Preliminary report on participation in international benchmarks

Deliverable D4.3

Project Number:	FP6 - 045389
Deliverable id:	D 4.3
Deliverable name:	Preliminary report on participation in international benchmarks
Date:	January 14, 2008

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	VITALAS
Project Full Name:	Video & image Indexing and Retrieval in the Large Scale
Document id:	D 4.3
Document name:	Preliminary report on participation in international benchmarks
Document type (PU, INT, RE)	PU
Version:	1.0
Date:	January 14, 2008
Authors:	Theodora Tsikrika Arjen P. de Vries
Organisation:	CWI
Email Address:	Theodora.Tsikrika@cwi.nl Arjen.de.Vries@cwi.nl

Document type PU = public, INT = internal, RE = restricted

ABSTRACT:
This deliverable is a premilinary report on our participation in an international evaluation benchmark. It presents our retrieval approaches based on a uniform generative probabilistic framework, and examines their retrieval effectiveness, in the context of INEX Multimedia. This benchmark for the evaluation of image retrieval is particularly suitable, since its test collection is very similar to the VITALAS environment. Our evaluation experiments indicate the value of linguistic evidence in the context of image retrieval, and raise research issues currently being investigated by the WP4 activities.

KEYWORD LIST:
cross-media retrieval, image retrieval evaluation, INEX Multimedia, generative probabilistic retrieval models, language models, evaluation experiments

MODIFICATION CONTROL			
Version	Date	Status	Author
0.1	01 November 2007	Draft	Theodora Tsikrika, CWI
0.2	12 November 2007	Draft	Theodora Tsikrika, CWI
0.3	21 November 2007	Draft	Theodora Tsikrika, CWI
0.4	10 December 2007	Draft	Theodora Tsikrika, CWI
0.5	03 January 2007	Draft	Theodora Tsikrika, CWI
0.6	07 January 2007	Draft	Theodora Tsikrika, CWI
0.7	07 January 2007	Draft	Arjen P. de Vries, CWI
0.8	09 January 2007	Draft	Theodora Tsikrika, CWI Christos Diou, CERTH-ITI Tasos Delopoulos, CERTH-ITI
0.9	11 January 2007	Draft	Theodora Tsikrika, CWI Daniel Schneider, FhG Nicolas Herve, INRIA
1.0	14 January 2007	Draft	Theodora Tsikrika, CWI Joost Geurts, INRIA
2.0		Final	

List of Contributors

- Tasos Delopoulos, CERTH-ITI
- Christos Diou, CERTH-ITI
- Joost Geurts, INRIA
- Nicolas Herve, INRIA
- Daniel Schneider, FhG
- Theodora Tsikrika, CWI
- Arjen P. de Vries, CWI

Contents

1 Introduction

The VITALAS project aims at providing advanced solutions for indexing, searching and accessing large scale digital audiovisual content.

The VITALAS Deliverable D4.3 is concerned with the **cross-media retrieval** approaches being developed in VITALAS WP4. It examines their *retrieval effectiveness*, in the context of an international benchmark for the evaluation of multimedia retrieval systems: INEX Multimedia[1]. We argue that this benchmark is particularly suited for evaluating cross-media retrieval in VITALAS. This evaluation is characterised as *preliminary*, given that it takes place in the first year of VITALAS, prior to the integration of the various components of the VITALAS system; in particular, prior to the integration of WP2, WP3, and WP4 which will actually realise the cross-media nature of the project. As a result, the retrieval component is currently a standalone system, rather than a part of an integrated VITALAS system, and it mainly exploits the textual features of the multimedia documents.

In this first phase of the project, the implemented retrieval techniques focus on (i) the combination of evidence derived primarily from textual features, and (ii) an initial investigation of the contribution of evidence derived from audiovisual features. The objective of the evaluation of this core cross-media retrieval system and its associated technologies is to act as a baseline for the evaluation of the retrieval module of the integrated VITALAS system, foreseen for the second phase of the project. Furthermore, the results reported here will be compared with the findings of the user studies to be performed in the context of the activities of VITALAS WP1 and of TRECVID Interactive Search Task, in order to validate whether improvements in the laboratory setting hold in practice, and vice versa.

The remainder of this VITALAS Deliverable D4.3 is structured as follows: Section 2 discusses cross-media retrieval in terms of the requirements of the VITALAS use cases, and of the retrieval approaches being developed in order to deal with them. Section 3 investigates the cross-media retrieval of images, by reporting on the participation in the INEX 2007 Multimedia track using the approaches developed in the context of the VITALAS retrieval component. Section 4 concludes this deliverable by highlighting its main contributions.

[1] http://inex.is.informatik.uni-duisburg.de/2007/mmtrack.html

2 Cross-Media Retrieval in VITALAS

The cross-media retrieval requirements of VITALAS are dictated by its use cases, which, in essence, determine the applicative orientations of the project. Therefore, we first introduce (some aspects of) these use cases, before presenting the cross-media retrieval approaches being developed in order to deal with them. Section 2.1 discusses the use cases by focussing only on their aspects that are pertinent to the research activities of VITALAS WP4 (i.e., it adopts an Information Retrieval (IR) viewpoint). Next, Section 2.2 introduces the retrieval framework underlying the retrieval techniques being developed. These retrieval techniques will be applied on data provided by the VITALAS content partners and will be evaluated in the context of the user studies that will take place in the second phase of the project. Since this deliverable presents an initial evaluation of their effectiveness on data provided by international benchmarks, Section 2.3 outlines the main components and characteristics of such evaluation environments.

2.1 Cross-media use cases

In VITALAS, the cross-media use cases specified in D1.1 aim at describing the typical information needs of users of archives of photographic and audiovisual content. These use cases have been expressed in the form of scenarios that detail the intentions of such users, the tasks they need to accomplish in their current work contexts, and the way they would approach and interact with a VITALAS system.

In the context of cross-media retrieval, the VITALAS use cases can be described in terms of the following aspects:

1. what users are searching for (i.e., images, audio, or video),
2. how they express their information needs as queries (i.e., by using text, image or audiovisual examples, concepts, or by specifying particular attributes), and
3. how they interact with the retrieval results (i.e., whether they provide relevance feedback, or request visualisation of results based on mono-media and cross-media similarities).

Table 1 briefly describes the cross-media VITALAS use cases; a more elaborated description is provided in D1.1 and in D4.1. Only the use cases that involve the cross-media retrieval component are presented here. Their description considers only the first two aspects of the above list, whereas the third aspect (interaction with retrieval results) is addressed in D4.2 and by components in other workpackages (mainly WP7 and its dependent workpackages). The following definitions apply in Table 1:

- *text*: (any combination of) natural language statements, keywords, phrases.
- *example*: an image, audio, or video file.

- *concept*: a (mono-media or cross-media) concept defined in the VITALAS lexicon (being specified in WP3).
- *attribute*: property of media items such as date, source, image orientation, etc.

Users' queries are also expressed as the AND and OR Boolean combinations of any of the above components. The aim of this brief description is to illustrate the **querying paradigms** employed in the VITALAS use cases: *query-by-text* (also referred to as *query-by-keywords* [39]), *query-by-example*, *query-by-concept*, and "*database-style*" *querying* (through the specification of attribute values).

Table 1: A IR-oriented description of the VITALAS use cases

	Users searching for ...	**express their queries by ...**			
Use Cases		text	example	concept	attribute
1.2	images, video, audio	X	X	X	X
2.1	images	X		X	
2.2	images	X	X	X	X
2.3	images	X	X		X
2.4	images	X			
2.5	images	X			X
3.1	video	X	X		X
3.2	video, audio				X
	images	X			X
3.3	video	X		X	X
4.2	video		X		X
4.3	video, audio	X	X		

The components of a user's cross-media query (i.e., the *text*, *example*, *concept*, and *attribute* components) are used by cross-media retrieval systems as evidence for retrieving media items relevant to the information need expressed by the query. Section 2.2 discusses how the *text*, *example*, and *concept* components are used (together with the media items' features) as uncertain evidence by the retrieval model in ranking the documents in the collection. The *attribute* component, on the other hand, is not necessarily treated as uncertain evidence, given that it can be interpreted either as a strict or as a loose/vague condition [13, 10]. For instance, consider a journalist looking for images of Antibes taken prior to 1990. By adopting a strict (database-centric) interpretation, the value of the date attribute is interpreted as a constraint; therefore, only images taken until the 31st of December 1989 are considered relevant. A loose (IR-centric) interpretation would treat this date as evidence (hint) of relevance, and could, for instance, also consider relevant images taken on January 1990. In VITALAS, WP4 adopts the strict interpretation of the *attribute* component, and, therefore, the database-style querying is handled outside the cross-media retrieval model. A more detailed discussion on this issue is beyond the scope of this deliverable; the issue is addressed in D4.1.

2.2 Cross-media retrieval approaches

In cross-media retrieval, the information content of documents and queries is represented by multiple modalities. A multimodal document is the basic retrievable unit of information, e.g., an image or a video segment. In VITALAS, we consider that each multimodal document D is represented as a tuple of a textual document T, a visual document V, an audio document A, and a concept-based document C: $D =< T,V,A,C >$.

