MeMAD

Methods for Managing Audiovisual Data

memad.eu
info@memad.eu

Twitter – @memadproject
LinkedIn – MeMAD Project

MeMAD Deliverable

D4.4 Report on Cross-Lingual Content Retrieval Based on Automatic Translation

Grant agreement number	780069
Action acronym	MeMAD
Action title	Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy
Funding scheme	H2020–ICT–2016–2017/H2020–ICT–2017–1
Version date of the Annex I against which the assessment will be made	23.6.2020
Start date of the project	1.1.2018
Due date of the deliverable	31.03.2021
Actual date of submission	25.03.2021
Lead beneficiary for the deliverable	University of Helsinki
Dissemination level of the deliverable	Public

Action coordinator's scientific representative
Prof. Mikko Kurimo
AALTO–KORKEAKOULUSÄÄTIÖ, Aalto University School of Electrical Engineering, Department of Signal Processing and Acoustics
mikko.kurimo@aalto.fi

MeMAD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780069. This document has been produced by theMeMAD project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Authors in alphabetical order		
Name	Beneficiary	e-mail
Maarit Koponen	University of Helsinki	maarit.koponen@helsinki.fi
Jorma Laaksonen	Aalto University	jorma.laaksonen@aalto.fi
Umut Sulubacak	University of Helsinki	umut.sulubacak@helsinki.fi
Jörg Tiedemann	University of Helsinki	jorg.tiedemann@helsinki.fi
Dieter van Rijsselbergen	Limecraft	dieter.vanrijsselbergen@limecraft.com

Internal reviewers in alphabetical order		
Name	Beneficiary	e-mail
Sabine Braun	University of Surrey	s.braun@surrey.ac.uk
Michael Stormbom	Lingsoft	michael.stormbom@lingsoft.fi

Abstract

In this deliverable, we report on our automatic content retrieval experiments and their implications for improving the discoverability of archive content, with a focus on cross-lingual retrieval, but also including our additional cross-modal retrieval tests.

First, we introduce the methods we used to simulate a realistic mixed-language media archive using the raw data from a publicly available collection of annotated images. We discuss the ways in which automatic content retrieval on this archive parallels or diverges from content search in the MeMAD prototype platform (Limecraft Flow), to clarify the extent to which they overlap. Afterwards, we describe how we further processed the data, drawing from our expertise in machine translation and image processing in order to enrich the archive, and to improve content retrieval performance. Next, we provide our experimental findings from using textual metadata translations and automatically-generated image captions to expand the metadata, as well as our tests on performing retrieval beyond using simple textual search queries. Our findings unequivocally validate the utility of metadata translations for cross-lingual content retrieval, and further encourage additional venues for cross-modal and multimodal retrieval methods. We describe these findings in detail alongside the empirical scores we have obtained from our own evaluations, and conclude the report with our general impressions and the lessons we have learned from this study.

Contents

1 Introduction

This report covers our efforts in exploring cross-lingual and cross-modal extensions of audiovisual content retrieval in the framework of the MeMAD project. One of the goals in the project is to provide efficient tools for content producers and consumers with support of multilingual sources. Machine translation (MT) is an essential tool that can facilitate the search of content in other languages. Below, we provide the background and related work as well as our own development in improving the performance of cross-lingual content retrieval.

The task for the MT module here is different from the other applications that we have reported in WP4. The result of the translation process is no longer intended for end-users or professional translators in post-editing workflows, but instead fed into a search engine with a content ranking algorithm. This means that the focus of the translation engine is to provide the important content to enhance search performance instead of providing fluent and grammatical output to be consumed by humans. Hence, adequacy becomes more important than fluency, and the MT process is hidden under the hood of the search engine.

Another essential task in our setup is the integration of features from different modalities. We focus on visual features combined with text-based retrieval, and report experiments on image retrieval as a proxy for video content retrieval. The setup was chosen due to the lack of appropriate benchmarks in cross-lingual video retrieval. Nevertheless, the effect of MT has also been tested as part of UC2.2 in the framework of the MeMAD prototype.

Below, we first summarise related work in cross-lingual content retrieval, and introduce our own methodology and integrations of machine translation and image captioning in image retrieval. Afterwards, we provide experimental results in the task of image retrieval, explain how they might relate to UC2.2 and the end-user experience when working with enhanced search facilities in the environment of Limecraft Flow, and conclude with our general remarks.

2 Background on cross-lingual information retrieval

Cross-lingual (or cross-language) information retrieval (CLIR) is a widely studied topic which focuses on approaches to make information available across language barriers in a situation where the language used for searching is different from the language in which the information is provided. As with most natural language processing tasks, much of the research on information retrieval has focused on English. However, unlike other NLP subfields, IR research included tasks in other languages already for a substantial amount of time because of the practical needs of search engine development for non-English languages. Such scenarios can be further divided into bilingual information retrieval, where queries are in one language while the content is in another, and cross-lingual retrieval, which can involve more than two languages, but these terms are often used interchangeably (Savoy and Braschler, 2019). According to Savoy and Braschler (2019), it is generally assumed that in the multilingual scenario, each language corresponds to a separate document collection, and the search must then be performed separately over several collections and languages.

In some scenarios, it may be possible to find some information on the basis of cognates and loan words, but such situations are generally limited to only closely related languages (see Savoy and Braschler, 2019). Normally translation is needed to cross the language barrier. The

basic approaches are to either translate the query into the language of the content (query translation), or to translate the content into the language used for querying (document translation). The first, simplest, approach to translation is the use of bilingual dictionaries which provide possible translations for each of the words in the query, often with some kind of weighting based on frequency or word sense disambiguation (Oard, 1998; Savoy and Braschler, 2019). However, ambiguity, unknown words and proper names may cause issues with dictionaries, and the comparison by Oard (1998) showed that MT approaches outperform dictionaries for precision. MT can translate free text and mitigate coverage problems in dictionaries, which has made it the dominant approach for CLIR, and research has often utilised free online MT systems (Khwileh et al., 2016).

Query translation is more commonly used than translating the content. The benefits of query translation are that it has a lower cost per query and can be implemented online for real-time querying with relatively short lag due to the shortness of the queries (Khwileh et al., 2016; Savoy and Braschler, 2019). On the other hand, the downside of the query translation approach is that information retrieval is very sensitive to the quality of the translated query (Khwileh et al., 2016). Translation errors are caused by ambiguity of polysemous and synonymous words, as well as proper names, and to tackle these issues, queries may be expanded with additional terms (e.g. morphological variants or related concepts) in the source language prior to translation (Savoy and Braschler, 2019). Translating the types of queries typical in information retrieval is particularly challenging because they are generally very short, lack sufficient context for MT, and do not form grammatical sentences (Khwileh et al., 2016; Saleh and Pecina, 2016). Some work has experimented with MT systems which are specifically adapted to the features of user queries and produce multiple alternative translation hypotheses which are then ranked and used for querying (Saleh and Pecina, 2016). In a cross-lingual scenario where the content being queried includes multiple languages, one approach is to translate the query into all target languages and to perform the search separately in each separate language collection. However, the translation into multiple languages may become difficult when the number of language pairs grows, and a further challenge is presented by merging the results from different languages (Savoy and Braschler, 2019).

The alternate approach of document translation has the benefit of providing more context for translation, which generally improves translation accuracy and makes the information retrieval less sensitive to translation errors (Savoy and Braschler, 2019). Document translation has indeed been found to improve the precision of search results over query translation (Oard and Hackett, 1997; Oard, 1998). Since document translation is done offline and no translation is needed at the time of querying, this approach also allows for the potential use of MT systems fine-tuned for the specific content and possibly even better retrieval performance (Khwileh et al., 2016). The downside of document translation is that it is computationally "costly", and translating the entire collection may not be feasible due to constraints on time and resources needed for translating particularly large amounts of content involving multiple languages (Khwileh et al., 2016; Savoy and Braschler, 2019). In a scenario with multilingual content, possible approaches are to translate all content from different source languages into one single target language (often English), which is then used for querying, or to translate all source languages to all target languages so that queries can be performed in any language (Savoy and Braschler, 2019). As noted by Savoy and Braschler (2019), scalability is a challenge particularly for the all-to-all scenario when the number of languages grows. One additional possibility is to use hybrid approaches which can involve translating both queries and

documents in parallel and then ranking results achieved with both, or a combination where a limited number of documents are first retrieved with translated queries, and then translated and re-ranked against the original source language query (Saleh and Pecina, 2016).

3 Cross-lingual retrieval of multimodal content

Most research on cross-lingual information retrieval has focused on textual information, but as the amount and importance of multimedia content has grown, research has also been increasingly focusing on multimodal content retrieval. On the other hand, various initiatives and evaluation challenges like TRECVID[1] and MediaEval[2] have addressed multimodal content retrieval, but these generally lack a cross-lingual aspect.

One initiative aiming to bring together the multimodal and cross-lingual aspects has been the CLEF evaluation initiative (originally Cross-Language Evaluation Forum, later Conference and Labs of the Evaluation Forum). According to Ferro (2019, p. 4), this series of evaluation campaigns and conferences was started in 2000 motivated by the challenge of developing "fully multilingual and multimodal information access systems" where the user would be able to enter queries in any language and be presented with "relevant information from a multilingual multimedia collection in any language and form". Although the ultimate goal is that, ideally, such system would integrate information retrieved from different modalities, in practice, research on multimodal content retrieval has mainly addressed them separately (Ferro, 2019). Various labs have focused on cross-lingual approaches involving images (e.g. Arni et al., 2008; Ionescu et al., 2019; Piras et al., 2019), speech (e.g. Pecina et al., 2008), and video (e.g. Larson et al., 2009). (See also Ferro, 2019, for an overview.)

The cross-lingual retrieval of multimedia content often also utilises textual information in the form of metadata. For example, in the ad-hoc image retrieval task (built on evaluation setups also used in the TREC campaigns) described by Arni et al. (2008), the multilingual aspect of the challenge was achieved by having a collection of photographs annotated with "semi-structured captions such as the title, location, description, date or additional notes" in both English and German, and then for one evaluation condition randomly masking the annotations in one language for each image. In their overview of various domain-specific image retrieval tasks organised as part of the ImageCLEF labs from 2007 to 2013, Piras et al. (2019) note that while approaches combining textual annotations and visual features became more common over the years, systems utilising visual features alone remained rare and were outperformed by systems using both. Due to the additional uncertainty introduced into the retrieval by translation (see Savoy and Braschler, 2019), higher precision is generally seen for monolingual than multilingual content retrieval, although Piras et al. (2019) note that overall the difference in performance was small. In addition to manually annotated images, some of the cross-lingual image retrieval tasks have also experimented with datasets using automatically collected data in the form of web images and the surrounding text on webpages. However, such data have been found to be problematic for retrieval purposes, because the relationship between the text and image is not necessarily clear, and much of the text may be unrelated (Piras et al., 2019).

