Project no. 033104

MultiMatch

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

D1.1 – State of the Art

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

Organization Name of Lead Contractor for this Deliverable: ISTI-CNR

Final Version

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

Document Information

Deliverable number:	D1.1
Deliverable title:	State of the Art Report
Due date of deliverable:	October 2006
Actual date of deliverable:	December 2006
Main Author(s):	Editor & Introduction: Carol Peters, ISTI-CNR Section 2: Johan Oomen, BandG; Contributions from Alinari Section 3: Carl Ibbotson OCLC PICA; Contributions from WIND, UniGE, Alinari, ISTI-CNR Section 4: Neil Ireson, USFD Section 5: Jaap Kamps, UvA Section 6: Gareth Jones, DCU; Contributions from UniGE, UvA Section 7: Paul Clough, USFD
Participant(s):	All Partners
Workpackage:	1
Workpackage title:	User Requirements & Functional Specification
Workpackage leader:	UNED
Dissemination Level:	RE (Restricted)
Version:	Final
Keywords:	cultural heritage, metadata, digital asset management, search engines, multilingual indexing and retrieval, multimedia indexing and retrieval, information classification, information extraction, user interaction, interface design

Abstract

MultiMatch aims at complex, heterogeneous digital object retrieval and presentation. The development of the system implies addressing a number of significant research challenges in a multidisciplinary context. This report describes the state of the art in the relevant areas of research, thus specifying the scientific and technology baseline from which the consortium partners start. We identified six main areas: existing technology for cultural heritage; search engines; information extraction and classification; multilingual/multimedia indexing; multilingual/multimedia retrieval; user interaction and interface design. Each area is reviewed in a separate chapter. Our aim has been to provide a complete panorama of the actual state-of-the-art in the areas of interest to MultiMatch, covering as far as possible all relevant aspects. The report will be monitored and, if necessary, revised and/or updated on delivery of the first and second prototypes, with particular reference to the results obtained in the project.

Table of Contents

Executive Summary

The objective of MultiMatch is to develop a multilingual search engine specifically designed for access, organization and personalised presentation of cultural heritage information. The development of the system thus implies addressing a number of significant research challenges in a multidisciplinary context. R&D expertise is required in a diverse set of system- and user-oriented research areas. On the system side, these relate to focused Internet crawling, information extraction and analysis, multilingual information access and retrieval, multimedia complex object management, and interface design. On the user side, relevant areas include user profiling, metadata and ontology studies, user/system interaction, and user-centred interface design. The technology in these areas tends to develop rapidly. For this reason, it was decided to prepare a detailed state of the art report in the initial phases of the project.

This report thus describes the state of the art in the principal sectors of research covered by MultiMatch in order to establish the scientific and technology baseline from which the consortium partners start. We identified six main areas: existing technology for cultural heritage, search engines, classification and information extraction, multilingual/multimedia indexing, multilingual/multimedia retrieval, and user interaction and interface design. Each area is reviewed in a separate chapter.

Technology for Cultural Heritage

A wide range of technologies are used in the different domains that can be classified under the general heading of cultural heritage. We review those of most direct interest for MultiMatch: metadata and encoding standards, and digital asset management systems. Of particular importance for efficient search and retrieval are decisions regarding the most suitable metadata schema(s) and conceptual reference framework(s) and consequent problems of interoperability over collections. The project recognises that content providers typically do not apply the same data model and conceptual schemas. However, the schemas adopted for MultiMatch will have to contain all the elements needed to describe the cultural heritage objects within the domain of the project. This chapter thus focuses in particular on providing an overview of the technology and standards used in this area; a more in-depth description can be found in Deliverable 2.1 which provides a detailed analysis of metadata and ontologies in the cultural heritage domain.

Focussed Search Engines

A search engine can be defined as a tool designed to retrieve information stored in some system. In the last decade or so, the web search engine has become of particular relevance and prominence. These search engines allow users to request content from the World Wide Web that meets specific criteria by supplying a set of search terms, usually in the form of words or phrases. In this chapter, we briefly survey current search engine technology focusing on the areas of main interest to MultiMatch: domain-specific or vertical engines specialising in multimedia and multilingual search and retrieval. We also give particular examples on the basis of the partners' own direct experience.

Classification and Information Extraction

Classification (also known as categorisation) and information extraction are part of the knowledge discovery process, which attempts to find "interesting" patterns in data, i.e. those which reveal some underlying meaning (semantics). This process incorporates a number of other sub-processes, including information retrieval, topic-tracking, summarisation and visualisation. Recent work in these areas encompasses a wide array of media types, such as text, images, audio and video.

In this chapter, we investigate the techniques currently being adopted in these areas for text, image and video, also providing references to the best-known systems providing various degrees of information classification and extraction functionality.

Multilingual/Multimedia Indexing

This chapter describes the state-of-the-art in the indexing of cultural heritage documents in various languages and of various media types. The special characteristics of cultural heritage documents are

first described. General approaches to indexing currently being developed are then discussed and the specific approaches available for each different media type are presented. Open problems and challenges that are of most direct relevance to indexing cultural heritage documents as envisioned by the MultiMatch project are indicated.

Multilingual/Multimedia Information Retrieval

The need to expand the scope of research in information retrieval (IR) beyond English text has been recognised in the last 15 years. Increasing amounts of work have been conducted and reported which explore non-English IR, cross-language information retrieval, multilingual information retrieval, and multimedia information retrieval. This work has greatly increased understanding of the issues of multilingual and multimedia information retrieval and access. A range of techniques have been proposed, explored, evaluated and refined. However, the techniques are imperfect and many challenges remain to improve effectiveness and to extend the scope of retrieval tasks. For example, significant issues arise with respect to translation between search topics and documents for cross-language and multilingual information retrieval. For multimedia IR, there are still problems related to the definition of retrieval units, i.e. what should we look for in an image or video, and the accuracy with which features can be detected automatically once they have been defined.

This chapter first provides a brief review of the relevant details and indexing assumptions of monolingual, cross-language and multilingual text IR. It then introduces multimedia IR and highlights some relevant experimental work. The final section looks toward future applications and challenges.

User-centred Interaction and Interface Design

The interface acts as the intermediary between users of information retrieval (IR) systems and the search system. This section reports on studies of users' information seeking behaviour in order to provide informative insight into user interface design. The focus is on understanding the user needs in a dynamic multilingual search context, and identifying system functionalities that support those needs.

Areas of relevance to the MultiMatch interface design include enabling the retrieval of multimedia objects (text, images, video, and audio) and then determining the best way of allowing the user to access this information (i.e. results visualisation). The interface should be interactive and adapt to meet a user's changing information needs.

In considering interface design, an important first step is to examine functionalities currently provided by existing systems. Therefore, a brief summary of related systems and their features is provided. These include online museum collections, cultural heritage websites, multimedia search engines, and other systems designed by academic research projects. Innovative experimental approaches to aspects of interface design and results visualisation are also mentioned. Conducting such a survey provides an overview of current practice and provides a basis upon which MultiMatch can expand. By examining and testing a variety of designs with potential user groups, MultiMatch can endeavour to build an interactive, innovative interface that is first and foremost successful at meeting its users' needs.

Our aim in this document has been to provide a complete panorama of the actual state-of-the-art in the areas of interest to MultiMatch, covering as far as possible all relevant aspects. The report will be monitored and, if necessary, revised and/or updated on delivery of the first and second prototypes, with particular reference to the advances made and results obtained in the project.

1. Introduction

The objective of MultiMatch is to develop a multilingual search engine specifically designed for access, organization and personalised presentation of cultural heritage information. The development of the system thus implies addressing a number of significant research challenges in a multidisciplinary context. R&D expertise is required in a diverse set of system- and user-oriented research areas including, on the system side, focused Internet crawling, information extraction and analysis, multilingual information access and retrieval, multimedia complex object management, interface design, and, on the user side, user profiling, metadata and ontology studies, user/system interaction, interface design from the user perspective. The technology in these areas tends to develop rapidly. For this reason, and as part of the project activity, it was decided to prepare a detailed state of the art report in the initial phases of the project.