- A textual document T is represented by the terms in the vocabulary as a vector of term counts (or simply as a bag of terms).
- A visual document V can be represented as a set or sequence of n-dimensional (low-level) visual feature vectors (see D2.0).
- An audio document A can be represented as a sequence of m-dimensional (low-level) audio feature vectors (see D2.0).
- A concept-based document representation C is a vector of the classification scores of the concepts in the WP3 lexicon.

The multimodal queries are represented in the same manner.

In VITALAS, we use generative probabilistic models for representing the multimodal documents and their constituent parts, i.e., we use the same probabilistic basis for all modalities and for their combination. In this uniform theoretical framework, retrieval corresponds to the likelihood of observing the query from each document's model. Given that in this first phase of the project cross-media retrieval focuses on the linguistic (i.e., textual and concept-based) evidence associated with the documents and queries, the focus is on the generative language models, presented in Section 2.2.1. For completeness, Section 2.2.2 discusses briefly an extension to generative visual and audio models. Section 2.2.3 discusses the combination of modalities, and Section 2.2.4 outlines the retrieval system that implements our retrieval approaches.

2.2.1 Generative probabilistic language models

Generative language models, known as statistical language models and simply referred to as *language models* (*LMs*), have a long history [37]; they were originally developed for speech recognition tasks and subsequently employed in other language technologies, such as machine translation, optical character recognition, spelling correction, and many more. In essence, a language model estimates the probability distribution $P(s)$ over all possible linguistic units s (e.g., sentences or whole documents) by applying statistical estimation techniques. In most language models, a linguistic unit is assumed to correspond to a sequence of words w_i. In principle, these sequences can be arbitrarily long, but, in practice, they are approximated by n-grams, $(w_1, w_2, \ldots, w_n)$ [41]. Then, the probabilities can be estimated by:

$$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2, w_1)\ldots P(w_n|w_{n-1}, \ldots, w_1) \quad (1)$$

Different independence assumptions lead to different language models. For instance, unigram language models assume that the probability of a word in a sequence does not depend on any of the previous words, i.e.,

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i) \quad (2)$$

In this deliverable, we are concerned with unigram language models (see [33, 41] for higher order n-gram models).

When the language modelling approach is applied to IR [36, 16, 33], a language model φ_D is inferred for each document D in the collection, i.e., the parameters of this language model are estimated from the document's text. Then, this language model is used for estimating probabilities of samples such as a query. Given a query Q, the ranking of the documents in the collection is produced by estimating the **likelihood of the query** (i.e., the probability of *generating* the query) $P(Q|\varphi_D)$, given φ_D the estimated language model for each document. For simplicity, the query likelihood is also denoted as $P(Q|D)$.

This *query likelihood* retrieval approach is used in the overwhelming majority of the language modelling approaches applied to IR [36, 16, 18, 41, 17, 34, 19]. However, there are alternative ways of using language models. For instance, by estimating language models for both queries and documents, documents are ranked by a distance (similarity) function between the query language model φ_q and each document language model φ_d, such as the Kullback-Leibler (KL) divergence: $-KL(\varphi_q||\varphi_d)$ [29].

For multimodal documents (and queries), language models are used for representing both their textual (e.g., image captions, speech transcripts, subtitles, etc.) components and their concept-based components. We first describe the textual language models, where documents are modelled using a *multinomial* distribution, and queries and documents are represented as sequences of terms [16]. Multinomial language models have been widely applied in other fields [37], and are also used in virtually all current investigations of language models in IR [19]. Queries are represented as sequences of k random variables each corresponding to a term, and the query likelihood is defined as:

$$P(\mathbf{q}|\varphi_{D_T}) = P(q_1, q_2, \ldots, q_k|\varphi_{D_T}) = \prod_{i=1}^{k} P(q_i|\varphi_{D_T}) \quad (3)$$

assuming that each q_i is generated independently from the previous ones given the document model. The language model is thus reduced to modelling the distribution of each single term.

The simplest estimation strategy for an individual term probability is the *maximum likelihood estimate* (*mle*). This corresponds to the relative frequency of a term t_i in a specific document d $P_{mle}(t_i|\varphi_{d_T}) = \frac{tf_{i,d}}{\sum_t tf_{t,d}}$, where $tf_{i,d}$, the term frequency of term t_i in document d, is normalised by the document's length (the sum of the term frequencies of all of its terms). However, this estimate is not suitable for IR, since it will assign zero query likelihood probabilities to documents missing even a single query term. This *sparse* estimation problem is addressed by *smoothing* techniques [58, 59]. In essence, smoothing redistributes some of the probability of the terms occurring in a document to those missing from it.

One of the most popular smoothing methods applied in IR [16] is the Jelinek-Mercer smoothing, which is a *mixture model* (a *linear interpolation*) of the document model with a background model (the collection model in this case):

$$P(\mathbf{q}|\varphi_{D_T}) = \prod_{i=1}^{k}(1-\lambda)P_{mle}(q_i|\varphi_{D_T}) + \lambda P_{mle}(q_i|\varphi_{C_T}) \tag{4}$$

where λ is a parameter and $P_{mle}(t_i|\varphi_{C_T}) = \frac{df_i}{\sum_t df_t}$, with df_i the document frequency of the term t_i in the collection. This equation clearly indicates that collection and term frequencies are integral parts of the language model and are not used heuristically as in many other approaches.

Concept-based language models can be treated similarly to the above standard language modelling approaches that are used for text retrieval. By assuming independence between concepts, $P(\mathbf{q}|\varphi_{D_C})$ is estimated in the same manner as Equation (3). The individual concept probabilities $P(q_i|\varphi_{D_C})$ can be derived from the classification scores estimated during the WP3 cross-media annotation, ensuring of course that the total probabilities of the concepts in the lexicon sum up to 1: $\sum_{j=1}^{|C|} P(c_j|\varphi_{D_C})$.

Finally, early work in multinomial language models for retrieval based on query likelihood [16, 18, 33], proposed to rank the documents in the collection using the posterior probability of a document D being relevant given a query:

$$P(D|Q) \propto P(D)P(Q|D) \tag{5}$$

where $P(D)$ is the prior probability of the document being relevant. Initially, these priors were considered to be uniform, but later research [19] used them to estimate the probability of relevance of a document given its feature(s) $\mathbf{f}$: $P(R|D_{\mathbf{f}})$. To estimate these priors, two approaches can be distinguished [28]: (i) direct estimation on some training data, and (ii) definition based on some general modelling assumptions. For direct estimation of the probability of relevance given a feature f (e.g., its length, similarly to [38]), the distribution of the feature in the relevant set and in the collection can be used as follows:

$$P(R|D_f) = \frac{P(D_f|R)P(R)}{P(D_f)} \propto \frac{P(D_f|R)}{P(f)} = \frac{\#(rel,f)}{\#(f)} \tag{6}$$

where $\#(rel,f)$ is defined as the number of relevant documents with feature f (e.g., a specific length), and $\#(f)$ as the total number of documents with that feature [55]. This estimate can then be used directly in Equation 5. Alternatively, one could make the general modelling assumption (with or without training data) that the a-priori probability of relevance is taken to be a linear function of that feature, e.g., the document length:

$$P(R|D_f) \propto C \times doclength(D) \tag{7}$$

where C a constant that can be ignored in the ranking. In the remainder of this report, for a given document d and query q, Equation 5 is considered to be the retrieval function $RSV_{LM}(d,q)$ for the language modelling approach to IR.

2.2.2 Alternative generative probabilistic models

The previous section has introduced how the language modelling approach to information retrieval can be applied to text and concepts. Alternatively, a multimedia document can also be represented by a generative probabilistic model of the features extracted from its audio-visual contents. For example, Gaussian mixture models may capture the density of (low-level) visual features [49, 51]. Retrieval can again be modelled by the probability that a document generates the (visual parts of the) query $P(\mathbf{q}|\varphi_{D_V})$ [51]. In previous research, CWI has investigated extensively the application of such generative probabilistic models both to image and to video retrieval [51, 56, 53, 54, 21, 52, 32]. In the case of video retrieval, we described the visual aspects using keyframes (a static model of the shot) as well as complete shots (a dynamic variant of the same model). Additional experiments investigated the integration of audio features into the same approach.

While having an integrated framework that spans text, concepts and features is desirable from a theoretical viewpoint, a drawback of using Gaussian mixture models to represent the audio or visual feature space is the relatively high computational cost of comparing the query to each of the document models. Because VITALAS is particularly concerned with scalability, retrieval based directly on (low-level) visual features has not been applied in scope of this deliverable.

