Int J Digit Libr
DOI 10.1007/s00799-016-0182-6

Evaluating unsupervised thesaurus-based labeling of audiovisual content in an archive production environment

Victor de Boer[1,2] · **Roeland J. F. Ordelman**[2,3] · **Josefien Schuurman**[2]

Received: 9 January 2016 / Revised: 7 June 2016 / Accepted: 17 June 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract In this paper we report on a two-stage evaluation of unsupervised labeling of audiovisual content using collateral text data sources to investigate how such an approach can provide acceptable results for given requirements with respect to archival quality, authority and service levels to external users. We conclude that with parameter settings that are optimized using a rigorous evaluation of precision and accuracy, the quality of automatic term-suggestion is sufficiently high. We furthermore provide an analysis of the term extraction after being taken into production, where we focus on performance variation with respect to term types and television programs. Having implemented the procedure in our production work-flow allows us to gradually develop the system further and to also assess the effect of the transformation from manual to automatic annotation from an end-user perspective. Additional future work will be on deploying different information sources including annotations based on multimodal video analysis such as speaker recognition and computer vision.

Keywords Audiovisual access · Information extraction · Thesaurus · Audiovisual archives · Practice-oriented evaluation

✉ Victor de Boer
v.de.boer@vu.nl

Roeland J. F. Ordelman
rordelman@beeldengeluid.nl

Josefien Schuurman
jschuurman@beeldengeluid.nl

[1] Department of Informatics, VU University Amsterdam, Amsterdam, The Netherlands

[2] Netherlands Institute for Sound and Vision, Hilversum, The Netherlands

[3] University of Twente, Enschede, The Netherlands

1 Introduction

Traditionally, audiovisual content in digital libraries is being labeled manually, typically using controlled and structured vocabularies or domain specific thesauri. From an archive perspective, this is not a sustainable model given (1) the increasing amounts of audiovisual content that digital libraries ingest (quantitative perspective), and (2) a growing emphasis on improving access opportunities for these data (qualitative perspective). The latter is not only addressed in the context of traditional search, but increasingly in the context of linking within and across collections, libraries, and media. Ultimately, search and linking is shifting from a document-level perspective towards a segment-level perspective in which segments are regarded as individual, 'linkable' media-objects. In this context, the traditional, manual labeling process requires revision to increase both quantity and quality of labels.

In earlier years, we investigated optimization of the labeling process from a "term suggestion" perspective (see e.g., [10]). Here the aim was to improve efficiency and inter-annotator agreement by generating annotation *suggestions* automatically from textual resources related to the documents to be archived. In [18] we defined collateral data[1] to refer to data that is somehow related to the primary content objects, but that is not regarded as metadata, such as subtitles, scripts and program-guide information. Previous work at our archive emphasized the ranking of possibly relevant terms extracted from the collateral text data, leaving the selection of the most relevant terms to the archivist [9]. The proposed term suggestion methods were evaluated in terms of Preci-

[1] This data is sometimes also referred to as 'context data' but as for example newspaper data can also be regarded as 'context' we prefer the term 'collateral data'.

Published online: 23 June 2016

sion and Recall by taking terms assigned by archivists as 'ground-truth'. The outcome was that a tf.idf approach gave the most optimal performance in combination with an importance weighting of keywords on the basis of a PageRank-type of analysis of keywords within the structure of the used thesaurus ($F@5 = 0.41$). One important observation of the study was that the inter-annotator agreement was limited, with an average agreement of 44 %. Although the results were promising, the evidence provided by the study was not conclusive enough to justify adaptations of the archival annotation work-flow and incorporate the suggested methodology. However, as the assumptions that drove the earlier study are still valid and have become even more clear and pressing, we recently took up the topic again. This time, however, from the perspective of fully *unsupervised* labeling. The main reason for this is that we expect that the efficiency gain of providing suggestions in a supervised labeling approach is too limited in the context of the increasing amounts of data that need labeling. Furthermore, instead of relying on topically condensed text sources such as program guide descriptions used in the previous study, we include a collateral text source more easily available in our production work-flow: subtitles for the hearing impaired. Finally, as inter-annotator agreement is expected to be limited, given the earlier study, we wanted to investigate how this agreement relates to an unsupervised labeling scenario that aims to generate labels for improving access to audiovisual collections. This makes our task different from more generic classification or tagging tasks such as that done in the MUMIS project [6].

In this paper, we present a two-stage evaluation of unsupervised labeling focusing on the practical usage of the method in an archive production environment. In Sect. 2 we overview the archival context of the labeling approach. In Sect. 3 we present the automatic term extraction framework. Section 4 describes the first stage of evaluations, focusing at determining parameter values. Section 5 then presents an evaluation of the framework after it was taken into production. Section 7 discusses and concludes the results from the evaluations, followed by some notes on future work.

2 Archival context

The implementation of innovative processes for automatic content annotation in an archive production work-flow needs to be addressed critically. A key requirement with respect to this type of innovation is that the archive remains in control of the quality of the automatically generated labels. Not only because of principals of archival reliability and integrity, but also from a service-level point of view. Media professionals use a broadcast archive to search for footage that can be re-used in new productions. The probability that their search process will get disturbed due to incorrect automatic labeling is undesired, despite the fact that the overall number of entry points generated by the automatic tool will increase, potentially having a positive effect on the search process.

Authority, being in control of the quality of the annotation tool, also means having control on parameters of the tool. In the case of automatic term labeling, two important variables are: (1) quality, specifically the balance between Precision (the number of true positives divided by the total number of elements labeled as belonging to the positive class) and Recall (the number of true positives divided by the total number of elements that actually belong to the positive class)[2] that controls the relation between quantity and quality of generated labels, and (2) the vocabulary that in an archival setting could be closely related to controlled vocabularies or thesauri that are used. In this work, the automatic labeling process is required to output terms that are defined in the Common Thesaurus for Audiovisual Archives[3] (GTAA). The GTAA is used in the Netherlands Institute for Sound and Vision (NISV)[4], which provides the context for this research.

The GTAA closely follows the ISO-2788 standard for thesaurus structures and consists of several facets for describing TV programs: subjects, people mentioned, named entities (Corporation names, music bands etc), locations, genres, producers and presenters. The GTAA contains approximately 160, 000 terms and is updated as new concepts emerge on television. For the implementation of unsupervised labeling in the archive's metadata enrichment pipeline, the balance between Precision and Recall, and the matching of candidate terms with the thesaurus have the main focus of attention.