This report thus describes the state of the art in the principal sectors of research covered by MultiMatch in order to establish the scientific and technology baseline from which the consortium partners start. We identified six main areas: existing technology for cultural heritage; search engines; information extraction and classification; multilingual/multimedia indexing; multilingual/multimedia retrieval; user interaction and interface design. In this Introduction, we summarise briefly the importance of these areas for MultiMatch. In the rest of the report, each of these topics is discussed in detail. As is to be expected, there is some overlapping between the arguments treated in the different chapters. For example, the question of metadata is addressed in Chapters 3, 4 and 5; but in each case from a different perspective. Similarly, indexing of multi-media data is discussed in both Chapters 4 and 5, with the focus of Chapter 4 on indexing for the purposes of information extraction whereas Chapter 5 is interested in indexing for the purpose of information access. Chapters 3 and 7 both talk about search engines, but while Chapter 2 describes the different types of existing search engines, Chapter 7 discusses the users' expectations and how they can interact with the functionality provided by the engines.

Our aim has been to provide a complete panorama of the actual state-of-the-art in the areas of interest to MultiMatch, covering as far as possible all relevant aspects. The report will be monitored and, if necessary, revised and/or updated on delivery of the first and second prototypes, with particular reference to the results obtained in the project.

1.1 Technology for Cultural Heritage

A wide range of technologies are used in the different domains that can be classified under the general heading of cultural heritage. In this report we focus on those of most direct interest for MultiMatch: metadata and encoding standards, digital asset management systems, and means to provide interoperability between objects in distributed collections.

Of particular importance for the project activity are decisions regarding the most suitable metadata schema(s) and conceptual reference framework(s). CH documents generally have rich metadata associated with them which reflect the provenance of the particular object, even when syntactically coded in a uniform format, such as Dublin Core, RDF, OWL. Making sense of heterogeneous metadata is one of the greatest challenges for today's cultural heritages institutions.

MultiMatch acknowledges the fact that current and future content providers will typically not apply the same data model and metadata schema. However, the metadata schema for MultiMatch will have to contain all the elements needed to describe the cultural heritage objects within the domain or scope of this project. Core MultiMatch metadata will thus be extracted from the metadata describing the selected cultural heritage objects, and converted into the central metadata schema. The rest of the metadata, contained in the possibly rich descriptions provided, will be admitted to the semantic background information of MultiMatch. Thus making it possible:

- For the user to read the content of these metadata, when viewing the search results, and
- For the metadata provided to play a useful role in associative searching.

Further research will make clear whether one of the standard metadata schemas described in this report can fulfil all the requirements of MultiMatch.

The metadata schema adopted by MultiMatch will probably also require an integrated, shared ontology for the information accumulated by archives, libraries, museums as well as by the other identified sub-domains. This shared ontology will make it possible for all the collections that the participants in this domain hold, and attribute to the vision of a 'digital continuum' with unrestricted, sustainable and reliable digital access to Europe's cultural heritage. The project is currently studying the possible role of several controlled vocabularies already widely used by cultural heritage institutions and their eventual adaptation and integration for the purposes of MultiMatch.

1.2 Focussed Search Engines

The MultiMatch project aims at developing an advanced, domain-specific search engine, with the following innovative features:

i) it will be the first search engine effectively combining automatic classification and extraction techniques with semantic web compliant markup;

ii) it will consider complex user profiles and search scenarios;

iii) it will be able to search across language boundaries and across different media;

iv) last but not least, it will provide extensive, scientific evaluation of every search component, in a field which is dominated by the (mostly US) industry and is, therefore, opaque from a scientific perspective.

MultiMatch aims to offer "complex object retrieval" through a combination of focused crawling, and semantic enrichment that exploits the vast amounts of metadata available in the cultural heritage domain. A major contribution of MultiMatch in this area will be to integrate focused crawling techniques to recognise and handle multilingual content.

1.3 Information Extraction and Classification

MultiMatch will use large scale information extraction from documents to identify entities and their relations in large Web corpora. This will enable classification and clustering of documents according to their content. A range of classifications, as well as various links to reviews, experience reports, and general background knowledge, will be provided. Documents will be classified on the basis of diverse dimensions, such as topical, geographical, and temporal and with respect to genre (review, experience report, background knowledge). MultiMatch will address the issue of how technologies, derived from the emerging field of Semantic Web based automatic annotation, can interact with and profit from the use of lighter weight strategies such as focused crawling, classification and IR in order to perform efficient and effective IE on a large scale.

1.4 Multilingual/Multimedia Indexing

Instead of returning documents in isolation, MultiMatch will provide complex search results that put documents of various media types into context. For the indexing-end of MultiMatch, complex object retrieval generates special challenges. First, documents of various media types (text, audio, image, video, or mixed-content) and accompanying metadata will be indexed. Existing generic standards such as MPEG-7 cater for such a data model by incorporating multimedia content and metadata in a single semi-structured document. Currently emerging XML databases provide a general framework for complex object retrieval. The development of the MultiMatch system will be greatly facilitated by the availability of such "complex object" databases, permitting a more comprehensive and meticulous indexing of documents compliant with CH metadata standards and themes

1.5 Multilingual/Multimedia Information Retrieval

For many years information retrieval research concentrated primarily on English language text documents. However, recent years have seen a significant increase in research activity extension to information retrieval techniques for multimedia and multilingual document collections. Unfortunately,

so far, there has been little transfer of research advances to real world applications. MultiMatch aims at bridging this gap.

Multimedia data can be classified according to its constituent media streams: audio, visual and textual. Research in audio retrieval has largely been concentrated in speech retrieval (SR), where the key challenge is accurate automatic content recognition. Research in visual information retrieval (VIR) for images and video data streams has similarly been underway for over 10 years. Problems of VIR relate to both recognition of visual content and the definition of visual content for IR. Images and video key frames are most often indexed using low-level features such as colour and texture, or recognising named individuals or objects based on specific trained models. Research is now underway exploring the automatic recognition of shapes and their use in retrieval. The long-term challenges of visual retrieval cannot be overstated. Many multimedia data sources comprise a combination of audio and visual data with textual metadata labels. Thus multimedia IR often combines retrieval using these separate data sources.

Multilingual information retrieval (MLIR) has also become an established area of research in recent years. MLIR focuses on the problem of using a request in one language to retrieve documents from a collection in multiple different languages. MLIR also introduces the problem of how to select documents in languages for presentation to the user. A range of approaches have been introduced and explored in recent years.

The development of MultiMatch will require limitations of existing work in both areas to be addressed. A major challenge will be to merge results from queries on language-dependent (text, speech) and language-independent material (video, image). Retrieving documents from collections of mixed media also introduces problems of consistent ranking across the different media.

The CH material to be used in MultiMatch will have a high degree of heterogeneity covering many different topics, from a variety of different resources of differing linguistic forms as well as different media, and potentially published over a long period of time. Again, this introduces significant problems for high quality IR. For example, it has been demonstrated that using general translation resources for documents in a specific domain is much less effective than using ones specialised for this domain. A second key research problem for MultiMatch will be to identify the domain of requests and documents, and to build, and then to identify and exploit suitable translation resources for the domains within the CH collection. Documents will also be sourced in different media. MultiMatch will thus need to address significant issues of document selection arising from document language, media and topic.

1.6 User-centred Interaction and Interface Design

Although there has been huge progress, content-based information retrieval (e.g. video and image retrieval by visual content) still faces significant barriers when attempting to create truly effective and comprehensive retrieval with respect to the user's needs. Low-level features can be automatically extracted by analysing the audio and video stream, but human intervention is still needed to add high-level features (i.e. metadata). However, recent advances in the areas of information retrieval and information extraction make it possible to automatically associate concepts to objects when text is available. The need for human intervention to annotate material is thus reduced. The MultiMatch user interface will integrate automatic techniques for low level feature extraction and automatic concept classification.

A key research problem for MultiMatch will be enabling the user to adequately formulate their query using the language of their choice and specify both low-level and high-level multimedia features.

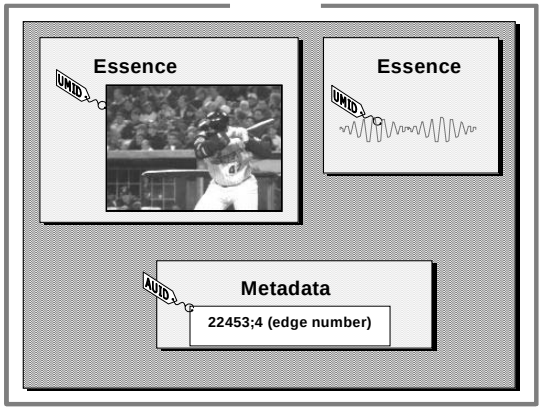
Accessing information through browsing has been demonstrated to be very effective in the domain of image retrieval. When image browsing is combined with text searching, users can chose their most preferred interaction mode and move between the two in a fluid way. MultiMatch intends to combine browsing and searching functionality in a multimedia context. The multimedia enriched ontology will be used to represent prototypical content.

