Project no. 033104

MultiMatch

Technology-enhanced Learning and Access to Cultural Heritage
Instrument: Specific Targeted Research Project
FP6-2005-IST-5

D7.3 Evaluation of Second Prototype

Start Date of Project: 01 May 2006
Duration: 30 Months

Organisation Name of Lead Contractor for this Deliverable
ISTI-CNR

Version: Final

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Document Information

Deliverable number:	D7.3
Deliverable title:	Evaluation of Second Prototype
Due date of deliverable:	September 2008
Actual date of deliverable:	November 2008
Authors:	Carol Peters, Eric Bruno, James Carmichael, Paul Clough, Gareth Jones, Martha Larson, Stephane Marchand-Maillet, Eamonn Newman
Participant(s):	ISTI-CNR, USFD, UniGE, Alinari, B&G, UA-BVMC
Workpackage:	7
Workpackage title:	Evaluation
Workpackage leader:	ISTI-CNR
Dissemination Level:	PU (Public)
Version:	Final
Keywords:	multilingual multimedia information retrieval component evaluation

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
V1	29.10.08	Outline	Carol Peters, ISTI-CNR	Circulated to partners involved
V2	3.11.08	Input for sections 3, 4, 5 & 6	Martha Larson, UvA, James Carmichael, USFD, Eric Bruno, UniGE	Waiting for input for Section 2
V3	3.12.08	Final		Final

Abstract

The results of the evaluation of the components of the second MultiMatch system prototype (P2) are reported and details of the experiments that have been conducted are given. The evaluation has focussed on laboratory-based assessment. Performance of individual components and of the overall system is discussed. Fields trials are described in Deliverable 7.4

Table of Contents

Executive Summary

The aim of the MultiMatch evaluation activity is to ensure that the individual components and the complete system prototypes are tested and that their development is assessed during the project life-cycle with respect to performance and usability (i.e. laboratory-based and user-centred evaluation). Component and system development is monitored to ensure that the functional specifications defined in WP 1 are respected. In this deliverable we describe component evaluation for the second system prototype (P2). Time constraints and the limited resources available have meant that evaluation has mainly focused on those areas where ground truth in some appropriate form is already existing.

The deliverable is divided into 5 main sections. In the first part (Section 2) the experiments conducted to test the cross-language text retrieval components are described. Evaluation of query translation is carried out on collections of user queries provided by CH organisations. The output of the hybrid translation system is compared to that of the commercial MT system. The user queries are translated by both systems and where the outputs differ, they are assessed by native speakers. Results of this experiment show that the hybrid system is much more effective at translating these queries for search.

Section 3 reports the evaluation of the Cross Modal Search Engine (CMSE), which provides the image search service for the second prototype. The CMSE builds a multi-modal index of the MultiMatch audiovisual data and delivers search facilities through the *query by text* combined to *query by examples* paradigms. The evaluation consists in testing its functional work flow as well as its retrieval performance using standard IR measure (*i.e.* MAP). Both the Corel image collection and the TRECvid video collection have been used in this assessment.

In order to evaluate the speech indexing component (Section 4), a set of 225 podcasts were gathered and a list of 60 queries were developed. Because of the effort involved with providing human relevance judgments to evaluate retrieval performance, the size and scope of the corpus needed to be restricted to Dutch-language podcasts drawn from the Kunststof podcast. The evaluation addressed the question of whether speech recognition transcripts provided viable indexing features for podcast retrieval.

Section 5 reports on the evaluation of the classification component on the Prototype 2 collection of video supplied by BandG. The choice was made to focus on this collection because of the availability of a substantial volume of ground truth in the form of subject class labels that had been assigned to the videos by archive professionals. Two sets of evaluations were conducted: in one labelled training data from the same video corpus was used; the other attacked the challenge of classifying video in the face of a lack of labelled training data.

Finally Section 6 describes the user-centred experiments performed to evaluate the PT audio and video components. This assessment has focused on the post-query retrieval process, i.e. what happens after the initial list of documents has been fetched from the server and presented to the user via the media-specific interfaces. The evaluation approaches adopted, therefore, essentially test the *intra-document* search functionality.

1 Introduction

The aim of the MultiMatch evaluation activity is to ensure that the individual components and the complete system prototypes are tested and that their development is assessed during the project life-cycle with respect to performance and usability (i.e. laboratory-based and user-centred evaluation). Component and system development is monitored to ensure that the functional specifications defined in WP 1 are respected.

With respect to the second system prototype (P2), a first stage of this activity has been the internal component and system debugging and testing work. This has been described in Deliverables 4.2, 5.2 and 6.2 with respect to the separate indexing, retrieval and interface components, and in Deliverable 3.5 for the integrated system. In particular, the user interface was tested manually by all partners in a series of formal evaluation sessions where evaluators executed a set of pre-specified operations designed for combinatorial testing of P2's various services (see Del 3.5). In this deliverable we describe more formal component evaluation exercises for Prototype 2..

Unfortunately, it has not been easy to perform much component evaluation in time to be reported in this deliverable. In fact, the development and integration of the components for P2 has required more effort and time than initially expected and this has clearly impacted on the amount of strict objective evaluation exercises that it has been possible to conduct within our deadline. The limited resources available have also meant that evaluation has mainly focused on those areas where ground truth in some appropriate form is already existing.

However, several groups have reported that they are currently planning or conducting more extensive evaluation experiments and these will be the subject of scientific papers for publication.. Examples of these are experiments to be conducted using the ImageCLEF test collection to assess the performance of the Image retrieval component, and a planned set of User Experiments to test user satisfaction with the interface.

The contents of this deliverable will be complemented by Deliverable 7.4 which will report on the results of the field trials with three separate user groups and which have been aimed at assessing user interest in the project objectives and user satisfaction with the services and performance offered by Prototype 2.

The deliverable is organised as follows: Section 2 describes the evaluation of the cross-language text retrieval component; Section 3 reports the evaluation of the image retrieval component, Sections 4 and 5 report the evaluation of the speech indexing and the classification components, respectively; finally Section 6 describes the experiments performed to evaluate the audio and video components.

2. Evaluation of Cross-language Text Retrieval Component

In Cross-Language Information Retrieval (CLIR) users try to find documents relevant to their information need in a language which is different to the one in which they expressed their query. Large amounts of work have appeared in the area of recent years which have explored a range of approaches to translation of queries or domains to cross the language barrier between the user's query and the potentially relevant documents. This work has typically adopted an approach of using some form of bilingual dictionary to translate query or machine translation (MT) to translate the query or the documents. While document translation has been shown to be effective for come CLIR tasks, the lower storage and set up computational overhead means that query translation is generally favoured. Dictionary based translation is using easy to set up, just requiring access ro a bilingual dictionary; however using multiple translation entries from a dictionary can introduce large amounts of ambiguity into a query and significantly degrade retrieval accuracy.

It has often been argued in CLIR research that MT is not an appropriate translation solution since queries are too short to provide sufficient context information and that adapting MT to address new tasks is too costly. However, despite their limitations in comparative CLIR evaluation experiments, standard MT often achieves among the best results in terms of retrieval accuracy. As such, MT has a prominent role in CLIR. It is worth considering that some of the aspects of MT which are important in other applications are less so in CLIR. For example, maintaining word order is not necessarily important, as the Information Retrieval (IR) system does not require a strict word order to find a document; however, finding a contextually-appropriate translation for a phrase or idiom is significantly important, since, as with regular MT tasks, a literal translation into the target language will often not be correct. A further important consideration is that CLIR applications are often specialised for a particular domain, and so domain-specific translation is necessary. Also since queries are typically short there is no redundancy in the text, and thus each concept is generally only stated once and this single occurrence needs to be translated correctly.

2.1 Domain-specific Query Translation Evaluation

In this section we evaluate the query translation methods we developed to deal with domain-specific language in the CH domain. We developed a query set which covers the languages of English, Spanish and Italian.

2.1.1 Test Collection Construction

The test collection was constructed from query logs provided to us by the owners of three CH websites. The queries were all provided by real users sending CH related queries to these websites. One of the sets consists of queries in Spanish, the second is in Italian and the third is in English. The set of Spanish queries came from a Digital Library based in Spain whose focus is on poetry and ancient and modern literature in the Spanish language. The set of Italian queries was taken from the “Cultural” section of a large Italian Internet Service Provider's website. The queries in English were extracted from the query logs of the website for a well-known art gallery based in London, U.K. There were 1423 Italian queries (with an average length of 2.49 terms), 1088 Spanish queries (3.39 terms on average) and 100 English queries (1.67 terms on average).

We translated the Spanish and Italian queries to English (and the English to Spanish and Italian) since we had bilingual evaluators available for these language pairs. These sets of translations are denoted *es-en*, *it-en*, *en-es* and *en-it*. There was a single evaluator for each set. The instructions given to each evaluator for the experiment are described in the following section.

Each query was translated by the Hybrid and the MT systems. Where both systems produced the same translation for a given text, the results were discarded since for this evaluation we are interested on the disagreements between the systems. The remaining translations were collated so that the evaluators could have a side-by-side comparison between the original text, the hybrid translation and the MT output. Some examples are given in Table 1.