2.1 Data

The general aim of the project for which the evaluation described in this paper was performed, is to label automatically the daily ingest of Radio and Television broadcasts. This data is quite heterogeneous: it contains news broadcasts, documentaries and talk shows but also sports and reality shows. As general-purpose named-entity extraction tools typically perform better for common entities as opposed to less common ones, we expect that the performance will differ for different genres.

For each program that is ingested also subtitles for the hearing impaired (TT888)—a verbatim account of the (Dutch) speech present in the data—is flowing into the archive. These TT888 files are used as input for the term-extraction pipeline described in Sect. 3. Instead of subtitles, other collateral data such as program guide information or production scripts and auto-cues could also be used. As the

[2] See also https://en.wikipedia.org/wiki/Precision_and_recall.

[3] http://datahub.io/dataset/gemeenschappelijke-thesaurus-audiovisuele-archieven.

[4] http://beeldengeluid.nl.

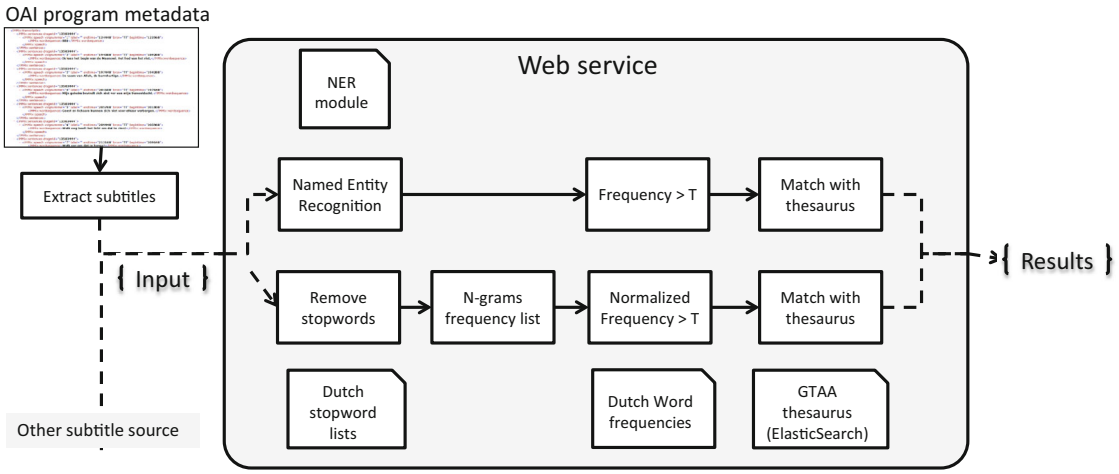

Fig. 1 Overview of the algorithm

availability of these data is less stable as is the case for subtitles, we focus on subtitles in the work-flow that forms the basis for the evaluations reported in this paper.

For evaluation purposes we selected one year of previously ingested programming for which we have manually generated labels, created by professional archivists. This set will be referred to as 'gold-standard' in our (pilot) experiments. However, as such a gold-standard implies exact matches or terminological consistency, we also asked professional archivist to assess the conceptual consistency (see also [12] about consistency, [10] for the approach that was taken earlier).

As discussed above, we use the internal thesaurus as a reference for extracted terms. The GTAA is available as Linked Open Data [3] and its concepts are identified through URIs. In the production system the extracted terms end-up as URIs identifying GTAA concepts unique IDs, which in turn can also be linked to and from using RDF relations. This allows us in the near future to reuse background information in the Linked Data cloud insofar as it is linked to or from those GTAA concepts. For the evaluation described here, the term-extraction pipeline only used the "subject" and "named-entities" facets of the thesaurus for validation.

3 Automatic term extraction

An overview of the term extraction pipeline is presented in Fig. 1. This shows the different steps performed in the algorithm, detailed below. The term-extraction pipeline is set up as a webservice. The webservice takes a single text, such as the subtitles for a television broadcast, as input and outputs a list of relevant thesaurus terms. This set-up allows the service to be re-used for other related tasks such as the extraction of terms from digitized program guides or other collateral data sources.

The web service is called through a HTTP post request, where the input text is passed in the body as a JSON string. At the same time, parameter settings can be passed in the same HTTP request to override default values for these parameters (see Sect. 3.4 for the parameters).

The output is a JSON object containing a list of thesaurus terms, on the basis of the parameter settings used (if not overridden, the default values are returned). For every term, also a *matching score* is returned (see Sect. 3.3). Within the archive production workflow, the service is called when new programs are ingested. The thesaurus terms provided by the service are then added to the program metadata without manual supervision.

For the experiments described below, the subtitles are derived from an OAI-PMH interface[5] to the archive's database. We retrieve for one or more programs the subtitle information from the OAI response (the program metadata) and remove the temporal metadata and other XML markup from the subtitles so that we end up with a single subtitle text per program. These are then presented one at a time to the service. As the extraction of subject terms and named entities require an individual tuning of parameters, the textual data is processed in two parallel tracks: one for subject terms and one for named entities (see Fig. 1).

3.1 Pre-processing and filtering

For the subject track, the first pre-processing step is to remove stopwords using a generic list of Dutch stopwords.[6] In the next step, frequencies for 1, 2, and 3-g are generated. For

[5] http://www.openarchives.org/pmh/.

[6] This was a list containing 104 words, retrieved from http://www.damienvanholten.com/blog/dutch-stop-words/. None of the words are labels of terms in the target thesaurus.

the uni-grams (single terms) also normalized frequencies are calculated using a generic list of Dutch word frequencies obtained from a large newspaper corpus.[7] In the filtering step, candidate terms (in the form of n-grams) above a certain threshold value of frequency scores are selected. Frequency scores are based upon both the absolute frequency (how often a term occurs in the subtitles) and a relative frequency (normalized by the frequency of the term in the Dutch language, only for 1-g). The frequency thresholds are parameters of the service. In the next phase, candidate n-gram terms are matched with terms in the thesaurus.

3.2 Named entity recognition

In the named-entity track of the algorithm, Named Entities (NEs) are extracted. Pilot studies as described in Sect. 4.1 determined that NEs—more so than non-entity terms—have a high probability of being descriptive of the program, especially if they occur in higher frequencies. For this track, we use a Named Entity Recognizer (NER). The NER is implemented as a separate module in the service and we experimented with different well-performing open-source NER systems for this module.