MultiMatch will offer the user access to multimedia content through query, browse and navigation facilities. We will make use of insights gained from previous interface design and interaction studies for multimedia and multilingual IIR research.

2. Technology for Cultural Heritage

by Johan Oomen

Defining the scope of the Technology for Cultural Heritage is not easy as it can include a broad range of topics. The IST Support Measure DigiCult[1] has identified topics such as: Customer Relationship Management; Digital Asset Management Systems; Smart Labels and Smart Tags; Virtual Reality and Display Technologies; Human Interfaces; Games Technologies; The Application Service Model; The XML Family of Technologies; Cultural Agents and Avatars, Electronic Programming Guides and Personalisation; Mobile Access to Cultural Information Resources; Rights Management and Payment Technologies; Collaborative Mechanisms and Technologies; Open Source Software and Standards; Natural Language Processing; Information Retrieval; Location-Based Systems; Visualisation of Data; Telepresence, Haptics, Robotics; Digital Durability. See also Annex 1 to this Chapter.
Some of these technologies are touched upon in different sections of this report, others are not yet used in daily practice and are not closely connected to the work in MultiMatch.



In this Chapter, we decided to focus on the primary working processes:

2.1 state of the art in metadata standardization

2.2 state of the art encoding standards

Media-industry terminology generally describes content as consisting of essence (video, audio, data) plus metadata (descriptive information about the essence). An asset is thought of as content and its associated rights (who owns and controls the content). Assets are managed by specialized software. Hence,

2.3 state of the art in Digital Asset Management systems is also included.

Finally, we conclude with a discussion on open access to distributed collections and CH collection and record description linking in

2.4 approaches to interoperability.

2.1 Metadata Standards

Metadata refers to “data about data”, in other words, information that describes information sources or objects, e.g. a Dublin Core record or a record from the catalogue of an archive. The format and structure of metadata is often dictated in a set of rules, called metadata schema. If this schema is in use nationwide or internationally throughout a sub domain, we speak of a metadata standard. A metadata schema consists of several metadata elements. For some elements the input is free (e.g. Title), for other elements the input is guided or even restricted by controlled vocabularies of all kinds (e.g. thesaurus for subject keywords).

It needs to be said that in many cases, the metadata schema does not follow an international standard and is rather dictated by the internal work processes it needs to support. In these cases, interoperability between collections can only be accomplished by making mappings between schemas; or to a common schema, such as Dublin Core. The schema listed below only documents internationally applied metadata standards.

Another important element to take into account is the Semantic Web. The Semantic Web intent is to enhance the usability and usefulness of the Web and its interconnected resources. Within MultiMatch the use of a Semantic Web-compatible markup will guarantee a rich use (mainly in retrieval functionality) of the metadata on Cultural Heritage Objects provided by the partners in combination

[1] http://www.digicult.info/pages/index.php

with several ontologies related to the CH domain. A domain ontology (or domain-specific ontology) models a specific domain, or part of the world. An ontology on arts can be used to say, for instance, that "Picasso" is a "Painter", and that a "Painter" is an "Artist". The combination of such ontologies together with the MultiMatch indexes automatically provides the end user with several extra ways to navigate through the MultiMatch collection. E.g. this combination can present all CH objects from museums in Spain, without the need for the content providing partners to manually add extra metadata to the descriptions of their objects.[2]

2.1.1 Metadata Cataloguing Standards

In order to systematically study current practice we use the sub-domain definition advocated by DEN, the Dutch Institute for Cultural Heritage[3], and ePSINet[4], European Public Sector Information Network:

- Archives
- Libraries
- Museums
- Educational sector
- Audiovisual sector
- Geospatial sector

For each of these domains, the 'state of the art' (in this instance: the most widely used) in metadata standards and controlled vocabularies are discussed in this document.

In addition to this certain controlled vocabularies are particularly popular and have already been used in many European countries:

- Getty Arts and Architecture Thesaurus
- The UNESCO thesaurus
- Library of Congress Subject Headings (LCSH)
- The HEREIN thesaurus
- The NARCISSE vocabulary and the EROS project
- ICONCLASS (in the field of iconographic description).

	Schema	Controlled vocabularies
Archives	EAD and ISAD(G)	IPTC thesaurus, ISAAR (CPF), Thésaurus architecture et patrimoine, UK Archival Thesaurus
Libraries	FRBR, MARC, MODS and METS	DDC, UDC, LCSH and RAMEAU
Museums	CDWA, Object ID, VRA	AAT, ULAN, TGN
Educational sector	IEEE LOM	ERIC thesaurus
Audiovisual sector	P_META and SMEF-DM	-
Geospatial sector	CSDGM and ISO 19115:2003	-

[2] A more detailed discussion of knowledge representation in the cultural heritage domain, i.e. metadata schemas and controlled vocabularies, can be found in Deliverable 2.1, which is publicly available of the MultiMatch website at http://www.multimatch.eu/publications.html

[3] http://www.icn.nl

[4] http://www.epsigate.org/a.htm

2.1.2 State of the Art Generic Identification Standards and Reference Models

CIDOC Conceptual Reference Model

Name	CIDOC Conceptual Reference Model
Acronym	CIDOC CRM
Status / version	version 4.2, model recently became a standard
Type	ISO/PRF 21127 in May 2006
Management	International Council of Museums (ICOM)
Short description	The CIDOC Conceptual Reference Model is an ontology for cultural heritage information. It describes, in a formal language, the implicit and explicit concepts and relations used in cultural heritage documentation. The model is specifically meant to integrate and exchange heterogeneous sources of information on cultural heritage in the context of the Semantic Web. In other words, CIDOC CRM is a basis for data exchange and for building integrated query tools. The CIDOC CRM is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. It is intended to be a common language for domain experts and implementers to formulate requirements for information systems and to serve as a guide for good practice of conceptual modelling. In this way, it can provide the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives. The CRM is thought to be primarily a tool for the museum community, which intellectually originates in the museum community, but enables an effective communication with the libraries and archives world. Also applicable in the sub domains archaeology and the preservation of monuments and historic buildings. " The CRM can be regarded as a model of history in the physical sense, as perceived by humans. As such, it contains very abstract concepts. " [DELOS]

Simple Knowledge Organization System

Name	Simple Knowledge Organization System
Acronym	SKOS Core
Status / version	Draft 2, November 2005 Review proposals every 2-3 months: http://www.w3.org/2004/02/skos/core/proposals
Type	recommendation/standard
Management	W3C SWBPD-WG
Short description	SKOS Core provides a model for expressing the basic structure and content of concept schemes (or knowledge organization systems) such as thesauri, classification schemes, subject heading lists, taxonomies, 'folksonomies', other types of controlled vocabulary, and also concept schemes embedded in glossaries and terminologies. The SKOS Core Vocabulary is an application of the Resource Description Framework (RDF), that can be used to express a concept scheme as an RDF graph. Using RDF allows data to be linked to and/or merged with other data, enabling data sources to be distributed across the web, but still be meaningfully composed and integrated. SKOS can be seen as a supplement to OWL Web Ontology Language (the semantic mark-up language for publishing and sharing ontologies on the WWW; http://www.w3.org/2004/OWL/ Viewed 2006-09-27).

Resource Description Framework

Name	Resource Description Framework
Acronym	RDF
Status / version	10 February 2004
Type	standard
Management	W3C
Short description	Graphing theory (i.e. arcs and nodes)-influenced, XML syntax-based metalanguage for expressing metadata about web resources. Designed to convey metadata for machine consumption.
XML encoding available y/n	y

2.1.3 The Semantic Web

The Semantic Web (SW) aims to enable documents to contain computer-readable meaning (semantics) with the goal that documents should become computer-understandable. This is achieved via the interaction of a number of complimentary markup languages and processing tools.