Table 2.1: Some translation examples

Original	Hybrid Translation	Machine Translation
Plinio il giovane	Pliny the Younger	Plinio the young person
Pittura a tempura	Egg tempera	Painting to moderates
Literatura infantil y juvenil	Children's literature	Infantile and youthful Literature
Al andalus	Islamic Spain	To andalus
Still life paintings	Bodegon pinturas	Pinturas de la vida inmovil

2.1.2 Experiments

In this section we describe the experiment in which we evaluate the effectiveness of different technologies for query translation using our test collection. Standard MT packages have achieved good retrieval effectiveness in CLIR evaluation campaigns such as CLEF. However, we anticipate that many words and phrases pertinent to the domain of CH will not be covered well by the general dictionaries within these systems. We compare our hybrid system with the off-the-shelf MT system.

MT: Machine Translation using the standard WorldLingo MT API. The text to be translated is passed to the WorldLingo system and the response is presented as the translation. There is no pre- or post-processing of the text. In this method, there is no provision for translation of domain-specific terms.

Hybrid: Worldlingo MT augmented by replacing the phrase translation from the domain-specific dictionaries. Here, the phrase which matches a dictionary entry is demarcated in the text passed to the MT system for easy identification. Thus the mis-translated phrase can be replaced with the appropriate translation taken from the dictionary. This should result in improved translation quality.

The MT hybrid translation strategies (and alternatives we investigated) are described in more detail in [Jones 2008].

2.1.3 Evaluation Methodology

Our experiments proceeded as follows:

- Establish the MT baseline by translating the original texts using the WorldLingo system.
- Translate the original texts using out hybrid system.
- Identify instances where the results of the systems differed.
- Bilingual evaluators determine which of the two translations is better.

For each query, evaluators were asked to mark which of the two translation results they "considered to be better". As there was only evaluator per set, we were not able to consider inter-annotator agreement on this subjective measure. Any possible bias due to a single evaluator will result in a skew of the results for one set, rather than the whole evaluation.

2.1.4 Experimental Results

Table 2 summarises the results of the experiments. There were 2711 queries to be translated in total, with both systems producing the same result for 1919 of them. As described above, these were not presented to our evaluators in this experiment, but will be evaluated for quality in future work.

Table 2.2: Translation Results

Language Pair	Number of Translations	Number of Disagreements	Hybrid Correct	Both Correct	MT System Correct	Unannotated
It – En	1423	482	288	63	75	56
Es – En	1088	281	222	0	58	1
En – It	100	15	9	1	2	3
En – Es	100	14	11	0	3	0

Of the remaining translations, we can see that for all language-pair sets, the hybrid translation system was generally regarded as providing a better translation. For Spanish-English, the hybrid translation was correct in 79% of the cases where there was a disagreement between the systems. For Italian-English, when we remove the unannotated instances and instances where both systems were deemed correct (leaving 482-(56+63) = 363 instances), we achieve a very similar score of 79.34% correctly translated by the hybrid system. For Italian-English instances in 63 cases both translations were deemed equally good. This raises the interesting point for CLIR (and indeed IR in general) since we would like to choose the translation which matches the one most likely to appear in a relevant document.

Because the set of English queries was an order of magnitude smaller, we cannot attach any significance to the results, however for the sake of completeness, we report correct translation rates of 81.82% for English to Italian and 78.5% for English to Spanish, which are similar to the results from the larger sets.

The similarity of these results, across different language pairs, different evaluators and different set sizes suggest that there was no bias inherent in any of the evaluations.

2.1.5 Conclusions

The results presented here show that our methods for enhancing an MT system by incorporating domain-specific dictionaries are successful. By identifying phrases and named entities which have a special meaning within the domain, we were able to improve upon the baseline translation in around 80% of cases.

Having native speakers as evaluators allows further analysis of the actual quality of the translations, rather than just comparing them to the baseline. The evaluators were also asked to highlight any translations which they thought were ``particularly good" or ``particularly bad". For example, the evaluator for translations between Spanish and English thought a translation of "poema del mio cid" was particularly good as it inserted the full name of the work (``Cantar de Mio Cid") into the translation (giving "poem of Cantar de Mio Cid") making it much better than the literal translation provided by the MT system ("poem of mine cid").

2.2 Other Related Experiments

The objective of our hybrid translation system is ultimately to improve CLIR accuracy. Since we did not have access to a suitable set of documents and corresponding relevance data for our user search topics, we conducted a preliminary set of CLIR experiments using a different IR test collection. We used the CLEF 2007 Cross Language Speech Retrieval (CL-SR) task. This consists of a small collection of about 8000 documents and 42 search topics with corresponding relevance data indicating which documents are relevant to each query. This provides an interesting test for search technologies within the project since it is a (non-CH) domain specific cross language multimedia retrieval task. However, the topic statements are generally rather longer than those typically entered into a web search engine. For the CLEF task we trained new bilingual dictionaries for the relevant in the domain of the CL-SR data set (issues relating to World War Two). These were then used in combination with a standard MT system to perform a set of comparative experiments exploring alternative translation strategies. The full results of these experiments were reported in [Zhang, 2007]. In summary, these showed that combining our domain-specific dictionaries with MT methods improves the cross language retrieval effectiveness in terms of Mean Average Precision (MAP) and Precision at rank 10 (P@10) for the CL-SR task. While our best submitted monolingual run was slightly less than (although not significantly) than the best submission, our submitted result for the cross language task was the best showing the lowest decrease relative to monolingual performance. These results are encouraging for our ongoing work, since they demonstrate that this approach can work well for ad hoc retrieval and when working with errorful transcribed output from speech recognition systems.

2.3 Additional Analysis

Users have differing levels of linguistic skills ranging from fluency to no knowledge at all. With this in mind, we undertook some analysis of the system in which users can interact with the hybrid translation service. In this, users with passive recognition language skills will be able to select a particular translation (or translations) of a phrase where multiple options exist within the dictionary, if they are not happy with the one provided by the hybrid service. For users with stronger language skills, it may be necessary for the user to provide their own translation, if none of the presented options are appropriate.

In this regard, the results given above were re-examined as if the alternative hybrid translations were also available to the evaluators. In many cases, one of the alternative hybrid translations matched the MT system translation exactly, or matched when stopwords were removed. Table 3 gives an updated set of results, incorporating the new findings.

Table 2.3: Translation Results

Language Pair	Number of Translations	Number of Disagreements	Hybrid Interactive Correct	Both Correct	MT System Correct	Unannotated
It – En	1423	482	353 (+65)	71 (+8)	2 (-73)	56
Es – En	1088	281	273 (+51)	0	7 (-51)	0
En – It	100	15	10 (+1)	2 (+1)	0 (-2)	3
En – Es	100	14	12 (+1)	2 (+2)	0 (-3)	0

The new results show that the allowing the user to interact with the translation service brings about a significant increase in correct translations via the hybrid system. As well as improving translation quality the potential for interaction and feedback also improves the user’s sense of control while using MultiMatch which results in increased user satisfaction.

3. Evaluation of Image Retrieval Component

The Cross Modal Search Engine (CMSE), as described in deliverable 5.3.2, provides the image search service for the second prototype. The CMSE builds a multi-modal index of the MultiMatch audiovisual data and delivers search facilities through the *query by text* combined to *query by examples* paradigms. The evaluation consists in testing its functional work flow as well as its retrieval performance using standard IR measure (*i.e.* MAP).

3.1 Functional Evaluation

The retrieval component has been fully integrated to the MultiMatch second prototype. Data were ingested correctly and retrieval tests have been successfully passed.

1. Data ingestion and indexing service
   The data load for the second prototype is around 43'00 images along with the corresponding meta-data files. Data preparation (e.g. feature extraction and indexing) takes around 12hrs after what the image search service becomes fully functional.
2. Data retrieval service
   Retrieval using both query by example and query by text is fully working. Various queries involving various image collections (ONB, Alinari, AISA) has been extensively tested without problems. .
3. Scalability issue
   The index growth rate is 4Ko/image, meaning that more than 250 million of images might be indexed on a standard Terabytes disk. The retrieval complexity is M*N+(M*N)log(N) (with N the size of the database and M the number of modalities indexed).

3.2 Retrieval Performance Evaluation

Experiments are conducted on two multimedia datasets widely used for benchmarking multimedia retrieval systems: The Corel data set, a famous collection of tagged images, and the NIST TrecVid video corpus consisting in hundred of hours of broadcast news.

Retrieval performance is given in terms of Mean Average Precision (MAP). Average Precision (AP) is the sum of the precision at each relevant hit in the retrieved list, divided by the minimum between the number of relevant documents in the collection and the length of the list. The MAP is simply the AP averaged over several classes. Additionally to the algorithm performance, a baseline consisting in retrieving randomly documents is always provided.

The CMSE is compared with a hierarchical SVM considered as an effective and standard technique for multimedia retrieval [Wu et al, 2004]. At the first level, base classifiers are trained in each mono-modal space. At the second level, a super classifier is used to fuse soft-outputs of all base classifiers. Base classifiers and super classifier are RBF SVM. Optimal classifier parameters have been determined through a leave-one-out cross validation

3.2.1 Corel image collection

The studied image collection is a subset of the Corel collection. It contains 1159 images annotated with 1 to 10 keywords per image (including some non-sense descriptions). The images are categorized into 49 classes

For each class, a set of 40 queries composed of 10 positive and 10 negatives examples is randomly defined. Mean Average Precision and execution time per query computed over all classes for multimodal retrieval and text-only search are provided below:

	multimodal	text-only	baseline	execution time
CMSE	**0.3**	**0.26**	0.04	**0.01s**
Hierarchical SVM	0.28	0.24	0.04	0.46s

The CMSE performance is slightly better than the SVM result, and the retrieval is speeded up by a factor 46. We also notice the multimodal retrieval improves the text-only search for both retrieval approaches.