2.2.3 Combination of modalities

So far, we have treated the various components of multimodal documents separately. To combine the different modalities in the generative probabilistic framework, we can simply compute the joint probability of observing the various query parts: $P(Q|\varphi_{D_T}, \varphi_{D_V}, \varphi_{D_A}, \varphi_{D_C})$. In this deliverable, we focus on the textual and concept-based representations of documents and queries:

$$P(Q|\varphi_{D_T}, \varphi_{D_C}) = P(Q_T, Q_C|\varphi_{D_T}, \varphi_{D_C}) = P(Q_T|\varphi_{D_T})P(Q_C|\varphi_{D_C}) \tag{8}$$

where the following two independent assumptions have been applied:

1. Textual terms and concepts are generated independently: $P(Q_T, Q_C|\cdot) = P(Q_T|\cdot)P(Q_C|\cdot)$.
2. The generation of documents in one modality is independent of the other modality, i.e., the generation of textual terms only depends on the language model and the generation of concepts only on the concept-based language model: $P(Q_T|\varphi_{D_T}, \varphi_{D_C}) = P(Q_T|\varphi_{D_T})$ and $P(Q_C|\varphi_{D_T}, \varphi_{D_C}) = P(Q_C|\varphi_{D_C})$.

Treating textual and concept-based information independently is a simplification, since these two types of information are dependent. As soon as a document is likely to be relevant based on the textual information, then the likelihood of observing a concept similar to the query concepts should increase. For example, if the name “Yasser Arafat” is mentioned, the likelihood of observing the concept associated with him (as this concept is defined in [40]) increases. In this first phase of the project, the independence assumptions listed above are used, while dependencies will be explored in cooperation with the annotation efforts of WP3. Furthermore, in this deliverable we are concerned with documents for which both textual and concept-based annotations exist. The issue of handling collections that contain some documents that are not associated with either of the representations is an open research question currently being investigated.

2.2.4 Retrieval system

The cross-media retrieval component that implements the above retrieval approaches is based on the `PF/Tijah` [2] system [20], which is briefly described in Appendix 5.2. `PF/Tijah` can process combined IR and DB XML queries; here, we focus on IR queries that are expressed in the system using NEXI (Narrowed Extended XPath I) [42, 43]. NEXI is an XML query language based on simplified XPath I [1] containing only the descendant axis, and extended with the about() function in order to perform IR queries on XML elements. NEXI queries can take one of the following forms:

```
//A[B]          Return A XML elements about B
//A[B]//C[D]    Return C XML elements about D,
                where C are descendants of A where A is about B
```

A and C are paths and B and D are clauses containing at least one about(). The about() function takes the following form:

[2] `http://dbappl.cs.utwente.nl/pftijah/`

```
about(<relative_path>, <query_component>)
```

Initially, the original NEXI query language [42] supported only textual query components, i.e.,

```
about(<relative_path>, text)
```

where *text* corresponds to its description in Table 1. For example:

```
//image[about(., cityscape)]
//image[about(.//caption, cityscape)]
```

Subsequently, multimedia extensions were added to NEXI to support querying by media *example* query components [48] and by *concepts* [57] (see Section 3.1)

```
about(<relative_path>, src:<example_file>)
about(<relative_path>, concept:<specific_concept>)
```

The **interepretation** of the about() function is tied to the retrieval model applied, but it is independent of NEXI. That provides flexibility given that the specification and implementation of the underlying retrieval techniques can be modified, without requiring adaptation of the NEXI queries.

2.3 Cross-media retrieval evaluation

In this first phase of the project, evaluation is performed by using data made available by international benchmarks, rather than user studies on VITALAS data (foreseen for the second phase of the project). This year, CWI participated independently of the other VITALAS partners in evaluation benchmarks and in particular in their automatic (rather than interactive) retrieval tracks. Such tracks evaluate retrieval systems using a test collection consisting of a set of documents, a set of topics and a set of relevance judgements. The documents are the basic media units to retrieve, the topics are descriptions of the information needs, and the relevance judgements list the set of relevant documents for each topic. Evaluation measures are the standard precision/recall and Mean Average Precision (MAP) IR metrics [47, 2]. An in depth discussion on the issue of evaluation for multimedia retrieval can be found in Chapter 6 of [51].

3 Evaluating Cross-Media Retrieval of Images

For the evaluation of cross-media retrieval of images, we participated in the INEX Multimedia evaluation benchmark. INEX Multimedia 2007 was organised by CWI [45], in coordination with the INEX organisers. CWI, in cooperation with University of Twente, participated in three tracks in INEX 2007 [44].

This section presents the participation in the INEX Multimedia track and is organised as follows. The justification of the suitability of INEX Multimedia as an evaluation benchmark is provided in Section 3.2, which compares INEX Multimedia to the VITALAS environment (where the selected data representation format is also XML) in order to highlight their similarities (and differences). This is preceeded by Section 3.1 which describes the INEX Multimedia test collection. Section 3.3 presents the evaluation results of our experiments.

3.1 INEX Multimedia test collection

The aim of the **Initiative for the Evaluation of XML Retrieval** (INEX)[3], launched in 2002, is to establish an infrastructure and provide means for the evaluation of content-oriented XML retrieval systems. To this end, it provides a large XML test collection and appropriate metrics for the evaluation of structured document retrieval approaches from XML documents. Such approaches aim at retrieving XML document fragments that contain relevant information. The aim of this retrieval paradigm is to reduce users' effort to locate relevant content by directing them not just to the documents containing the relevant information, but to their most relevant parts. This is of particular benefit for information repositories containing long documents or documents covering a wide variety of topics. The main INEX Ad Hoc task focuses on text-based XML retrieval.

Although text is dominantly present in most XML document collections, other types of media can also be found in those collections; those media form the focus of **INEX Multimedia**[4]. Existing research on multimedia IR has already shown that it is far from trivial to determine the combined relevance of a document that contains several multimedia objects [30]. The objective of INEX Multimedia is to exploit the XML structure that provides a logical level at which multimedia objects are connected, to improve the retrieval performance of an XML-driven multimedia information retrieval system. To this end, it provides an evaluation platform for the retrieval of multimedia documents (corresponding to images and their metadata) and XML document fragments (corresponding to XML elements or passages that contain images and text). INEX Multimedia ran a pilot evaluation study in 2005 [48] and has been established as an INEX track in 2006 [57] and 2007 [45].

This section introduces the main parts of the INEX Multimedia 2006-2007 test collection: documents, tasks, topics, and relevance assessments (Sections 3.1.1–

[3] `http://inex.is.informatik.uni-duisburg.de`

[4] `http://inex.is.informatik.uni-duisburg.de/2007/mmtrack.html`

3.1.4). More detailed information can be found in the reports on the INEX Multimedia track for 2006 [57] and 2007 [45].

3.1.1 Wikipedia collections and additional resources

In 2006 and 2007, INEX Multimedia employed two Wikipedia-based collections:

Wikipedia XML collection: This is a structured collection of 659,388 Wikitext pages from the English part of Wikipedia, the free content encyclopedia (`http://en.wikipedia.org`), that have been converted to XML [9]. This collection has been created for the Ad Hoc track[5]. Given, though, its multimedia nature (as indicated by its statistics listed in Table 2), it is also being used as the target collection for a multimedia task that aims at finding relevant XML fragments given a multimedia information need (see Section 3.1.2).

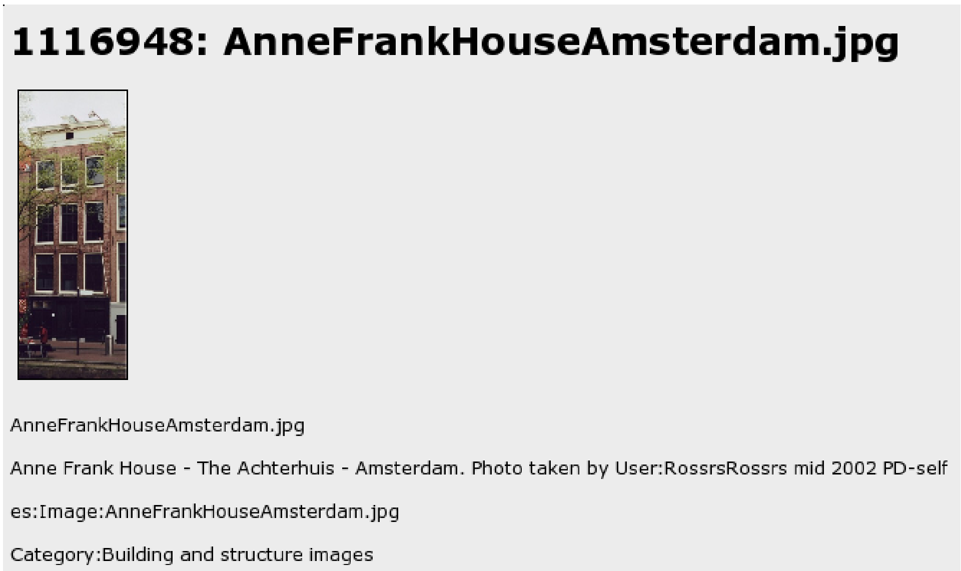
Table 2: Wikipedia XML collection statistics

Total number of XML documents	659,388
Total number of images	344,642
Number of unique images	246,730
Average number of images per document	0.52
Average depth of XML structure	6.72
Average number of XML nodes per document	161.35

Wikipedia image XML collection: This is a collection consisting of the images in the Wikipedia XML collection, together with their metadata. These metadata, formatted in XML, usually contain a brief caption or description of the image, the Wikipedia user who uploaded the image, and the associated copyright information. Figure 1 shows an example of such a document consisting of an image and its associated metadata. Some images from the Wikipedia XML collection have been removed due to copyright issues or parsing problems with their metadata, leaving us with a collection of 170,370 images with metadata. This collection is used as the target collection for a multimedia/image retrieval task that aims at finding images (with metadata) given a multimedia information need (see Section 3.1.2).

