around the entity. To retrieve the set of potential candidates, we submit a Lucene search query with the entity name. The top-5 most relevant Wikipedia pages are considered as the potential candidates. Next, we extract the entities (using Entityclassifier.eu) from the paragraph where the entity occurs, and we also extract entities from the corresponding DBpedia abstract for each of the Wikipedia candidate pages. Finally, the entity is linked with the page with the highest number of overlapping entities.

5.2.4.3 Surface Form Index Entity Linking This approach approach uses a custom entity candidate index. *The candidate index* contains all surface forms S^W found in Wikipedia articles together with their candidates E_s (this corresponds to links and their anchor texts extracted from Wikipedia dump files). Structure of the index is denoted in Figure 21. Together with each candidate e, n_e^s keeps the record of how many times the candidate occurred under the given surface form s.

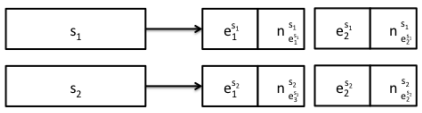

Figure 21: Structure of the candidate index for $s_1, s_2 \in S^W$, $e_1^s, e_2^s \in E_s$. Each record contains sorted set of candidates together with counts of their occurrences under a given surface form.

We experimented with various representations of surface forms in the index and methods of entity linking taking into account co-occurrence of entities in the same paragraphs. The details of the final algorithm are present in [10].

5.2.5 Entity Classification

5.2.5.1 Mappings Based Classification In this approach, we assume an entity is classified with a DBpedia Ontology v3.9 fine-grained class and our task is to find an appropriate mapping to one of the four entity types: Person (PER), Organization (ORG), Geo-political Entities (GPE) and Miscellaneous (MISC). In this approach we manually established mappings between all 537 DBpedia classes, from the DBpedia Ontology v3.9 to the four coarse grained types. Mappings to entity types are created according to the "TAC KBP 2014 - Entity Linking Query Development Guidelines".[31]

5.2.5.2 Supervised Classifier This approach uses a machine learning technique to classify the entities into one of the four entity types: Person (PER), Organization (ORG), Geo-political Entities (GPE) and Miscellaneous (MISC). As the training data we harvested DBpedia dataset using DBpedia SPARQL endpoint[32] so as to have a high number of three balanced entity types: PER $536\times$, ORG $451\times$ and GPE $501\times$.[33] Each DBpedia resource was represented as a vector of term frequencies (TF) of words from the dbpedia-owl:abstract. Due to the high size of abstracts for DBpedia resources we applied 25% periodic prunning of the original abstracts and eliminated common English words using stopwords. As a result, the training dataset consists of 1488 instances represented by vectors of their TFs for 2403 attributes.

For entity classification we trained a supervised classifier based on the Support Vector Machines (with linear kernel) using Weka wrapper for LibSVM[34] library. Our classification model achieved 94% accuracy in ten-fold cross-validation setting.

5.2.5.3 Stanford NER Based Classifier This approach uses the Stanford Named Entity Recognizer [16] to classify the entity into one the four classes. StanfordNER distinguishes four coarse grained types: Person (PER), Organization (ORG), Location (LOC) and Miscellaneous (MISC). While the PER and ORG types are defined in the TAC 2014 the TAC 2014 Entity Discovery and Linking task, the LOC was not present. Therefore, each entity of type LOC was mapped to the GPE type.

5.2.6 Submission Description

For the TAC KBP 2014 Entity Discovery and Linking task we have submitted three runs. Additionally, after the submission, we experimented with variations of the methods and we evaluated two additional runs. The descriptions of these runs follow.

Run #1. This run used the pattern based approach described in Sec 5.2.4.1 to detect entity mentions. Each entity mention for further linked with the Lucene index approach described in Sec 5.2.4.1. To link the entity we submit a Lucene search query with the entity name and the first non-disambiguation page is considered as the correct entity link. If the entity is successfully linked, then our supervised classifier described in Sec 5.2.5.2 assigns the entity to one of the four defined classes (PER, ORG, GPE or MISC). If the model failed to classify into one of the four classes, we further processed the entity mention with the StanfordNER and performed the classification. Only entity mentions which were classified as PER, ORG, or GPE were included in the output.