Work on cross-lingual retrieval of video content is also dependent on metadata, which may include information such as titles or free text descriptions provided in either professional col-

[1] https://trecvid.nist.gov/
[2] http://www.multimediaeval.org/

lections like news archives or user-generated collections like YouTube (Khwileh et al., 2015; Küçük and Yazıcı, 2013; Braslavski et al., 2016). Practical constraints on manual labelling of videos can, however, lead to metadata being inconsistent and sparse, particularly in the case of user-generated data, where the quantity and reliability of metadata varies considerably, as does the robustness of different metadata fields (e.g. title vs free text) for cross-lingual retrieval purposes (Khwileh et al., 2015, 2016). To respond to the practical constraints of manual annotation, the use of automatic information extraction from the audiovisual components has been explored. However, Küçük and Yazıcı (2013) note that the automatically-extracted data are generally not sufficient for extracting the semantic information needed, and a gap remains between the automatic data and how users interpret the content of videos. To address these issues, some work has therefore explored the use of audio transcripts for indexing and retrieval of video content and MT of the transcripts to further enable cross-lingual retrieval (Küçük and Yazıcı, 2011, 2013; Khwileh et al., 2015, 2016). The work by Küçük and Yazıcı (2011, 2013) also utilises named entity recognition on the audio transcripts. Combining automatic speech recognition to produce the transcripts with MT for translation naturally adds further noise to the outputs, as ASR quality varies depending on the language as well as audio quality and the speaking style of the individual speakers (Khwileh et al., 2015). In many cases, the video alone will be useful as an answer to the query, like in the "how-to" question use case examined by Braslavski et al. (2016), but Küçük and Yazıcı (2011) also note that document translation when applied to ASR of the foreign language videos has the additional benefit of giving the information seeker also the gist of what is being said.

4 Content retrieval methodology

In the context of the use case UC2, the way in which we perform retrieval has been formulated as a search through the textual metadata annotated on the items that constitute a media archive. The cross-lingual element in these searches are realised by use of search queries in one language to retrieve videos in other languages from a representative media archive established on the MeMAD prototype platform (Limecraft Flow). In the corresponding evaluations with participants from interest groups, we use a sampling of videos from the providing partners Yle and INA, with Finnish (or bilingual Finnish/Swedish) and French metadata attached, respectively. We assume that any such reasonably large archive would contain media (at least partially) in languages other than the primary language(s) for metadata annotation. Retrieval based on matches in textual metadata could then be expected to run into language barriers, only managing to return off-language results occasionally on surface similarities like cognate words or proper names in search queries. To extend the coverage of our search queries, the main method that we have agreed on is to enrich media archives with translations of the metadata in other languages for which to enable cross-lingual content retrieval.

In parallel with the UC2.2 archive search use case evaluations, we have conducted a series of experiments using a basic automatic retrieval system in order to investigate the theoretical utility of various retrieval methods. Our experimental settings, discussed in detail in Sections 4.2 and 4.3, simulate searches through a set of different media archives, each one making use of metadata in a different way. We base our simulations of media archives on a collection of annotated images rather than videos, primarily because datasets of images with textual metadata are more plentiful and accessible than those of videos (Sulubacak et al., 2020). An-

⋮	⋮	⋮
105	DE	chinesische Schriftzeichen
	EN	chinese characters
	FR	caractères chinois
106	DE	Familienstammbaum
	EN	family tree
	FR	arbre généalogique
107	DE	Nahaufnahme einer Sonneblume
	EN	sunflower close up
	FR	gros plan de tournesol
108	DE	Karneval in Rio
	EN	carnival in Rio
	FR	carnaval de Rio
⋮	⋮	⋮

Table 1: Topic examples with trilingual descriptions from the ImageCLEF Wikipedia Image Retrieval Dataset.

other important reason is that we needed to be able to apply the image processing expertise in MeMAD to our tests, exploring additional venues for multimodal search, while our experience with automatic video processing has not been comparable. Furthermore, the large body of scientific research on content retrieval spearheaded by the CLEF initiative, and especially the long-running ImageCLEF shared tasks on cross-lingual image retrieval, have also had a strong influence on our decision.

The ImageCLEF Wikipedia Image Retrieval Dataset (Tsikrika et al., 2012) was created as an exclusive test set for the ImageCLEF 2010 Wikipedia Retrieval Task (Popescu et al., 2010), and expanded further for the rerun of the task in 2011 (Tsikrika et al., 2011). It is a fairly large collection of 237 434 images from Wikipedia[3], annotated via crowdsourcing for relevance or non-relevance to a diverse set of 50 topics. The topics have short descriptions in English, French and German attached to them, each formulated concisely as shown in Table 1, making them ideal for use as search queries. Each topic also has a longer, English-only description called the "narrative", which mainly served as instructions for the crowdworkers annotating topic relevances for the images. Further textual annotations in the same three languages, such as captions, descriptions and comments, along with the English Wikipedia articles containing the image, are attached to each image as metadata. The dataset also contains low-level visual features of the images, as well as 5 additional held-out representative images for each of the 50 topics, in order to facilitate retrieval based on the visual modality. The release contains only a small, high-confidence subset of image-topic pairs explicitly annotated for relevance (see Table 2), while the majority of pairs are left without explicit relevance annotations due to ambiguity. As for the other annotations, roughly half of the images have textual metadata in only one language, and only about a fifth of the remaining images have annotations in all three. Overall, the dataset is quite heterogeneous, but not so noisy as to be lacking structure, which makes it the perfect collection for simulating a realistic media archive.

[3] https://www.wikipedia.org

topic ID	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**	**81**	**82**	**83**
relevant	164	229	229	156	157	179	136	154	159	144	135	189	144
non-relevant	3660	3239	3840	4024	3614	4085	3582	3738	4133	3850	4183	4494	3990
topic ID	**84**	**85**	**86**	**87**	**88**	**89**	**90**	**91**	**92**	**93**	**94**	**95**	**96**
relevant	111	156	136	166	137	124	190	164	107	215	311	241	217
non-relevant	3666	4020	3928	3888	4328	3735	3467	4396	4515	3833	3649	4286	4380
topic ID	**97**	**98**	**99**	**100**	**101**	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**
relevant	179	137	212	68	159	45	66	34	341	108	36	72	43
non-relevant	3888	4228	4165	1634	1756	1885	1907	2321	1873	2138	2106	1464	1950
topic ID	**110**	**111**	**112**	**113**	**114**	**115**	**116**	**117**	**118**	**119**	**120**		
relevant	618	72	120	293	49	137	134	41	32	121	46		
non-relevant	2230	1394	1768	1825	1760	2095	2730	1773	1863	1742	2099		

Table 2: The numbers of images explicitly annotated as relevant or as non-relevant for each of the 50 topics in the ImageCLEF Wikipedia Image Retrieval Dataset, out of the total number of 237 434 images.

4.1 Content retrieval in the MeMAD prototype platform

In this section, we describe the search and retrieval methodology employed in Limecraft Flow, the platform of the MeMAD prototype, in detail. While searches on this platform are multi-faceted, complex, and optimised for the retrieval of videos rather than images, they are still based on search indices built from textual metadata attached to the media. The details provided below are intended to facilitate comparisons to the automatic image retrieval methods explained later in Sections 4.2 and 4.3, and clarify the extent to which the findings from our experiments would be applicable for searches on this platform.

The search functions within the prototype platform are based on an Apache Lucene/Solr[4] search index system with a customised indexing scheme. The search index supports full text search and advanced search queries on all aspects of the data — comments, fixed and custom fields, but also more extensive data such as transcripts and subtitles. Essentially, all data that can be reduced to some textual form can be indexed and then searched through. This means that all content features to be searched through are reduced to a text form before they can be indexed. The nature of the system also dictates a natural preference for document translation approaches to content retrieval, as the bulk of the processing is done during document indexing with the aim of making the query process as lightweight and responsive as possible.

The search index operates by means of data denormalisation to provide full-text search capabilities with reasonable response times. All relevant data that should match a search query is grouped into a single 'document' of the smallest useful granularity (e.g., a temporal subclip or a transcript paragraph). This relevant data includes both data applicable to that segment's granularity (e.g., the transcript's speaker and text) but also those data inherited from broader levels or granularity (e.g., metadata that belong to the entire encompassing clip). By collecting this data into single index documents, search queries can be made to match in full to single documents without requiring complex joining operations that would severely slow down the search result retrieval process.

[4]Further details and documentation can be found at https://solr.apache.org/

4.1.1 Populating the search index

All data eligible to be searched is indicated as such and is then put into an inverted search index. This inverted index maps unique values of text (or other types of data) to each indexed document in which the search term occurs. Search queries are executed by matching the query's terms against all documents in this inverted index, which produces a list of matching documents as the search result. This search result is then interpreted by client applications (including the platform's web GUI) to determine which subclips, transcripts, or entire clip's metadata actually matched the given query.

We summarise how this general concept is implemented in practice in the search engine set up on the platform, (a) to enable the indexing of very heterogeneous metadata on the one hand, and (b) to support the retrieval of multilingual metadata on the other hand.

1. All metadata in the platform is stored as an *Annotation* object type. An annotation is linked to an audiovisual clip and is optionally delimited in time (i.e., has a temporal start and duration along the audiovisual content's time line) and describes on aspect of the content. The annotation finds its origin in the storage of comments and segment descriptions, but is now used for all kinds of metadata. Within the *Annotation*, generic metadata fields (custom named fields) are available for a variety of purposes. These can be added by users at will and are all indexed to be queried. Non-generic content related specifically to the *Annotation*'s metadata type, e.g., an dialogue transcript from an ASR system, a person identification from a facial recognition system, etc. are stored as in the *Annotation*'s structured content field (which internally stores a schema-less structured JSON object).
2. The following metadata was produced by the final version of the integrated prototype (refer also to D6.8 for a more in-depth discussion of all components involved):

   - 'Legacy' metadata sourced from the original archive systems and data sets provided by Yle and INA. Both data about programs in their entirety as data describing temporal segments of this content were provided and imported into the platform. This data typically represents the authoritative description of the content, curated by professional archivists.

     Legacy metadata is typically available in a single language, often the native language used at the institution that provided that data, in this case French for INA and Finnish for Yle. Additionally, some fragments of Legacy metadata can be in other languages, e.g., the language of the source material if it was acquired from an external source.
   - Dialogue transcript metadata obtained from an ASR process that converted the audio signal into a literal textual representation of the spoken dialogue. This data includes information on speaker turns, per-word timing and (depending on the algorithms used) accuracy scores and transcript alternatives.