3.2.2 TrecVid video corpus

We now consider the TRECVID 2005 benchmark. In our setup, videos are segmented into around 89’500 segments using the common shot reference [Petersohn, 2004]. These shots are considered as individual and independent documents. This means that no contextual information is taken into account and that shot description is restricted to its audiovisual content (e.g. visual, audio and speech information). The Search Task, as defined in TRECVID-05, consists in retrieving shots that are relevant to some predefined queries (called topics). There are 24 topics concerning people (person-X queries), objects (specific or generic), locations, sports and combinations of the former. For each topic, keywords, pictures and several video shots (4-10) are provided as positive examples. Further details about the Search Task may be found in [Smeaton et al, 2006]. During the experiments, we only considered video shots as positive examples. The positive examples are completed with ten negative examples randomly selected within the test set. Starting with this initial query, a *relevance feedback* loop is initiated by adding to the query up to 10 new positive and negative examples returned in the 1000-entries hit-list. The process is repeated ten times. Following the TRECVID evaluation protocol, the performance was measured at each iteration by MAP at 1000. Additionally to the algorithm performance, a baseline consisting of retrieving randomly documents is always provided.

RF iteration n°	0	5	10
CMSE	**0.01**	**0.23**	**0.33**
SVM	0.009	0.20	0.30
Random	2.10-4	0.07	0.15

The CMSE outperforms the SVM-based retrieval on this large scale experiment. Again the retrieval speed-up factor is 46 when compared to the SVM.

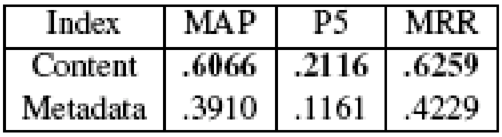
▪

4. Evaluation of Speech Indexing Component

In order to evaluate the speech indexing component, a set of 225 podcasts were gathered and a list of 60 queries were developed. Because of the effort involved with providing human relevance judgments to evaluate retrieval performance, the size and scope of the corpus needed to be restricted. We chose to focus on Dutch-language podcasts drawn from the Kunststof podcast.[1] The queries were designed to represent five categories: 1) person name 2) title, 3) quotation, 4) general topic, and 5) current issue or event. These categories emerged as representing the predominant user information needs as determined by an extensive user study designed to determine search goals in podcast retrieval [Besser 2008].

The evaluation addressed the question of whether speech recognition transcripts provided viable indexing features for podcast retrieval. A comparison was carried out between retrieval using indexing features derived from metadata (i.e., title and description field of the podcast feed) and retrieval using indexing features derived from spoken content (i.e., speech transcripts.) Results were reported in terms of Mean Average Precision (MAP), Precision at 5 (P5) and Mean Reciprocal Rank (MRR). As can be seen from Table 4.1, the context-derived features outperformed the metadata-derived features according to all three measures.

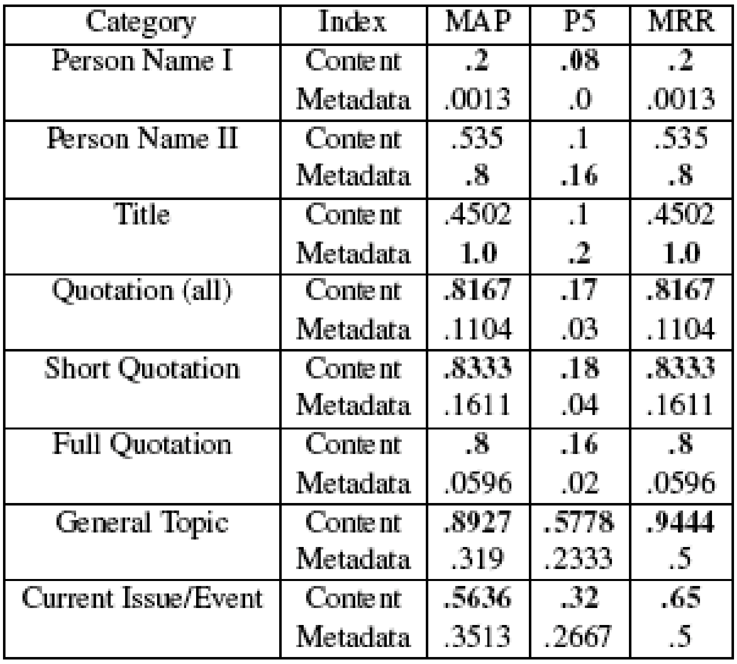
Table 4.1: Retrieval results for podcast retrieval based on content and metadata indexing features



Index	MAP	P5	MRR
Content	**.6066**	**.2116**	**.6259**
Metadata	.3910	.1161	.4229

Table 4.2: Retrieval results for podcast retrieval based on content
and metadata indexing features broken down over categories reflecting query type

[1] http://www.omroep.nl/nps/kunststof/



Category	Index	MAP	P5	MRR
Person Name I	Content	**.2**	**.08**	**.2**
	Metadata	.0013	.0	.0013
Person Name II	Content	.535	.1	.535
	Metadata	**.8**	**.16**	**.8**
Title	Content	.4502	.1	.4502
	Metadata	**1.0**	**.2**	**1.0**
Quotation (all)	Content	**.8167**	**.17**	**.8167**
	Metadata	.1104	.03	.1104
Short Quotation	Content	**.8333**	**.18**	**.8333**
	Metadata	.1611	.04	.1611
Full Quotation	Content	**.8**	**.16**	**.8**
	Metadata	.0596	.02	.0596
General Topic	Content	**.8927**	**.5778**	**.9444**
	Metadata	.319	.2333	.5
Current Issue/Event	Content	**.5636**	**.32**	**.65**
	Metadata	.3513	.2667	.5

Table 4.2 reports the performance results broken down into the different types of query categories.

Here, it is possible to see that for certain types of queries, metadata-derived indexing features are more appropriate and for other types of queries, content (i.e., speech recognition)-derived indexing features are more appropriate. The conclusion of the evaluation is that both metadata-derived and speech-derived features should be used to index podcasts and that, in order to cover the range of queries related to the different types of information needs of users, both metadata-based and speech-based features should be used for audio indexing. Further details on the evaluation are available in [Besser, 2008]

5. Evaluation of Classification Component

The classification component was evaluated on the Prototype 2 collection of video supplied by BandG. The choice was made to focus on this collection because of the availability of a substantial volume of ground truth in the form of subject class labels that had been assigned to the videos by archive professionals. We carried out two sets of evaluations, one in which we used labelled training data from the same video corpus and one in which we attacked the challenge of classifying video in the face of a lack of labelled training data.

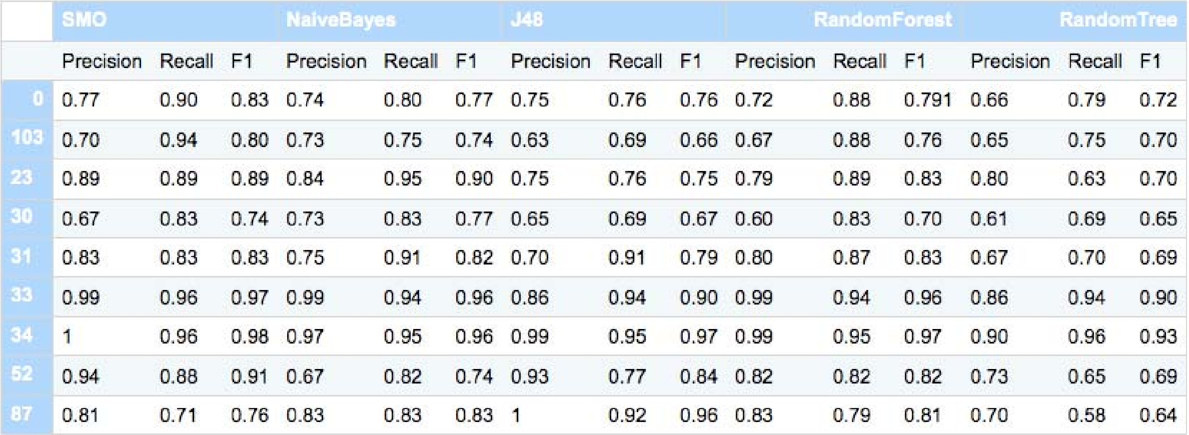
In the first evaluation (Evaluation I), we chose to address the research question of whether features derived from speech recognition transcripts would be helpful in classifying video into subject label classes. Evaluation was performed on a set of 440 videos, which were labelled with the 9 classes drawn from the Cultural Heritage domain, shown in Table 5.1.