Run #1v2. This run also uses the pattern-based approach to detect entity mentions and uses the Lucene index to link with the first non-disambiguation page. The main purpose of this run is to evaluate the supervised classifier. To this end, in this run we used the manual classification approach explained in Sec 5.2.5.1. Since the manual classification requires DBpedia Ontology classes, the most specific DBpedia Ontology class as returned by the Entityclassifier.eu NER was used to map to one of the three required classes. If we failed to classify the entity, then the StanfordNER was used to perform the classification.

Run #1v3. This run also uses the pattern-based approach to detect entities. The main purpose of this run is to evaluate the quality of the Lucene-based linking approach used in run#1, which as a correct link considers the first *non-disambiguation* page. Therefore, in this run, as the correct entity link we did not skip the disambiguation pages, and the Wikipedia page with the highest rank in the Lucene index was

[31] TAC 2014 EDL Query Development Guidelines - `http://nlp.cs.rpi.edu/kbp/2014/annotation.html`

[32] `http://dbpedia.org/sparql`

[33] Entity is classified as MISC type if classifier cannot classify entity with sufficiently high confidence to other three entity types.

[34] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

considered as correct. For classification, this run uses the supervised classifier together with a fallback to the StanfordNER classifier.
Run #2. This run uses StanfordNER to extract the entity mentions. Further, it uses approach based on a surface form index (described in Sec 5.2.4.3) to perform entity linking and StanfordNER for the entity classification.
Run #3. This run uses the pattern-based approach to detect entity mentions. For this run we developed a more advanced entity linking described in Sec 5.2.4.2 which considers also the context around the entity when choosing the right candidate. In this run, for classification we used the supervised classifier with a fallback to the StanfordNER classifier.

5.2.7 Evaluation

5.2.7.1 Metrics The TAC 2014 KBP Entity Discovery and Linking challenge evaluated the performance of the entity detection, linking, classification and clustering. Bellow we provide brief description of the evaluation metrics.
Strong Mention Match. A micro-averaged metric for evaluation of the entity mentions. The begin and the end offsets of the entity must exactly match with the ground-truth to be counted as correct.
Strong Typed Mention Match. A micro-average metric for evaluation of the entity detection and classification. In addition to the begin and end offsets, also the type must match with the ground-truth to be counted as correct.
Strong All Match. A micro-average metric for evaluation of the entity linking. A mention is counted as correct if the link (KB link or NIL link) matches the ground-truth link.
Mention CEAF. A metric for evaluation of the entity clustering. It is based on a one-to-one alignment between system and ground-truth clusters (KB and NIL). It computes the optimal mapping based on the overlap between system-gold pairs. The entity mention offsets must match the ground-truth spans and incorrect matches affects the precision and the recall.

5.2.7.2 Results We report all four metrics for each of our main three submissions (run #1-3) and the results of the two additional runs (run #1v2 and run #1v3).

Table 34: Results from the entity mention detection evaluation - Strong Mention Match metric.

Id	Precision	Recall	F1
run #1	0.383	0.433	0.407
run #2	0.589	0.602	**0.595**
run #3	0.390	0.434	0.411
run #1v2	0.408	0.212	0.279
run #1v3	0.388	0.414	0.400

Table 34 summarizes the results from the evaluation of entity mention detection. The highest F1-score 0.595 was achieved by the StanfordNER (run #2) followed by the submission based on Lucene and enhanced with the entity context (run #3), which achieved F1-score 0.411.

The results from the evaluation of the entity linking are summarized in Table 35. It can be observed that the most-frequent-sense approach, which uses the surface form index (run #2) performed the best achieving F1-score 0.369, while second best results were achieved by the Lucene index 0.269 (run #1). Our assumption for the poorer performance of the run using the Lucene index might be due to i) the poor performance of preceding pattern-based entity spotting approach (0.407 F1 compared to 0.595 of the StanfordNER), and/or ii) our dated Lucene index, created from a Wikipedia snapshot as of 8/9/2012. For the most-frequent-sense approach, which uses the surface form index (run #2) we used more recent dataset based on Wikipedia snapshot as of 4/6/2013.