Currently documents generally contain markup which facilitates the presentation of the contents in a human-readable form, i.e. font-types, positional information, etc. It is possible to infer some of the meaning behind the content, for example the summary of a document's contents can be assumed to be given by its title or in a section headed summary or abstract, given a table with headers labelled item and price, the rows can be assumed to provide the relative price for each specified item. Web scraping attempts to make use of such regularities in HTML documents to extract such information; with varying degrees of success. However the SW aims to make the meaning behind the contents of a page explicit. Thus an item would be represented in a standardised form (i.e. an item number) which might be linked to a repository giving further description and specification and each item would be linked to a given price (i.e. a numerical representation in Euro).

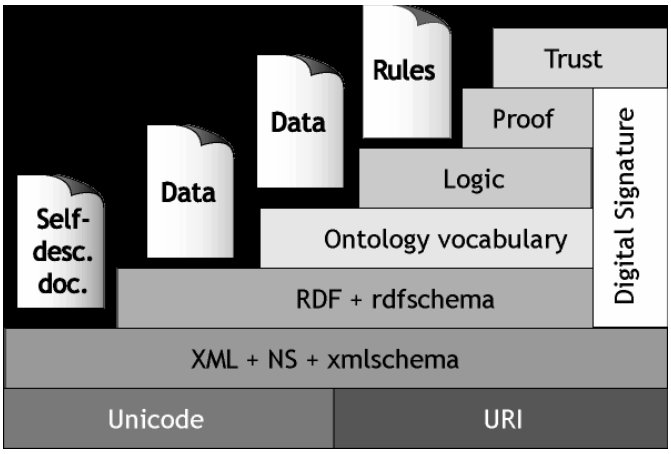
Within the SW, the contents of a document are marked-up using the Extensible Markup Language which provides the syntax for structuring the contents, with XML Schemas providing restrictions for the structuring of this syntax. This identifies the entities (resources) within a document, but does not impose a semantics on those entities.

A Resource Description Framework (RDF) model is used to provide a data model describing how the entities (resources) are related; in RDF each resource is described by a Uniform Resource Identifier (URI). RDF Schema (RDFS) is a vocabulary for describing groups of related resources and the relationships between them. RDFS uses resources to determine characteristics of other resources, such as the domains and ranges of properties. RDFS (or the more expressive Web Ontology Language (OWL)) is used to define an ontology which provides a conceptualisation of a given domain.

The Semantic Web in MultiMatch

As part of MultiMatch, documents within the Cultural Heritage domain, will be marked-up with semantic information (or metadata) from a common vocabulary. One criticism levelled at the SW is the cost associated with providing this markup; the project will examine the use of classification and information extraction techniques to alleviate this problem. The SW is also concerned with the interoperability between different vocabularies (and ontologies); an issue which will also have to be addressed within MultiMatch.

There are also issues which relate to the SW, such as "trust" and the provenance of information, privacy and censorship and the provision of Web services which, whilst not central, will be examined in the project.



- XML provides a surface syntax for structured documents, but imposes no semantic constraints on the meaning of these documents.
- XML Schema is a language for restricting the structure of XML documents.
- RDF is a simple data model for referring to objects ("resources") and how they are related. An RDF-based model can be represented in XML syntax.
- RDF Schema is a vocabulary for describing properties and classes of RDF resources, with a semantics for generalization-hierarchies of such properties and classes.
- OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

Whilst there is no specific aim to direct the results of the MultiMatch project towards providing material for the SW, there is an obvious relationship between the goals of MultiMatch and the SW. Much of the technology examined in MultiMatch will consider issues relevant to the development of the SW. Thus the project will both add to and benefit from SW technologies and research, and provide tools and materials which are exploitable in the context of the SW.

2.2 Encoding Standards

2.2.1 Audiovisual Encoding standards

In this paragraph, the most important high-resolution encoding standards are mentioned. We look at three families: SMPTE, MPEG, ITU-T. As an alternative Motion JPEG is listed. Further, audio file formats are briefly presented. The lower bitrate video compression schemes (MPEG-4 video (e.g., H.264, XviD and DivX), RealVideo, Windows Media Video) should not be used for the long-term storage of video material and are thus excluded from this discussion.

SMPTE

The Society of Motion Picture and Television Engineers is an international professional association, based in the United States of America, of engineers working in the motion imaging industries. An internationally-recognized standards developing organization, SMPTE has over 400 standards, Recommended Practices and Engineering Guidelines for television, motion pictures, digital cinema, audio and medical imaging.[5]

[5] http://www.smpte.org/smpte_store/standards/

SMPTE D10

The SMPTE D10 standard is currently most widely used by leading broadcast archives. This is the specification for a professional video format. SMPTE D10 is composed of MPEG Video 4:2:2 I-frame only and 8 channel AES3 audio streams. These AES3 audio usually contain 24 bit PCM audio samples. It is possible to find video bitrates of 30, 40 and 50 MBit/s. D10 is also called IMX by Sony.

The application of D10 is closely connected to the MFX format.

MXF is a "container" or "wrapper" format that supports a number of different streams of coded "essence", encoded with any of a variety of codecs, together with a metadata wrapper which describes the material contained within the MXF file. MXF has been designed to address a number of problems with non-professional formats. MXF has full timecode and metadata support, and is intended as a platform-agnostic stable standard for future professional video and audio applications.

MXF has been developed to essentially carry a subset of the Advanced Authoring Format (AAF) data model, under a policy known as the Zero Divergence Directive (ZDD). This enables MXF/AAF workflows between non-linear editing systems using AAF and cameras, servers, and other devices using MXF. MXF is in the process of evolving from standard to deployment. The breadth of the standard can lead to interoperability problems as vendors implement different parts of the standard.

SMPTE 421M (VC-1)

VC-1 is the informal name of the SMPTE 421M video codec standard. On April 3, 2006, SMPTE announced the formal release of the VC-1 standard as SMPTE 421M. Its most popular implementation is Windows Media Video 9.

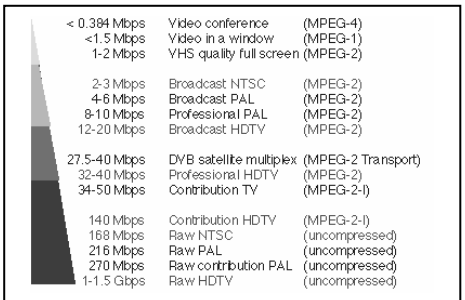
It is an evolution of the conventional DCT-based video codec design also found in H.261, H.263, MPEG-1, MPEG-2, and MPEG-4. It is widely characterized as an alternative to the latest ITU-T and MPEG video codec standard known as H.264/MPEG-4 AVC. VC-1 contains coding tools for interlaced video sequences as well as progressive encoding. The main goal of VC-1 development and standardization is to support the compression of interlaced content without first converting it to progressive, making it more attractive to broadcast and video industry professionals.

Although widely considered to be Microsoft's product, there are actually 15 companies in the VC-1 patent pool (as of 17 August 2006). As a SMPTE standard, VC-1 is open to implementation by anyone, although implementers are hypothetically required to pay hefty licensing fees to the MPEG LA, LLC. licensing body (or directly to its members who hold essential patents on the format, since it is a non-exclusive licensing body).

Both HD DVD and Blu-ray Disc have adopted VC-1 as a mandatory video standard, meaning their video playback devices will be capable of decoding and playing video-content compressed using VC-1. Windows Vista will partially support HD DVD playback by including the VC-1 decoder and related components needed for playback of VC-1 encoded HD DVD movies.

ISO: MPEG-2

MPEG is an acronym for Moving Picture Experts Group, a committee formed by the ISO (International Organization for Standardization) to develop this standard. MPEG was formed in 1988 to establish an international standard for the coded representation of moving pictures and associated audio on digital storage media.



Summary of MPEG Compression Capability

MPEG-2 is the designation for a group of coding and compression standards for Audio and Video (AV), agreed upon by MPEG, and published as the ISO/IEC 13818 international standard. MPEG-2 is typically used to encode audio and video for broadcast signals, including direct

broadcast satellite and Cable TV. MPEG-2, with some modifications, is also the coding format used by standard commercial DVD movies. Where software patentability is upheld, the use of MPEG-2 requires the payment of licensing fees to the patent holders via the MPEG Licensing Association.[6]

MPEG-2 includes a Systems part (part 1) that defines two distinct (but related) container formats. One is Transport Stream, which is designed to carry digital video and audio over somewhat-unreliable media. MPEG-2 Transport Stream is commonly used in broadcast applications, such as ATSC and DVB. MPEG-2 Systems also defines Program Stream, a container format that is designed for reasonably reliable media such as disks. MPEG-2 Program Stream is used in the DVD and SVCD standards.