Table 5.1: Classes used for Classification in Evaluation I

index	Subject Class Label	English Translation
0	schilderkunst	painting
103	musea	museums
23	geschiedenis	history
30	kunstschilders	artists
31	schrijvers	writer
33	poëzie	poetry
34	dichters	poets
52	literatuur	literature
87	muziek	music

Evaluation was carried out using five different classifiers, Support Vector Machine (SMO), a Naïve Bayes classifier and three tree classifiers (J48, RandomForest and Random Tree). Results of the experiment runs (10-fold cross validation) are given in Table 5.2. The features used for the classification were words derived from the speech recognition transcripts of the videos.

Table 5.2: Results of classification for Evaluation I



	SMO			NaiveBayes			J48			RandomForest			RandomTree		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
0	0.77	0.90	0.83	0.74	0.80	0.77	0.75	0.76	0.76	0.72	0.88	0.791	0.66	0.79	0.72
103	0.70	0.94	0.80	0.73	0.75	0.74	0.63	0.69	0.66	0.67	0.88	0.76	0.65	0.75	0.70
23	0.89	0.89	0.89	0.84	0.95	0.90	0.75	0.76	0.75	0.79	0.89	0.83	0.80	0.63	0.70
30	0.67	0.83	0.74	0.73	0.83	0.77	0.65	0.69	0.67	0.60	0.83	0.70	0.61	0.69	0.65
31	0.83	0.83	0.83	0.75	0.91	0.82	0.70	0.91	0.79	0.80	0.87	0.83	0.67	0.70	0.69
33	0.99	0.96	0.97	0.99	0.94	0.96	0.86	0.94	0.90	0.99	0.94	0.96	0.86	0.94	0.90
34	1	0.96	0.98	0.97	0.95	0.96	0.99	0.95	0.97	0.99	0.95	0.97	0.90	0.96	0.93
52	0.94	0.88	0.91	0.67	0.82	0.74	0.93	0.77	0.84	0.82	0.82	0.82	0.73	0.65	0.69
87	0.81	0.71	0.76	0.83	0.83	0.83	1	0.92	0.96	0.83	0.79	0.81	0.70	0.58	0.64

The results show that a relatively high level of classification efficacy can be attained by making use of features derived from speech recognition transcripts. In general, recall levels are higher than precision levels. Performance is relatively stable across classifiers.

The second evaluation (Evaluation II) was carried out in the framework of the VideoCLEF 2008 pilot benchmark test. This evaluation addressed the question of whether classification is possible when no labelled video data is available for training classifiers. The video corpus used for Evaluation II consisting of 40 videos of television documentaries from BandG. Archeology (archeologie), Architecture (architectuur), Chemistry (chemie), Dance (dansen), Film (film), History (geschiedenis), Music (muziek), Paintings (schilderijen), Scientific research (wetenschappelijk onderzoek) and Visual arts (beeldende kunst). We investigated 3 experimental conditions designed to allow us compare the use of archivist-generated metadata (title and description) with content features (from speech recognition transcripts). For this set of experiments, a Support Vector Machine (LS-SMV) was chosen as a classifier. The classifier was trained using data collected from Wikipedia by submitting the class label as a query and using the returned documents as the training set. The classification results presented in .3le 5 are lower than those achieved in Evaluation I. Apparently, classification without training data from the video domain is quite challenging. Notice that the results of Evaluation I and Evaluation II cannot be directly compared since the data and class label sets are different and since Evaluation I uses 10-fold cross validation.

Table 5.3: Results of classification for Evaluation II

	Test data represented using metadata features only	Test data represented using speech recognition transcript features only	Test data represented using metadata and speech features
Macro-average precision	0.11	0.44	0.44
Macro-average recall	0.38	0.38	0.46
Macro-average F1	0.17	0.41	0.41

Evaluation II allowed us to draw two conclusions. First, although it is challenging to train classifiers without data from the video domain, it is not impossible. Particularly encouraging was the performance of top-performing individual classes. The “music” classifier, for instance, achieves a precision of 0.57 and a recall of 0.36. Second, we were able to establish that speech recognition transcripts are helpful for classification. The results led us to draw the conclusion that in cases where metadata and speech transcripts are both available, both should be exploited in order to classify the documents. Further information about Evaluation II can be found in [He 2008]

6. Evaluation of Audio and Video Components

The design and implementation of the audio and video interfaces for Prototype 2 was a particularly challenging problem for a number of reasons. For example, in the case of the audio component, it was important that any web page-embedded audio playback functionality should be able to accommodate a variety of audio file formats (both streaming and non-streaming). For this reason, audio playback has been designed to be generic, i.e. capable of working with the most popular media player plug-ins (e.g. Windows Media Player, RealPlayer, and QuickTime). The decision was taken, therefore, to build a standard JavaScript-based "wrapper" application which served as an interface for any type of specialist audio plug-in installed on the computer of the end-user. This wrapper software was equipped with its own customised playback controls (e.g. click-and-play tag clouds) which directly access the underlying audio application. A technical drawback, however, is that since it is not possible to pre-determine the format of any retrieved audio file, the wrapper software for the audio interface must first correctly detect the end-user's computer operating system and default audio web browser plugin (since not all JavaScript audio playback commands are valid for all types of audio software). Despite such challenges, however, this approach has shown to be fully compatible with approximately 75% of the client-side web browser/audio plug-in combinations tested.

Unlike its audio counterpart, the video playback technology does not have to cope with a variety of multimedia file formats, since all of the video material in the MultiMatch collection originates from a single source: the archives of the Dutch Institute for Sound and Vision (B&G). Furthermore, all of the video material used in Prototype 2 has been converted to the RealVideo (.rv) format, thus eliminating the problems of heterogeneous data. Nonetheless, video retrieval in the MultiMatch context poses certain technical challenges. For example, correct speech-to-text processing of the soundtracks for those videos in the collection which are multilingual (i.e. featuring substantial amounts of speech which are in some language other than Dutch – the officially designated language of the video corpus). This has resulted in substantially elevated word error rates, a partial solution would appear to be using phonetic similarity to alert the reader to the possibility that the query word/phrase may have been mis-transcribed (e.g. the French "Bonjour" erroneously rendered as the English "Bond your").

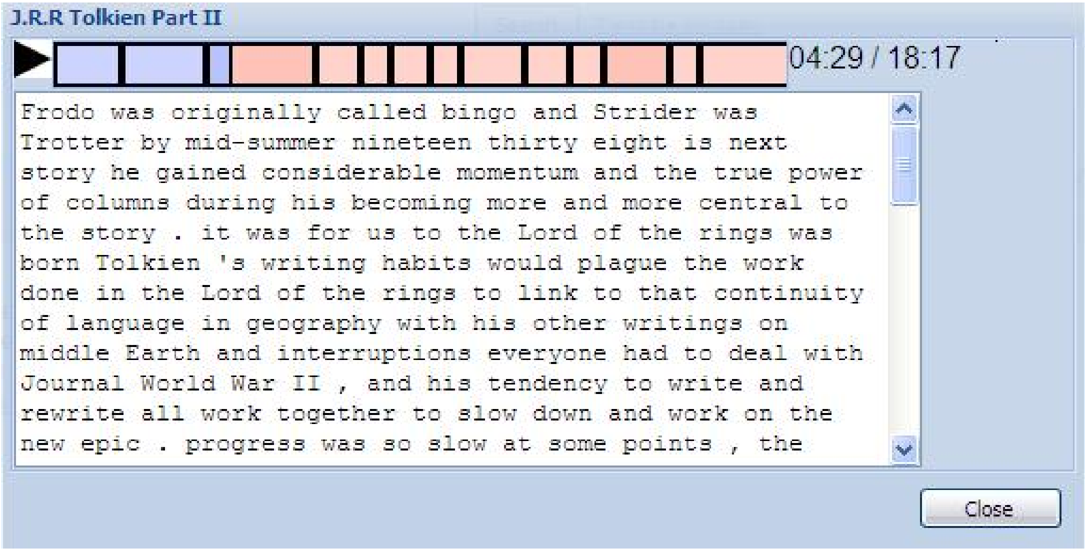
In addition to attempting to overcome problems with multilingual speech processing, the video interface has also undergone a substantial re-design from the first prototype. The video user interface has now been divided into two distinct components: (i) the *inter-document* video search interface (in which a user can search the speech transcripts of all the video documents returned for a given query) and (ii) the *intra-document* search interface – a version of which had already been implemented for PT1. The PT2 intra-document video retrieval sub-system, now features a more compact but information-rich interface which is even better suited to the professional user.

6.1 Specific Evaluation of the Audio/Video Interfaces

As discussed in Deliverable 7.2, the search algorithms employed to retrieve audio and video documents from the MultiMatch collection are essentially no different from those used for retrieving any other document type such as text (html) or image files, i.e. key terms are extracted from the user's text-based search criteria and these terms are then used to query metadata records which describe the MultiMatch document repository. The assessment of PT2 audio and video functionality has, therefore, focused on the post-query retrieval process, i.e. what happens after the initial list of documents has been fetched from the server and presented to the user via the media-specific interfaces. The evaluation approaches adopted, therefore, essentially test the *intra-document* search functionality, the details of which are presented in the following sections.