The results from the evaluation also show that run #1 which skips the disambiguation pages when performing linking achieved better results than the run #v3, which does not skip the disambiguation pages.

Table 36 summarizes the results from the evaluation of entity classification. The highest F1-score 0.429 achieved the submission which relies on StanfordNER classifier. The results also show that the submission #1 which uses the supervised classifier achieved better F1-score 0.351, compared to the submission #v2 based on the manual mappings. On the other hand, the manual mappings based submission achieved higher precision 0.368 than the supervised model 0.319.

Table 35: Results from entity linking evaluation - Strong All Match metric.

Id	Precision	Recall	F1
run #1	0.241	0.272	0.255
run #2	0.365	0.373	**0.369**
run #3	0.221	0.246	0.232
run #1v2	0.258	0.134	0.176
run #1v3	0.261	0.278	0.269

Table 36: Results from the entity classification evaluation - Strong Types Mention Match metric.

Id	Precision	Recall	F1
run #1	0.319	0.361	0.338
run #2	0.550	0.561	**0.555**
run #3	0.314	0.350	0.331
run #1v2	0.368	0.191	0.252
run #1v3	0.338	0.360	0.348

Table 37 presents the results from the clustering evaluation. The best F1-score for the CEAF clustering metric was achieved by the run #2 0.429, which uses the method based on surface form index linking and the "exact name" NIL clustering technique. Second best F1-score 0.351 was achieved by submission #1, which uses the Lucene-based linking approach.

Table 37: Results from clustering evaluation - Mention CEAF metric.

Id	Precision	Recall	F1
run #1	0.331	0.374	0.351
run #2	0.425	0.434	**0.429**
run #3	0.330	0.368	0.348
run #1v2	0.361	0.188	0.247
run #1v3	0.333	0.355	0.344

5.2.8 Lessons Learned

We hereby summarize the lessons learned from the evaluation.

- Accurate entity mention detection is highly required. Since incorrectly spotted entity directly influences the entity linking and classification, the mention spotting is a crucial step. Therefore, in the future we should also focus our efforts on developing more precise entity mention detection methods.
- Most-frequent-sense or context based entity linking. We evaluated also more sophisticated approaches to entity linking based on their context and co-occurrences with other entities. This approach usually performs better for rare meanings of entities. However, for the general case of TAC 2014 KBP Entity Discovery and Linking dataset the most-frequent-sense method provided best results. Also we observed that surface forms normalization in our indexes improved the results.
- Learning classification from knowledge graphs. The results from the evaluation shows that open knowledge graph data is mature enough and can be also useful for learning entity classification. This year, we used the multi-domain knowledge graph DBpedia to learn a model for entity classification which showed promising results. In the future we would like further to explore and leverage open data from additional knowledge graphs such as YAGO[35].

5.3 #Microposts 2015 NEEL challenge

Microposts are a highly popular medium to share facts, opinions or emotions. They compose an invaluable wealth of data, ready to be mined for training predictive models. The NEEL challenge at the Microposts workshop series consists of three consecutive steps: 1) extraction and typing of entity mentions within a tweet; 2) link of each mention to an entry in the English 2014 DBpedia representing the

[35] http://yago-knowledge.org/

same real world entity, or NIL in case such an entry does not exist; and 3) clustering of all mentions linked to NIL. Thus, the same entity, which does not have a corresponding entry in DBpedia, will be referenced with the same NIL identifier.

5.3.1 Basic Concepts

An entity, in the context of the NEEL challenge, is used in the general sense of being, not requiring a material existence but requiring to be an instance of a class in a taxonomy. Thus, a mention to an entity in a tweet can be seen as a proper noun or an acronym referring to an entity. The extent of an entity is the entire string representing the name, excluding the preceding definite article (i.e. "the") and any other pre-posed (e.g. "Dr.", "Mr.') or post-posed modifiers. Compound entities should be annotated in isolations.