     Due to the inherent nature of the ASR process, the literal transcription implies that the language of the content dictates the language of the transcript, and this language is known at all times because current state-of-the-art ASR systems still need to be instructed which language to 'transcribe' in.
   - Person identification metadata, obtained from a facial recognition process, which returned person names and spatial and temporal coordinates of when and where that person occurs in the audiovisual content.

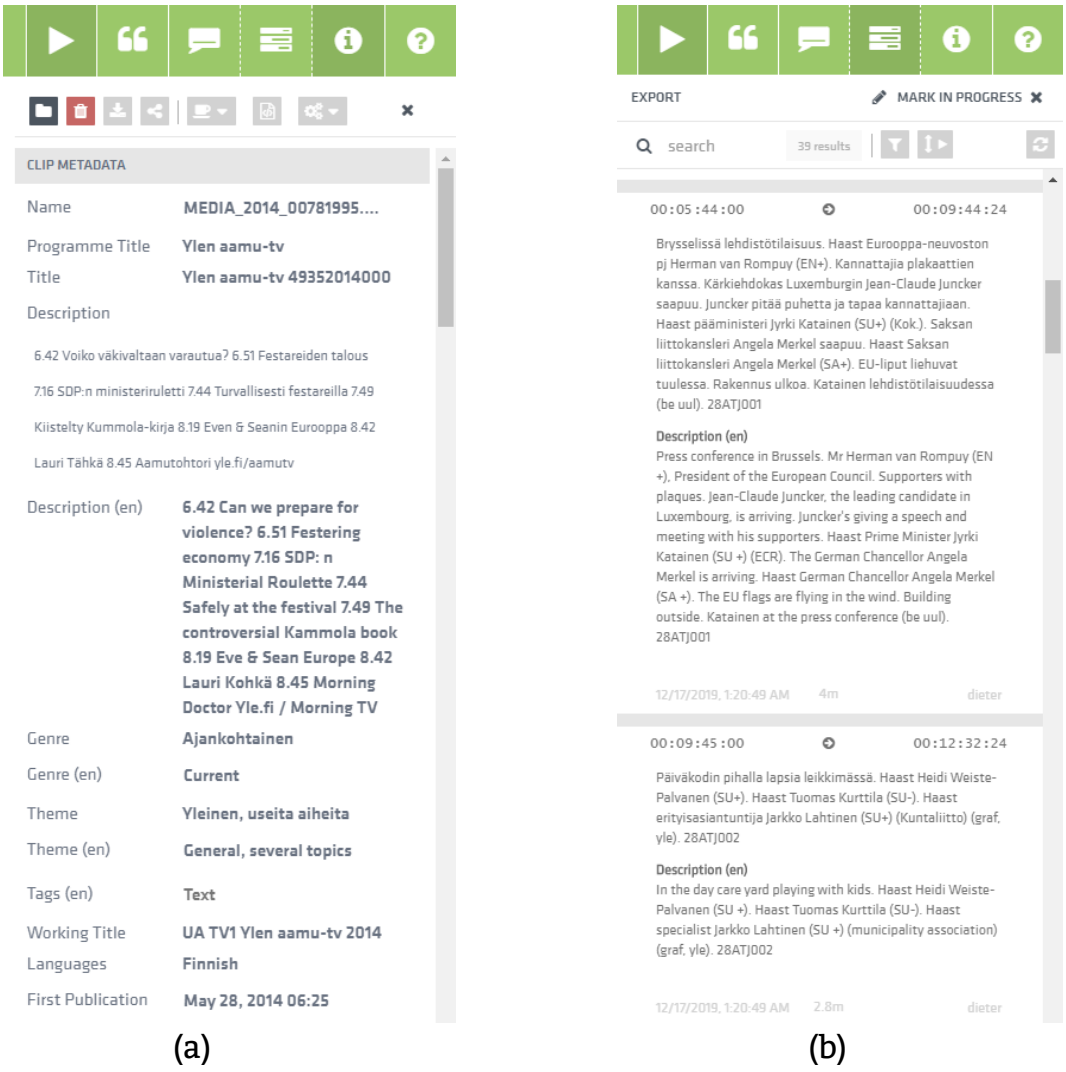

Figure 1: Translated metadata visualised in the Limecraft Flow prototype platform, with per-clip metadata in panel (a) and segmentation metadata in panel (b).

Person identification data is language-agnostic to a certain extent, and depends on the labelling assigned by the recognition system, which typically consists of not much more than a textual string. If that is the case, these labels can be disambiguated at a later stage (cf. NER data below).

- Textual metadata obtained from OCR processes that analyse each image in a video content item and turns that back in the textual representation that was originally recorded as part of the image acquisition.

  Metadata obtained from out-of-the-box OCR systems are often assigned language labels, and as such, this can be taken into account for classifying OCR results.

- Named entity metadata obtained from a NER process, which in turn uses either former form of metadata to identify text elements that represent a known and named entity of a set of categories (e.g., persons, places, countries, etc).

  Disambiguated named entities can often benefit from knowledge stored in central data sources such as linked open data knowledge bases for information and translations of the entity's name into different languages if applicable. This information is stored whenever it is made available by the NER process.

3. Each relevant metadata field from the above list was translated into matching counterparts by a collection of MT tools, chosen based on the input and output language pairs of the translation in question. Tools included the OPUS-MT tools (see D4.3) and commercial offerings such as the Deepl translation service. Translations ranged from single concepts (e.g., a program genre), short descriptions (e.g., one-liner descriptions of a content segments) and multi-sentence transcript paragraphs. The result is illustrated in Figure 1.

   Overall, we ensured that each element was translated in such a way that in total, all elements were available in English (as the de facto lingua franca used in the project), Finnish (representing the content contributed by Yle and one of the low-resource languages addressed by the project) and French (representing the content contributed by INA). An exception was made for translations of named entity labels, which were source from Wikidata and for which we used all multilingual labels available for a given entity. Essentially, we follow the document translation approach (cf. Savoy and Braschler (2019)), but limit the computational requirements by only machine translating to a limit set of language relevant for the use cases and content of the project.

4. Once all items are enriched with metadata and this metadata is subsequently translated, the results (all stored as *Annotations*) are indexed by the system. This involves the following steps.

   (a) An index function is applied to each *Annotation* to obtain a set of index documents that can be added to the search index:

   $$\{d \mid d \in document_{index}[\, fields_{denorm}, timing, ID_{lang}, ID_{content}\,]\} = f_{index}(Annotation) \quad (1)$$

   This function (which is configured using a custom Solr indexing schema) maps fields from a variety of *Annotation* subtypes to an index document that will serve as a potential search result. Common *Annotation* fields are handled identically for each *Annotation* type, while the schema also specifies how to deal with specific structured content, e.g., the structure of NER entities or transcript data. The index document also contains identifiers of the content to which it relates, as well as timing information on where it applies.

   (b) Denormalisation of fields such that type information is included in the index, with and without language identifies, which enables drill-down type searches, e.g., faceted search to narrow in on search results even with limited query input.

   (c) The values of the denormalised fields added to index documents by the index function are then injected into the search index. These are the values that will eventually be matched by the search query. This step includes another optional transformation, potentially, to map language or script-specific constructs to more common expressions that will make querying easier. Examples include replacing accented characters with accent-less characters, or compensating for specific per-language spelling constructs.

Using the mechanism described above the heterogeneous data stored in the platform can be indexed for searching using a common pipeline, while at the same time cross-lingual retrieval is ensured by (a) inserting all source data and derivative translations in the index, and thereby (b) ensuring that language identifiers are properly encoded in the index document such that language filtering is possible at all stages of the retrieval process.

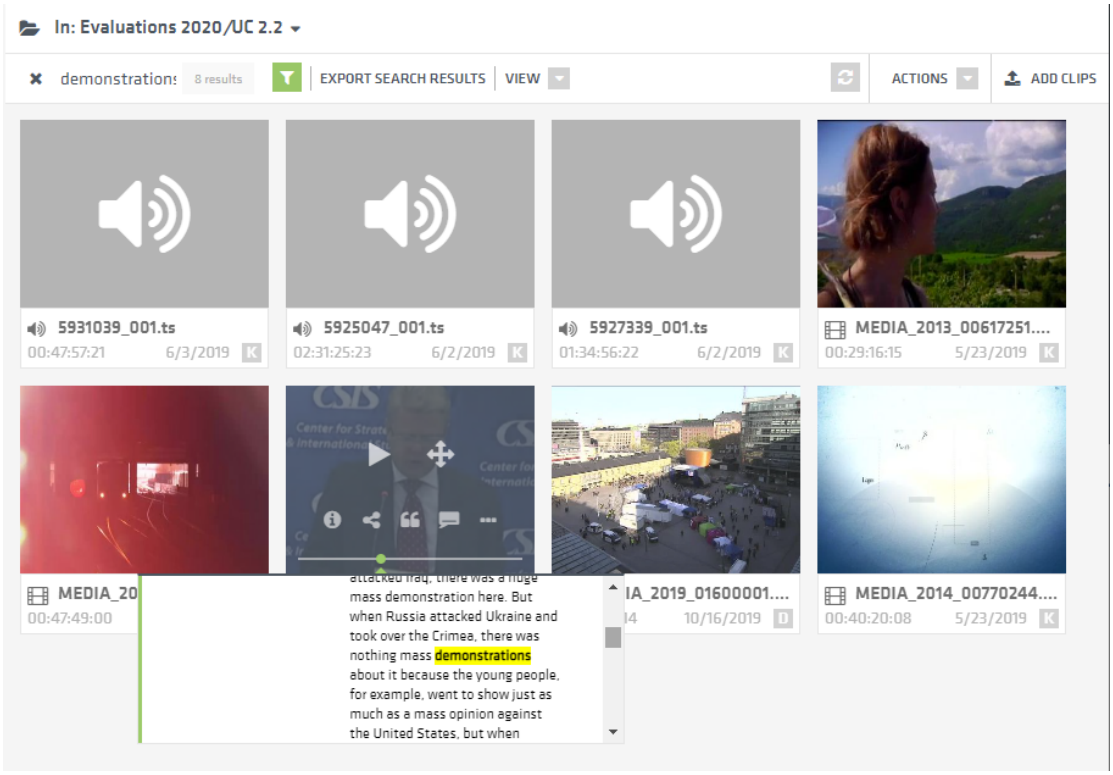

Figure 2: Free-text search across all indexed metadata, with visualised search results, including dialogue transcript fragments that match the query.

4.1.2 Querying the search index

Content retrieval is a combination of the query process against the index and the subsequent visualisation and interaction with (intermediate) search results, and the two go hand-in-hand to deliver an effective content retrieval system.