(1) **XTAS**. The NER tool from the open-source xTAS text analysis suite.[8]
(2) **CLTL**. An open-source NER module developed at the CLTL group.[9]

In the current Web service, the NER module to be used is a parameter of the method and can be set to "XTAS" or "CLTL" for the respective services. Both modules are implemented as wrappers around existing services which take as input a text (string) and as output a JSON list of entities and their types. The types used by the web service are *person*, *location*, *organization* or *misc*. Internal NE types from the individual modules are mapped to these four types

3.3 Vocabulary matching

The previous phases yield candidate terms to be matched against the thesaurus of five categories: subjects (from the subject track) and persons, places, organizations, and miscellaneous (from the NE track). The next step in the algorithm identifies the concepts in the thesaurus that match these terms.

As there can be many candidate terms at this stage and the GTAA thesaurus is fairly sizable with some 160,000 concepts, we need to employ a method for matching terms to thesaurus concepts that is scalable.

For this, the thesaurus has been indexed in an ElasticSearch instance.[10] ElasticSearch is a search engine that indexes documents for search and retrieval. In our case, thesaurus concepts are indexed as documents, with preferred and alternative labels as document fields. The concept schemes (facets or "axes" in the GTAA) are represented as different ElasticSearch *indices* which allows for fast search for term matches across and within a concept scheme. When searching for concepts matching a candidate term, ElasticSearch will respond with candidate matches and a *score* indicating the quality of the match between candidate term and the document.[11] In our algorithm, we employ a threshold on this score, resulting in an additional parameter. In this final step, the different categories of candidate terms are matched to a specific concept scheme. Specifically, PERSONs are matched to the "Persoonsnamen" (Person names) concept scheme in the GTAA thesaurus; both the subject terms and MISC terms are mapped to the "Onderwerpen" (Subject) concept scheme; PLACEs are mapped to "Geografische Namen" (Geographical Names); ORGANIZATIONs are mapped to the "Namen" concept scheme, which includes names for organizations.

3.4 Parameters

The algorithm parameters are shown in Table 1. This table shows the parameter name, the default value and the description. All default values can be overridden in the HTTP POST request. These default values were determined in pilot experiments (Sect. 4.1) and the experiment described in Sect. 4.2 was used to determine optimal values for a number of these parameters for a specific task.

4 Experiments

4.1 Pilot experiments

We performed a number of pilot experiments to fine-tune the setup of the main experiment. In one of these pilot experiments, we compared the output of an earlier version of the algorithm to a gold-standard of existing manual anno-

[7] We used the list provided by the OpenTaal society: http://www.opentaal.org/naslagwerken which contains frequencies of over 1 Million words.

[8] http://xtas.net/. Specifically, the FROG module was used using default settings.

[9] http://www.cltl.nl/. Here the OpenNER web service was used in combination with the CLTL POS tagger.

[10] http://www.elastic.co/products/elasticsearch.

[11] This score is the result of traditional TF.IDF measure and additional matching features as explained in https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html. Note that this score is not independent from (the size of) the corpus, so these values are not transferable to other situations.

Table 1 Parameters and default values for the service

nr.	Parameter name	Default	Description
P1	tok.min.norm.freq	4×10^{-6}	threshold on normalized freq for 1-g
P2	tok.max.gram	3	Maximum N for topic N-grams
P3	tok.min.gram	2	Minimum N for topic N-grams (excl. 1)
P4	tok.min.token.freq	2	threshold on absolute freq for 1-gram
P5	repository	cltl	NER module (xtas or cltl)
P6	ne.min.token.freq	2	Threshold on absolute freq for all NEs
P7	ne.organization.min.score	8	Threshold on ElasticSearch matching score
P8	ne.organization.min.token.freq	2	Threshold on absolute freq for
P9	ne.person.min.score	8	Threshold on matching score for persons
P10	ne.person.min.token.freq	1	Threshold on absolute freq for persons
P11	ne.location.min.score	8	Threshold on matching score for locations
P12	ne.location.min.token.freq	2	Threshold on absolute freq for locations
P13	ne.misc.min.score	8	Threshold on matching score for misc
P14	ne.misc.min.token.freq	2	Threshold on absolute frequency for misc

tations (see Sect. 2.1). The results showed that although there was some overlap,[12] comparing to this gold standard was not deemed by the experts to be an informative evaluation, since many "false positives" identified by the algorithm were identified to be interesting nonetheless. Therefore, in subsequent experiments, we presented the extracted terms to domain experts for evaluation. In this way, only precision of the suggested terms can be determined (no "recall"). This pilot also suggested that the correctness of suggested terms should be determined on a scale rather than correct or incorrect.

In a second pilot experiment, we presented extracted terms for random programs to four in-house experts and asked them to rate this on a five point Likert-scale [14]. The results were used to improve the matching algorithm and to focus more on the named entities rather than the generic terms since the matching here seemed to result in more successful matches. Lastly, in feedback to this pilot the experts indicated that for some programs the term extraction was considerably less useful than for others. This was expected but to reduce the amount of noise from programming that from an archival perspective has a lesser degree of annotation priority, we selected programs with a high priority.[13] For the main experiment we sampled from this subset rather than from the entire collection. From this evaluation we derived default parameter values shown in Table 1 which result in a limited amount of obvious errors including for example the value for P4, P6, P8, P10, P12 and P14 (minimum frequencies for terms to be considered a candidate term).

[12] For this non-optimized variant, recall was 21 %.

[13] This prioritization is done by archivists independently of this work. It is in use throughout the archive and mostly determined by potential (re)use by archive clients.

In the main experiment, the goal was twofold: (1) to determine the quality of the algorithm and (2) to determine optimal values for other system parameters.

4.2 Experimental setup

For the main experiment, we randomly selected 18 individual broadcasts from five different Dutch television shows designated as being of high-priority by the archivist. The individual broadcasts were randomly selected from the pool items that make up these shows. These shows are the evening news broadcast (4 videos), two talk shows (3 + 4 videos), a documentary show (4 videos) and a sports news show (3 videos). In Table 2, we list statistics for the duration, word count and number of terms generated for the 18 broadcasts. For each factor, we list the minimum, the maximum, average and standard deviation.