The Video part (part 2) of MPEG-2 is similar to MPEG-1, but also provides support for interlaced video (the format used by analogue broadcast TV systems). MPEG-2 video is not optimized for low bit-rates (less than 1 Mbit/s), but outperforms MPEG-1 at 3 Mbit/s and above. All standards-conforming MPEG-2 Video decoders are fully capable of playing back MPEG-1 Video streams.

With some enhancements, MPEG-2 Video and Systems are also used in most HDTV transmission systems.

ITU-T: H.264

The ITU Telecommunication Standardization Sector (ITU-T) coordinates standards for telecommunications on behalf of the International Telecommunication Union (ITU) and is based in Geneva, Switzerland.

H.264, MPEG-4 Part 10, or AVC, for Advanced Video Coding, is a digital video codec standard which is noted for achieving very high data compression. It was written by the ITU-T Video Coding Experts Group (VCEG) together with the ISO/IEC Moving Picture Experts Group (MPEG) as the product of a collective partnership effort known as the Joint Video Team (JVT). The ITU-T H.264 standard and the ISO/IEC MPEG-4 Part 10 standard (formally, ISO/IEC 14496-10) are technically identical. The final drafting work on the first version of the standard was completed in May of 2003.[7]

Both of the major candidate next-generation DVD rival formats planned for product deployment in 2006 include the H.264/AVC High Profile as a mandatory player feature — specifically:

- The HD DVD format of the DVD Forum
- The Blu-ray Disc format of the Blu-ray Disc Association (BDA)

H.264 is a name related to the ITU-T line of H.26x video standards, while AVC relates to the ISO/IEC MPEG side of the partnership project that completed the work on the standard, after earlier development done in the ITU-T as a project called H.26L. It is usual to call the standard by H.264/AVC (or AVC/H.264 or H.264/MPEG-4 AVC or MPEG-4/H.264 AVC) to emphasize the common heritage. The name H.26L, harkening back to its ITU-T history, is far less common, but still used. Occasionally, it has also been referred to as "the JVT codec", in reference to the JVT organization that developed it. (Such partnership and multiple naming is not unprecedented, as the video codec standard known as MPEG-2 also arose from a partnership between MPEG and the ITU-T, and MPEG-2 video is also known in the ITU-T community as H.262.)

The intent of the H.264/AVC project was to create a standard that would be capable of providing good video quality at bit rates that are substantially lower (e.g., half or less) than what previous standards would need (e.g., relative to MPEG-2, H.263, or MPEG-4 Part 2), and to do so without so much of an increase in complexity as to make the design impractical (excessively expensive) to implement. An additional goal was to do this in a flexible way that would allow the standard to be applied to a very wide variety of applications (e.g., for both low and high bit rates, and low and high resolution video) and to work well on a very wide variety of networks and systems (e.g., for broadcast, DVD storage, RTP/IP packet networks, and ITU-T multimedia telephony systems).

[6] http://www.chiariglione.org/mpeg/standards/mpeg-2/mpeg-2.htm

[7] http://en.wikipedia.org/wiki/H.264

Alternative: Motion JPEG

Motion JPEG (M-JPEG) is an informal name for multimedia formats where each video frame or interlaced field of a digital video sequence is separately compressed as a JPEG image[8]. Unlike the video formats specified in international standards such as MPEG-2 and the format specified in the JPEG still-picture coding standard, there is no document that defines a single exact format that is universally recognized as a complete specification of "Motion JPEG" for use in all contexts.

Motion JPEG uses intraframe coding technology that is very similar in technology to the I-frame part of video coding standards such as MPEG-1 and MPEG-2, but does not use interframe prediction. The lack of use of interframe prediction results in a loss of compression capability, but eases video editing, since simple edits can be performed at any frame when all frames are I-frames. Video coding formats such as MPEG-2 can also be used in such an I-frame only fashion to provide similar compression capability and similar ease of editing features.

Using only intraframe coding technology also makes the degree of compression capability independent of the amount of motion in the scene, since temporal prediction is not being used. (Using temporal prediction can ordinarily substantially improve video compression capability, but makes the compression performance dependent on how well the motion compensation performs for the scene content.)

M-JPEG is frequently used in non-linear video editing systems. Reproduction of this format at full speed requires fast JPEG decoding capability.

Specialized audio file format: WAVE

WAV (or WAVE), short for WAVE form audio format, is a Microsoft and IBM audio file format standard for storing audio on PCs. It is a variant of the RIFF bitstream format method for storing data in "chunks", and thus also close to the IFF and the AIFF format used on Macintosh computers. It takes into account some peculiarities of the Intel processor such as little-endian byte order. The RIFF format acts as a "wrapper" for various audio compression codecs. It is the main format used on Windows systems for raw audio.[9]

Though a WAV file can hold audio compressed with any codec, by far the most common format is pulse-code modulation (PCM) audio data. Since PCM uses an uncompressed, lossless storage method which keeps all the samples of an audio track, professional users or audio experts may use the WAV format for maximum audio quality. WAV audio can also be edited and manipulated with relative ease using software.

As file sharing over the Internet has become popular, the WAV format has declined in popularity, primarily because uncompressed WAV files are quite large. More frequently, compressed but lossy formats such as MP3, Ogg Vorbis and Advanced Audio Coding are used to store and transfer audio, since their smaller file sizes allow for faster transfers over the Internet, and large collections of files consume only a conservative amount of disk space. There are also more efficient, lossless codecs available, such as Monkey's Audio, TTA, WavPack, FLAC, Shorten, Apple Lossless and WMA Lossless.

The WAV format is limited to files that are less than 2 GiB in size, due to the way its 32-bit file size header is read by most programs. Although this is equivalent to more than 3 hours of CD-quality audio (44.1 kHz, 16-bit stereo), it is sometimes necessary to go over this limit.

[8] http://www.siggraph.org/education/materials/HyperGraph/video/codecs/MJPEG.html and http//www.jpeg.org

[9] http://www.digiwik.org/wiki/index.php/WAV

2.2.2 Photograph and Page Oriented Encoding Standards

JPEG 2000

JPEG 2000 is a wavelet-based image compression standard. It was created by the Joint Photographic Experts Group committee[10] with the intention of superseding their original discrete cosine transform (DCT) based JPEG standard. Common filename extensions include .jp2 and .j2c, while the MIME type is image/jp2.

JPEG 2000 can operate at higher compression ratios without generating the characteristic 'blocky and blurry' artefacts of the original DCT-based JPEG standard. It also allows more sophisticated progressive downloads.

TIFF

The Tagged Image File Format (abbreviated TIFF)[11] is a file format for mainly storing images, including photographs and line art. Originally created by the company Aldus, jointly with Microsoft, for use with PostScript printing, TIFF is a popular format for high colour depth images, along with JPEG and PNG (portable network graphics).

The TIFF format is widely supported by image-manipulation applications such as Photoshop by Adobe, GIMP, Ulead PhotoImpact and Paint Shop Pro by Jasc, by desktop publishing and page layout applications, such as QuarkXPress and Adobe InDesign, and by scanning, faxing, word processing, optical character recognition, and other applications. Adobe Systems, which acquired the PageMaker publishing program from Aldus, now controls the TIFF specification.

Page oriented: Portable Document Format

Portable Document Format (PDF) is a file format proprietary to Adobe Systems for representing two-dimensional documents in a device independent and resolution independent fixed-layout document format. Each PDF file encapsulates a complete description of a 2D document (and, with the advent of Acrobat 3D, embedded 3D documents) that includes the text, fonts, images, and 2D vector graphics that compose the document. PDF files do not encode information that is specific to the application software, hardware, or operating system used to create or view the document. This feature ensures that a valid PDF will render exactly the same regardless of its origin or destination (but depending on font availability).

Anyone may create applications that read and write PDF files without having to pay royalties to Adobe Systems; Adobe holds a number of patents relating to the PDF format and claims that it is an open standard, licensing them on a royalty-free basis for use in developing software that complies with its PDF specification. PDF files are most appropriately used to encode the exact look of a document in a device-independent way. While the PDF format can describe very simple one page documents, it may also be used for many pages, complex documents that use a variety of different fonts, graphics, colours, and images.