- No query translation is performed at present. The queries are interpreted as is by the Solr index system. This does include a specific query syntax to allow for combination of search criteria, as explained in more detail in D6.5.
- All index documents matching a query (where a match is defined by the application of the search operators and query) are returned by the search system. It is then up to the client application to interpret (and possibly post-process) the results and display them to the user or instruct further queries that can lead to a more accurate end result.
- We provide users with multiple search paradigms for cross-lingual retrieval. We illustrate them here.
  - **Free-text querying:** As illustrated in Figure 2, users can search for the occurrence of any string of free text in any of the indexed metadata by entering that string in the search box. The example shows a search for “demonstrations” and displays the search results below.
    Note in particular that the search returned cross-lingual results since the English search query “demonstrations” was found exclusively in machine-translated transcripts derived from metadata for the original content that was available only in either French or Finnish.

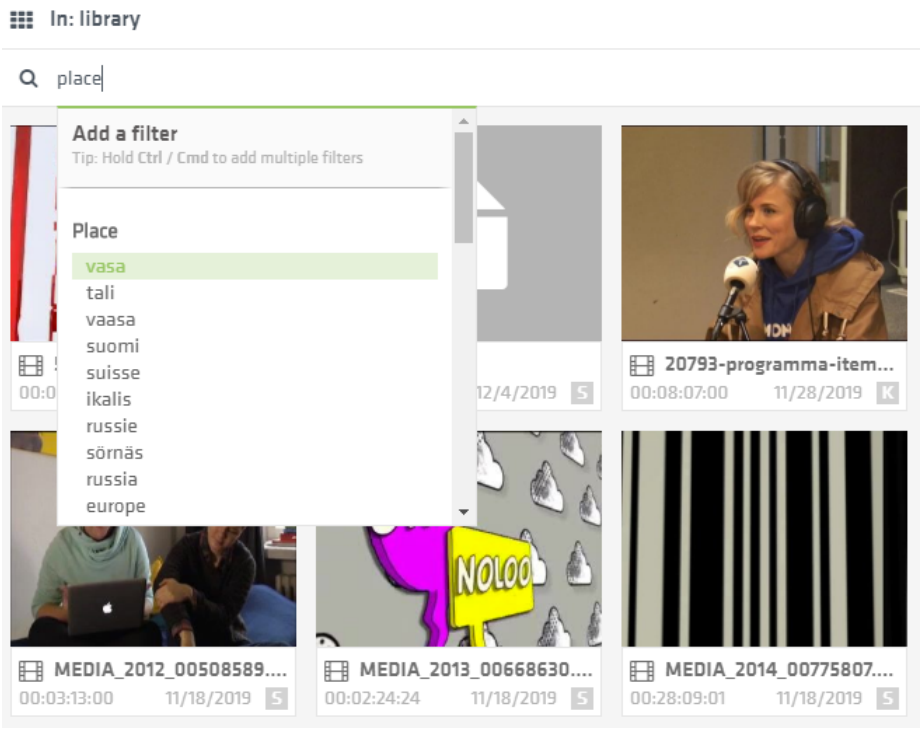

Figure 3: Auto-suggest search shows faceted results with multilingual labels all mapped to the '**Place**' field.

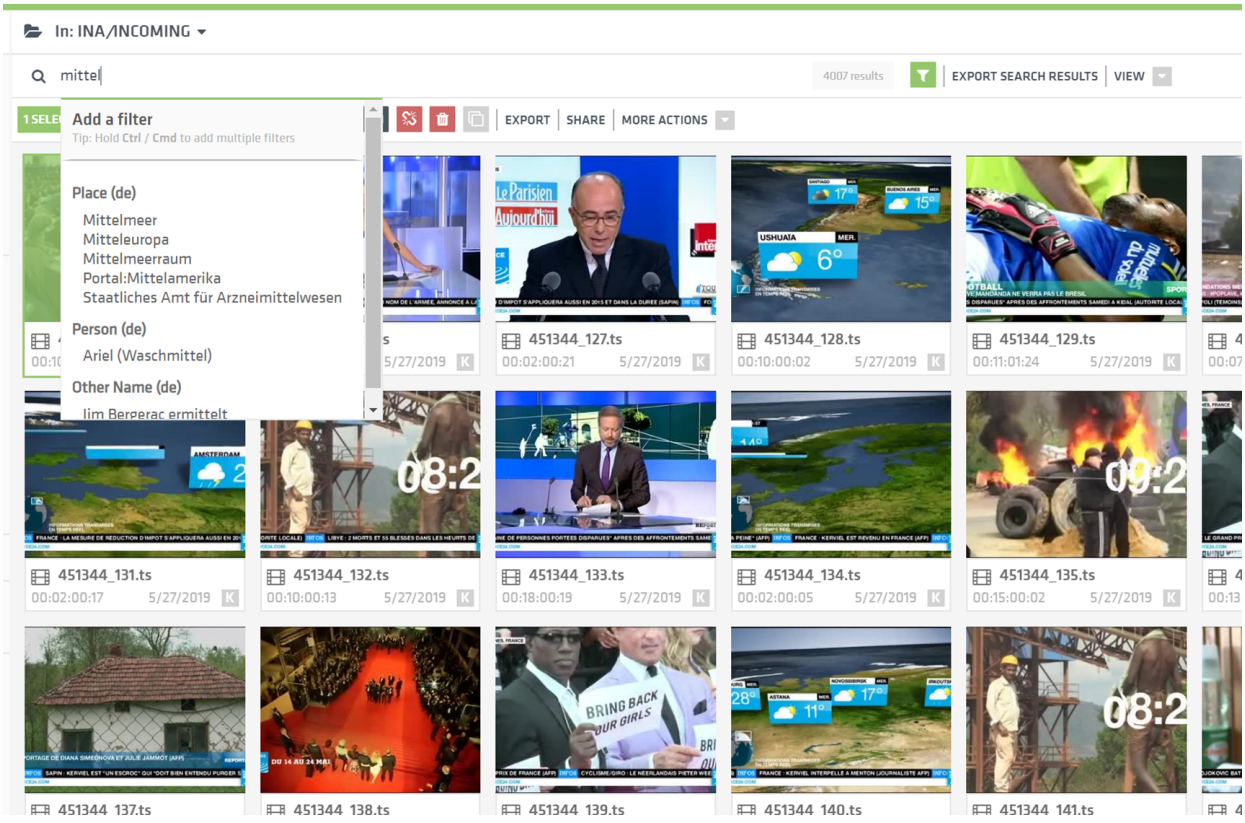

Figure 4: Auto-suggest search shows faceted results with multilingual labels across different metadata fields.

– **Context-dependent search result visualisation**. The one highlighted results in Figure 2 also depicts how the client application interprets and visualises the results depending on the source metadata. In this case, the match was found in a dialogue transcript (visualised in the pop-up panel) and its temporal position is indicated by an interactive dot on the clip's timeline. This allows users to interact with the search results in a more intuitive fashion. More information on this topic is available in D6.5.

– **Complex search queries**. Free text searches can be combined with metadata field names and search operators to construct complex search queries that provide users with more flexibility than free-text searches. Please refer to D6.5 for more details on this feature.

– **Faceted search**. In addition to free-text searches and complex search queries, users can also interactively narrow down on search results using a faceted search mechanism. This is exposed, a.o., via the auto-suggestion interface when users type terms in the search box, as depicted in Figures 3 and 4. When users enter the specific field they would like to search in (e.g., "Place"), the autosuggestion returns a set of multilingual suggestions as shown in Figure 3. Reversely, they can also enter the query phrase, which is then broken down in suggestions according to the field the value can be found in, cf. Figure 4. Again, this feature works cross-language as the inverse index locates the query regardless of language, and then the match can be inverted to the source document and metadata field that originated from the indexing phase.

4.2 Text based automatic retrieval

To construct our simulated media archives for the automatic retrieval experiments, we start by collating the available metadata from the Wikipedia Image Retrieval Dataset. Our motivation is to build a minimal acceptable setting for retrieval, on which to explore the effects of various enrichments and retrieval strategies, rather than identify a single best strategy and fine-tune the setting to maximise automatic retrieval performance. With this in mind, our aim is to extract a bare-bones archive using the metadata sparingly and without applying any special processing, focusing instead on having a balanced distribution of available metadata between English, French and German.

Altogether, the dataset organises metadata in several strata: Full texts of the Wikipedia articles in which each image appeared, short captions, longer descriptions, and contributor comments for the image, and the topic(s) for which it was annotated as relevant and non-relevant, each with a set of descriptions. The Wikipedia articles typically have a larger scope than the part that the associated image relates to, which causes a mismatch between the topic of the image and that of most of the article's textual contents, making the full text a severely noisy source of metadata for use in retrieval. Although it might have been possible to use heuristic approaches to extract parts of the text most likely to be related to the image, this is not a straightforward task, and comes with the risk of error propagation. For these reasons, we do not use the full texts from the articles in our automatic retrieval experiments. We make use of the other strata in order to create a variety of contrastive archive *settings*, each with a different composition of metadata, and perform our evaluations on each *setting* to observe the individual effect of each component.

```
"243658": {
    "metadata-de": {
        "topic-description": [
            "herzförmig"
        ]
    },
    "metadata-en": {
        "image-caption": "Two hands forming the outline of a heart shape.",
        "image-description": "Two left hands forming an outline of a heart
           shape against a blue sky. Both hands are wearing a similar wedding
            ring.",
        "topic-description": [
            "heart shaped"
        ],
        "topic-narrative": [
            "Photos or paintings of any kind of object in heart shape are
               relevant. Photos or drawings of the human heart are not
               relevant."
        ]
    },
    "metadata-fr": {
        "topic-description": [
            "forme de coeur"
        ]
    },
    "non-relevant-topics": [
        96
    ],
    "relevant-topics": [
        79
    ]
}
```


Figure 5: Metadata of an example image shown in our JSON format, with one relevant (#79, *"heart shaped"*) and one non-relevant (#96, *"shake hands"*) topic annotation, descriptions of the relevant topic in three languages and a *"narrative"* in English, and an image caption along with a description in English, from `setting-original`.

For ease of processing, we start from a maximally inclusive setting by collating all image captions, descriptions and comments, topic relevance annotations, and descriptions of the relevant topic(s) together in a JSON-formatted file, as shown in Figure 5. We use the name `setting-original` for this initial setting. Since it directly incorporates the topic descriptions that we also use as search queries in our experiments, searches through an index built from this setting are guaranteed to find exact matches of the queries. This is unrealistic to expect in a real media archive, since users are in practice allowed to submit freeform search queries, and it is impossible for media to have such comprehensive metadata as to guarantee matches with all possible forms that a relevant query may take. Rather, the point of this setting is to investigate how well the automatic retrieval system is theoretically able to perform, under the assumption that it always has access to *some* form of directly-identifying metadata. The expectation is that this ideal performance will still be imperfect due to limitations of the search engine, which will be useful to know in assessing the utility of other setting variants.

```
"243658": {
    "metadata-de": {},
    "metadata-en": {
        "image-caption": "Two hands forming the outline of a heart shape.",
        "image-description": "Two left hands forming an outline of a heart
           shape against a blue sky. Both hands are wearing a similar wedding
            ring."
    },
    "metadata-fr": {},
    "non-relevant-topics": [
        96
    ],
    "relevant-topics": [
        79
    ]
}
```


Figure 6: Metadata for the example image in Figure 5 after it has been filtered for `setting-masked`.