For these videos, we presented evaluators with (a) the video, (b) the existing metadata (which did not include descriptive terms) and (c) the terms generated by the algorithm using different parameter settings. The evaluators were asked to indicate the relevance of the terms for the video on a five-point Likert scale:

Table 2 Statistics for the 18 broadcasts

	Duration (min)	Subtitle word count	Terms generated
Min.	6.9	995	1
Max.	124.4	5771	35
Average	34.1	3237.6	15.7
St. dev	28.9	1638.0	9.5

0: Term is totally irrelevant or incorrect,
1: Term is not relevant,
2: Term is somewhat relevant,
3: Term is relevant,
4: Term is very relevant

4.2.1 Evaluators

The terms were evaluated by four different evaluators. These are Media Managers (archivists) at our institute and as such are very familiar with the material, with the manual annotation practice and with the thesaurus used. For our evaluation purpose, we are mainly interested in comparing the results of the automatic term extraction to in-house domain experts. Therefore, it is not possible to scale up the number of evaluators considerably.

4.2.2 Parameter settings

Parameters P1-P4 were set to their default values as listed in Table 1 as established in the pilot experiments for this specific task. For P5, we used both values, so both NER modules are evaluated. Some terms were found by both modules, and other terms were found by only one of the two. Evaluators did not see the origin of the terms. P6 was fixed to 2, as were the thresholds on the NE specific frequencies (P7, P9, P11, P13). For the Elasticsearch matching scores, we used a bottom threshold of 9.50 and presented all terms with a score higher than that value to the evaluators. We retain the scores so that in the evaluation we can compare the quality for threshold values of 9.50 and higher. The pilot studies showed that with thresholds below 9.50, mostly incorrect terms were added. The scores were also not available to the evaluators to avoid an evaluation bias. For the 18 videos, a total of 289 terms for XTAS and 222 terms for CLTL were presented to the evaluators.

4.3 Results

One of the evaluators (Eval4) finished all 18 videos. Table 3 shows the statistics for the four evaluators including the average score given for all terms. To measure inter-annotator agreement, we calculated the Pearson-coefficient between the pairs of evaluators. We did this only for those items for which both evaluators of the pair entered an evaluation. We used Pearson product-moment correlation coefficient since we here deal with evaluations on an ordered scale for which we assume a continuous scale (as opposed to for example Cohen's κ which is used for non-ordered categorical evaluations or Spearman's ρ measure which does not assume a continuous scale).The results are shown on the right in Table 4. The agreement between Eval2 and Eval3 is relatively low at 0.63, but for the other pairings it is 0.78 or higher indicating a strong agreement. For most of the subsequent evaluations, we use the average score for an extracted term given by the evaluator.[14]

Table 3 Evaluator results

Evaluator	Evaluated	Avg. score
Eval1	8	1.31
Eval2	14	2.21
Eval3	6	1.57
Eval4	18	1.64

Table 4 Inter-annotator agreement matrix

	Eval2	Eval3	Eval4
Eval1	0.81	0.79	0.84
Eval2	0.67	0.80	
Eval3	0.78		

4.3.1 Named entity modules

To determine the difference in quality of the two NER modules, we separated the scores for the two values (CLTL and XTAS) and determined the average score. If all terms are considered (respectively 289 and 222 terms for XTAS and CLTL), the average score for XTAS is 1.79 and that for CLTL is slightly higher at 1.94. We can also plot the average scores of the two modules, given a single threshold on the matching scores for the terms (in this case we use a single value for the threshold parameters P7, P9, P11 and P13). This is shown in Fig. 2.

This figure shows that the performance of the two modules is very comparable. It shows that at very low thresholds (between 9.5 and 10), the performance for both modules indeed drops considerably. Investigation of the data shows that below 10, mostly terms with average score 0 are added, which corresponds with findings from the pilot study. Furthermore, the graph shows that increasing the threshold, increases the average evaluation score for both modules. However, there is only a slight gain between 10 and 16. Based on these results, we concluded that the choice of NER module is of no great consequence to the overall quality of the results

4.3.2 Global precision values

Other than averages, we also determined precision values by setting cutoff points to the average score. Specifically, we

[14] For full transparency, we have published the raw evaluations and analyses for this experiment online at https://dx.doi.org/10.6084/m9.figshare.3187337.v1.

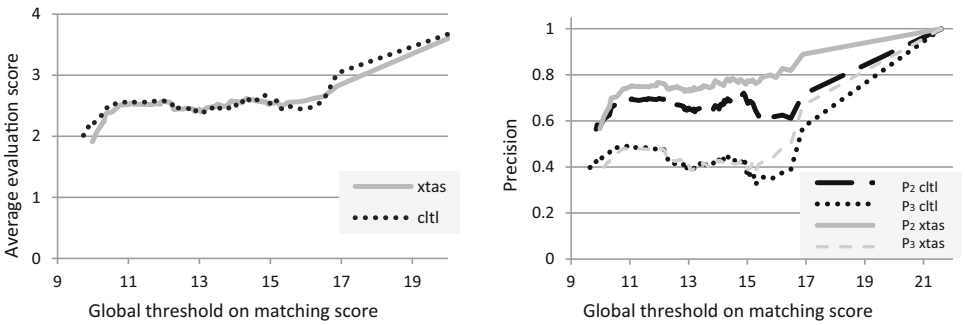

Fig. 2 Average scores (*left*) and precision graphs (*right*) for the global threshold values on matching score for the two NER modules

Table 5 Frequencies of terms in average evaluation bins for six threshold values

Score bin	Threshold											
	10		10.5		11		12		14		16	
	cltl	xtas	cltl	xtas	cltl	xtas	cltl	xtas	cltl	xtas	cltl	xtas
0–1	42	62	26	31	21	23	18	18	8	5	2	1
1–2	16	20	15	19	13	16	12	16	8	13	3	4
2–3	40	48	37	42	37	41	37	41	22	26	10	10
3–4	81	88	73	78	68	75	62	70	29	33	9	14
Total	179	218	151	170	139	155	129	145	67	77	24	29

calculate P_N which we define as the precision, given that a term with a score of N or higher is considered "correct". We calculate this for $N = 2$ and $N = 3$, which corresponds to minimum scores of "somewhat relevant" and "relevant" respectively. Figure 2 shows these values for the different global threshold values. Here, we can see that the P_2 values are around 0.7–0.8 for most threshold values (not considering very high values where very few terms are added. The more strict version of P_3 hovers around 0.4, which is considerably low. To get an even better insight in the hits and misses of the two versions of the algorithm, for different values of the threshold we list the number of terms evaluated in four bins (0–1, 1–2, 2–3, 3–4) . These are shown in Table 5 for both CLTL and XTAS. This table shows for example that given a threshold on the matching score of 11, the algorithm extracts a total of 155 terms when using the XTAS tool. In that case, 18 extracted terms receive an evaluation between 0–1 and 116 receive an average evaluation between 2 and 4 (41 + 75).