Mentions and Typification. In this task we consider that an entity may be referenced in a tweet as a proper noun or acronym if:

1. it belongs to one of the categories of the #Microposts2015 NEEL challenge taxonomy (Thing, Event, Character, Location, Organization, Person, Product).

2. it can be linked to a DBpedia entry or to a NIL reference depending on the context of the tweet.

Knowledge Base. The #Microposts2015 NEEL challenge is based on the English 2014 DBpedia snapshot[36]) as the Knowledge Base for linking. DBpedia is a widely available Linked Data dataset and is composed of a series of RDF (Resource Description Framework) resources. Each resource is uniquely identified by a URI (Uniform Resource Identifier). A single RDF resource can be represented by a series of triples of the type <S,P,O> where S contains the identifier of the resource (to be linked with a mention), P contains the identifier for a property and O may contain a literal value or a reference to another resource. In this challenge, a mention in a tweet should be linked to the identifier of a resource (i.e. the S in a triple). In this challenge, only the final IRI describing a real world entity (i.e. containing their descriptive attributes as well as relations to other entities) are considered for linking. Thus, if there is a redirection chain given by the property wikiPageRedirects, the correct IRI is the one at the end of this redirection chain.

5.3.2 Dataset

The dataset contains tweets extracted from a collection of over 18 million tweets. The dataset includes event-annotated tweets provided by the Redites[37] project covering multiple noteworthy events from 2011 to 2013 (including the death of Amy Winehouse, the London Riots, the Oslo bombing and the Westgate Shopping Mall shootout) and tweets extracted from the Twitter firehose from 2014. Since the task of this challenge is to automatically recognise and link entities, we have built our dataset considering both event and non-event tweets. While event tweets are more likely to contain entities, non-event tweets enable us to evaluate the performance of the system in avoiding false positives in the entity extraction phase. The training set is built on top of the entire corpus of the #Microposts NEEL 2014 Challenge. We have further extended it for typing the entities and adding the NIL references.

5.3.3 Gold Standard (GS) Generation Procedure

The GS was generated with the help of 3 annotators. The annotation process followed three phases. In the first one, an unsupervised annotation of the GS has been performed, with the intent to extract candidate links which were meant as inputs of the second stage. In the second stage annotations were performed by two annotators using GATE[38]. The annotators were asked to analyze the entity mentions, categories and links provided in the first stage and to add and to remove any others. The annotators were also asked to mark any problematic case when encountered. In the third phase, a third annotator went through the problematic cases and, involving the two initial annotators, refined the annotation procedures. An iterative process has then taken place looping on stage 2 and 3, until all problematic cases were resolved.

[36] http://wiki.dbpedia.org/Downloads2014
[37] http://demeter.inf.ed.ac.uk/redites
[38] https://gate.ac.uk

5.3.4 Evaluation

Participants are required to implement their systems as a publicly accessible web service following a REST based protocol and to submit their contending entries (up to 10) to a registry of the #Microposts2015 NEEL challenge services. Upon receiving the registration of the service, calls to the contending entry will be scheduled in two different time windows, namely D-Time (meant to test the APIs) and T-Time for the final evaluation and metric computations. In the final stage, each participant can submit up to 3 final contending entries.

We will use the metrics proposed by TAC KBP 2014[39] and in particular, we will focus on:

- *tagging*: strong_typed_mention_match (check entity name boundary and type)
- *linking*: strong_link_match
- *clustering*: mention_ceaf (NIL detection)
- *latency*: it estimates the computation time distribution

[39] https://github.com/wikilinks/neleval/wiki/Evaluation

6 Summary and outlook

We have presented the final linked media layer architecture produced by the LinkedTV Work Package 2. While the tools and the motivation for them in the LinkedTV pipeline have been previously presented in details in the previous deliverables (or their description is available in the referenced scientific publication), this deliverable focused on presenting recently performed evaluations, which were performed either directly on the "LinkedTV" content, or on annotated datasets or within relevant international competitions.

Chapter 2 briefly described the final architecture of the Linked Media Layer in LinkedTV. A more in-depth description of the architecture is covered in D5.5 Final Integration Platform and D5.6 Final End-to-end Platform.