The most recent PDF version also offers a rich metadata capability known as the Extensible Metadata Platform (XMP), which is based on the XML and Resource Description Framework (RDF) specifications of the World Wide Web Consortium (W3C).

Readers for many platforms are available, such as Xpdf, Foxit and Adobe's own Adobe Reader; there are also front-ends for many platforms to Ghostscript. PDF readers are generally free. There are many software options for creating PDFs, including the PDF printing capability built in to Mac OS X, the multi-platform OpenOffice, numerous PDF print drivers for Microsoft Windows, and Adobe Acrobat itself. There is also specialized software for editing PDF files.

Proper subsets of PDF have been, or are being, standardized under ISO for several constituencies. Within Cultural heritage, PDF/A is widely accepted. PDF/A is a constrained form of Adobe PDF

[10] http://www.jpeg.org/

[11] http://home.earthlink.net/~ritter/tiff/

version 1.4 intended to be suitable for long-term preservation of page-oriented documents for which PDF is already being used in practice. The proposed standard is being developed by an ISO working group with representatives from government, industry, and academia and active support from Adobe Systems Incorporated.[12] The official name is: ISO 19005-1. Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF (PDF/A) (the standard was published on October 1, 2005).[13]

Some advantages of a PDF archive over a TIFF or a paper-based archive are:

- PDF stores objects (e.g. text, graphics), allowing for an efficient full-text search in an entire archive. TIFF is a raster format and must first be scanned with an OCR (optical character recognition) engine.
- PDF files require only a fraction of the memory space of original or TIFF files, without loss of quality. The smaller file size is especially advantageous for electronic file transfers (FTP, e-mail attachment etc.)
- The PDF format can be optimized. The optimization can be focused on images (e.g. scanned checks) or extracting structured data (e.g. voucher information). TIFF treats all file information the same.

2.3 Digital Asset Management Systems

2.3.1 Use of DAMS in the Cultural Heritage Domain

Seamus Ross, Director of Humanities Computing and Information Management at the University of Glasgow, has written an influential position paper on the use of DAMS in the cultural heritage sector.[14] In it, he states:

Digital assets have the very unique characteristic of being both product and asset. Some digital assets exist only in digital form while others are created through the digitization of analogue materials such as text, still images, video and audio. Content has the same value to institutions as other assets such as facilities, products and knowhow.

Just as an organization seeks to make efficient and effective use of its financial, human and natural resources, it will now wish to use its digital assets to their full potential without reducing their value. Digital Asset Management Systems (DAMS) provide mechanisms to enable institutions to manage their digital resources. When associated with suitable policies, procedures and licensing arrangements, DAMS provide institutions with a way to facilitate the exploitation of their digital assets without depleting the value of the asset itself.

At a basic level Digital Asset Management systems use technology, such as commercial-off-the-shelf database management tools, to manage resources in ways that enable users to discover them and owners to track them. This may consist of either media catalogues with pointers to where the assets are stored or asset repositories, or a combination of both. These can be made accessible for use only in-house by staff in the content originating organization, for restricted use by others or made more widely available o specific communities or the public through online access.

Digital Asset Management involves the creation of a digital archive to hold resources, the provision of an infrastructure that will help to keep the entities from becoming obsolete, and a range of discovery and browsing tools to enable potential users to be able to identify, locate and retrieve the digital entities held by the DAMS.

A DAMS can serve a range of functions including:

[12] See also: http://www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfa.pdf

[13] http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml

[14] http://www.digicult.info/downloads/thematic_issue_2_021204_low_resolution.pdf#search=%22Digital%20asset%20management%20systems%20digicult%22

- providing support for content acquisition of born digital entities, and digitized materials such as text, still images, audio and video, and its cataloguing, management and storage can be enhanced through the use of DAMS;
- mechanisms to manage metadata associated with digital entities;
- a foundation for services to manage the delivery of digital content
- the foundation for the storing, managing and migrating of digital entities across time. These provide the basic building blocks for long-term digital preservation systems.

Generally, when we think of a DAMS we consider it as managing the entire process from acquisition (ingest) of a digital entity through its retrieval, delivery and use to its long-term archiving. Commercial off-the-shelf DAMS support (some of them are mentioned below) these functions, although not all with the same degree of sophistication. For example, some DAMS are better able to handle time-based media (e.g. audio and moving image material) than others. Off-the-shelf packages, although often expensive, represent a lower risk for most organizations than writing software from scratch. In addition, they benefit from having other users and a support network.

DAMS bring many advantages for heritage institutions. For example, they:

- support the centralisation of discovery and access;
- provide mechanisms to enable institutions to create coherent services from disparate projects
- enable mechanisms for tracking the authenticity and integrity of digital entities
- give organizations the ability to implement effective and easily manageable authorisation, security and tracking systems
- support the implementation of organization-wide mechanisms for managing intellectual property rights
- can generate savings by reducing the duplication of effort and resources
- produce time savings for the creators and users through organizational structure and centralization of digital resources;
- enable institutions to put in place asset browsing and querying tools
- provide organizations with the tools to monitor the types of entities they hold, how users discover
- and select entities, and what types or specific entities attract the most attention from users.

2.3.2 DAMS Vendors and their Products.

Below, eight leading vendors of DAMS are listed. There are other options available, but added together, these vendors cover a major part of asset management systems in use at cultural heritage organizations throughout the world.

An in-depth assessment of the technological details of a cost/benefit comparison would be excessive in the light of the MultiMatch project, as the project does not aim to develop a DAMS, it develops devices for online retrieval. This overview, however, does provide the state of the art in terms of products available for managing content, among which also cultural heritage objects.

IBM Content Manager

International Business Machines Corporation (IBM, or, colloquially, Big Blue) is a multinational computer technology corporation headquartered in Armonk, New York, USA. IBM manufactures and sells computer hardware, software, infrastructure services, hosting services, and consulting services in areas ranging from mainframe computers to nanotechnology. With almost 330,000 employees worldwide and revenues of $US91 billion annually (figures from 2005), IBM is the largest information technology company in the world, and holds more patents than any other technology company.

From the IMB website: “IBM DB2 Content Manager delivers a comprehensive content management solution in one easy-to-install, easy-to-manage package. Built on open standards, DB2 Content Manager Express delivers document management, production imaging and workflow capabilities and lets you share digitized content among diverse applications and across processes- securely and cost-effectively. IBM DB2 Content Manager has strong capabilities to store, organize, and easily access many types of information, enabling informed business decisions.”

Blue Order

Blue Order (based in Germany) supplies DAMS installations for many broadcasters. From their website: "Blue Order provides turnkey Media Asset Management (MAM) solutions, helping customers improve workflow efficiency, enhance product quality and generate new business opportunities. Media and Entertainment companies, Corporations and Public Institutions use Blue Order's Digital Asset Management product suite to support their entire digital production workflow incorporating live ingest and live logging to retrieval, browsing, editing and cataloguing of any digital audio-visual and multi-format content."

Their flagship product is Media Archive. “Media Archive professional is a complete, self-contained, easy to maintain media management solution, designed to support the management of assets in media cantered businesses such as, but not limited to, broadcasters, post houses, and advertising agencies. The strength of Media Archive professional is its elaborate support for audiovisual assets, i.e., video and audio objects. The unique feature set supports all major workflows within content creation, production, archiving and cataloguing. Being derived from Blue Order’s enterprise media management solution, Media Archive enterprise, this feature set has been designed together with major customers and has been applied in a large number of Media Asset Management projects world-wide.”[15]

EMC Documentum

EMC Corporation is an American manufacturer of software and systems for information management and storage. It is headquartered in Hopkinton, Massachusetts, USA. EMC produces a range of enterprise storage products, including hardware disk arrays and storage management software.