Although the above two Wikipedia-based collections are the main search collections, additional sources of information are also provided to help participants in the retrieval tasks. These resources are:

[5] The main retrieval task performed at INEX is the Ad Hoc retrieval of XML documents. Ad Hoc retrieval is described as a simulation of how a library might be used, and it involves the searching of a static set of XML documents using a set of queries that may contain both content and structural conditions. In response to a query, arbitrary XML elements may be retrieved from the library.



1116948: AnneFrankHouseAmsterdam.jpg

AnneFrankHouseAmsterdam.jpg

Anne Frank House - The Achterhuis - Amsterdam. Photo taken by User:RossrsRossrs mid 2002 PD-self

es:Image:AnneFrankHouseAmsterdam.jpg

Category:Building and structure images

Figure 1: Example Wikipedia image and metadata document from the Wikipedia image XML collection.

Image classification scores: For each image, the classification scores for the 101 different MediaMill concepts are provided by UvA [40]. The UvA classifier is trained on manually annotated TRECVID video data and the concepts are selected for the broadcast news domain.

Image features: For each image, the set of the 120D feature vectors that has been used to derive the above image classification scores is available [46]. Participants can use these feature vectors to custom-build a CBIR system, without having to pre-process the image collection.

The above resources are beneficial to researchers who wish to exploit visual evidence without performing image analysis.

3.1.2 Retrieval tasks

The aim of the retrieval tasks in INEX Multimedia is to retrieve relevant (multimedia) information, based on an information need with a (structured) multimedia character. To this end, a structured document retrieval approach should be able to combine the relevance of different media types into a single ranking that is presented to the user.

For INEX 2006 and 2007, two tasks are defined:

MMfragments task: Find relevant XML fragments in the **Wikipedia XML collection** given a multimedia information need. These XML fragments can correspond to XML elements or passages. This is similar to the direction taken by the INEX Ad Hoc track. The difference is that MMfragments topics ask for multimedia fragments (i.e., fragments containing at least one image) and may also contain visual hints (see Section 3.1.3).

MMimages task: Find relevant images in the **Wikipedia image XML collection** given a multimedia information need. Given an information need, a retrieval system should return a ranked list of documents (i.e., images and their metadata) from this collection. Here, the type of the target element is defined, so basically this is closer to an image retrieval (or a document retrieval) task, rather than XML element or passage retrieval. Still, the structure of (supporting) documents, together with the visual content and context of the images, could be exploited to get to the relevant images (and their metadata).

All track resources (see Section 3.1.1) can be used for both tasks.

3.1.3 Topics

Two sets of topics, one for each task, have been created for INEX Multimedia. These topics have been developed and subsequently assessed by the participants in the benchmark.

The INEX Multimedia topics are descriptions of (structured) multimedia information needs that may contain not only textual, but also structural and multimedia hints. The structural hints specify the desirable elements to return to the user and where to look for relevant information, whereas the multimedia hints allow the user to indicate that results should have images similar to a given example image or be of a given concept. These hints are expressed in the NEXI query language [42].

The original NEXI specification determines how structural hints can be expressed, but does not make any provision for the expression of multimedia hints. These have been introduced as NEXI extensions during the INEX 2005 and 2006 Multimedia tracks [48, 57]:

- To indicate that results should have images similar to a given example image, an *about* clause with the keyword *src:* is used. For example, to find images of cityscapes similar to the image at `http://www.bushland.de/hksky2.jpg`, one could type:

```
//image[about(.,cityscape) and
             about(.,src:http://www.bushland.de/hksky2.jpg)]
```

- To indicate that the results should be of a given concept, an *about* clause with the keyword *concept:* is used. For example, to search for cityscapes, one could decide to use the concept “building”:

```
//image[about(.,cityscape) and about(.,concept:building)]
```

This feature is directly related to the concept classifications that are provided as an additional source of information (see Section 3.1.1). Therefore, terms following the keyword *concept:* are obviously restricted to the 101 concepts for which classification results are provided.

The INEX 2007 Multimedia topics consist of the following parts:

<title> The topic `<title>` simulates a user who does not know (or does not want to use) the actual structure of the XML documents in a query and who does not have (or want to use) example images or other visual hints. This profile is likely to fit most users searching XML digital libraries and also corresponds to the standard web search type of keyword search.

<castitle> A NEXI expression with structural hints.

<mmtitle> A NEXI expression with structural and visual hints.

<description> A brief, matter of fact, description of the information need in natural language.

<narrative> A clear and precise description of the information need. The narrative unambiguously determines whether or not a given document or document part fulfils the given need. It is the only true and accurate interpretation of a user's needs. Precise recording of the narrative is important for scientific repeatability - there must exist, somewhere, a definitive description of what is and is not relevant to the user. To aid this, the `<narrative>` should explain not only what information is being sought, but also the context and motivation of the information need, i.e., why the information is being sought and what work-task it might help to solve.

For example, one of the INEX MMimages 2007 topics is the following:

<inex_topic topic_id="27">

<title>cities by night**</title>**

<castitle>`//article[about(.,cities by night)]`**</castitle>**

<mmtitle>`//article[about(.,cities by night) and about(.,concept:building) and about(.,src:http://www.bushland.de/hksky2.jpg)]`**</mmtitle>**

<description>Find photos of cities by night.**</description>**

<narrative> I am decorating my flat and as I like photos of cities at night, I would like to find some that I could possibly print into posters. Photos of a specific building at night and photos of cities (or the earth) from space are not relevant. I would like to find photos of skylines or photos that contain parts of a city at night (including streets and many buildings).**</narrative>**

```
</inex_topic>
```

In 2006, both structural and visual/multimedia hints were expressed in the `<castitle>` field. In 2007, the `<castitle>` contains only structural hints, while the `<mmtitle>` is an extension of the `<castitle>` that also incorporates the additional visual hints (if any). The introduction of a separate `<mmtitle>` is particularly useful, since it makes it easier for systems to compare runs using structural hints to those using both structural and visual hints, without having to modify the query expression. For MMimages, in particular, the queries in the `<castitle>` and `<mmtitle>` fields are restricted to: `//article[X]`, where `X` is a predicate using one or more *about* functions with textual and/or multimedia hints. This is justified given that MMimages requires retrieval at the document level, rather than elements or passages.

3.1.4 Relevance assessments

The INEX 2007 MMfragments task was run in parallel with the Ad Hoc track, with the relevance assessments for this task being arranged by the Ad Hoc track organization. Since XML retrieval requires assessments at a sub-document level, a simple binary judgement at the document level is not sufficient. Still, for ease of assessment, retrieved fragments are grouped by document. Once all participants have submitted their runs, the top N fragments for each topic are pooled and grouped by document. The documents are alphabetised so that the assessors do not know how many runs retrieved fragments from a certain document or at what rank(s) the fragments were found. Assessors then look at the documents in the pool and highlight the relevant parts of each document. The assessment system stores the relevance or non-relevance of the underlying XML elements and passages.

The MMimages task is a document retrieval task. A document, i.e., an image with its metadata, is either relevant or not. For this task, the INEX Multimedia organisers adopted TREC style document pooling of the documents and binary assessments at the document (i.e., image with metadata) level. Again, the top N documents from the submitted runs were pooled per topic. One assessor per topic (i.e., the participant who had created the topic) looked at the pooled documents (which were ranked in alphabetical order) and indicated each document's relevance. In 2006, the pool depth was set to 500 for the MMimages task, with post-hoc analysis showing that pooling up to 200 or 300 would have given the same system ordering [57]. This led to the decision to pool this year's submissions up to rank 300, resulting in pools of between 348 and 1865 images per topic, with both mean and median around 1000 (roughly the same size as 2006).

3.2 INEX Multimedia test collection vs. the VITALAS environment

This section compares the INEX Multimedia test collection with the VITALAS image retrieval environment, i.e., the images in the archives of Belga News Agency[6] and the queries users would submit in contexts similar to the VITALAS image retrieval use cases (see Table 1). The similarities arising from this comparison justify the choice of INEX Multimedia as a suitable evaluation benchmark. The differences, on the other hand, could aid us in understanding any dissimilarities between the effectiveness of our retrieval approaches when applied to INEX Multimedia and their effectiveness when applied to VITALAS data.

The INEX Multimedia image collection (Wikipedia image XML collection) is an heterogeneous collection covering various topics, rather than being restricted to a narrow domain. This makes it comparable to Belga's image collection which contains a wide variety of editorial and creative pictures; obviously, Belga's archive is much richer and the quality of its photos is much higher, given that they are taken by professional photographers. In addition, even though INEX Multimedia does not contain as many images as the Belga archives do, it is still a sizable collection, particularly when compared with the photographic collections in other benchmarks (e.g., ImageCLEF[7]). Finally, the images in INEX Multimedia are associated with user-generated metadata; this might not be the case in the complete set of Belga data to be made available, but the quality of the manual metadata in Belga's pictures will be very high, given that they are annotated by professional archivists.