Chapter 3 presented the evaluation of the content annotation services. The results indicate that the consortium developed effective tools for entity expansion and entity salience detection, which significantly outperform baseline algorithms. The results of THD entity spotting and linking suggest that there is no single method that would provide the best results over all datasets and entity types. Finally, the results of Linked Hypernyms Dataset evaluation show that the accuracy of the most reliable partition of the dataset is on par with the accuracy of statements in the DBpedia knowledge base. This finding justifies the use of this dataset as a complement to DBpedia in the THD tool.

Chapter 4 presented the evaluation of some content enrichment services, namely: News Enricher for the LinkedNews scenario, IRAPI crawler/retrieval engine and the Solr module to retrieve related chapters within the same collection of programs. In most cases, the WP2 tools were able to suggest in response to a query at least one enrichment content item judged as relevant by Sound&Vision or RBB. Consistently, the best results were obtained for web page media type with nearly two saved enrichment per query. The statistics for video and image retrieval were brought down by the fact that for a number of queries there were no hits. If these queries were not considered, than the average number of saved enrichments would meet or exceed one for all media types and tools. The Solr-based related chapter retrieval worked also well, producing at least one saved enrichment for both RBB and Sound&Vision.

Chapter 5 gave an account of LinkedTV participation in the MediaEval'14 and TAC'14 contests. As one of 9 participants of MediaEval, 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. We also took part in the highly competitive TAC contest, organized by the U.S. National Institute for Standardization (NIST), which provided a benchmark of of our linking algorithms with those of 19 other institutions from around the world.

Most of the individual components evaluated in this deliverable have been made available under an open source license or open source release is in progress. This ensures that the software is available after the end of the project, and gives the wide possibilities for future extensions and improvements. For example, the THD system will be made available as a plugin for GATE, which is a leading open natural language processing framework. Since IRAPI is primarily an extension of the Apache Nutch crawling software, its reusable components will be released as open source and announced to the Apache Nutch community. The Recognyze system will be further evaluated using the Gerbil platform. Additional plans for further improvements and usage of those software modules are described in the deliverable D8.8.

References

[1] 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.

[2] Maria-Luiza Antonie and Osmar R. Zaïane. Text document categorization by term association. In *ICDM '02*, Washington, DC, USA, 2002. IEEE Computer Society.

[3] 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.

[4] Christian Borgelt and Rudolf Kruse. Induction of association rules: A priori implementation, 2002.

[5] 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.

[6] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the International World Wide Web Conference (WWW)*, 2013.

[7] W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.

[8] Milan Dojchinovski and Tomas Kliegr. Entityclassifier.eu: Real-Time Classification of Entities in Text with Wikipedia. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 654–658. Springer Berlin Heidelberg, 2013.

[9] Milan Dojchinovski, Tomáš Kliegr, Ivo Lašek, and Ondřej Zamazal. Wikipedia search as effective entity linking algorithm. In *Text Analysis Conference (TAC) 2013 Proceedings*. NIST, 2013.

[10] Milan Dojchinovski, Ivo Lašek, Ondřej Zamazal, and Tomáš Kliegr. Entityclassifier.eu and semitags: Entity discovery, linking and classification with wikipedia and dbpedia. In *Text Analysis Conference (TAC) 2014 Proceedings*. NIST, 2014. to appear.

[11] 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.

[12] 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, 2013.

[13] 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*, ECIR'12, pages 170–181, Berlin, Heidelberg, 2012. Springer-Verlag.

[14] Slim Essid, Marine Campedel, Gaël Richard, Tomas Piatrik, Rachid Benmokhtar, and Benoit Huet. *Machine learning techniques for multimedia analysis*. John Wiley & Sons, Ltd, July 2011.

[15] Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[16] 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 Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[17] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1):89–108, 2002.

[18] Dan Gillick and Jesse Dunietz. A new entity salience task with millions of training examples. In *Proceedings of the European Association for Computational Linguistics*, 2014.

[19] 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*, ICMR '13, pages 89–96, Dallas, Texas, USA, April 2013.