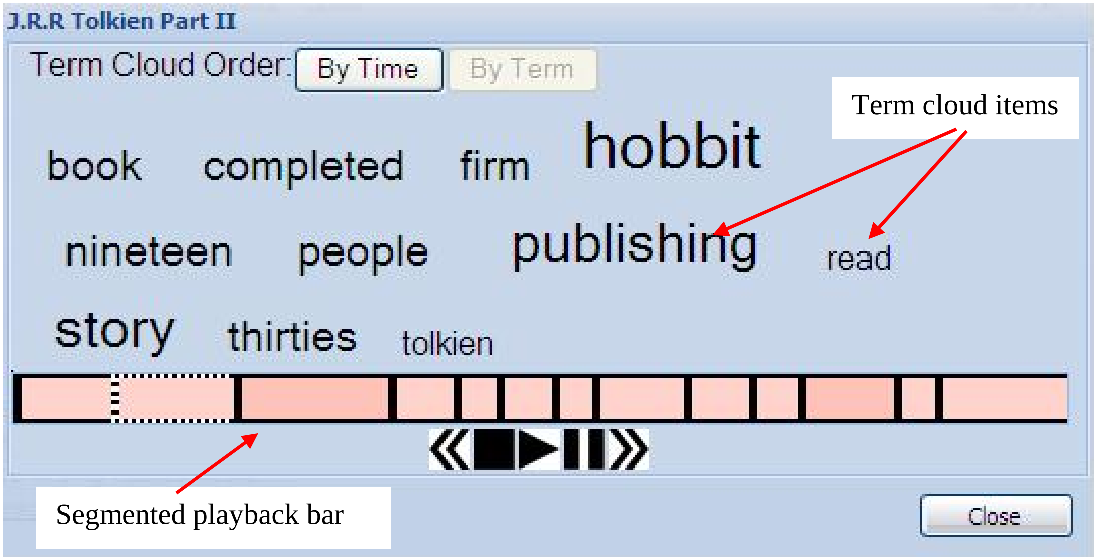
6.1.1 PT2 Audio Evaluation

For PT2, two audio interfaces have been developed using two substantially different techniques for presenting transcripts (see Figures 6.1 and 6.2) of any speech content detected in the audio stream, these presentation techniques are: (i) key-word term clouds and (ii) displaying the “raw” (i.e. word-for-word) automatically generated speech transcripts.



Frodo was originally called bingo and Strider was Trotter by mid-summer nineteen thirty eight is next story he gained considerable momentum and the true power of columns during his becoming more and more central to the story . it was for us to the Lord of the rings was born Tolkien 's writing habits would plague the work done in the Lord of the rings to link to that continuity of language in geography with his other writings on middle Earth and interruptions everyone had to deal with Journal World War II , and his tendency to write and rewrite all work together to slow down and work on the new epic . progress was so slow at some points , the

Figure 6.1: Transcript Browse Audio UI (transcript for 3rd segment presented)



Term cloud items

Segmented playback bar

Figure 6.2: Term Cloud Audio UI (term cloud for 2nd segment presented)

The PT2 audio evaluation was specifically formulated to determine which of the two presentation techniques is more effective. Accordingly, users participating in the evaluation study were presented with four search scenarios, two for each audio UI. A "Latin square" arrangement was used to remove order bias of the topics-systems (shown in Table 6.1).

Table 6.1: Latin square experiment design for 12 users and 4 tasks (search scenarios)

	Search Scenario 1	Search Scenario 2	Search Scenario 3	Search Scenario 4
User 1	ASR	ASR	Term Cloud	Term Cloud
User 2	Term Cloud	Term Cloud	ASR	ASR
User 3	ASR	ASR	Term Cloud	Term Cloud
User 4	Term Cloud	Term Cloud	ASR	ASR
User 5	ASR	ASR	Term Cloud	Term Cloud
User 6	Term Cloud	Term Cloud	ASR	ASR
User 7	ASR	ASR	Term Cloud	Term Cloud
User 8	Term Cloud	Term Cloud	ASR	ASR
User 9	ASR	ASR	Term Cloud	Term Cloud
User 10	Term Cloud	Term Cloud	ASR	ASR
User 11	ASR	ASR	Term Cloud	Term Cloud
User 12	Term Cloud	Term Cloud	ASR	ASR

Twelve users (7 male and 5 female) were recruited for the study, four from Dublin City University (DCU) and eight from the University of Sheffield. The distribution of participants per age group is detailed in Table 6.2.

Table 6.2: Ages of participants recruited for audio evaluation

Age Group	Number of Participants
20-24	1
25-29	3
30-34	2
35-39	3
40-44	2
45-49	1

All of the participants spoke fluent English and worked with (or used) audio data at least once a week. Nonetheless, the users did not generally search for Cultural Heritage data and seldom searched for audio data on the Internet. Only one of these participants had some acquaintance with the PT2 audio interfaces being tested. After completing a training exercise to familiarise themselves with the various audio UI playback controls and information presentation techniques, the participants completed four search tasks (see Appendix B). As they manipulated the audio UIs, observers logged their various interactions, such as the number of times they clicked on the segmented audio playback bar (see Figures 6.1 and 6.2) or closed a top-level window.

Upon completing the task, users filled in a post-experiment questionnaire eliciting feedback on the audio UIs under evaluation, particularly in relation to the effectiveness of the data visualisation components (the term clouds). These opinions were expressed as ratings on a 7-point Likert Scale, ranging from “Strongly Agree - 1” to “Strongly Disagree - 7” (see Appendix C).

PT2 Audio Evaluation Results

Responses to the post-experiment questionnaires revealed a slight majority preference for the “Transcript Browse” feature, due to the fact that the key words featured in the term-cloud audio UI (see Figure 6.2) were often deemed not relevant in relation to the search task at hand. Conversely, the full-transcript audio UI allowed users to thoroughly peruse any available speech transcriptions to determine what items were relevant. Four of the twelve participants suggested incorporating both tag/term cloud and full-transcript presentation techniques into a single interface (an implementation option which should be explored for future investigations).

As mentioned previously, users’ interaction with both audio UIs was logged for the purposes of measuring interface usability in terms of time-motion measurements. Figure 6.3 compares the total number of minutes spent across users when completing each task on either interface (note: TB in this graph denotes “Transcript Browse” and B denotes “Browse”). On average, users spend more time searching when they use term clouds as opposed to perusing the “raw” speech transcript. Paradoxically, a time-motion analysis of the users’ interactions with the two audio UIs (Figure 6.4) reveals that, more often than not, more actions were required in order to get to relevant information on the “Transcript Browse” interface than on the “Browse” interface. This is because to examine all the contents of the text box containing the verbatim speech transcripts, it was usually necessary to perform repeated scrolling actions (a manoeuvre which was neither necessary nor supported on the “Browse” interface).

It should be highlighted, however, that term cloud summarisations did not always prove more efficient with respect to interaction than simply presenting the speech transcripts in a text box. For example, an increase in the number of user actions was noted for Task 3 when executed using the term-cloud interface, an observation which is not surprising since none of the query terms were present in the term clouds and the users were therefore obliged to perform more click-and-play actions to locate relevant information.

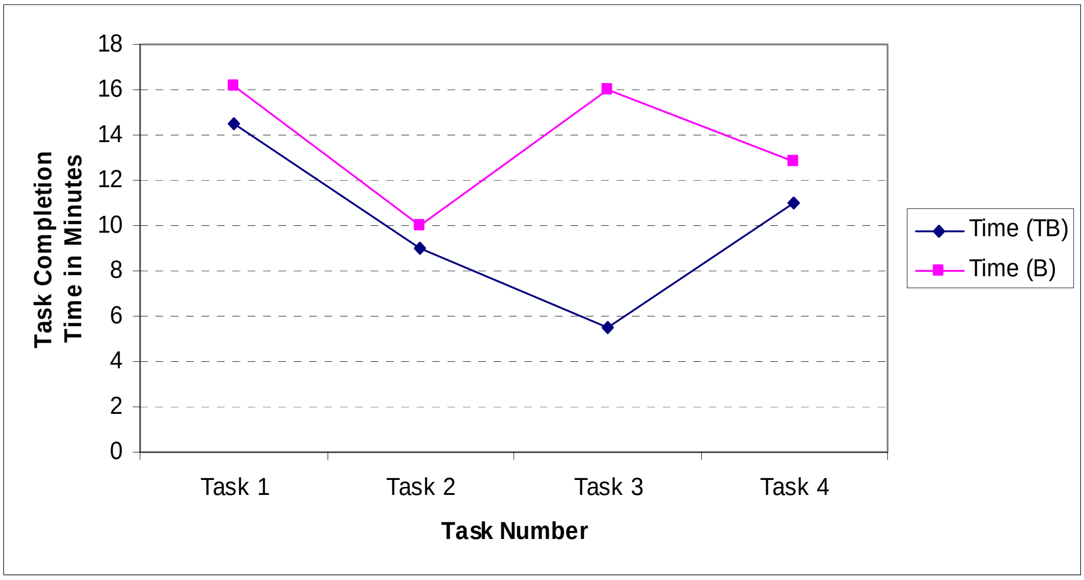


Figure 6.3: Task completion times per search scenario (TB = Transcript Browse; B = Browse)