The nature of the image retrieval task performed by (foreseen) VITALAS users and by (current) Belga users is similar to INEX's MMimages task. The MMfragments task, on the other hand, is more representative of users of archives that contain documents that are mainly text-based and possibly illustrated by images (e.g., newspapers, magazines, etc.). Nevertheless, the techniques developed for the MMfragments task could also be used in other cross-media retrieval tasks. For instance, retrieval approaches for the combination of evidence from different modalities can be useful for the MMimages task, whereas techniques that determine the level of granularity to return to the user when there is no predefined retrieval unit[8] can be useful in video retrieval for identifying the appropriate segment to retrieve.

The VITALAS use cases require that the VITALAS system will support users in employing (combinations of) various querying paradigms when submitting their queries. Apart from query-by-text and query-by-example which are supported by most (if not all) multimedia evaluation benchmarks, INEX Multimedia also provides topics that contain explicit query-by-concept components. Table 3 shows a summary of the distribution of the types of queries[9] over the last two years, whereas the Tables in Appendix 5.1 provide a more detailed view of the same information

[6] Belga News Agency (`http://www.belga.be`) is a VITALAS content provider partner.

[7] `http://ir.shef.ac.uk/imageclef/`

[8] The identification of the appropriate level of granularity of the retrieval results constitutes the main research question in the field of XML IR.

[9] All topics include a query-by-text component, which is not explicitly mentioned in Table 3.

(presented in a format identical to that of Table 1). Moreover, Table 3 indicates that not all topics contain visual/multimedia hints; this corresponds well with realistic scenarios, since users who express multimedia information needs do not necessarily want (or have the ability) to employ visual hints.

Table 3: Querying paradigms employed in the INEX MMimages topics

	INEX MMimages		
	2006	**2007**	**2006-2007**
Number of topics	13	20	33
Number of topics with multimedia hints	7	10	17
Number of topics with *src:* (*query-by-example*)	6	7	13
Number of topics with *concept:* (*query-by-concept*)	2	6	8
Number of topics with both *src:* and *concept:*	1	3	4

Query length is one of the important features in retrieval and its effect has been widely investigated (e.g., [4]). Table 4 shows some statistics on the MMimages topics' length. Each topic's `<title>` corresponds to query-by-keywords (the most popular form of querying and one of the VITALAS requirements), while the `<description>` is its more verbose version in the form of a natural language statement. The average length of `<title>` is between 2-3 terms, which is consistent with both general and image request queries on the Web. In particular, analyses of search (transaction) logs of general purpose Web search engines report that the average query length on the Web is 2.21 terms [22, 24], whereas the average length of image/multimedia request queries is slightly higher: 3.74 [14] and 4 [23] terms per query. This higher mean number of terms in image/multimedia searching can be attributed to the fact that the above studies analysed search logs of general-purpose search engines (such as Excite[10] and Altavista[11]); therefore, the users who submitted the queries had to perform image/multimedia searching using the standard interface methodology for general Web searching, i.e., a text box, where they had to include at least one term in the query to indicate the type of desired retrieval results (e.g., "images of Albert Einstein"). By excluding such terms, the average length drops to 2-3 terms per query, similar to that of the INEX MMimages topics.

Table 4: Statistics for the INEX MMimages topics

	INEX MMimages		
	2006	**2007**	**2006-2007**
Number of topics	13	20	33
Average number of terms in `<title>`	3	2.35	2.61
Average number of terms in `<description>`	7.08	7.65	7.42

We also analysed the length of queries submitted to Belga's image search web

[10] `http://www.excite.com`

[11] `http://www.altavista.com`

interface during 3 months (June-September 2007). Belga's interface currently supports only query-by-text (together with specification of various attribute values). Table 5 indicates that the average query length is much lower than that of the INEX MMimages topics and that of queries submitted to general-purpose Web search engines. However, it is close to the findings of the analysis of a similar log: two small samples (1,439 and 1,371 queries) of the search logs from a commercial image provider, where the mean length is 1.84 and 1.9 terms per query, respectively [25]. This suggests that the VITALAS system should be able to handle effectively even shorter queries, and that query expansion techniques[12] might prove crucial in improving the retrieval effectiveness.

Table 5: Statistics for queries submitted to Belga's image search

	Belga search logs			
	Month 1	**Month 2**	**Month 3**	**Total**
Number of queries	34,362	32,488	34,435	101,285
Average number of terms	1.44	1.42	1.45	1.43

The above discussion clearly indicates that INEX Multimedia is particularly well suited for evaluating the research objectives investigated by VITALAS. It further guides us in determining the setting of our evaluation experiments so that it resembles (as much as possible) the VITALAS environment. For instance, our evaluation focuses on the INEX MMimages task, with the `<title>` and/or the `<mmtitle>` parts of the topics being the most suitable querying paradigms.

3.3 Evaluation experiments and results

We participated in both tasks of the INEX Multimedia 2007 track [44]. Here, we only present the MMimages task, where the aim is to retrieve documents (images and their metadata) from the Wikipedia image XML collection, and, thus, it is closer to VITALAS' retrieval requirements. For our official submissions, we focussed on the textual components of documents and queries. Each image is represented either by its textual metadata in the Wikipedia image XML collection, or by its textual context when that image appears as part of a document in the (Ad Hoc) Wikipedia XML collection. We used the image retrieval approach based on language models described in Section 2.2, with λ set to 0.8 based on training in other test collections.

To be more specific, we submitted the following three runs:

title_MMim We represent each image by the metadata accompanying it in the Wikipedia image XML collection. We create a stemmed index from these representations and perform retrieval using only the MMimages topics' title field: `//article[about(.,$title)]`.

[12] Query expansion/modification techniques are investigated in D4.2 in the context of relevance feedback approaches.

article_MMim For this run, we do not use the Wikipedia image XML collection. We create a stemmed index of the articles in the Ad Hoc Wikipedia XML collection, and rank these articles using each topic's title field. Then, we retrieve the images that these articles contain and filter the results, so that only images that are part of the Wikipedia image XML collection are returned.

figure_MMim For this run, we also do not use the Wikipedia image XML collection. We represent the figures in the Ad Hoc Wikipedia XML collection using their captions. We rank these figures by performing retrieval using each topic's title field: `//figure[about(.,$title)]`. We return the images of these figures (ensuring that these images are part of the Wikipedia image XML collection).

The *article_MMim* and *figure_MMim* runs represent each image by using evidence from its **usage context**, i.e., the surrounding text (and possibly other media) when an image is placed within a document. This technique has also been applied on the Web (e.g., [15]); however, its application on the VITALAS environment might be difficult given that the usage context of the images (e.g., when a image is included as part of a magazine article) is not currently provided.