We derive our next setting from `setting-original` first by removing all topic descriptions, and thus resolving the case against directly-identifying metadata. The absence of these strata forces the automatic retrieval to be made based on matches of search queries with other types of metadata that we might more realistically expect a media archive to contain, such as image captions and descriptions. Afterwards, we mask the remaining strata on all images with multilingual metadata, so that each would only retain a portion of their metadata in a single language, using a greedy approach. To accomplish this, we go through all the images in `setting-original` in order, keeping a tally of how many times each language was represented across what we have traversed. When we encounter an image with metadata in more than one language, we only keep metadata in the language with the least representation so far, and mask the others. The result is a reduced setting which we call `setting-masked`, representing a simulation of a mixed-language media archive with a roughly equal distribution of languages, and strictly monolingual metadata attached to each item. Figure 6 shows how the example in Figure 5 looks after this masking procedure. We consider this as a minimal baseline on which to evaluate the performance of automatic retrieval before we apply any textual enrichments.

We introduce two different types of enrichments as applied over both `setting-original` and `setting-masked`: Automatic translations of existing textual metadata, and automatically-generated image captions. To produce metadata translations, we use six general-purpose MT models released earlier as part of the MeMAD subtitle translation pipeline[5], trained to translate in all directions between English, German, and French. With this setup, we take each individual stratum of metadata (e.g. an image description in French), and translate it from its original language to the other two languages (e.g. produce English and German translations of the image description). We add these translations as additional metadata under new strata (see Figure 7), marking the setting with the extension `.translations` (e.g. `setting-masked.translations`). Enriching a setting with metadata translations is a unimodal procedure, but it also establishes a potentially crucial layer of multilingual metadata to facilitate cross-lingual searches. Therefore, we expect it to make a substantial impact on

[5]Deliverable D4.3 contains more information on the models, as well as instructions for translating non-subtitle data with them.

```
"243658": {
    "metadata-de": {
        "image-caption-from-en": "Zwei Hände bilden den Umriss einer Herzform.",
        "image-description-from-en": "Zwei linke Hände bilden einen Umriss einer Herzform gegen einen blauen Himmel, beide Hände tragen einen ähnlichen Ehering."
    },
    "metadata-en": {
        "image-caption": "Two hands forming the outline of a heart shape.",
        "image-description": "Two left hands forming an outline of a heart shape against a blue sky. Both hands are wearing a similar wedding ring."
    },
    "metadata-fr": {
        "image-caption-from-en": "Deux mains formant le contour d'une forme cardiaque.",
        "image-description-from-en": "Deux mains gauche formant un contour d'une forme de cœur contre un ciel bleu, les deux mains portent une bague de mariage similaire."
    },
    "non-relevant-topics": [
        96
    ],
    "relevant-topics": [
        79
    ]
}
```


Figure 7: Metadata for the example image in Figure 6 enriched with automatic translations.


```
"243658": {
    "metadata-de": {},
    "metadata-en": {
        "auto-caption": "a hand holding a pair of scissors in front of its face",
        "image-caption": "Two hands forming the outline of a heart shape.",
        "image-description": "Two left hands forming an outline of a heart shape against a blue sky. Both hands are wearing a similar wedding ring."
    },
    "metadata-fr": {},
    "non-relevant-topics": [
        96
    ],
    "relevant-topics": [
        79
    ]
}
```


Figure 8: Metadata for the example image in Figure 6 enriched with automatically-generated image captions.

```
"243658": {
    "metadata-de": {
        "auto-caption": "Eine Hand, die eine Schere vor dem Gesicht hält.",
        "image-caption-from-en": "Zwei Hände bilden den Umriss einer Herzform.",
        "image-description-from-en": "Zwei linke Hände bilden einen Umriss einer Herzform gegen einen blauen Himmel, beide Hände tragen einen ähnlichen Ehering."
    },
    "metadata-en": {
        "auto-caption": "a hand holding a pair of scissors in front of it s face",
        "image-caption": "Two hands forming the outline of a heart shape.",
        "image-description": "Two left hands forming an outline of a heart shape against a blue sky. Both hands are wearing a similar wedding ring."
    },
    "metadata-fr": {
        "auto-caption": "Une main tenant une paire de ciseaux devant son visage.",
        "image-caption-from-en": "Deux mains formant le contour d'une forme cardiaque.",
        "image-description-from-en": "Deux mains gauche formant un contour d'une forme de cœur contre un ciel bleu, les deux mains portent une bague de mariage similaire."
    },
    "non-relevant-topics": [
        96
    ],
    "relevant-topics": [
        79
    ]
}
```


Figure 9: Metadata for the example image in Figure 6 enriched with both captions and translations.

overall retrieval performance across searches using different query languages. The procedure we use to automatically generate image captions is based on a cross-modal neural architecture that condenses the major visual elements in the input image into a short verbal description (Shetty et al., 2018; Laaksonen and Guo, 2020). Our system produces original outputs in English, which we later translate to French and German using the same pipeline as for the other metadata translations. We add these automatically-generated captions as added metadata under strata of their own (see Figure 8), though only in the language(s) in which each image already had prior metadata (e.g. if an image only had French metadata, we only add French auto-captions). We use the extension `.autocaps` for the resulting enriched setting (e.g. `setting-masked.autocaps`). Unlike translated metadata, automatically-generated captions comprise a multimodal enrichment, but they purposefully do not affect multilinguality, allowing us to gauge the relative utilities of these methods.

```
<DOC>
  <DOCNO>243658</DOCNO>
  <TEXT>
    Eine Hand, die eine Schere vor dem Gesicht hält.
    Zwei Hände bilden den Umriss einer Herzform.
    Zwei linke Hände bilden einen Umriss einer Herzform gegen einen blauen
       Himmel, beide Hände tragen einen ähnlichen Ehering.
    a hand holding a pair of scissors in front of it s face
    Two hands forming the outline of a heart shape.
    Two left hands forming an outline of a heart shape against a blue sky.
       Both hands are wearing a similar wedding ring.
    Une main tenant une paire de ciseaux devant son visage.
    Deux mains formant le contour d'une forme cardiaque.
    Deux mains gauche formant un contour d'une forme de cœur contre un ciel
       bleu, les deux mains portent une bague de mariage similaire.
  </TEXT>
</DOC>
```

Figure 10: Metadata for the example image in Figure 9, converted to the TREC format for indexing in Zettair.

Finally, we generate one last variant of our metadata settings by adding together the contributions from the translations of existing metadata and from automatically-generated captions, in order to observe their combined effects. This results in a both multilingually- and multimodally-enriched setting (see Figure 9), which we mark with the extension `.fully-enriched` (e.g. `setting-masked.fully-enriched`). In this variant, we ensure that each image has at least *some* amount of raw or translated metadata, including automatically-generated captions with their translations, in all three languages. Altogether, the combined enrichments in this variant make it the most information-packed of all metadata settings, which would presumably lend itself to our best retrieval performances. Metadata translation and automatic image caption generation are both imperfect procedures that occasionally produce erroneous or noisy outputs (see e.g. the automatically-generated caption in Figure 8). When such processes are cascaded together, error propagation typically causes these errors to be amplified, which may cause more false positives than before to appear in our retrieval results. Therefore, it is hard to predict the effect of this final enriched setting on the performance of automatic retrieval with high confidence.

We use the free software Zettair[6] to index the image metadata collected in our contrastive settings, and later make text searches through these indices to retrieve ranked lists of matches. Zettair is a lightweight and fast search engine much like the Apache Lucene/Solr search engine used in the MeMAD prototype platform for the search use case evaluations. We have opted for Zettair as it is easier to set up, and followed a retrieval approach comparable to the solution in the MeMAD prototype, expecting that our experimental findings would translate into similar trends in the integrated framework. To build the search indices, we convert our JSON-formatted settings each into the input format that Zettair expects, which is an XML file describing lists of documents using the TREC tag set (see Figure 10). Once built, the indices are then used to run search queries, and return a list of image IDs ranked by how well their metadata match the search query, as shown in Figure 11.

[6]Zettair is available for download from http://www.seg.rmit.edu.au/zettair under a BSD licence.

```
$ zet -f setting-original.zettair-index -n 3 "heart shaped"
>   1. 243658 (score 6.995820, docid 229698)
>   2. 80732 (score 6.952699, docid 75825)
>   3. 80869 (score 6.942686, docid 75946)
>
>   3 results of 506 shown (took 0.229290 seconds)
>   229352 microseconds querying (excluding loading/unloading)
```

(1) **(2)** **(3)**

Figure 11: An example Zettair search through the index built from `setting-original` to retrieve the top 3 matches for the query *"heart shaped"*. In each match, `score` indicates confidence, and `docid` is the image ID.

4.3 Visual similarity based automatic retrieval

Image retrieval based on visual similarity to query images typically uses an approach where visual features with fixed dimensionality of real values are extracted from all query and database images (Veltkamp and Tanase, 2002). The mutual similarity of two images, I_i and I_j, can then be defined based on a selected *similarity measure* $s(\cdot,\cdot)$ applied to the extracted features as

$$s(I_i, I_j) = s(\mathbf{f}[I_i], \mathbf{f}[I_j]) = s(\mathbf{f}_i, \mathbf{f}_j) \;, \tag{2}$$

where the notation $\mathbf{f}[\cdot]$ is used to denote the feature extraction process and the f's stand for the resulting features. Typical choices for the similarity measure are the *dot* or *inner product* $p(\cdot,\cdot)$ and *cosine* similarities $c(\cdot,\cdot)$ defined as

$$p(\mathbf{a}, \mathbf{b}) = \mathbf{a} \cdot \mathbf{b} = \sum_{k=1}^{d} a_k \, b_k \tag{3}$$

$$c(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|} = \frac{\mathbf{a} \cdot \mathbf{b}}{\sqrt{\mathbf{a} \cdot \mathbf{a}} \sqrt{\mathbf{b} \cdot \mathbf{b}}} \;, \tag{4}$$

where d is the dimensionality of the extracted visual features. Alternatively, it is likewise possible to base visual example based retrieval on a *distance* between the feature vectors instead of their similarity. In that case, the *Euclidean distance* would be the most straightforward choice.