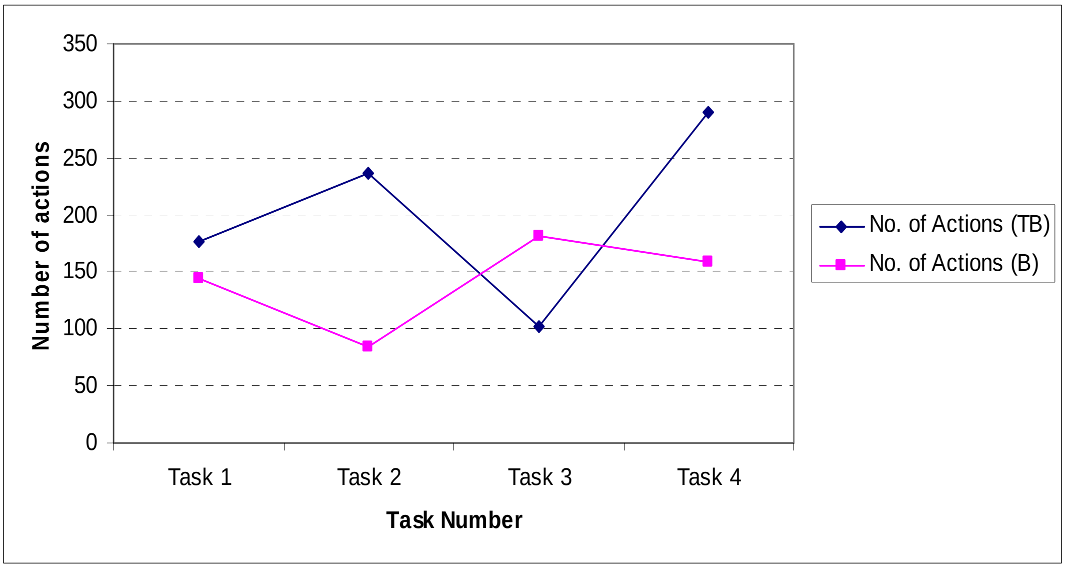


Figure 6.4: Number of user actions required for task completion per search scenario

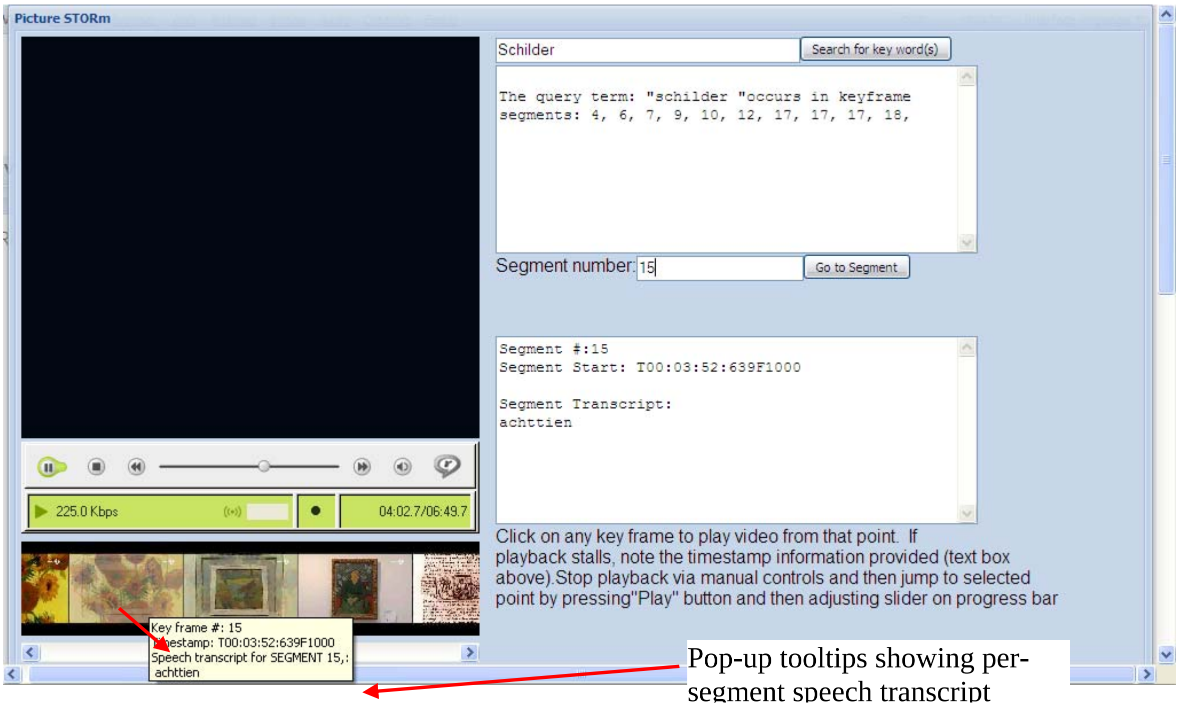
Based on participants' ratings (see Table 6.3), both audio UI interfaces proved easy to master as information retrieval tools, with the only major drawback being some instances of audio plug-in incompatibility[2]. All users found the term cloud presentation technique to be generally helpful, but a substantial number of the participants felt that these term clouds should be *query-sensitive*, i.e. the items in a term cloud should be selected and vary in size and font-colour in direct relation to user-defined query terms. Furthermore, two of the users observed that the update method for the term clouds was sometimes confusing since it was possible for the audio UI to be playing a different segment to that which the term clouds were representing. For the most part, users were neutral in their preferences concerning the possibility of ordering term clouds chronologically or alphabetically; furthermore, it was observed that only three of the eight participants from Sheffield actually used the click-and-play functionality from the term cloud, preferring instead to manipulate playback via the segmented progress bar. Table 2 gives an average rating for each statement across all users (1 = Strongly Agree, 2 = Quite Agree, 3 = Little Agree, 4 = Neutral, 5 = Little Disagree, 6 = Quite Disagree, 7 = Strongly Disagree).

[2] Wherever possible, the audio UI testing was conducted on the users' computer systems; unfortunately, there were four occasions when the audio UIs playback functionality failed to work due – in all probability – to some incompatibility with the client's web browser audio plug-in.

Table 6.3: Average of ratings supplied by users for post-experiment questionnaire statements

	Average User Rating (TB)	Average User Rating (B)
The system is easy to use	1.5	1.8
Learning how to use the system was easy	1.6	1.7
The system response time was fast enough	2.2	2.2
The system interface allowed me to do the task efficiently	2.8	2.6
I liked the look and feel of the interface	2.7	2.8
I liked the colours used.	3	3
I liked the presentation of term clouds		1.5
I liked the ordering of the term clouds		3 (Alphabetical) 2.5 (Chronological)
I found the term clouds helpful in finding relevant information		2.6
I found the term clouds helpful in navigating through podcasts		2.5
There were few erroneous terms in the term clouds/transcript	4.7 (1 answer missing)	2.75
There were few irrelevant terms in the term clouds/transcript	3.2 (1 answer missing)	2.5
Using term clouds to find information is more effective than using an audio player alone		2

In summary, it would appear – judging from the users' comments – that the ideal speech data presentation technique would be a combination of verbatim transcript display together with query-sensitive term clouds. The single most important audio UI shortcoming is that of its incompatibility with certain audio plug-ins, a non-trivial technical issue which constitutes a good starting point for future work in this area.



Picture STORm

Schilder

Search for key word(s)

```
The query term: "schilder "occurs in keyframe
segments: 4, 6, 7, 9, 10, 12, 17, 17, 17, 18,
```

Segment number: 15

Go to Segment

```
Segment #:15
Segment Start: T00:03:52:639F1000

Segment Transcript:
achttien
```

Click on any key frame to play video from that point. If playback stalls, note the timestamp information provided (text box above).Stop playback via manual controls and then jump to selected point by pressing"Play" button and then adjusting slider on progress bar

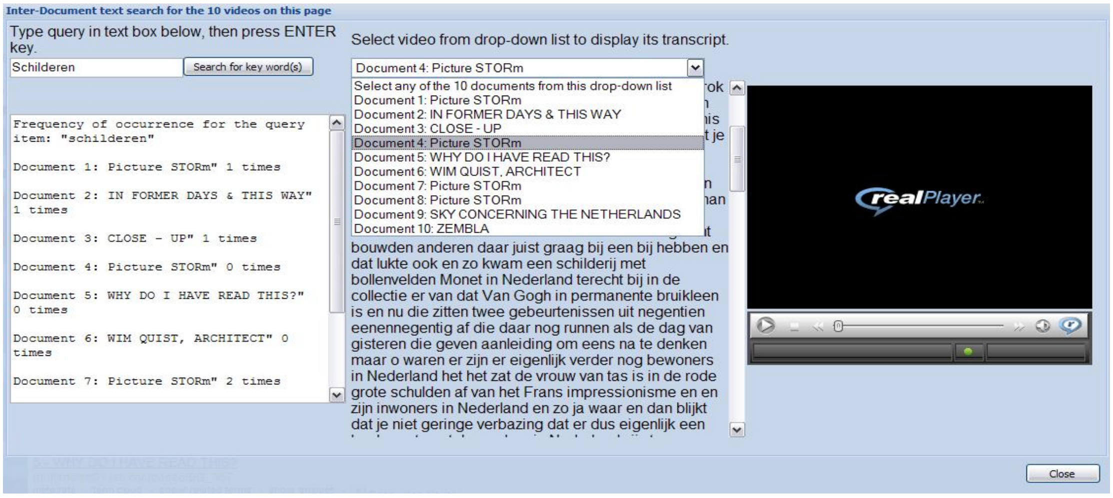
Key frame #: 15
Timestamp: T00:03:52:639F1000
Speech transcript for SEGMENT 15,:
achttien

Pop-up tooltips showing per-segment speech transcript

Figure 6.5: Intra-document video interface

6.1.2 PT2 Video Evaluation

An *intra*-document evaluation procedure – previously documented in Deliverable 7.2 and in a study by Carmichael et al [2008] – had been established for the first prototype and this was retained, with a few modifications, for evaluation of the PT2 video interface. Users were provided with pre-defined search tasks and asked to locate clip-length segments (i.e. video clips up to 3 minutes' duration) within a given video or audio file. For audio-only material, the search criteria were entirely verbal, i.e. users will be required to find sub-document segments which feature some key word(s) or phrase(s). For video documents, users were also required to identify video footage depicting some scenario that matches any pre-specified search parameters/descriptions (e.g. "*2006 Paris riots*"). An example video evaluation questionnaire is included in Appendix A.