[20] Thomas Hain, Asmaa El Hannani, Stuart N Wrigley, and Vincent Wan. Automatic speech recognition for scientific purposes-webasr. In *Interspeech*, pages 504–507, 2008.

[21] 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.

[22] 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.

[23] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

[24] Tomás Kliegr and Ondřej Zamazal. Towards linked hypernyms dataset 2.0: complementing dbpedia with hypernym discovery. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 3517–3523, 2014.

[25] Tomáš Kliegr. Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery. *Web Semantics*, 2015. In press.

[26] Tomáš Kliegr, Jaroslav Kuchař, Davide Sottara, and Stanislav Vojíř. Learning business rules with association rule classifiers. In Antonis Bikakis, Paul Fodor, and Dumitru Roman, editors, *Rules on the Web. From Theory to Applications*, volume 8620 of *Lecture Notes in Computer Science*, pages 236–250. Springer International Publishing, 2014.

[27] 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.

[28] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[29] 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. Association for Computational Linguistics, 2003.

[30] 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, July 2006.

[31] 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*. NIST, USA, 2012.

[32] 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*, MULTIMEDIA '04, pages 508–511, New York, NY, USA, 2004. ACM.

[33] José Luis Redondo Garcia, Laurens De Vocht, Raphael Troncy, Erik Mannens, and Rik Van de Walle. Describing and contextualizing events in tv news show. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 759–764, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[34] Giuseppe Rizzo, Marieke van Erp, and Raphaël Troncy. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In *9th International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

[35] M. Röder, R. Usbeck, S. Hellmann, D. Gerber, and A. Both. N3 - a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. In *The 9th edition of the Language Resources and EvaluationConference, 26-31 May, Reykjavik, Iceland*, 2014.

[36] Anthony Rousseau, Paul Deléglise, and Yannick Estève. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC 2014*, 2014.

[37] 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 KINGDOM, 04 2014.

[38] Mathilde Sahuguet, Benoit Huet, Barbora Cervenkova, Evlampios Apostolidis, Vasileios Mezaris, Daniel Stein, Stefan Eickeler, Jose Luis Redondo Garcia, and Lukas Pikora. LinkedTV at MediaEval 2013 search and hyperlinking task. In *MEDIAEVAL 2013, Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[39] 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)*, Melbourne, Australia, September 2013.

[40] 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(8):1163–1177, August 2011.

[41] 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(8):1163–1177, August 2011.

[42] 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.

[43] 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*, MULTIMEDIA '96, pages 87–98, New York, NY, USA, 1996. ACM.

[44] 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.

[45] 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.

[46] Václav Zeman Stanislav Vojíř. Easyminer – next development, 2014. Technical report, University of Economics, Prague.

[47] 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.

[48] 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. Association for Computational Linguistics, 2003.

[49] Dorothea Tsatsou, Stamatia Dasiopoulou, Ioannis Kompatsiaris, and Vasileios Mezaris. Lifr: A lightweight fuzzy dl reasoner. In *11th ESWC 2014 (ESWC2014)*, May 2014.

[50] Dorothea Tsatsou and Vasileios Mezaris. Lumo: The linkedtv user model ontology. In *11th ESWC 2014 (ESWC2014)*, May 2014.

[51] Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. GERBIL – general entity annotation benchmark framework. In *24th WWW conference*, 2015.

[52] 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(9):1582–1596, September 2010.

[53] 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*, CIKM '07, pages 691–700, New York, NY, USA, 2007. ACM.

[54] Milan Šimůnek. Reuters-21578 text categorization test collection automated analysis using the LISp-Miner system for purpose of the LinkedTV project, 2014. Technical report.

[55] Albert Weichselbraun, Daniel Streiff, and Arno Scharl. Linked enterprise data for fine grained named entity linking and web intelligence. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, WIMS '14, pages 13:1–13:11, New York, NY, USA, 2014. ACM.

[56] William E Winkler. Overview of record linkage and current research directions. In *Bureau of the Census*, 2006.

[57] Minerva Yeung, Boon-Lock Yeo, and Bede Liu. Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, 71(1):94–109, July 1998.

[58] Ondřej Zamazal and Tomáš Kliegr. Type inference in DBpedia from free text. Under review, 2015.