Let us assume that we have a set of n query images $Q = \{I_1^q, I_2^q, \ldots, I_n^q\}$ and the corresponding visual features $\{\mathbf{f}_1^q, \mathbf{f}_2^q, \ldots, \mathbf{f}_n^q\}$. Similarly we can assume to have a database of N images $B = \{I_1^b, I_2^b, \ldots, I_n^b\}$ and the corresponding visual features $\{\mathbf{f}_1^b, \mathbf{f}_2^b, \ldots, \mathbf{f}_n^b\}$ among which the similar images are being retrieved. A straightforward way of defining the similarity between a query image I_i^q and *any* image in database B is to use the maximum of the similarities as

$$s(I_i^q, B) = \max_{j=1}^{N} s(I_i^q, I_j^b) \;. \tag{5}$$

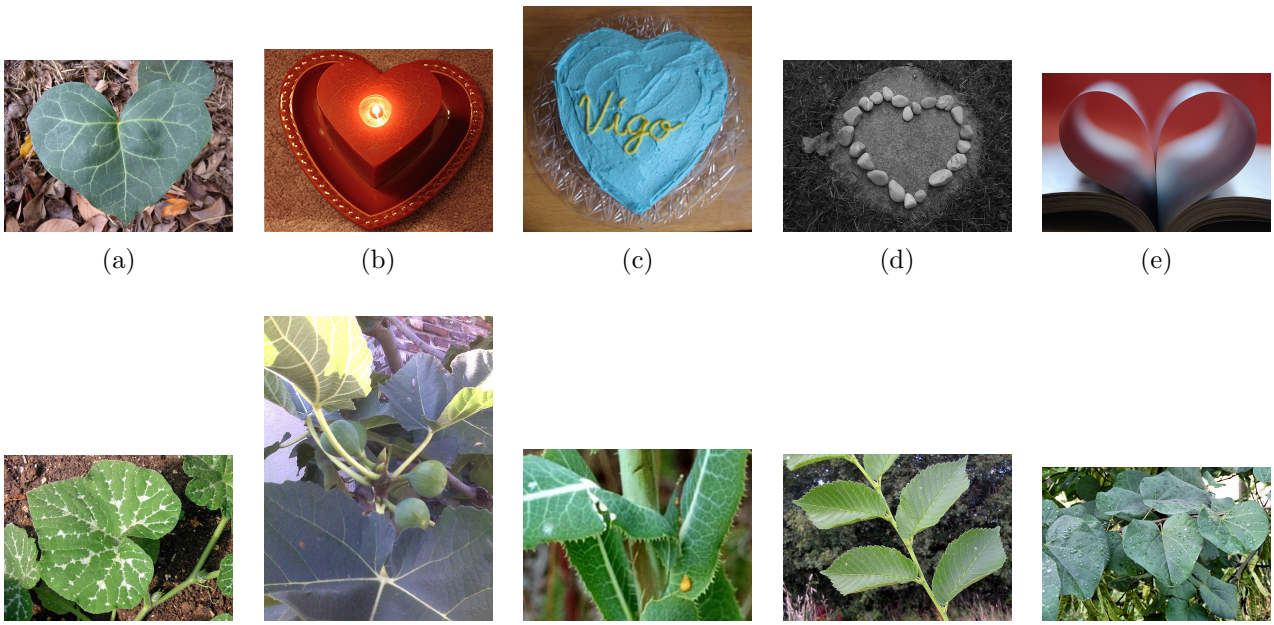

Figure 12: (a)–(e): query examples for topic #79 *heart shaped*, (f)–(j): the best five visual retrieval results for topic #79 together with their scores according to Eq. (7).

This can then be naturally extended to the case of many query images Q by writing

$$s(Q,B) = \max_{i=1}^{n} \max_{j=1}^{N} s(I_i^q, I_j^b) = \max_{j=1}^{N} \max_{i=1}^{n} s(I_i^q, I_j^b) \; . \tag{6}$$

With the same principle, the whole database can be permuted in the order of decreasing similarity to the set of query images. The retrieval result in database B with respect to query Q can then be expressed as the series

$$B_Q = \{I_{o(k)}^b \mid \max_{i=1}^{n} s(I_i^q, I_{o(k)}^b) \geq \max_{i=1}^{n} s(I_i^q, I_{o(k+1)}^b) \, , k = 1, 2 \ldots, N-1\} \; , \tag{7}$$

where the series $\{o(1), o(2), \ldots, o(N)\}$ is thus the permutation of the image indices of the database that orders them by decreasing similarity to the set of query images.

In our experiments, we used the ResNet-152 visual features (He et al., 2016) as the feature extractors $f[\cdot]$ in Eq. (2) and the cosine similarity Eq. (4) for ordering the database images in Eq. (7). Instead of returning the whole image database B of N images, we truncated the list of retrieval results to $N' \ll N$ images. In our case, $N = 237\,434$ and we used $N' = 1\,000$ most similar images. In that way it was feasible to run the visual retrieval experiment once and store the 1 000 retrieval results for each topic for later use.

Figure 12 shows in its top row the five query images for topic #79 *heart shaped* and in the bottom row the five most similar images found in the database together with their cosine similarity scores of Eq. (4) calculated between their ResNet-152 features and ordered according to Eq. (7). All the retrieval results (f)–(i) are actually best matched with the first query image (a). We can see that the first (f) and last (j) retrieved images match very well with the topic description whereas the other three images must be regarded as false positives.

4.4 Fusion of retrieval results

In addition to performing automatic retrieval based on textual searches and visual similarity, we also experiment with combining results from multiple systems. For the combination of strictly textual searches in different languages, there would be easy ways such as by submitting concatenated queries to the search engine. However, since our visual similarity based retrieval method is different, and does not use queries in the same way, it is less straightforward to achieve multimodal combinations. Rather than having to implement a hybrid retrieval system that accepts both textual and visual input from the ground up, we seek to exploit other commonalities between the two systems. One notable fact is that textual and visual forms of automatic retrieval have at least their output in common, formatted as a list of matching items ranked by confidence. Following this observation, we use reciprocal rank fusion (Cormack et al., 2009) to combine ranked lists of results into a single list that is a fusion of its components. This is a very straightforward method that assigns a fusion score to each item equal to the sum of the inverses of its rank positions in each input, and reorders the items by this fusion score to generate the fused ranking. We use reciprocal rank fusion to create fusions of rankings in our baseline metadata setting (discussed later in Section 5.3) so that it would inform us on the potential utility of using multilingual and/or multimodal searches.

5 Automatic retrieval experiments

The intention behind our automatic retrieval experiments is for them to complement our findings around UC2.2 ("Discoverability of archive content"). The UC2.2 archive search experiments conducted during the third round of MeMAD evaluations involved six different search tasks for different types of content, and also used translations of metadata to support cross-lingual content retrieval. The participants assessed their experience related to search tasks and the quality of different metadata types, including MT, using a questionnaire with 7-point Likert scales $[-3, +3]$ and open questions regarding the metadata, as well as semi-structured interviews. Although the evaluations mainly focused on the search task and the platform used, some observations can be made about the participants' comments on MT. Cross-lingual information and MT were potentially relevant in three tasks (out of a total six different search tasks) where the participants were asked to find video clips containing discussions on specific topics in a collection of Finnish and French videos. In the questionnaire, the participants assessed the quality of the MT output to be very good for two of the tasks with a more neutral evaluation given in one task. In the open questions included in the questionnaire, one participant specifically mentioned that the MT was good. When asked about what other metadata would have been useful in the task they completed, the participants made a total of five comments stating that metadata in other languages would have been useful. Feedback was also given in interviews carried out after the experiment. One of the seven participants in this study commented several times that the MT quality was high and that having the MT outputs was useful for providing access to multilingual material. Two other participants also noted that the quality of the machine translated audio transcripts was good. Of the remaining participants, one only made a passing mention of seeing MT output as part of the metadata but did not discuss its quality or use, and three did not comment on MT at all.

The UC2.2 experiments are reported in full detail in deliverable D6.9, and we therefore

leave more elaborate discussions of its results out of the scope of this report. In the following subsections, we present our setup and evaluation strategy for our automatic content retrieval experiments, explain the results we have obtained, and discuss our findings and interpretations pertaining to each experiment. In particular, we discuss the cross-lingual aspect in our findings and their implications for performing searches on a media archive, as well as our observations from some additional cross-modal and multimodal retrieval tests. Section 2 has further information about the technical details of content retrieval, evaluations and results, which may provide some useful comparisons to our work.

5.1 Experimental setup

Our priority in designing our experimental setting for automatic content retrieval has been to keep it as simple as possible, while capturing enough parallels with the retrieval process in the MeMAD prototype platform to ensure that our findings would be applicable. Concerning this question of applicability, it is important to be conscious of the similarities and differences between these systems. Both retrieval systems make use of standard search engines to index metadata attached to individual items within a collection. This provides a computationally efficient way to assess how well a given search query matches the metadata describing an item. On every search, items get sorted by a score assigned to each of them based on such an assessment, forming a ranked list of results. Both retrieval systems follow the same procedure, and return a sampling of the top-ranking items from the produced list.

One major difference is in the type of media that underlie each retrieval system, as the MeMAD prototype platform is a video archive, while we use a dataset of images for our experiments. Although this does create a discrepancy, it makes little difference for the retrieval processes, which are both conditioned on textual metadata regardless of the type of media they describe. For example, image captions and video descriptions contribute to the search in the same way, and generating or translating these are largely analogous methods of metadata enrichment. However, a video is still fundamentally different from an image in that it spans a sequence, and that it is naturally able to encode language (typically spoken language). Therefore, using images for our experiments precludes evaluating certain enrichment processes that might have been relevant if the subject matter were videos, such as automatic subtitling.

Another difference comes from how search queries are formulated in each system, because it is not possible to automatically generate human-like freeform search queries based on ambiguous prompts. More precisely, the accuracy of retrieval upon submitting an arbitrary search query correspondingly requires human arbitration, and cannot be evaluated in a fully automatic fashion. Our solution to this problem has been to avoid formulating our own search queries for the topics annotated on our data, but rather to use the given topic descriptions directly as our queries. In this case, evaluation can be performed automatically against the gold standard of relevant images exactly as annotated, since the original annotators also had the same descriptions to work with. As discussed previously in Section 4, these descriptions already bear some similarities to conventional search queries—for example, they are typically short, neither vague nor highly detailed, and they often feature keywords.

Our motivation for running these tests in parallel with the UC2.2 evaluations has been to complement the findings (see D6.9) with further projections, especially for variables that were impossible to test with human participants due to budget constraints. For instance, while the

evaluations indicate that metadata translations in the archive has been essential in allowing monolingual searches through mixed-language content, there is no data to be drawn for other query languages, and there has been no direct comparison of searches with and without the translations. Furthermore, archive searches in the MeMAD prototype platform are incompatible with cross-modal input (i.e. text searches are conditioned strictly via text), which gets in the way of exploring the implications of multimodal search. For these reasons, our experimental setting for automatic content retrieval has been designed as a good enough approximation of its analogue for UC2.2, in order to reinforce and diversify our findings on the use case through rapid experimentation.