4.3.3 Individual score parameters

In the previous paragraphs, we have used a global value for the parameters P7, P9, P11 and P13. We now look at optimal values for each of these. For this, we weigh the Precision for each axis (Named Entity class corresponding to one of the four parameters) against an estimated recall. For this estimated Recall we assume that the total number of correct items for that NE class is the total number to be found. This means that the maximum Recall is 1.0 (which is found at threshold values 9.5). This is of course an incorrect assumption but it does give us a gradually increasing Recall when the threshold is lowered and a reasonable estimate for the true Recall. After calculating the Recall, we then calculated the F1 measure, which is the weighted average between Precision and Recall. All three values are calculated with the assumption that an average evaluation of 2 or higher is "correct", we therefore, get P_2, $R_{est,2}$ and $F1_{est,2}$. The maximum value for $F1_{est,2}$ is an indication for the optimum value of the threshold. These optimal values are presented in Table 6. This shows that the optimal threshold values are approximately 10 for person and 12 for locations and miscellaneous (regardless of the NER module). For organizations, the two modules present different values. This might reflect an artifact in the data

4.4 Result summary

The evaluation results indicate that the agreement between evaluators quite strong. Using their assessments as ground-truth we saw that precision values of around 0.7–0.8 are obtained in a less strict evaluation where terms should minimally "somewhat relevant" (P_2). When we apply a stricter evaluation that requires a term to be "relevant", performance drops to around 0.4. Concerning parameter settings, thresholds in the range of 10 for person and 12 for locations and miscellaneous provides optimal results. With respect to the two NER modules we have seen that the choice of NER module is of no significant consequence to the overall quality of the results.

Table 6 "Optimal" values for the threshold parameters for the four NE categories for both NER modules

	Threshold		P_2		$R_{est,2}$		$F1_{est,2}$	
	cltl	xtas	cltl	xtas	cltl	xtas	cltl	xtas
P7 (person)	10.12	10.12	0.58	0.54	0.88	0.83	0.7	0.65
P9 (organization)	10.56	12.05	0.8	0.76	0.89	0.85	0.84	0.8
P11 (location)	12.19	12.19	0.82	0.79	1.00	1.00	0.90	0.88
P13 (misc)	12.15	12.15	0.75	0.83	1.00	1.00	0.86	0.91

At these values the $F1_{est,2}$ is maximized

5 Experiments in the production environment

On the basis of the results of the experiments as described in the previous section, the term extraction module was taken into production. A number of small adaptations were implemented in a Java-based environment and adapted to comply with security and performance requirements.[15] This resulted in a production version of the Term Extraction Service (labeled TESS 2.0) which then ran with the default parameter values as described below for a number of months. The web service was called as part of the program ingestion service. The extracted terms are added to the program metadata in the multimedia catalog used by the archive [17]. To assess the performance of the production version, a second round of evaluations was performed. For this evaluation, we focused on programs selected by the evaluators. We were specifically interested in performance variance per type of term and per title. The intuition behind the latter analysis is that different types of titles would have different types of spoken—and therefore subtitled—text (interviews, news reports, sports broadcasts etc). Insight into this allows us to adjust parameter settings for each program type or in extreme cases do not accept generated terms for a specific title.

5.1 Evaluation setup

This evaluation was performed by two of the Media Managers (archivists) from the archive for terms extracted for a number of different television programs. The terms were extracted for programs that ran in July and August of 2015. The programs to be evaluated were chosen by the archivists. They selected programs of three title categories: evening news broadcasts (*8 uur Journaal*), sporting news broadcasts (*Sportjournaal*) and miscellaneous programs. These were selected as they represented a broad range of titles and are considered to have high priority by the archivists.

The generated terms were exported from the multimedia catalog in blocks of 2 weeks. The extracted terms, their program title and program ID where stored in a spreadsheet table and presented to the evaluators for the purpose of retrieving the original program and subtitle file to aid them in their evaluation of the extracted terms.

For each term/program combination, the evaluators were asked to determine two aspects:

Correctness: Is the extracted term correct? Does the term actually occur in some form in the subtitles? This aspect is evaluated on a binary scale (either "correct" or "incorrect").

Relevance: Is the extracted term relevant with respect to the topic of the program or item. This aspect is evaluated on a three point scale where *1 = "relevant", 2 = "term is not a main topic but is not disruptively irrelevant", 3 = "term is disruptively irrelevant"*. For incorrect terms, the evaluators were told they could leave out the relevance evaluation (a value of 3 is assumed).

The evaluators were asked to provide values for correctness and accuracy and to write any comments in the same row as the term-program pairs. They did not receive further training or instructions, but rather were instructed to judge the terms based on their normal annotation practice and experience.

5.2 Results

5.2.1 Data cleaning

In total, 2735 term/program pairs were evaluated.[16] However, of these, 684 were mistakenly evaluated on a different (five point) scale and therefore discarded. Some further data cleaning was performed. For the accuracy, we corrected typos removed unclear entries leaving 2051 evaluations. For correctness, four terms were evaluated as "unclear" or "?". We mapped these values to "incorrect".

[15] The term extraction service is developed as Open Source software and is made available through the Sound and Vision GitHub repository at https://github.com/beeldengeluid/term-extract. At the moment, the running web service is only accessible from inside the archive network.

[16] For full transparency, we have published the raw evaluations and analyses for this experiment online at https://dx.doi.org/10.6084/m9.figshare.3187337.v1.

As the evaluation task was nearly identical to the previous evaluation and performed by two of the same evaluators, we chose not to have overlapping evaluations between the evaluators. We therefore, did not perform inter-annotator agreement analysis.