Inter-Document text search for the 10 videos on this page

Type query in text box below, then press ENTER key.

Schilderen

Search for key word(s)

```
Frequency of occurrence for the query
item: "schilderen"

Document 1: Picture STORm" 1 times

Document 2: IN FORMER DAYS & THIS WAY"
1 times

Document 3: CLOSE - UP" 1 times

Document 4: Picture STORm" 0 times

Document 5: WHY DO I HAVE READ THIS?"
0 times

Document 6: WIM QUIST, ARCHITECT" 0
times

Document 7: Picture STORm" 2 times
```

Select video from drop-down list to display its transcript.

Document 4: Picture STORm

Select any of the 10 documents from this drop-down list
Document 1: Picture STORm
Document 2: IN FORMER DAYS & THIS WAY
Document 3: CLOSE - UP
Document 4: Picture STORm
Document 5: WHY DO I HAVE READ THIS?
Document 6: WIM QUIST, ARCHITECT
Document 7: Picture STORm
Document 8: Picture STORm
Document 9: SKY CONCERNING THE NETHERLANDS
Document 10: ZEMBLA

bouwden anderen daar juist graag bij een bij hebben en dat lukte ook en zo kwam een schilderij met bollenvelden Monet in Nederland terecht bij in de collectie er van dat Van Gogh in permanente bruikleen is en nu die zitten twee gebeurtenissen uit negentien eenennegentig af die daar nog runnen als de dag van gisteren die geven aanleiding om eens na te denken maar o waren er zijn er eigenlijk verder nog bewoners in Nederland het het zat de vrouw van tas is in de rode grote schulden af van het Frans impressionisme en en zijn inwoners in Nederland en zo ja waar en dan blijkt dat je niet geringe verbazing dat er dus eigenlijk een

Figure 6.6: Inter-document video interface

To evaluate the success of the interfaces, users were asked to decide for themselves which video clips were relevant to the search task:

"*…all of the participants independently viewed the video document in its entirety.... During this final viewing, each assessor was requested once again to select all video segments which were pertinent to the search criteria. The number and quality of segments in this final selection is then regarded as the "ground truth" relevance measure, i.e. the assessor's ideal choice of snippets against which to compare the initial selection made by the same individual when restricted to searching via [the MultiMatch video retrieval sub-system's] summarisation techniques. Any change in an evaluator's initial and final selection of clips is expressed as a shot difference (SD) measure, whereby the actual shots which comprise the clips selected in the first and second rounds are compared to determine the extent of overlap. The following example illustrates the computation of the SD measure: after the second viewing of a selected video, an evaluator chooses two clips, each consisting of ten specific shots. In the first viewing (with the assistance of [the MultiMatch video retrieval sub-system]), the same evaluator had originally selected one clip made up of fifteen shots. If the clip selected during the first round has only five shots which differ from those clips selected in the second round, then the shot difference percentage would be 25.0% (representing the five out of twenty-five shots which differ between the user's selection of clips for the first and second rounds).*" [Carmichael et al, 2008, p. 98].

Apart from a general user evaluation, a second user study was conducted in May 2008, at the B&G headquarters in Hilversum. In this particular investigation, time-motion studies – similar to those described in the section for the audio UI evaluation – were conducted to determine the system's usability and efficiency by professional users of video retrieval systems.

PT2 Video Evaluation Results

In terms of correctness of automatic classification, the key frame extraction techniques employed and ASR accuracy would appear to be adequate. For the selected video from which 156 key frames were extracted, only three per cent (5 frames) were duplicates; this percentage compares well with state of the art in this domain [3].The key terms spoken in the video appeared in the transcript two out of three times. Neither duplicate shots, nor dropped key terms, had a significant negative effect on the usefulness of how shot level documents were represented in the retrieval process. Seven out of the ten evaluators did not alter selection of segments even after viewing the video file in its entirety. For the three evaluators who did make alterations, their differences in shot selection were 3.8%, 12.6% and 19.0% respectively.

Users' responses to the questionnaire indicated that the provision of speech transcripts played an integral role in locating and identifying relevant clip-length segments. The ten scores of the individual evaluators to the second question of the questionnaire averaged 3.6; the mode score being "5". The comments of one of the evaluators regarding the usefulness of the shot-aligned speech transcripts typify the sentiments of his colleagues: "*…for instance when you see a talking head, the transcript will disclose what this person is talking about*". Representation by the key frame of the entire contents of the video was also judged to be satisfactory and receiving an average rating of 3.9 in response to question 5.1 (b) – see appendices. The only area where there was a notable level of dissatisfaction concerned the speed of the video playback, as evidenced by the evaluators' mean average rating of 2.3 for question 5.1(c). On some occasions, commencement of video playback – when initiated by clicking on a key frame – was excessively delayed due to poor connectivity with the remote streaming server. These bandwidth problems were, in all likelihood, also responsible for occasional deterioration in playback quality (dropping of frames, pixilation etc) which reduced user satisfaction with the video UI's application.

Overall, the qualitative and quantitative assessments of the video UI system conducted in this investigation indicate that the task of selecting relevant clips from some video document is indeed facilitated by the presenting the user with both shot-aligned speech transcripts and a series of key frames which comprehensively summarise the visual contents of a video. There is, however, much scope for further research in the area of improving the ASR accuracy, especially in the case of videos featuring multilingual speech content.

For the ten B&G professional users who participated in a specific exercise to compare the effectiveness of the PT2 video UI with that of an existing video retrieval system called the *Catalogue*, the average number of mouse clicks clearly indicate that – when searches are successful – the PT2 video UI proves the more effective IR tool. When evaluating overall usability, however, 40% of the B&G professional user group considered both the Catalogue and PT2 video UI interfaces to be too cluttered, suggesting that most of the text boxes could be removed and any information therein presented in the form of tooltip-style pop-up boxes appearing only if the mouse pointer is placed over some other graphical component, such as a key frame image. These user requests to condense available information into a smaller UI footprint are, paradoxically, accompanied by demands for the provision of a greater range of multi-faceted data. The B&G researchers report that it is not unusual for them to receive searchrequests involving personalities or events with which they are unfamiliar, thus it is necessary to conduct a preliminary search – often using the more popular search engines such as Google or Wikipedia – to procure still images and/or audio clips in order to know what the investigated person or object looks and sounds like. Such multimodal preliminary searching is supported by the MultiMatch video UI, which features a specialist image-search interface. Half of the users considered, however, that this image-search interface could be integrated into the principal search page in order to minimise the number of task-related mouse clicks [Carmichael et al, 2008b].

In summary, the PT2 video UI would appear to have fulfilled it's stated functional objectives, i.e. the realisation of a more efficient intra-document multimodal search methodology via the integration of key frame visual summarisation and automatic speech recognition techniques. The principal shortcomings are the occasional slowness of system response for key frame click-and-play functionality along with the speech transcription inaccuracies which can result in the user forming the impression that the information being searched for is not present in the video document when this is not necessarily the case.

Table 4: User **t**ime-motion measurements for search tasks
(average times for each 5-member subgroup as a whole)

Search Engine used	**Intra-video search Task Description**	**Avg. No. of Mouse clicks**	**Avg. Time Taken (sec.)**	**Failed Searches**
PT2 Video UI	Locate spoken instance of term “decoratie” [**correct ASR**]	7.4	17.6	0
Catalogue	Locate spoken instance of term “Maria Callas” [ASR unavailable]	29.7	58.3	1
PT2 Video UI	Locate spoken instance of term “Maria Callas” [**incorrect ASR**]	41.3	77.2	4
Catalogue	Locate spoken instance of term “decoratie” [ASR unavailable]	33.5	43.5	0

References

Buckland, M.., Gey, F., "The relationship between Recall and Precision", Journal of the American Society for Information Science, Volume 45 Issue 1, pp. 12 – 19, January, 1994.

Carmichael, J., Larson, M., Marlow, J., Newman, E., Clough, P., Oomen, J., Sav, S., "Multimodal Indexing of Digital Audio-Visual Documents: a case study for Cultural Heritage Data", *Proceedings of the Sixth International Workshop on Content-Based Multimedia Indexing*, pp. 93 – 100, June, 2008.