5.2 Evaluation and metrics

The ranked lists of matches predicted by the retrieval processes we have explained so far lend themselves to various types of performance evaluation with the potential to provide insights on different aspects of the retrieval. The precision measure is among the most straightforward and useful metrics, and also the most common, as demonstrated in our recap of related scientific literature in Sections 2 and 3. This is because it is possible to minimise but not completely eliminate mistakes in such probabilistic systems, and we would like to know, first and foremost, the extent to which our results are relevant matches rather than false positives. We calculate precision over the top N results ($p@N$, or "precision at N") as the number of relevant matches within the top N results, divided by N, which yields a score in the range $[0, 1]$. Recall measures are also commonly reported alongside precision as a complementary metric, which indicate the proportion of relevant results the retrieval system has been able to detect and return. We calculate recall over the top N results ($r@N$, or "recall at N") as the number of relevant matches within the top N results, divided by the total number of relevant items K in the collection, which similarly yields a score in the range $[0, 1]$. In a list of retrieval results ranked by confidence, evaluating $p@N$ with increasing values of N cause gradually lower-confidence samples being included in the list, which results in an expected lower precision measure. Conversely, $r@N$ with increasing values of N leaves more room for lower-confidence relevant results to make it into the increasingly longer list of results, which results in $r@N$ behaving like a monotonically non-decreasing function. In this way, precision and recall are in a trade-off relation, where lower values of N can be interpreted to lead to more conservative retrieval that favours precision, and higher values to more inquisitive retrieval that favours recall. In our evaluations of automatic retrieval performance, we provide both precision and recall scores for each run at different values of $N \in \{5, 10, 20, 50\}$ to demonstrate this relation in practice.

To be clear, we remain strictiy in the domain of quantitative evaluation with our automatic retrieval experiments, to complement the more qualitative evaluation for the UC2.2 archive search use case. While we report a limited array of scores, we do recognise that there are other studies describing more rigorous evaluations with a wider selection of metrics to draw insights from. Perhaps the most obvious examples come from large-scale shared tasks on content retrieval (discussed earlier in Section 3), which have conventionally made use of additional metrics (e.g. $p@N$ for up to $N = 1\,000$, precision at certain recall score thresholds, and other well-known measures such as mean average precision). In contrast, the scope for content retrieval in MeMAD is limited, and our priority is to detect and interpret differences between our own contrastive systems, rather than comparing our best performance with the state of the art. However, we would at least like our scores to be comparable to those reported

Index	Query	p@5	p@10	p@20	p@50	r@5	r@10	r@20	r@50
`setting-original`	(de)	0.96	0.95	0.88	0.73	0.19	0.36	0.56	0.82
	(en)	1.00	0.98	0.89	0.69	0.20	0.37	0.57	0.82
	(fr)	0.95	0.94	0.86	0.68	0.18	0.35	0.55	0.81
(only explicit relevance)	(de)	0.97	0.95	0.88	0.74	0.19	0.36	0.56	0.82
	(en)	1.00	0.98	0.89	0.71	0.20	0.37	0.57	0.82
	(fr)	0.95	0.94	0.86	0.69	0.18	0.35	0.55	0.81
`setting-masked`	(de)	0.21	0.18	0.16	0.14	0.04	0.06	0.09	0.11
	(en)	0.32	0.27	0.20	0.14	0.06	0.09	0.13	0.18
	(fr)	0.34	0.28	0.21	0.14	0.07	0.11	0.15	0.18
(only explicit relevance)	(de)	0.23	0.20	0.19	0.18	0.04	0.07	0.09	0.12
	(en)	0.30	0.25	0.19	0.13	0.05	0.09	0.12	0.17
	(fr)	0.33	0.25	0.20	0.15	0.07	0.10	0.14	0.17

Table 3: Precision (p) and recall (r) scores over the top N ranked results (@N) using topic descriptions as textual search queries in `setting-original` and `-masked`, with or without allowing ambiguous images in results.

by the ImageCLEF shared tasks that featured the Wikipedia Image Retrieval Dataset, as our experiments are based on the same data. To achieve this, we selectively report scores marked as *"only explicit relevance"*, meaning that the images for which no topic was marked as relevant or non-relevant have been removed from the search index for that run. This ensures that such images never appear in search results, reducing risks of false positive results and improving both precision and recall, in line with what has been done in ImageCLEF. In other runs without the *"only explicit relevance"* marking, these ambiguous images have been considered implicitly non-relevant for all topics, but allowed to appear in search results. We hold that this way of evaluation makes for a better simulation of searching in a real media archive, where relevance to the search prompt is a less clear-cut and more fuzzy property.

5.3 Results and discussions

To obtain our empirical results, we perform retrieval using a combination of different search queries on a selection of metadata settings, following the experimental setup we have established so far. We base our investigation of cross-lingual content retrieval in particular as a series of searches conditioned on monolingual queries in each of the three languages (English, French and German) represented in our metadata. We do not report averages over the scores resulting from querying in, for example, English and French, mainly to demonstrate how different query languages are affected in different settings, but also because averages over a sample size of three languages would not be very meaningful. Conversely, all scores we report are effective averages over the scores from separately retrieving images relevant for each of our 50 topics, which serves to iron out variations on the ambiguity or difficulty of retrieval across different topics.

We start with a basic side-by-side comparison of automatic retrieval scores calculated from

Index	Query	p@5	p@10	p@20	p@50	r@5	r@10	r@20	r@50
setting-masked *(only explicit relevance)*	(de)	0.23	0.20	0.19	0.18	0.04	0.07	0.09	0.12
	(en)	0.30	0.25	0.19	0.13	0.05	0.09	0.12	0.17
	(fr)	0.33	0.25	0.20	0.15	0.07	0.10	0.14	0.17
	(de+en)	0.31	0.25	0.20	0.14	0.06	0.09	0.13	0.19
	(de+fr)	0.33	0.27	0.20	0.15	0.06	0.11	0.15	0.20
	(en+fr)	0.34	0.30	0.23	0.16	0.07	0.11	0.16	0.22
	(de+en+fr)	0.36	0.30	0.24	0.17	0.07	0.11	0.16	0.23
	(vi)	0.58	0.52	0.44	0.35	0.09	0.15	0.22	0.36
	(de+vi)	0.48	0.43	0.38	0.31	0.08	0.14	0.22	0.36
	(en+vi)	0.51	0.43	0.38	0.29	0.09	0.14	0.22	0.36
	(fr+vi)	0.52	0.45	0.39	0.29	0.09	0.15	0.23	0.36
	(de+en+vi)	0.48	0.40	0.35	0.27	0.09	0.13	0.21	0.34
	(de+fr+vi)	0.50	0.41	0.36	0.27	0.09	0.14	0.22	0.35
	(en+fr+vi)	0.50	0.43	0.36	0.27	0.09	0.14	0.22	0.35
	(de+en+fr+vi)	0.50	0.41	0.34	0.26	0.09	0.14	0.21	0.35

Table 4: Extended retrieval results in `setting-masked` with ambiguous images left unindexed. The language "vi" indicates visual similarity based retrieval, and multiple languages delimited by '+' indicate rank fusion.

queries in `setting-original` and `setting-masked` along with their *"only explicit relevance"* variants, shown in Table 3. The scores from `setting-original` approximate the theoretical upper limits of automatic retrieval performance (using our system) in the presence of maximally useful text metadata. The precision scores for this case show that the top 5 to 10 retrieved results have consistently been relevant matches, regardless of the query language. However, the scores drop quickly for the top 20 to 50 results (even though the data contains 100+ relevant samples for most topics, cf. Table 2). Even in this optimal case, the average recall across all topics becomes capped out at around 0.8 over the top 50 results, at which point precision stays around 0.7. This suggests that, due to miscellaneous limitations in our data and automatic retrieval system, we cannot expect to go beyond these scores. In contrast, the scores from `setting-masked` establish our baseline scores, where we cannot rely on relevant images having metadata that would reasonably match our queries. With these measures established, we can now interpret the corresponding scores from enriched variants of `setting-masked` in terms of how much progress they have stimulated from the baseline to the optimal case.

Our next set of scores comes from our experiments using the visual modality for search, as discussed previously in Section 4.3. To recap briefly, this type of retrieval uses the representative images for each topic (rather than the textual topic descriptions) to drive the search, circumventing the search engine, and instead generating a ranking based on visual similarity. We tabulate results from visual retrieval along with text-based retrieval in Table 4, where the query language "(vi)" denotes that the retrieval has been conditioned on visual input instead of textual queries. The table also displays scores obtained from the reciprocal rank fusion of all combinations of runs using different search queries (e.g. the query language "(de+en)"

Index	Query	p@5	p@10	p@20	p@50	r@5	r@10	r@20	r@50
`setting-masked`	(de)	0.21	0.18	0.16	0.14	0.04	0.06	0.09	0.11
	(en)	0.32	0.27	0.20	0.14	0.06	0.09	0.13	0.18
	(fr)	0.34	0.28	0.21	0.14	0.07	0.11	0.15	0.18
(only explicit relevance)	(de)	0.23	0.20	0.19	0.18	0.04	0.07	0.09	0.12
	(en)	0.30	0.25	0.19	0.13	0.05	0.09	0.12	0.17
	(fr)	0.33	0.25	0.20	0.15	0.07	0.10	0.14	0.17
`setting-masked .translations`	(de)	0.33	0.29	0.24	0.18	0.06	0.10	0.14	0.16
	(en)	0.34	0.32	0.26	0.18	0.07	0.12	0.17	0.25
	(fr)	0.35	0.31	0.25	0.17	0.08	0.13	0.19	0.24
(only explicit relevance)	(de)	0.35	0.30	0.26	0.21	0.06	0.10	0.15	0.17
	(en)	0.30	0.29	0.24	0.18	0.06	0.11	0.16	0.23
	(fr)	0.34	0.30	0.26	0.20	0.08	0.13	0.19	0.24
`setting-masked .autocaps`	(de)	0.20	0.18	0.15	0.13	0.04	0.06	0.09	0.11
	(en)	0.30	0.25	0.19	0.13	0.05	0.09	0.13	0.18
	(fr)	0.32	0.27	0.21	0.13	0.07	0.11	0.16	0.19
(only explicit relevance)	(de)	0.25	0.22	0.20	0.19	0.04	0.08	0.10	0.12
	(en)	0.32	0.27	0.21	0.16	0.06	0.10	0.13	0.19
	(fr)	0.34	0.29	0.23	0.18	0.07	0.11	0.16	0.19
`setting-masked .fully-enriched`	(de)	0.30	0.27	0.22	0.17	0.05	0.10	0.14	0.16
	(en)	0.36	0.31	0.26	0.17	0.07	0.12	0.17	0.23
	(fr)	0.35	0.30	0.25	0.16	0.08	0.13	0.19	0.24
(only explicit relevance)	(de)	0.35	0.31	0.27	0.23	0.06	0.11	0.16	0.18
	(en)	0.34	0.33	0.27	0.20	0.07	0.12	0.18	0.24
	(fr)	0.40	0.34	0.29	0.23	0.08	0.14	0.20	0.25