From the EMC website: “The EMC Documentum family helps you manage all types of content across multiple departments within a single repository. With a unified repository, various groups can easily share and reuse their content with other areas of the business that would benefit from access to this valuable information. Our product family also allows your business to share its content safely with outside organizations, including partners, vendors, and customers.” [16]

For managing Cultural Heritage collections, the Documentum Digital Asset Manager product is of most relevance. “EMC Documentum Digital Asset Manager exposes a set of powerful transformations and enhanced content previews, enabling companies to fully leverage the value of their digital assets. Users are provided with the complete set of content management capabilities offered by the Documentum platform, so organizations can manage all their content through one Web-based interface”.[17] Major customers include: Wolters Kluwer and Blue Star Print Group

Autonomy: VS Archive

Autonomy Corporation plc is an enterprise software company based in Cambridge, United Kingdom and San Francisco, USA. It develops a variety of knowledge management applications using adaptive pattern recognition techniques centred on mathematical Bayesian analysis. Autonomy's software is used by various large global corporations and public sector agencies. In 2002, Autonomy acquired

[15] http://www.blue-order.com/customers.html

[16] http://software.emc.com/products/product_family/documentum_family.htm

[17] http://software.emc.com/products/software_az/digital_asset_manager.htm?hlnav=T

Softsound, a company developing speech recognition software. In December 2005 Autonomy acquired Verity, one of its main competitors.

Autonomy acquired Virage two years ago. Virage has the 'VS Archive' DAMS in its product portfolio. From the Virage website: "As a leader in visionary Rich Media Management Technology, Virage offers a wide range of solutions which enable customers in all industries to maximize the value of their rich media assets. Virage's world leading technology provides advanced solutions for Enterprise Rich Media, Security and Surveillance, Video Search and IPTV and Broadcast environments."

VS Archive used by organizations such as Deutsche Bank and Boeing, is a content management solution to store, categorize, manage, retrieve and distribute audio, video and other rich media content fast and efficiently. The product suite not only streamlines the process of retrieving archived content for broadcasters but it is also invaluable for organizations such as intelligence agencies, educational establishments and corporations. These organizations hold a wealth of information in recorded lectures, interviews, presentations and broadcasts and frequently need to access this content for the purposes of training, marketing or investigation. VS Archive gives organizations the security of knowing that valuable assets will be preserved for the future.

Corbis Media Management

Corbis is a digital imaging/stock photography company founded by Bill Gates in 1989. Its headquarters are located in Seattle, Washington. Among other company operations, Corbis archives over 11,000,000 photographs and other media. In 2005, Corbis acquired eMotion, and henceforth created the "Corbis Media Management" product line[18].

From the Corbis website: "Corbis Media Management is a powerful set of tools for improving workflow and managing, distributing, deploying, and archiving digital assets of all kinds – photography, video, audio, presentations, logos, PDFs, sales materials and more. Corbis Media Management is the world's leading provider hosted applications for managing digital media content. Our solutions are used daily by users around the world to power business cases such as Digital Asset Management libraries, Marketing Extranets and Brand Portals. The core of Corbis Media Management's hosted technology platform is a sophisticated web-based digital asset management engine that helps you to manage and store your valuable digital media files, as well as get them into the hands of the people who need them, when they need them, in exactly the right format."

Open Text: Artesia TEAMS

Open Text Corporation is a Canadian high-tech company based in Waterloo, Ontario, Canada. It produces and distributes computer software applications designed to enable enterprise content management solutions for large corporate systems. Its flagship product, Livelink, is a Web-based content management system, with integrated business process management capabilities.

In 2004, Open Text acquired Artesia. The Artesia TEAMS DAM solution is based on the real-world workflows and requirements of over 150 customers from diverse industries and departments. Artesia's partnership with its customers results in features and functions that complement the way in which users work.

Artesia's Digital Asset Management solution does away with traditional process inefficiencies. It establishes a digital asset portal that serves as an organized storage hub for all of your digital publishing assets, no matter what stage of production or distribution, and no matter the location. Assets can include manuscripts, page layouts and their components, graphics, photographs, audio, video, as well as most file types. An easy to use web-interface provides secure access to all participants in the publishing process. These participants can use Artesia TEAMS to easily collaborate by securely accessing digital assets via the Internet, an extranet, or an intranet.[19]

[18] http://pro.corbis.com/creative/services/mediamanagement/default.aspx

[19] http://www.artesia.com/html/solutions_publishing.html

By centralizing storage and allowing efficient access to valuable digital assets, Artesia enables collaboration from a central location. To support this, it offers essential functionality like version control, usage tracking, and the capability to link rights and permissions information directly to the assets. In addition, Artesia is committed to, and develops according to, open industry-recognized standards. These standards contribute to more efficient production processes and a simplified exchange of digital assets. As such, Artesia's TEAMS DAM software can easily integrate with Quark and a variety of other authoring tools and delivery applications.

Oracle: Oracle Content DB

Oracle Corporation is one of the major companies developing database management systems, tools for database development, middle-tier software (Fusion Middleware), enterprise resource planning software (ERP), customer relationship management software (CRM) and supply chain planning (SCM) software. Oracle was founded in 1977, and has offices in more than 145 countries around the world. As of 2005, it employs over 50,000 worldwide.

From the Oracle Content DB white paper: "The content management market has changed. Oracle Content Database (Oracle Content DB) is the next generation of content management, based on the industry-leading Oracle Database. To control the rapid growth of unstructured content that typically makes up 80% of business information, organizations need solutions that enable enterprise-wide adoption. Oracle Content DB uniquely offers easy-to-use content management capabilities for true enterprise deployment. Oracle Content DB provides unmatched scalability, security, and availability of unstructured content in your Oracle Database. It includes a library of ready-to-use Web services to seamlessly integrate content management capabilities into the business processes and applications you use every day. With Oracle Content DB and Oracle Records DB, your organization can cut costs and improve productivity while reducing risk and enabling compliance."[20]

OCLC PICA

OCLC PICA is a library automation systems and services company which originated from a co-operation of the Dutch Pica foundation (Stichting Pica) and the U.S. non-profit library company OCLC Online Computer Library Center.

The portfolio of OCLC PICA includes a DAMS, called 'Digital Archive'. From the OCLC website: "OCLC's Digital Archive offers real-world solutions for the challenges of archiving and preservation in the virtual world. This flexible system allows you to archive assets in two ways. Use Web archiving for item-by-item harvesting and submission of Web pages and Web-based documents, or Batch archiving to submit your collections on various storage media for ingest and automated metadata creation at OCLC."[21]

OCLC PICA software is used by the Netherlands union catalogue, several German library consortia (including GBV, Hebis and SWB), the Australian national library, the French union catalogue SUDOC and many other libraries. Sisis and Fretwell Downing also have many notable customers in Germany, the UK and worldwide.

2.3.3 Open Source DAMS

Next to these commercial platforms listed above, several open source alternatives have risen to the surface in the past yeas: Dspace and Fedora are the leading ones.

DSpace

DSpace[22] is an open source software package which provides the tools for management of digital assets, and is commonly used as the basis for an institutional repository. It is also intended as a platform for digital preservation activities. Since its release in 2002, as a product of the HP-MIT

[20] http://www.oracle.com/technology/products/contentdb/pdf/contentdb-bus-whitepaper.pdf

[21] http://www.oclcpica.org/dasat/index.php?cid=100886&conid=0&sid=77544628ceba50ab09ab7041b4e7eb0a

[22] http://www.dspace.org/

Alliance, it has been installed and is in production at over 100 institutions around the globe, from large universities to small higher education colleges and research centres. It is shared under a BSD licence.

The first version of DSpace was released in November 2002, following a joint effort by developers from MIT and HP Labs in Cambridge, Massachusetts. Recently version 1.4 was released in July 2006.

DSpace is written in Java and JSP, using the Java Servlet Framework. It uses a relational database, and supports the use of PostgreSQL and Oracle. It makes its holdings available primarily via a web interface, but it also supports the OAI-PMH v2.0, and is capable of exporting METS (Metadata Encoding and Transmission Standard) packages also. Future versions are likely to see increasing use of web services, and changes to the User Interface layer. (text taken from Wikipedia)

Fedora

Fedora[23] (Flexible Extensible Digital Object Repository Architecture) (not to be confused with Fedora Core) is a modular architecture built on the principle that interoperability and extensibility is best achieved by the integration of data, interfaces, and mechanisms (i.e., executable programs) as clearly defined modules. Fedora is a digital asset management (DAM) architecture, upon which many types of digital library systems might be built. Fedora is the underlying architecture for a digital repository, and is not a complete management, indexing, discovery, and delivery application.