Figure 2: Official CWI submissions to INEX MMimages 2007

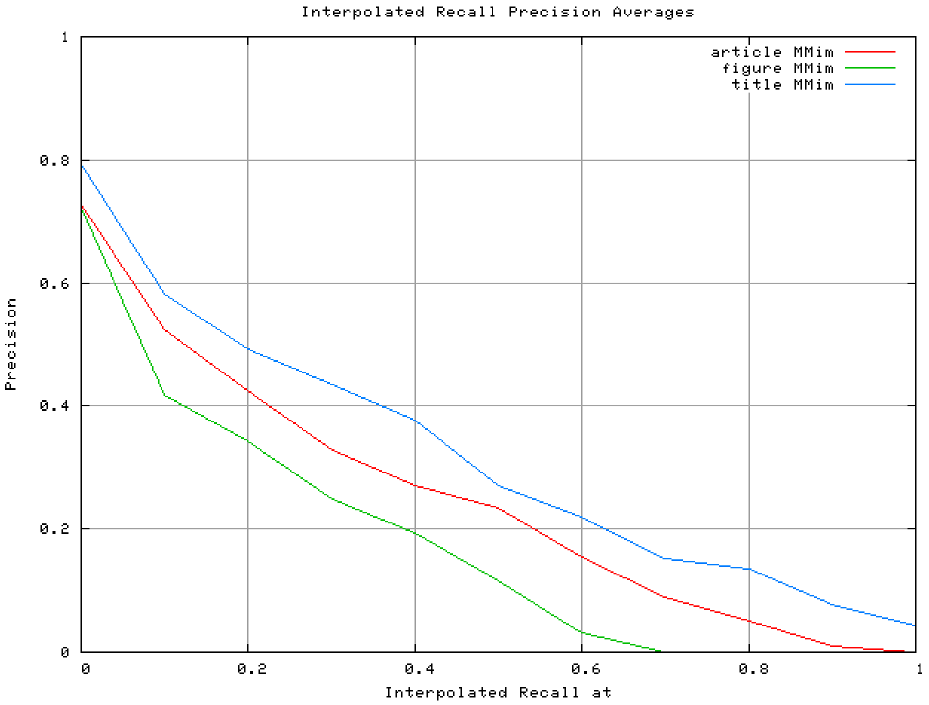


Figure 3: All official submissions to INEX MMimages 2007

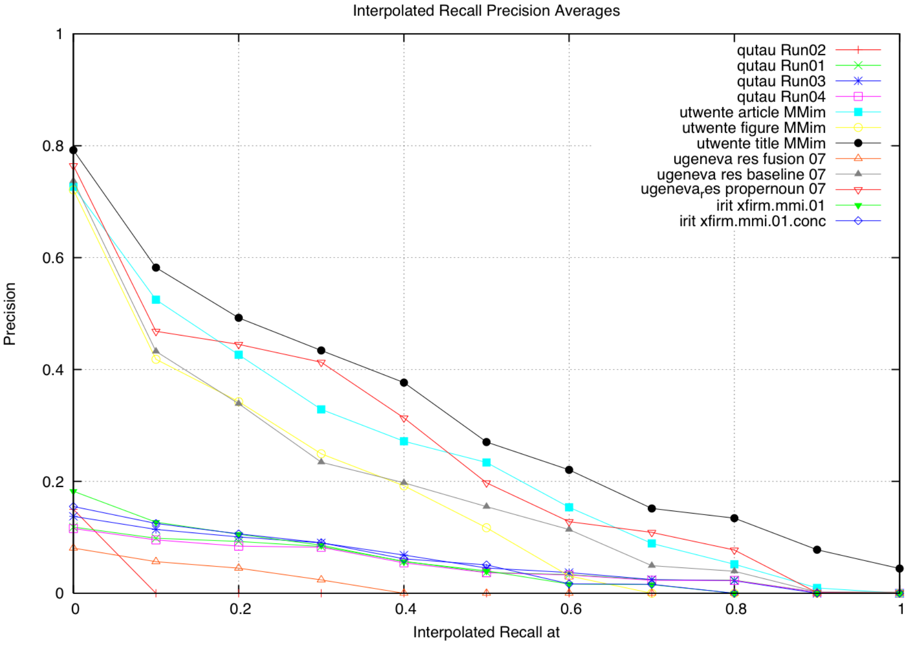


Table 6: Mean Average Precision (MAP) for CWI's official MMimages 2007 runs

Retrieval run	**MMimages 2007**
article_MMim	0.2240
figure_MMim	0.1551
title_MMim	0.2998

Table 6 presents the Mean Average Precision (MAP) of the official runs, and Figure 2 plots their precision/recall graphs. Figure 3[13] compares these runs against all the runs submitted to the MMimages task. Our experimental results indicate that these text-based runs give a highly competitive performance on the MMimages task when compared with other participants' runs that make use of other (beyond-textual) sources of evidence. In particular, *title_MMim*, which relies solely on the images' metadata, is the most effective. Motivated by this finding and in order to resemble the VITALAS environment, our additional runs are also only based on the linguistic evidence accompanying the images in the collection.

Our additional runs are named as *X_Y_Z_W* where:

- *X* denotes the part of the INEX MMimages topics used in the run, i.e., `<title>`, `<castitle>`, `<mmtitle>`, `<description>`, or `<narrative>`. The multimedia parts of `<mmtitle>`, i.e., the *concepts* and the *examples*, can also be individually considered.

- *Y* denotes the resources used to represent the documents, i.e., (1) textual resources: (i) the Wikipedia XML collection (wikiXML) and (ii) the Wikipedia image XML collection (wikiIMG), (2) concept-based resources: image classification scores (Concepts), and (3) visual resources: image features (Features).

- *Z* denotes the retrieval model, i.e., smoothed language model (LMS - Equation (4)) for text-based retrieval, concept-based language model (LMC) for concept-based retrieval, and Gaussian mixture models (GMM) for visual-based retrieval. The combination of evidence from different modalities is performed according to Equation (8).

- *W* denotes the document prior employed in retrieval (see Equation (5)).

For instance, the official *title_MMim* run is represented in the new notation as: title_wikiIMG_LMS_noPrior.

Table 8 (p. 29) presents a summary of our runs, where the resources these runs use are indicated. As mentioned above, these runs focus on the textual and concept-based parts of queries and documents. Table 7 presents the MAP of the additional runs, when these are performed against the INEX MMimages datasets of the last two years. We consider the title_wikiIMG_LMS_noPrior (i.e., the official *title_MMim*) run as our baseline.

First, we performed a run similar to the baseline, with the only difference that we used the `<description>` part of the topic instead of the `<title>`. Although the `<title>` part of the INEX topics is closer to the VITALAS image queries in terms of its length compared to the `<description>` part (see discussion in Section 3.2),

[13] In this figure, the official CWI runs are denoted as *utwente_run*. This is due to the fact that CWI cooperates with University of Twente when participating in international benchmarks, and *utwente* is usually adopted as our id.

Table 7: Mean Average Precision (MAP) for runs in all MMimages datasets

	MMimages dataset		
Retrieval run	**2006**	**2007**	**2006-2007**
title_wikiIMG_LMS_noPrior	0.3864	0.2998	0.3339
title_wikiIMG_LMS_lengthPrior	0.4006	0.3094	0.3453
title_wikiIMG_LMS_logLengthPrior	0.3984	0.3066	0.3428
description_wikiIMG_LMS_noPrior	0.3325	0.3022	0.3141
description_wikiIMG_LMS_lengthPrior	0.3239	0.3081	0.3143
description_wikiIMG_LMS_logLengthPrior	0.3297	0.3072	0.3160
title+concepts_wikiIMG_LMS_noPrior	0.3811	0.2740	0.3162
title+concepts_wikiIMG_LMS_lengthPrior	0.3960	0.2832	0.3276
title+concepts_wikiIMG_LMS_logLengthPrior	0.3946	0.2812	0.3259
mmtitle_wikiIMG+Concepts_LMS+LMC_noPrior	0.3843	0.2790	0.3212

we wanted to investigate how longer queries affect the retrieval effectiveness. If longer queries would improve the performance, it could be useful to encourage searchers to enter longer queries by proper support via the user interface [3].

For MMimages 2007, description_wikiIMG_LMS_noPrior slightly improves the effectiveness, whereas for the MMimages 2006 dataset the results are worse than the baseline. This can be attributed to the fact that the `<description>` part of topics is a more verbose version of the `<title>`, rather than offering more information on the user's need. For instance, for the MMimages 2006 topic with `<title>` "barcelona", its `<description>` is "I'm looking for pictures of the Barcelona city". Therefore, even though the `<description>` is longer, its does not necessarily offer more evidence to the retrieval function, but possibly more noise.

Overall, the results of the description_wikiIMG_LMS_noPrior run are incoclusive in terms of the effect of longer queries (i.e., queries that offer more evidence about a user's information need). Further experiments with truly longer user queries are needed. The effect of longer queries will also be investigated in the context of query expansion techniques applied by (blind and interactive) relevance feedback approaches in D4.2.

One way to add more evidence to the textual `<title>` part of the query is to extract the concepts from the `<mmtitle>` part (where available) and add them to the `<title>` by treating them as terms. For instance, a topic with `<title>` "cityscape" and `<mmtitle>` `//image[about(.,cityscape) and about(.,concept:building)]` would become "cityscape building". The retrieval effectiveness of this run, which is denoted as title+concepts_wikiIMG_LMS_noPrior, is slightly worse than the baseline for both datasets, indicating that simply treating the concepts as terms against the index created by the images' metadata is not particulalry useful.

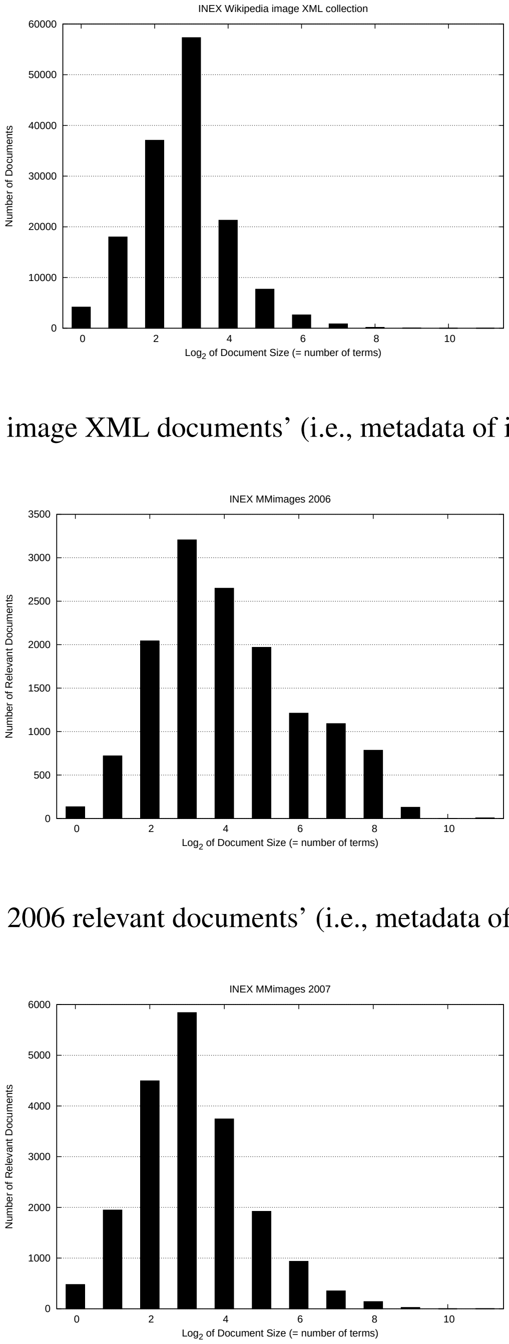
Next, we treat the concepts from `<mmtitle>` as a different modality and combine it with the textual evidence. This combination of modalities is denoted as mmtitle_wikiIMG+Concepts_LMS+LMC_noPrior. The effectiveness is again worse then the baseline, but still better than simply treating concepts as terms. Further investigation is needed by using query expansion methods and by taking into account the dependencies between modalities.