5.2.2 Overall accuracy

In total, 1535 terms were evaluated as "correct" and 512 terms were evaluated as "incorrect", resulting in an accuracy of 0.75. When compared to the results found in Sect. 4.3, we see that this score for "correct and part of subtitle" terms is comparable to the precision scores found for P_2 where a positive a term needed to be both correct and "somewhat relevant".

5.2.3 Relevance

Of the 2051 evaluated terms, 1713 received an explicit relevance evaluation. This means that some incorrect terms still were assigned a relevance value by the evaluators (mostly with "3"). As described above, for the other incorrect terms we assume a value of 3 (irrelevant). In total, 1193 terms received value 1, 263 terms received value 2 and 595 terms received the value 3. This brings the average relevance for all 2051 terms to 1.71 ($\sigma = 0.89$). This means that the average relevance of all assigned tags is somewhere between "relevant" and "term is not a main topic but is not disruptively irrelevant".

The average relevance of all terms that were identified as being "correct" is 1.27 ($\sigma = 0.74$).

5.2.4 Results per concept scheme

To evaluate the performance per term type, we can use the SKOS Concept Scheme of the GTAA thesaurus that the term is a part of. TESS extracts terms from four GTAA Concept schemes (*Geographic Names*, which lists geographical entities, *Names*, which lists named entities such as organizations, *Topics*, which has general concepts, and *Persons*, which lists person names. Table 7 lists the results per Concept Scheme.

The results show that the accuracy for Names is the highest and for Topics the lowest. The latter is very low, compared to the other Concept Schemes even though the related parameter value (P13) is set relatively high and not many terms are extracted in this category. Still the majority is incorrect. The Topic terms also score worst in terms of relevance (at nearly 2.5), even if only the correct Topic terms are considered, the score is still only 1.5.

For the other Concept Schemes, the average relevance is around 1.5, which is between "relevant" and "term is not a main topic but is not disruptively irrelevant". The best scoring Concept Scheme in terms of both relevance and accuracy is "Geographic Names".

Table 7 Results of the second evaluation per concept scheme

	All	Geographic names	Names	Topics	Person names
Total evaluated	2051	560	181	89	1222
Total correct	1515	470	141	33	872
Accuracy	0.74	0.84	0.78	0.37	0.71
Avg. relevance	1.71	1.48	1.66	2.45	1.77
σ relevance	0.89	0.77	0.87	0.89	0.90
nr of 1's	1193	387	109	24	673
nr of 2's	263	76	23	1	163
nr of 3's	595	97	49	64	386

5.2.5 Results per title

We also analyze the term evaluation per Title. The term-program pairs come from 63 different titles, with 130 different programs (individual broadcasts). Some titles, especially the evening news and the sports news had evaluations for multiple programs from different days. We grouped the results per title.

Table 8 shows the results for the titles with 20 or more evaluated terms.

This table shows that the top-5 titles have more than half of the evaluated terms. These titles have programs that appear on television multiple times a week and are considered important titles, in the sense that they often contain material that can be reused and therefore, it is important they can be retrieved.

The top-5 contains a sports program and a news program as well as "magazine"-like programs with reports and interviews on current events. For these programs, the accuracy is above 0.73, with the news program (Journaal) having a very high accuracy and relevance of 0.91 and 1.21, respectively. Other news-like titles such as "Nieuwsuur" or "Eenvandaag" also have very high scores. This can be explained by the observation that in many cases when persons, geographic entities and such are mentioned, they are likely on topic of the program and therefore relevant.

Sports programs like "Studio Sport" and "Bureau Sport" perform less. This can partly be explained by the specific type of voice-over that these programs ending up in the subtitles. One example is the use of geographic names to denote sports teams ("Germany passes the ball around quickly"), here these geographic names end up being extracted, while the geographic entity is not really a topic of relevance.

Other low scoring programs include "Max TV Wijzer", "Van Moslimbroeders tot IS" and "Katholiek Nederland Kerkt in. . .". The former is a nostalgia-driven program about historical TV programs. Here, the format of the program,

Table 8 Results of the second evaluation per title for titles with 20 or more evaluated terms

Title	Evaluated terms	Total correct	Accuracy	Avg. relevance	σ relevance
All titles	2051	1515	0.74	1.71	0.89
Studio sport	489	358	0.73	1.75	0.89
Journaal	277	252	0.91	1.29	0.65
De Wereld Draait Door	220	177	0.80	1.65	0.82
Nieuwsuur	203	173	0.85	1.39	0.73
Eenvandaag	115	97	0.84	1.50	0.77
Sportjournaal	46	34	0.74	1.65	0.85
Max TV Wijzer	45	28	0.62	2.00	0.88
Jeugdjournaal	44	28	0.64	1.86	0.93
Bureau sport	43	27	0.63	1.88	0.93
Sterren.nl	38	30	0.79	1.82	0.80
Van Moslimbroeders tot IS	31	11	0.35	2.42	0.92
Katholiek Nederland Kerkt in...	27	7	0.26	2.52	0.85
De Beste Zangers van Nederland	22	16	0.73	2.18	0.66
Profiel	22	15	0.68	1.95	0.90
Spangas	22	15	0.68	1.77	0.92

with frequent references to other programs can be the cause of the low scores. The latter two titles include religiously-themed programs which have different formats as the more news-item programs. Analysis of the errors here showed that in both titles, there was one program where one ambiguous word was erroneously matched to many thesaurus terms. ("Music" to "Arabian Music", "Elektronic Music" etc.). This type of 'compound word matching' can be improved with relatively easy methods (see next section).

5.3 Discussion

For further analysis we looked at the incorrect entries and specifically at the notes the evaluators added to their evaluations. We did look at incorrect terms, but also at terms that were deemed 'correct', but were considered less relevant (relevance score 2) or irrelevant (relevance score 3).

5.3.1 Incorrect terms

Incorrect terms are mostly mismatches (false positives). These mismatches can occur at a number of stages in the algorithm. One source of mismatches can occur in the Named Entity Recognizer module. For example, the word "Person" was indicated as a person in one occasion and the word "tip" in another. In both cases, these false positives were in a later stage matched to the thesaurus, returning an extracted term.

Another, more frequent source of errors, is in the disambiguation step, where extracted terms cannot be matched with the thesaurus. For example, for one program, the political party with the abbreviation "PVV" was matched to the Flemish party, rather than to the Dutch party with the same name. In other cases, person names were often disambiguated to incorrect persons. This type of error can be addressed by more intelligent term disambiguation techniques such as the one described in [16]. Especially for geographical and person terms, we aim to implement specific disambiguation algorithms that improve on the current version. We discuss this further in Sect. 5.4 .