7 Annex I: List of software

This annex presents a list of LinkedTV-supported software used in the final WP 2 workflow, which was employed in the trials reported in D6.5. It also presents the evaluation instructions for the Topic Labelling experiment of ch. 3.6.

Table 38: WP 2 workflow software

Acronym	Description	Availability	License
THD	performs entity detection, classification and salience computation	Source code available at https://github.com/entityclassifier-eu	GNU GPLv3
IRAPI	crawler/search engine built on top of Apache Nutch and Solr	Source code available at https://github.com/KIZI/IRAPI	Apache License 2.0
LHD	framework for generating the LHD dataset	Source code available at https://github.com/KIZI/LinkedHypernymsDataset, generated datasets at http://ner.vse.cz/datasets/linkedhypernyms/	GNU GPLv3
NERD	a framework compiling 12 NER extractors (AlchemyAPI, dataTxt, DBpedia Spotlight, Lupedia, OpenCalais, Saplo, SemiTags, TextRazor, THD, Wikimeta, Yahoo! Content Analysis Framework and Zemanta) that detects automatically the language of the text being analyzed and provide annotations in the form of entity type and entity links anchored in a knowledge base	Source code available at https://github.com/NERD-project	Apache License 2.0
TV2RDF	a REST API and an online application that converts both legacy metadata and Exmaralda metadata into RDF using the LinkedTV ontology	http://linkedtv.eurecom.fr/tv2rdf/api	Apache License 2.0
TVNewsEnricher	a REST API for enriching TV newscast with online articles and media available in the Web	http://linkedtv.eurecom.fr/newsenricher/api/	Apache license 2.0. This service makes use of the search capabilities from Google CSE and Twitter API.

8 Annex II: Evaluation instructions for Topic Labelling

Precision

Here you will have to provide a score that represents how much you think *all* the topics detected for an article are relevant to the article's text. Provide *one* score for the entire set of topics per article.

Summation: *rate only what you see*.

You will have to take also into account the degree of each topic into your judgement. If you think that the topics are relevant to the text and their corresponding degrees also match the relevance with the content, even if there are a couple of 'outliers' topics that are mostly irrelevant but have a low degree, then give a better precision rating (e.g. 4) for this article. If you think that the relevant topics detected are too few compared to existing irrelevant ones and also have a somewhat weak degree although they should have a stronger one, then give a lower rating for this article.

How to interpret evaluation scores for *precision*:

- 0: the topics of the article, in combination with their respective degrees, have no relevance to the text for this article.
- 1: the topics of the article, in combination with their respective degrees, have little relevance to the text for this article, and the few relevant ones also have a low degree.
- ...
- 5: the topics of the article, in combination with their respective degrees, are almost perfectly relevant to the text for this article.

Recall

Here you will have to provide a score that represents how much you think the most prominent topics that *should* have described the text for an article, are in fact presented for each article. Provide *one* score for the entire set of topics per article.

Summation: *rate what you think you should have seen.*

Again, you will have to take also into account the degree of each topic into your judgement. If you think that the more prominent topics of the text were not detected at all (= are missing from the form), or few of them were detected with a really low degree, then give a lower recall rating for this article. If most of the topics presented, are in fact the most prominent ones that describe the text in your opinion and have an analogous degree, then give a better recall rating (e.g. 4) for this article. If most relevant topics were detected, but with a really weak degree (e.g. 0.2), then give a medium rating.

How to interpret evaluation scores for *recall*:

- 0: there are several topics that I can think of that represent the text of the article at hand, but none of them is present in the detected topics.
- 1: there are several topics that I can think of that represent the text of the article at hand, but only 1-2 of them are present in the detected topics - and those who are there have a very low degree. E.g. I can think of 8 topics that pertain to the text. Out of those 8, only 1 is detected OR out of those 8, two of them are in detected but with an extremely low degree (e.g. 0.1-0.2)
- ...
- 5: from all the topics that I can think of that represent the text in of the article at hand, most - or almost all - are present in in the detected topics and with an analogous (usually strong, e.g. 0.8 - 0.9 - 1.0) degree.