Carmichael, J., Clough, P., Newman, E., Jones, G., "Multimedia Retrieval in MultiMatch: The Impact of Speech Transcript Errors on Search Behaviour", Proceedings of the 2008 Information Access to Cultural Heritage (IACH '08) workshop, Aarhus, Denmark, September 2008.

Cox, S., Rose, R., "Confidence Measures for the SWITCHBOARD database" *Proceedings of the International Conference on Acoustics, Acoustics, Speech, and Signal Processing* (ICAASSP), **1**, 509-511, 1996.

Jones, G.J.F., Fantino, F., Newman, E., Zhang, Y., "Domain-Specific Query Translation for Multilingual Information Access Using Machine Translation Augmented With Dictionaries Mined From Wikipedia", *Proceedings of the 2nd International Workshop on Cross Lingual Information Access (CLIA 2008),* Hyderabad, India, pp 34-41, January, 2008.

Zhang, Y., Jones, G.J.F., Zhang, K., "Dublin City University at CLEF 2007: Cross-Language Speech Retrieval (CL-SR) Experiments", *Proceedings of the CLEF 2007: Workshop on Cross-Language Information Retrieval and Evaluation*, Budapest, Hungary, 2007

APPENDIX A: Video Evaluation Protocol Document

TASK 1: TO BE DONE WITH B&G CATALOGUE

Find a video

You are looking for some material from a documentary that focuses on the career, life and death of a man named Pier Paolo Pasolini.

This is a Dutch-language video that was first shown on 31st August 1981.

- What is the title of this video?

Find a segment

Your client would like a 5-second clip of Pier Paolo Pasolini talking (in any context.) Please locate an appropriate segment and write down the approximate timestamps of this segment.

- What are the timestamps?

Find a part of the video (visual)

Use the display of video keyframes to find the part near the beginning of the video showing a newspaper headline that reads "Pasolini assassinato." (Approximate timestamp 03:31)

Play the video from this point, and see if you can read the name of the newspaper (written in the upper left hand corner.)

- What is the newspaper called?

Find a part of the video (spoken)

Now you would like to find a part in the video that mentions a person called Maria Callas. This is between timestamps 28:00 and 30:00.

- At which point in time is the name "Maria Callas" mentioned?

Now you would like to find a part in the video in which the word “Renaissance” is spoken. Again, this is between timestamps 28:00 and 30:00.

- At which point in time is the word “Renaissance” spoken?

Final Questionnaire (Pasolini)

How easy were the following tasks?

Very easy Not at all easy

Finding the video 1 2 3 4 5 6 7

Finding a clip in the video 1 2 3 4 5 6 7

Identifying the writing in the keyframe (newspaper title) 1 2 3 4 5 6 7

Finding spoken words (Maria Callas) 1 2 3 4 5 6 7

Finding spoken words (Renaissance) 1 2 3 4 5 6 7

TASK 2: TO BE DONE WITH MULTIMATCH SYSTEM

Find a video

You are looking for some material from a news program that discusses a recent work by the architect Rem Koolhaas.

This was broadcast on 15th November of 2003.

- What is the title of the program?

Find a segment

Your client would like a 5-second clip of Koolhaas talking (in any context.) Please locate an appropriate segment and write down the approximate timestamps of this segment.

- What are the timestamps?

Find a part of the video (visual)

Click on the link below the video you want to watch that says "play video." You may have to wait some time for it to load, so please be patient and do not click this link more than once if the yellow "searching" message is displayed.

Once the video has loaded, use the sliding bar of images from the video (keyframes) to find four keyframes showing a map of Turkey. Click on the final keyframe of this series (around timestamp 01:50.)
If you click on this keyframe, the video will start playing at this point. If you are prompted for a username and password, enter
Username: multimatch2
Password: f64ty1

Once the video has started to play from that point, you will see two boxes with black arrows pointing to the sites of 2 explosions.

- What are the names of the two sites?

Find a part of the video (spoken)

Now you would like to find a part in the video in which Rem Koolhaas is being interviewed and he says the word “crisis.” This is between timestamps 20:00 and 23:00.

To do this, you can search through the automatically generated transcript of the video's audio content. Enter your query term (crisis) into the box above the video and press enter.

Then if you scroll down to the third box below the video, you will see the segment(s) in which this term is mentioned. You can then manually navigate to that segment in the filmstrip and click on it to start playing from that point.

- In which segment of the video is the word “crisis” spoken?

- At what point in time does the segment start?

Now, please play the part of the video mentioning “crisis”
by clicking on the appropriate keyframe (in case you did not find it
earlier, it is at Timestamp 0:22:11, the image showing a
closeup of Koolhaas talking

Click this keyframe so the video starts playing at this point.
Listen to the words being spoken by the narrator and compare them
The words that are written in the transcript box below the keyframes
("van het Duitse rendement een opeenvolging werven crisis is een het
gebouw ligt de lopende plek van de stad te die eigenlijk alle lagen
van de sinds de heeft meegemaakt de ontvangstruimte van de ambassadeur”)

- How closely do the spoken words correspond to the written words?

□ Not at all (0% overlap)

□ Not well (~25% correct)

□ Somewhat (~50% correct)

□ Mostly (~75% correct)

□ Totally (~100% correct)

Now you would like to find a part in the video in which the word “Potsdamer Platz” is spoken. Again, it is between timestamps 20:00 and 23:00.

- In which segment is “Potsdamer Platz” mentioned?
- At what timestamp does the segment start?

Final Questionnaire (Koolhaas)

How easy were the following tasks?

	Very easy	Not at all easy
Finding the video	1 2 3 4 5 6 7	
Finding a clip in the video	1 2 3 4 5 6 7	
Identifying the writing in the keyframe (explosion sites)	1 2 3 4 5 6 7	
Finding spoken words (crisis)	1 2 3 4 5 6 7	
Finding spoken words (Potsdamer Platz)	1 2 3 4 5 6 7	

How useful were the following features?

	Very useful	Not at all useful
Click and play keyframe functionality	1 2 3 4 5 6 7	
Being able to <u>view</u> the automatically generated transcript of the spoken words	1 2 3 4 5 6 7	
Being able to <u>search</u> the automatically generated transcript of the spoken words	1 2 3 4 5 6 7	

How helpful would the following features be?

Very helpful Not at all helpful

Having tag clouds (key terms) 1 2 3 4 5 6 7
instead of pure audio transcripts to
represent the spoken content

Being able to manually 1 2 3 4 5 6 7
fix incorrect parts of the audio transcript
for future reference

Being able to identify and signal 1 2 3 4 5 6 7
the language being spoken
(for multilingual videos)

What other features would help you locate video material in your daily work?

APPENDIX B: Audio Evaluation Protocol Document I

Multimatch Audio Browser Tests -

Please read the following instructions carefully. If you encounter any problems, please do not hesitate to ask.

<u>Participant 2:</u>

Complete the following 4 tasks. When answering, please write down the title of the relevant podcast/result, and – in the case of "Transcript Browse" – the timing information, e.g. "*found answer/information in Podcast entitled "St. George's Day*" at 4 min. 13 seconds". There is no time-limit on this experiment.

Please go to http://homer.multimatch.hostedbyfdi.net:8090/audio/ to begin your tasks.

<u>Using Browse Only:</u>

Task 1:

Find out information about the movie adaptation of the life of Beethoven

Task 2:

Find out information about women joining the Order of the Garter.

<u>Using Transcript Browse:</u>

Task 3:

Find out what the original names for Frodo and Strider were in J.R.R. Tolkien's book "Lord of the Rings".

Task 4:

Find information about any Russian composers who were suppressed during the Soviet regime (information you find must include the names of composers discussed).

APPENDIX C: Audio Evaluation Protocol Document II

POST-TEST QUESTIONNAIRE

Multimatch Audio Search Experiments 2008, Dept. of Information Studies, The University of Sheffield

▶ **What is your overall impression of the system you have used?**

▶ **Which feature of the interface did you find most useful?** Transcript Browse ___ Browse ___

Why?

▶ Please rate (i.e., check an appropriate box) agreement or disagreement with the following statements:

		Strongly agree	Quite agree	A little agree	Neutral	A little disagree	Quite disagree	Strongly disagree
The system is easy to use	Transcript Browse							
	Browse							
Learning how to use the system was easy	Transcript Browse							
	Browse							
The system response time was fast enough	Transcript Browse							
	Browse							
The system interface allowed me to do the task efficiently	Transcript Browse							
	Browse							
I liked the look and feel of the interface	Transcript Browse							
	Browse							
I liked the colours used.	Entire Interface							
I liked the presentation of term clouds	Browse							
I liked the ordering of the term clouds	Alphabetically Ordered							
	Time Ordered							
I found the term clouds helpful in finding relevant information	Browse							
I found the term clouds helpful in navigating through podcasts	Browse							
There were few erroneous terms in the term clouds/transcript	Transcript Browse							
	Browse							
There were few irrelevant terms in the term clouds/transcript	Transcript Browse							
	Browse							
Using term clouds to find information is more effective than using an audio player alone	Transcript Browse							

▶ Please add comments on what you liked BEST either version you used:

▶ Please add comments on what you liked WORST about either version you used:

▶ Any other comments you want to tell us?

Thank you very much!