In some cases, the correct term does not occur in the thesaurus, but an incorrect term with the same label does. This happens for example with the topic "Radar", which does not occur as a topic in the thesaurus, but does occur as a name of a television program with the same name.

5.3.2 Correct but less relevant terms

In this evaluation, we were also interested in identifying what makes terms relevant according to the annotators. We therefore, looked at terms that were "correct" but received a relevance score of 2 or 3. These values were assigned to 262 and 595 extracted terms respectively.

These values were mostly assigned to terms that were not the main topic of a program. A large number of these are persons, places or topics that are mentioned (sometimes multiple times) by speakers in the video. One example is an interviewed speaker talking about his favourite singer. Here the interviewee is very important, but his or her name might not be said out loud (rather occurring in on-screen text). Another example is of sports coaches being named in reports of sports matches of the teams they coached. Even though the coach is named, the specific program itself is not about the coach.

More generally, relevance values of 2 were given often to terms that are mentioned in the subtitles, but cannot be seen in the video itself. This is definitely one of the downsides of using the subtitle text as the only source for term labeling, especially when the video images could be considered "more important" than the sound track.

For sports programs specifically, we found a number of cases where sports teams or players that were going to be discussed in a next broadcast were mentioned at the end of a show or where the presenter discusses sports teams that "will play tomorrow". By taking into account more of the context the algorithm might be amended to avoid terms that occur in such blocks of speech. However, these types are highly dependent on the type of title.

In some cases, ambiguous terms received lower relevance scores. One example are the geographic names to denote sports teams as described in Sect. 5.2.5. This could be partially solved by better matching of compound words. In some cases, compound words ("German team"), the 2-word term should not be split up and count towards a higher frequency of the word "German".

The frequency threshold is an effective but very simplistic way of filtering topical terms. In some cases, unimportant terms are repeated a lot. This can happen for rhetorical reasons (a speaker repeats one line with uncommon words multiple times, causing the words to cross the boundary). This can happen a lot in situations where song lyrics or poetry is part of the speech. These might be detected by specialized text analysis tools. A simple way might be to filter out exact duplicates of sentences. In one interesting example, a program about singers in the Dutch town of Volendam had the sentence "People from Volendam are not better singers than people from let's say Gorinchem or Veghel?[17] occurred three times, resulting in the —obviously not relevant— extracted geographical term Gorinchem.

5.4 Further improvements

Based on the error and relevance analysis, for the upcoming version of the TESS algorithm, we are working on a number of improvements.

As discussed above, matching of person names should be improved. Too many errors occur when only a first or last name is extracted and matched to the wrong person. In this case, we will implement the restriction that for a person to be matched, we need (at least once) a first and last name. This will remove some true positives, but will most certainly drastically reduce false positives.

The topic extraction as it is, does not perform up to the standard. For now, extracted terms from the Topic Concept Scheme are discarded. However, still there is a need for these topical keywords to be added as terms in the metadata. We are therefore, currently investigating techniques to reliably add good topical terms. These include latent semantic indexing techniques, or algorithms that use structured background knowledge such as Wordnet [8] for keyword expansion.

6 Related work

Our work is related to that described in [13]. In that paper, the authors describe how the collection of the British Broadcasting Corporation (BBC) was published as Linked Data [2]. To establish links between the different collections as well as to external datasets on the Web of Data. For this purpose, they use a legacy auto-categorization system called CIS. The entities found by CIS are then linked to external datasets, including DBPedia [1] using a vocabulary alignment approach. This is very similar to our task. However, here we did not have such a well-developed system in place and this was part of our effort described here. At the same time, as in the BBC, we are currently making an effort to link (part of) our collection to the Web of Data, using the GTAA as a stepping stone.

In [19], the Linked Media Framework is presented. This is a set of tools that allow for the easy development and deployment of media content and metadata using Linked Data principles. It includes an annotation and interlinking module, which can be directly used to establish links between the content and external datasets. Such a framework could be used to expose and further enrich the content.

We, here describe two experiments on automatic term extraction with professional archivists. In [21], the authors describe a number of experiments evaluating two metadata schemes developed for Moving Image Collections (MIC),[18] an integrated online catalog of moving images. Here, they focus specifically on the usefulness of metadata. In this paper, we focus on correctness and relevance. However, further studies could be done to assess how the automatically derived metadata scores on usefulness, extending previous research done at our institute [11].

7 Discussion and conclusion

In this paper we reported on the two-stage evaluation of automatic labeling of audiovisual content in an archive production environment. The aim was to evaluate if an unsupervised labeling approach based on subtitles using off-the-shelf NER tools and a baseline thesaurus matching approach would

[17] Translated from Dutch: "Volendammers kunnen toch niet beter zingen dan pakweg mensen uit Gorinchem of Veghel"

[18] http://imtcdrupal.imtc.gatech.edu/content/moving-image-collections-mic.

yield results that match archival production requirement with respect to quality, authority and service levels to external users. On average, accuracy levels of 0.75 are reached, with relevancy being evaluated as on average as being between 'relevant' and 'not main topic'. This is achieved with parameter settings that are optimized using a strict evaluation approach, allowing only terms when they are relevant as opposed to somewhat relevant. Precision given these parameter settings is sufficiently high not to disturb the archival quality requirements but the downside is that Recall is rather low as professional archivists label content with some labels that are not found by the automatic approach. However, given the pressure on manual resources in the traditional workflow, the current automated set-up is a useful starting point. Furthermore, having a stable production work-flow running allows us to (1) monitor the longitudinal behavior of the approach, among others by asking for feedback from external users, allowing us to assess the effect of the change also from an end-user perspective, and (2) work on incremental improvements, gratefully deploying the experimentation framework that was set-up during the research described here. We have seen that the NER modules used do not differ much, so that considerations such as stability, speed and resource use may be the most important factors for choosing a module. However, we note that we only tested two modules and there are many others around such as the Stanford NLP toolkit [15] or GATE [5]. It is likely that NER modules that are trained specifically on the type of input (in our case speech transcriptions) would improve performance both in terms of recall and precision.