For the above retrieval approaches, we also integrate surface features as another source of evidence in the form of document priors. Surface features are those properties of (multimedia) documents that do not describe their content. Examples include the length of a document, a reference to where the document is located, and the production date of a document. Although these features do not directly relate to the document's content, they can be valuable additional sources of information in a retrieval setting. In text retrieval for example, the length of a document is often used as an indicator of relevance (longer documents are more likely to be relevant). Similarly, the number of hyperlinks pointing to a document is an indicator of the importance of a document [35, 6, 27]. These surface features are typically ignored in multimedia retrieval (for an exception see [55]).

We study the influence of a very basic feature: the size of each document's metadata. Size priors have played an important role in information retrieval [38, 16, 28]. Kamps et al. [26] studied length normalization in the context of XML retrieval and ad hoc INEX collections, and found that the size distribution of relevant elements differed significantly from the general size distribution of elements. Emphasizing longer elements by introducing, linear, quadratic or even cubic length priors improved the retrieval results significantly on the INEX 2002-2005 (IEEE) collection. Of course, the nature of the textual components of multimedia collections is different from that of text-based document collections, but the intuition that (in some applications) users might consider more useful images that are accompanied by more information still holds.

We incorporated a document prior based on length (defined as the number of terms in the metadata) and the log of this length. By defining the priors in that manner, we are able to apply them without performing any training. Our results indicate that both priors improve over the corresponding baselines, with the length prior improving the most.

These runs are based on the assumption that the distribution of document size is different for relevant and non-relevant images. We perform a retrospective analysis of the distribution of length in the MMimages collection (Figure (4a)), and the relevant documents for both 2006 (Figure (4b)) and 2007 (Figure (4c)). While the collection contains many small documents, these are rarely relevant. If we would not pay attention to document length and just use a retrieval model that does not have a bias for documents of any size, we would retrieve too many small documents. Simply giving a bias towards longer documents in the context of the INEX MMimages task has the potential of improving the retrieval result, which is confirmed by our evaluation experiments.



(a) Wikipedia image XML documents' (i.e., metadata of images) sizes

(b) MMimages 2006 relevant documents' (i.e., metadata of images) sizes

(c) MMimages 2007 relevant documents' (i.e., metadata of images) sizes

Figure 4: Size distribution of metadata in Wikipedia image XML collection and in relevant images indicating the relevance bias towards longer documents.

Table 8: Runs on the INEX MMimages 2007 dataset and the resources they used.

		Wikipedia resources used			
Retrieval run	**INEX runID**	image metadata	image classification scores	image features	textual context
title_wikiXML_LMS_noPrior	article_MMim				X
title_wikiXML_LMS_noPrior	figure_MMim				X
title_wikiIMG_LMS_noPrior	title_MMim	X			
title_wikiIMG_LMS_lengthPrior		X			
title_wikiIMG_LMS_logLengthPrior		X			
title_wikiIMG_LMS_logNormalPrior		X			
description_wikiIMG_LMS_noPrior		X			
description_wikiIMG_LMS_lengthPrior		X			
description_wikiIMG_LMS_logLengthPrior		X			
description_wikiIMG_LMS_logNormalPrior		X			
title+concepts_wikiIMG_LMS_noPrior		X			
title+concepts_wikiIMG_LMS_lengthPrior		X			
title+concepts_wikiIMG_LMS_logLengthPrior		X			
title+concepts_wikiIMG_LMS_logNormalPrior		X			
mmtitle_wikiIMG+Concept_LMS+LMC_noPrior		X	X		

4 Conclusions

This deliverable is a preliminary report on our participation in the INEX Multimedia image retrieval evaluation benchmark. INEX Multimedia is a particularly well suited benchmark, since the characteristics of its test collection are very similar to the VITALAS retrieval environment. Our retrieval approaches are based on a uniform generative probabilistic framework. The results of our evaluation experiments indicate the value of linguistic evidence in the context of image retrieval. The use of the textual query components against representations based on the images' metadata, together with the integration of document priors, appears to an effective retrieval strategy.

Given that not all images are associated with such metadata, other sources of evidence need to be considered. So far, we only considered concept-based representations of documents, which we used against the concept-based query components. This has however limited applicability, because not all query expressions contain a query-by-concept component. A possible solution would be to associate textual query components to the concept-based document representations. This can be achieved by considering dependencies in the combination of the modalities, which can be determined either within the same generative probabilistic framework employed for retrieval, or by employing the cross-media annotation techniques developed in WP3, where textual features are associated with concepts. These, together with query expansion techniques developed in the context of relevance feedback approaches and which can be applied for text-only, concept-only, and text+concept query components are currently being investigated by the activities of WP4 task 4.2.

For the next phase of the project, we foresee participation in two international benchmarks: TRECVID and ImageCLEF (which will probably host the INEX Multimedia test collection), in order to further investigate our research objectives.

References

[1] XML Path Language (XPath) Version 1.0, 1999. W3C Recommendation [16 November 1999].

[2] R. Baeza-Yates and B. Ribeiro-Neto, editors. *Modern Information Retrieval*. ACM Press and Addison Wesley, 1999.

[3] N. J. Belkin, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and C. Cool. Rutgers interactive track at TREC 2002. In Voorhees and Buckland [50].

[4] N. J. Belkin, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and C. Cool. Query length in interactive information retrieval. In Clarke et al. [7], pages 205–212.

[5] P. Boncz, T. Grust, M. van Keulen, S. Manegold, J. Rittinger, and J. Teubner. MonetDB/XQuery: a fast XQuery processor powered by a relational engine. In S. Chaudhuri, V. Hristidis, and N. Polyzotis, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 479–490. ACM Press, June 2006.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In P. Thistlewaite and H. Ashman, editors, *Proceedings of the 7th international conference on World Wide Web (WWW 1998), Computer Networks, 30(1-7)*, pages 107–117. Elsevier Science Publishers, April 1998.

[7] C. Clarke, G. Cormack, J. Callan, D. Hawking, and A. Smeaton, editors. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, July 2003.

[8] W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, September 2001.

[9] Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[10] N. Fuhr and K. Großjohann. XIRQL: A query language for information retrieval in XML documents. In Croft et al. [8], pages 172–180.

[11] N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors. *Advances in XML Information Retrieval and Evaluation, Proceedings of the 4th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005), Revised Selected Papers*, volume 3977 of *Lecture Notes in Computer Science*. Springer, March 2006.

[12] N. Fuhr, M. Lalmas, and A. Trotman, editors. *Preproceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, December 2007.

[13] N. Fuhr and T. Rölleke. A probabilistic relational algebra for the integration of information retrieval and database systems. *ACM Transactions on Information Systems*, 15(1):32–66, 1997.

[14] A. Goodrum and A. Spink. Image searching on the Excite Web search engine. *Information Processing and Management*, 37(2):295–311, 2001.

[15] V. Harmandas, M. Sanderson, and M. D. Dunlop. Image retrieval by hypertext links. In N. J. Belkin, A. D. Narasimhalu, P. Willett, and W. Hersh, editors, *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 296–303. ACM Press, July 1997.

[16] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In C. Nikolaou and C. Stephanidis, editors, *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL 1998)*, volume 1513 of *Lecture Notes in Computer Science*, pages 569–584. Springer, September 1998.

[17] D. Hiemstra and A. P. de Vries. Relating the new language models of information retrieval to the traditional retrieval models. Technical Report TR-CTIT-00-09, University of Twente, CTIT, The Netherlands, May 2000.

[18] D. Hiemstra and W. Kraaij. Twenty-One at TREC7: Ad-hoc and cross-language track. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of 7th Text Retrieval Conference (TREC-7)*, pages 227–238. NIST Special Publication 500-242., 1998.

[19] D. Hiemstra and W. Kraaij. A language-modelling approach to TREC. In E. M. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, Digital Libraries and Electronic Publishing, chapter 16, pages 373–395. MIT Press, 2005.

[20] D. Hiemstra, H. Rode, R. van Os, and J. Flokstra. PF/Tijah: text search in an XML database system. In M. Beigbeder, W. Buntine, and W. G. Yee, editors, *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR) (held in conjunction with SIGIR 2006)*, pages 12–17, August 2006.