Table 5: Comparative retrieval results in `setting-masked` and all of its enriched variants, using only textual queries and no rank fusion, with or without ambiguous images for each variant.

indicates the fusion of the ranked results from separately querying in German and in English). Visual retrieval alone is identical between `setting-original` and `setting-masked` (shown in Table 4), while rank fusion with visual retrieval in `setting-original` has a net negative effect on the score (not shown in the table) since text queries already get near-perfect results. The results in `setting-masked` show that visual retrieval is significantly better than using text queries, presumably because textual metadata are not very reliable in this baseline setting. Our interpretation of this fact is that the utility of visual/multimodal search can be fairly high when the availability of textual metadata is limited, which is an encouraging finding for MeMAD. Furthermore, the scores show that fusions of different textual queries also lead to consistently better precision and recall than if a single textual query were used. While this is a useful observation, it may be difficult to exploit in practice, since the translation of isolated search queries is likely to produce noisy output due to the limited textual context.

Furthermore, we provide Table 5 to summarise how our various enrichments affect the automatic performance of retrieval. We make these comparisons exclusively on our baseline `setting-masked`, but provide scores from both the *"only explicit relevance"* and unfiltered versions of the setting. The positive effects of adding translations of existing metadata (`.translations`) are immediately obvious from the scores, showing significant increases in both precision and recall for all query languages. The contribution from metadata translations becomes more noticeable with larger numbers of results returned, coming to a roughly 35% relative improvement in p and r at 50. For some reason, only when querying in German, we observe that p and r at 5 and 10 improve even more, up to an approximately 60% relative increase. Next, the effect of adding automatically-generated captions (`.autocaps`) remains fairly ambiguous compared to metadata translations. On the one hand, the added captions appear to have a small but meaningful positive effect on the retrieval of media with explicit relevance annotations. On the other hand, when searching through ambiguous media as well, they rather seem to confuse the retrieval system, leading to scores that are slightly worse, or about the same as without the added captions. Finally, the results show that the effect of using both enrichment types combined (`.fully-enriched`) also depends on the existence of ambiguous images. When searching through all media, this setting seems to yield slightly worse precision scores overall, while recall scores remain largely unchanged. However, limiting the search to media with explicit relevance annotations appears to result in a slight increase in both scores. This increase becomes more apparent with larger numbers of results returned, clearly mirroring the effect of using automatically-generated captions over the baseline `setting-masked`.

6 Conclusion

In this report, we have presented our various fully-automatic experiments on improving the performance of content retrieval, along with a detailed comparison of retrieval methodologies between these experiments and the MeMAD prototype platform, as well as how our experimental findings might relate to improving content retrieval in a media archive. Based on our simulations of video archives with text metadata represented by a collection of images with filtered textual annotations, our experiments have allowed us to get concrete scores for various content retrieval methods that we have envisioned within the MeMAD project. Our results clearly demonstrate the utility of metadata enrichments through machine translation in facilitating cross-lingual searches, supporting the participant feedback from the qualitative evaluations conducted as part of use case UC2.2 on the discoverability of archive content. Further experiments we have discussed indicate that metadata enrichment via automatically-generated image captions (which might be generalised to cross-modal content descriptions) yields ambiguous results, but may still be useful in certain contexts.

As part of our study on supporting cross-lingual content retrieval, we have only investigated the effects of translating metadata, but not of translating search queries. All of our MT models are trained to translate large units such as sentences and utterances with ample linguistic context (as are the vast majority of translation systems), and using them to translate short search queries inevitably leads to disappointing results. Furthermore, due to the limitations of our dataset, we were unable to investigate the ways in which the outcome might have been different if our languages were less closely-related. Regardless, we have observed a fairly promising

outcome from our tests on the fusion of retrieval results using textual queries in multiple languages (e.g. "en+fr"), which has encouraged us to re-evaluate the possible benefits. While it is probably unrealistic to expect archivists to type the same search query in as many languages as they can in order to improve search accuracy, we propose that this process could be automated by use of prospective MT models that specialise in query translation. According to our results, assuming one could produce reasonable translations for search queries, a background process could translate submitted search queries, retrieve the lists of results from the original query as well as its translations, and finally display a fusion of these results for the user. Although we have not experimented with this particular flow of search, we believe that it may be a worthwhile subject to investigate in the future.

The results from our experimentation with visual similarity based image retrieval are very optimistic, and may indicate that implementing this kind of search for content retrieval in media archives. This type of search requires the archivist to supply related media in order for the system to retrieve other media with similar content, which comes with its own limitations on searching. Such media samples may often be unavailable in practice, or make the retrieval process slower than if the archivist used typed queries instead, potentially overpowering any observed improvements in retrieval accuracy. Currently, the MeMAD prototype platform does not support this type of search, which would require the implementation of some highly involved changes in the platform. While this precludes qualitative evaluations of the pertinent effects within MeMAD, it might still prove useful to offer this technology as an option for users in future endeavours.

References

Thomas Arni, Paul Clough, Mark Sanderson, and Michael Grubinger. 2008. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In *CLEF 2008: Evaluating Systems for Multilingual and Multimodal Information Access*, pages 500–511.

Pavel Braslavski, Suzan Verberne, and Ruslan Talipov. 2016. Show me how to tie a tie: Evaluation of cross-lingual video retrieval. In Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, volume 9822 of *Lecture Notes in Computer Science*, pages 3–15. Springer International Publishing, Cham.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Nicola Ferro. 2019. What Happened in CLEF ... For a While? In *CLEF 2019: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 11696 LNCS, pages 3–45. Springer International Publishing.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, Nevada, USA. IEEE.

Bogdan Ionescu, Henning Müller, Renaud Péteri, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Dzmitri Klimuk, Aleh Tarasau, Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, Duc Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh Triet Tran, Mathias Lux, Cathal Gurrin, Obioma Pelka, Christoph M. Friedrich, Alba Garcìa Seco de Herrera, Narciso Garcia, Ergina Kavallieratou, Carlos Roberto del Blanco, Carlos Cuevas, Nikos Vasillopoulos, Konstantinos Karampidis, Jon Chamberlain, Adrian Clark, and Antonio Campello. 2019. ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature. In *CLEF 2019: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 11696 LNCS, pages 358–386.

Ahmad Khwileh, Debasis Ganguly, and Gareth J.F. Jones. 2015. An investigation of cross-language information retrieval for user-generated internet video. In *Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9283, pages 117–129.

Ahmad Khwileh, Debasis Ganguly, and Gareth J.F. Jones. 2016. Utilisation of metadata fields and query expansion in cross-lingual search of user-generated internet video. *Journal of Artificial Intelligence Research*, 55:249–281.

Dilek Küçük and Adnan Yazıcı. 2011. Multilingual Video Indexing and Retrieval Employing an Information Extraction Tool for Turkish News Texts: A Case Study. In *Flexible Query Answering Systems. FQAS 2011*, volume 7022 of *Lecture Notes in Computer Science*, pages 128–136, Heidelberg. Springer.

Dilek Küçük and Adnan Yazıcı. 2013. A semi-automatic text-based semantic video annotation system for turkish facilitating multilingual retrieval. *Expert Systems with Applications*, 40(9):3398–3411.

Jorma Laaksonen and Zixin Guo. 2020. PicSOM experiments in TRECVID 2020. In *Proceedings of the TRECVID 2020 Workshop*, Gaithersburg, MD, USA.

Martha Larson, Eamonn Newman, and Gareth J. F. Jones. 2009. Overview of VideoCLEF 2009: New perspectives on speech-based multimedia content enrichment. In *Proceedings of the 10th International Conference on Cross-Language Evaluation Forum: Multimedia Experiments*, CLEF'09, page 354–368, Berlin, Heidelberg. Springer-Verlag.

Douglas W. Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, volume 1529 of *AMTA '98*, pages 472–483, Berlin, Heidelberg. Springer-Verlag.

Douglas W Oard and Paul G Hackett. 1997. Document Translation for Cross-Language Text Retrieval at the University of Maryland. In *The Sixth Text REtrieval Conference (TREC-6)*, pages 687–696.

Pavel Pecina, Petra Hoffmannová, Gareth J. Jones, Ying Zhang, and Douglas W. Oard. 2008. *Overview of the CLEF-2007 Cross-Language Speech Retrieval Track*. Springer-Verlag, Berlin, Heidelberg.

Luca Piras, Barbara Caputo, Duc-Tien Dang-Nguyen, Michael Riegler, and Pål Halvorsen. 2019. Image retrieval evaluation in specific domains. In *Information Retrieval Evaluation in a*

Changing World: Lessons Learned from 20 Years of CLEF, 1, pages 275–305. Springer International Publishing, Cham.

Adrian Popescu, Theodora Tsikrika, and Jana Kludas. 2010. Overview of the Wikipedia Retrieval Task at ImageCLEF 2010. In *Cross-Language Evaluation Forum (CLEF) 2010 Working Notes*, Padua, Italy.

Shadi Saleh and Pavel Pecina. 2016. Reranking hypotheses of machine-translated queries for cross-lingual information retrieval. In Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, volume 9822 of *Lecture Notes in Computer Science*, pages 54–66. Springer International Publishing.

Jacques Savoy and Martin Braschler. 2019. Lessons learnt from experiments on the ad hoc multilingual test collections at CLEF. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, pages 177–200. Springer International Publishing, Cham.

Rakshith Shetty, Hamed R.-Tavakoli, and Jorma Laaksonen. 2018. Image and video captioning with augmented neural architectures. *IEEE MultiMedia*, 25(2):34–46.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34(2):97–147.

Theodora Tsikrika, Jana Kludas, and Adrian Popescu. 2012. Building reliable and reusable test collections for image retrieval: The Wikipedia Task at ImageCLEF. *IEEE Annals of the History of Computing*, 19(03):24–33.

Theodora Tsikrika, Adrian Popescu, and Jana Kludas. 2011. Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011. Amsterdam, The Netherlands.

Remco C Veltkamp and Mirela Tanase. 2002. A survey of content-based image retrieval systems. In *Content-based image and video retrieval*, pages 47–101. Springer.