In the experiments described in this paper, we employ archivists, who are extremely familiar both with the task and the content to provide high-quality evaluations and feedback. However, an interesting opportunity lies in using people other than experts to (continuously) assess the quality of the extracted terms. This could be done using crowdsourcing or nichesourcing [4]. User-provided content (annotations) matched to thesaurus terms can be combined with automatically extracted terms and added to the metadata. One important hurdle is that most of the content in the archive is access-restricted and therefore, cannot be exposed to the general public.

Other improvements in recall can be achieved through clustering of synonyms, using (external) structured vocabularies or by improving the named entity reconciliation (identifying the occurrence of the same entity in a text even though spelling variants are used).

The second round of evaluations in the production environment shows the value of differentiating across titles and types of titles. This indicates that term extraction from subtitles is especially successful for news-type titles and that for other often-occurring program types (e.g., sports shows), developing specific threshold values or extraction rules is likely beneficial. The analysis showed that for high-priority titles, and for specific types of terms (person, places and other named entities), the term extraction works with an accuracy of 0.75 or higher. For concept extraction, the tool needs to be further refined. For this, we are currently investigating the use of structured background knowledge for term expansion and other techniques, as introduced in the previous section.

One other direction for improvement we are also currently investigating is the use of other collateral data sources such as program guides and scripts, and combinations of data sources, potentially also coming from multimodal analysis components such as speaker recognition and computer vision [20]. When we can effectively combine evidence from multiple, redundant sources, it is likely we can counter the errors that stem from the biases of the specific sources (an example of which we saw in the "Gorinchem" case in our data). One way of combining evidence from multiple algorithms is by separately running these enrichment algorithms and then combining the results using, for example, a weighted average of scores. This results in a form of ensemble method for which machine learning techniques can be used to optimize the weighting [7]. Such a method is also likely to improve recall, as we are no longer constrained to what is said during a program, but also what is shown in the image. As these types of multimodal information extraction algorithms move from academic research to re-usable components, heritage institutions such as audio-visual archives can incorporate them successfully in their processing workflows.

Acknowledgments This research was funded by the MediaManagement Programme at the Netherlands Institute for Sound and Vision, the Dutch National Research Programme COMMIT/ and supported by NWO CATCH program (http://www.nwo.nl/catch) and the Dutch Ministry of Culture.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A Nucleus for a Web of Open Data. Springer (2007)
2. Berners-Lee, T.: Linked data-design issues (2006)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data—the story so far. Int. J. Semantic Web Inf. Syst. **5**(3), 1–22 (2009). doi:10.4018/jswis.2009081901
4. de Boer, V., Hildebrand, M., Aroyo, L., De Leenheer, P., Dijkshoorn, C., Tesfa, B., Schreiber, G.: Nichesourcing: harnessing the power of crowds of experts. In: Knowledge Engineering and Knowledge Management, pp. 16–20. Springer (2012)

5. Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H.: Evolving gate to meet new challenges in language engineering. Nat. Lang. Eng. **10**, 349–373 (2004). doi:10.1017/S1351324904003468. http://journals.cambridge.org/article_S1351324904003468
6. Declerck, T., Kuper, J., Saggion, H., Samiotou, A., Wittenburg, P., Contreras, J.: Contribution of nlp to the content indexing of multimedia documents. In: Enser, P., Kompatsiaris, Y., Connor, N., Smeaton, A., Smeulders, A (eds.) Image and Video Retrieval, Lecture Notes in Computer Science, vol. 3115, pp. 610–618. Springer Berlin Heidelberg (2004). doi:10.1007/978-3-540-27814-6_71
7. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple classifier systems, pp. 1–15. Springer (2000)
8. Fellbaum, C.: WordNet. Wiley Online Library (1998)
9. Gazendam, L., Malaisé, V., Schreiber, G., Brugman, H., et al.: Deriving semantic annotations of an audiovisual program from contextual texts. In: Proceedings of First International workshop on Semantic Web Annotations for Multimedia (SWAMM 2006), vol 23 (2006)
10. Gazendam, L., Wartena, C., Malaisé, V., Schreiber, G., de Jong, A., Brugman, H.: Automatic annotation suggestions for audiovisual archives: Evaluation aspects. Interdisciplinary Sci. Rev. **34**(2–3), 172–188 (2009). doi:10.1179/174327909X441090
11. Huurnink, B., Hollink, L., Van Den Heuvel, W., De Rijke, M.: Search behavior of media professionals at an audiovisual archive: a transaction log analysis. J. Am. Soc. Inf. Sci. Technol. **61**(6), 1180–1197 (2010)
12. Iivonen, M.: Consistency in the selection of search concepts and search terms. Inf. Process. Manage. **31**(2), 173–190 (1995)
13. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R.: Media meets semantic web—how the bbc uses dbpedia and linked data to make connections. In: The Semantic Web: Research and Applications, pp. 723–737. Springer (2009)
14. Likert, R.: A technique for the measurement of attitudes. Arch. Psychol. (1932)
15. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
16. Maynard, D., Ananiadou, S.: Acquiring contextual information for term disambiguation. In: Proc. of 1st Workshop Computational Terminology, Computerm98. Citeseer (1998)
17. Oomen, J., Ordelman, R.: Accessing audiovisual heritage: a roadmap for collaborative innovation. MultiMedia IEEE **18**(4), 4–10 (2011)
18. Ordelman, R., Heeren, W., Huijbregts, M., de Jong, F., Hiemstra, D.: Towards affordable disclosure of spoken heritage archives. J. Digit. Inf. **10**(6) (2009)
19. Schaffert, S., Bauer, C., Kurz, T., Dorschel, F., Glachs, D., Fernandez, M.: The linked media framework: Integrating and interlinking enterprise media content and data. In: Proceedings of the 8th International Conference on Semantic Systems, pp. 25–32. ACM (2012)
20. T. Tommasi, R. Aly, K. McGuinness, K. Chatfield, R. Arandjelovic, O. Parkhi, R. Ordelman, A. Zisserman, T.T.: Beyond metadata: searching your archive based on its audio-visual content. In: IBC 2014. Amsterdam, The Netherlands (2014). doi:10.1049/ib.2014.0003
21. Zhang, Y., Li, Y.: A user-centered functional metadata evaluation of moving image collections. J. Am. Soc. Inf. Sci. Technol. **59**(8), 1331–1346 (2008)