[21] T. Ianeva, L. Boldareva, D. Hiemstra, T. Westerveld, R. Cornacchia, and A. P. de Vries. Probabilistic approaches to video retrieval. In *TREC Video Retrieval Evaluation Online Proceedings*, 2004.

[22] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: a study of user queries on the Web. *SIGIR Forum*, 32(1):5–17, 1998.

[23] B. J. Jansen, A. Spink, and J. Pedersen. An analysis of multimedia searching on AltaVista. In N. Sebe, M. S. Lew, and C. Djeraba, editors, *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval (MIR 2003)*, pages 186–192. ACM Press, November 2003.

[24] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing and Management*, 36(2):207–227, 2000.

[25] C. Jörgensen and P. Jörgensen. Image querying by image professionals. *Journal of the American Society for Information Science and Technology*, 56(12):1346–1359, 2005.

[26] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. Length normalization in XML retrieval. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–87. ACM Press, July 2004.

[27] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[28] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In K. Järvelin, M. Beaulie, R. Baeza-Yates, and S. H. Myaeng, editors, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, August 2002.

[29] J. D. Lafferty and C.-X. Zhai. Document language models, query models, and risk minimization for information retrieval. In Croft et al. [8], pages 111–119.

[30] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Tranactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.

[31] J. List, V.Mihajlovic, G.Ramirez, A.P. de Vries, D. Hiemstra, and H.E. Blok. TIJAH: Embracing IR Methods in XML Databases. *Information Retrieval*, 8(4):547–570, 2005.

[32] V. Mihajlović, G. Ramírez, T. Westerveld, D. Hiemstra, H. E. Blok, and A. P. de Vries. TIJAH scratches INEX 2005: Vague element selection, image search, overlap, and relevance feedback. In Fuhr et al. [11], pages 72–87.

[33] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In F. Gey, M. Hearst, and R. Tong, editors,

Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 214–221. ACM Press, August 1999.

[34] P. Ogilvie and J. Callan. Language models and structured document retrieval. In N. Fuhr, N. Gövert, M. Lalmas, and G. Kazai, editors, *Proceedings of the 1st International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2002)*, December 2003.

[35] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[36] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281. ACM Press, August 1998.

[37] R. Rosenfeld. Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278, 2000.

[38] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In H. P. Frei, D. Harman, P. Schaüble, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29. ACM Press, August 1996.

[39] C. G. M. Snoek, M. Worring, D. Koelma, and A. W. M. Smeulders. Learned lexicon-driven interactive video retrieval. In H. Sundaram, M. R. Naphade, J. R. Smith, and Y. Rui, editors, *Proceedings of the 5th International Conference on Image and Video Retrieval (CIVR 2006)*, volume 4071 of *Lecture Notes in Computer Science*, pages 11–20. Springer, July 2006.

[40] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In K. Nahrstedt, M. Turk, Y. Rui, W. Klas, and K. Mayer-Patel, editors, *Proceedings of the 14th ACM International Conference on Multimedia*, pages 421–430. ACM Press, 2006.

[41] F. Song and W. B. Croft. A general language model for information retrieval. In S. Gauch, editor, *Proceedings of the 8th ACM International Conference on Information and Knowledge Management (CIKM 1999)*, pages 316–321. ACM Press, November 1999.

[42] A. Trotman and B. Sigurbjörnsson. Narrowed Extended XPath I (NEXI). In Fuhr et al. [11], pages 16–40.

[43] A. Trotman and B. Sigurbjörnsson. Nexi, now and next. In Fuhr et al. [11], pages 41–53.

[44] T. Tsikrika, P. Serdyukov, H. Rode, T. Westerveld, R. Aly, D. Hiemstra, and A. P. de Vries. Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah. In Fuhr et al. [12], pages 273–286.

[45] T. Tsikrika and T. Westerveld. Report on the INEX 2007 Multimedia Track. In Fuhr et al. [12], pages 410–422.

[46] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders. Robust scene categorization by learning image statistics in contexti (abstract). In T. Huang, J. Luo, and M. Naphade, editors, *International Workshop on Semantic Learning Applications in Multimedia (in association with the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - CVPR 2006)*, pages 105–105, June 2006.

[47] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, UK, 2nd edition, 1979.

[48] R. van Zwol, G. Kazai, and M. Lalmas. The INEX 2005 Multimedia Track. In Fuhr et al. [11], pages 497–510.

[49] N. Vasconcelos. *Bayesian models for visual information retrieval.* PhD thesis, Massachusetts Institute of Technology, 2000.

[50] E. M. Voorhees and L. P. Buckland, editors. *Proceedings of 11th Text Retrieval Conference (TREC-2002).* NIST Special Publication 500-251., 2002.

[51] T. Westerveld. *Using generative probabilistic models for multimedia retrieval.* PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2004.

[52] T. Westerveld, R. Cornacchia, A. P. de Vries, J. C. van Gemert, and D. Hiemstra. An integrated approach to text and image retrieval- the Lowlands team at TRECVID 2005. In *TREC Video Retrieval Evaluation Online Proceedings*, 2005.

[53] T. Westerveld and A. P. de Vries. Experimental result analysis for a generative probabilistic image retrieval model. In Clarke et al. [7], pages 135–142.

[54] T. Westerveld, A. P. de Vries, T. Ianeva, L. Boldareva, and D. Hiemstra. Combining information sources for video retrieval. In *TREC Video Retrieval Evaluation Online Proceedings*, 2003.

[55] T. Westerveld, A. P. de Vries, and G. Ramírez. Surface features in video retrieval. In M. Detyniecki, J. M. Jose, A. Nürnberger, and C. J. van Rijsbergen, editors, *Adaptive Multimedia Retrieval: User, Context, and Feedback, Third*

International Workshop (AMR 2005) Revised Selected Papers, volume 3877 of *Lecture Notes in Computer Science*, pages 180–190. Springer, July 2006.

[56] T. Westerveld, A. P. de Vries, and A. van Ballegooij. CWI at the TREC 2002 Video Track. In Voorhees and Buckland [50].

[57] T. Westerveld and R. van Zwol. The INEX 2006 Multimedia Track. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *Comparative Evaluation of XML Information Retrieval Systems, Proceedings of the 5th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2006), Revised Selected Papers*, volume 4518 of *Lecture Notes in Computer Science*, pages 331–344. Springer, March 2007.

[58] C.-X. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In Croft et al. [8], pages 334–342.

[59] C.-X. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.

5 Appendix

5.1 INEX Multimedia MMimages topics

Table 9: INEX Multimedia 2006 MMimages topics

	Users searching for ...	**express their queries by ...**			
MMimages topics		text	example	concept	field
1	images	X			
2	images	X			
3	images	X			
4	images	X	X		
5	images	X	X		X
6	images	X	X		
7	images	X			
8	images	X	X	X	
9	images	X	X		
10	images	X	X		
11	images	X		X	
12	images	X			
13	images	X			
All 13 topics	**images**	**13**	**6**	**2**	**1**

Table 10: INEX Multimedia 2007 MMimages topics

	Users searching for ...	**express their queries by ...**			
MMimages topics		text	example	concept	field
23	images	X			
24	images	X			
25	images	X			
26	images	X			
27	images	X	X	X	
28	images	X			
29	images	X			
30	images	X			
31	images	X			
32	images	X	X		
33	images	X		X	
34	images	X		X	
35	images	X	X	X	
36	images	X			
37	images	X	X		
38	images	X			
39	images	X	X		
40	images	X	X		
41	images	X	X	X	
42	images	X		X	
All 20 topics	**images**	**20**	**7**	**6**	**0**

5.2 The PF/Tijah System

PF/Tijah, a research project run by the University of Twente, aims at creating a flexible environment for setting up search systems. It achieves that by including out-of-the-box solutions for common retrieval tasks, such as index creation (that also supports stemming and stopword removal) and retrieval in response to structured queries (where the ranking can be generated according to any of several retrieval models). Moreover, it maintains its versatility by being open to adaptations and extensions.

PF/Tijah is part of the open source release of MonetDB/XQuery (available at `http://www.sourceforge.net/projects/monetdb/`), which is being developed in cooperation with CWI, Amsterdam and the University of München. PF/Tijah combines database and information retrieval technologies by integrating the PathFinder (PF) XQuery compiler [5] with the Tijah XML information retrieval system [31]. This provides PF/Tijah with a number of unique features that distinguish it from most other open source information retrieval systems:

- It supports retrieval of arbitrary parts of XML documents, without requiring a definition at indexing time of what constitutes a document (or document field). A query can simply ask for any XML tag-name as the unit of retrieval without the need to re-index the collection.
- It allows complex scoring and ranking of the retrieved results by directly supporting the NEXI query language.
- It embeds NEXI queries as functions in the XQuery language, leading to ad hoc result presentation by means of its query language.
- It supports text search combined with traditional database querying.

The above characteristics also make PF/Tijah particularly suited for environments like INEX and TRECVID, where search systems need to handle highly structured XML collections with heterogenous content. Information on PF/Tijah, including usage examples, can be found at: `http://dbappl.cs.utwente.nl/pftijah/`.