Fedora provides a general-purpose management layer for digital objects. Object management is based on content models that represent data objects (units of content) or collections of data objects. The objects contain linkages between datastreams (internally managed or external content files), metadata (inline or external), system metadata (including a PID – persistent identifier – that is unique to the repository), and behaviours that are themselves code objects that provide bindings or links to disseminators (software processes that can be used with the datastreams). Content models can be thought of as containers that give a useful shape to information poured into them; if the information fits the container, it can immediately be used in predefined ways. (text taken from Wikipedia)

2.4 Interoperability

Cultural heritage is distributed. Material is owned by different museums, galleries, picture libraries and so on, all over the world. There are all sorts of reasons for this:

- It may be due to where the objects were discovered.
- It may be due to who actually bought and collected works of art from the original artist.
- It may be due to the effect of wars and other political factors.
- There are general geographic issues - for example with large artifacts that are fixed in place, like buildings and archaeological dig sites.

Current technology can now overcome some of these issues. Much emphasis is currently placed on integrating local and scattered resources, whether in museums, libraries or archives.

There are several angles to this challenge of establishing interoperability. We discuss the two most common: ‘Supporting distributed networked information’ (linking collections), and semantic interoperability (linking record descriptions).

2.4.1 Supporting Distributed Networked Cultural Heritage Information

Cultural heritage institutions and photographic libraries are rich content resources, depicting people, objects, events, places and monuments. Making this material accessible requires rich metadata structures, able to capture the diversity of the media, the subject matter and the historical context around each information asset. This information tends to be 'locked away' in internal legacy systems, each with its own metadata format that has been designed to deal with a specific collection or set of objects. The first hurdle to be overcome is thus enabling networked information to be linked together. There are several approaches to this problem. Most common is the Open Archives Initiative Protocol

[23] http://www.fedora.info/

for Metadata Harvesting, that enables metadata to be harvested to a central index. Other approaches such as Z39:50 and SRU are also used.

Archives Initiative Protocol for Metadata Harvesting

The Open Archives Initiative (OAI, http://www.openarchives.org/) is an attempt to build a "low-barrier interoperability framework" for digital archives ("institutional repositories") containing digital content ("digital libraries"). It allows people (Service Providers) to harvest metadata (from Data Providers). This metadata is used to provide "value-added services", often by combining different data sets. Initially, the initiative has been involved in the development of a technological framework and interoperability standards specifically for enhancing access to e-print archives, in order to increase the availability of scholarly communication; OAI is, therefore, closely related to the Open Access movement. The developed technology and standards, though, are applicable in a much broader domain than scholarly publishing alone.

The OAI technical infrastructure, specified in the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), currently in version 2.0, defines a mechanism for data providers to expose their metadata. This protocol mandates that individual archives map their metadata to the Dublin Core, a simple and common metadata set for this purpose.

Commercial search engines have started using OAI-PMH to acquire more resources. Google has started to accept OAI-PMH as part of their Sitemap Protocol, and they are using OAI-PMH to harvest information from the National Library of Australia Digital Object Repository. In 2004, Yahoo! acquired content from OAIster (University of Michigan) that was obtained through metadata harvesting with OAI-PMH.

Examples of OAI directories and applications are:

- OAIster[24] is a project of the University of Michigan Digital Library Production Service. Their goal is to create a collection of previously difficult-to-access, academically-oriented digital resources that are easily searchable by anyone. 10 million records are currently harvested (December 2006)
- *Open*DOAR [25] is an authoritative directory of academic open access repositories. Each *Open*DOAR repository has been visited by project staff to check the information that is recorded here. This in-depth approach does not rely on automated analysis and gives a quality-controlled list of repositories.
- PictureAustralia[26] is an example of OAI-PMH in action. PictureAustralia harvests image data from Australian libraries, universities, museums and galleries. It then provides a single search system to access all the images.

Accessing remote databases: Z39.50 and SRU/SRW

Z39.50 is a client server protocol for searching and retrieving information from remote computer databases. It is covered by ANSI/NISO standard Z39.50, and ISO standard 23950. The standard's maintenance agency is the Library of Congress. Z39.50 is widely used in library environments and is often incorporated into integrated library systems and personal Bibliographic Reference software. Interlibrary catalogue searches for interlibrary loan are often implemented with Z39.50 queries.

In practice, however, the functional complexity is limited by uneven implementations by developers and commercial vendors. The syntax of Z39.50 is abstracted from the underlying database structure; for example, if the client specifies an author search (Use attribute 1003), it is up to the server to determine how to map that search to the indexes it has at hand. This allows Z39.50 queries to be formulated without having to know anything about the target database; but it also means that results

[24] http://oaister.umdl.umich.edu/

[25] www.opendoar.org/

[26] www.pictureaustralia.org/

for the same query can vary widely among different servers. One server may have an author index; another may use its index of personal names, whether they are authors or not; another may have no suitable index and fall back on its keyword index; and another may have no suitable index and return an error.

Z39.50 is a pre-Web technology, and various working groups are attempting to update it to fit better into the modern environment. These attempts fall under the designation ZING (Z39.50 International: Next Generation), and pursue various strategies. The most important are the twin protocols SRU/SRW[27], which drop the Z39.50 communications protocol (replacing it with HTTP) but attempt to preserve the benefits of the query syntax. SRU is REST based and enables queries to be expressed in URL query strings; SRW uses SOAP. Both expect search results to be returned as XML. Since these projects allow the relatively small market for library software to benefit from the web service tools developed for much larger markets, they have a much lower barrier to entry for developers than the original Z39.50 protocol.

The European Library[28] uses a combination of OAI and SRU.[29]

OAI and SRU combined

SRW and OAI clearly complement each other. Although the two protocols have chosen different answers to certain questions, this does not prevent them from being stacked up like building blocks into very different and interesting configurations. OAI's lower barrier to entry and specific goal make it easy to recommend for anyone to implement, whereas SRW is somewhat more complicated but aims to reproduce the essential functions of Z39.50 in facilitating distributed searching rather than harvesting.[30] [Sanderson, 2005]

Apart from the typical inverted pyramid metasearch model, there are also great benefits to be had from implementing OAI as a gateway interface to an SRW server. This progresses to having both protocols available and interlinked in the same server, such that records selected with a search can then be harvested at leisure. Not only can regular databases of records have value added to them by these protocols, the protocols can also be used to maintain registries. It is important to have service and collection description documents available so that appropriate routes to information can be taken, but also important are the internal identifiers within the protocols which could be usefully maintained in registries.

Other approaches: MICHAELplus

The MICHAELplus project has developed an electronic system to access, manage and update existing digital records of Europe's collections, including museum objects, archaeological and tourist sites, music and audiovisual archives, biographical materials, documents and manuscripts.

MICHAELplus culminates several progressive efforts under the eEurope Action Plan to harmonise EU Member States' programs to scan, photograph and otherwise enter cultural records into digital databases. This inventory, set up by public institutions, will use a distributed and Open Source platform suitable to be extended to any other country.

The technical results of the MICHAELplus project are:

- National inventories on a common meta-data model, data model and service model
- National portals running on a common open source technical platform, localized as necessary
- Trans-national inventory portal
- Sustainable, flexible extensible model based on XML technologies

[27] http://www.loc.gov/standards/sru/

[28] www.theeuropeanlibrary.org

[29] http://www.dlib.org/dlib/may06/vanveen/05vanveen.html

[30] http://www.dlib.org/dlib/february05/sanderson/02sanderson.html

- Open source solution built on Apache Tomcat, Cocoon, XtoGen, etc.
- Methodology and model, which is easy to deploy and replicate in additional countries.

2.4.2 Semantic Interoperability (at record level)

Methods described above enable searching though (remote) collections. Cultural-heritage collections are typically indexed with metadata derived from a range of different vocabularies, such as AAT, Iconclass and in-house standards. This presents a problem when one wants to use multiple collections in an interoperable way. In general, it is unrealistic to assume unification of vocabularies. Vocabularies have been developed in many sub-domains, each with their own emphasis and scope. Still, there is significant overlap between the vocabularies used for indexing.

To achieve the level of understanding usually implied by the term semantic interoperability requires the use of a knowledge representation language that is sufficiently expressive to describe all the nuances of meaning that are significant to the task at hand. Cultural heritage documentary systems do not use explicit formal meaning: the controlled vocabularies in use are not ontologies. As already mentioned in paragraph 2.1.3, the Semantic Web technology proposes solutions to the CH world regarding semantic interoperability for « Meta-Language » standardization: for metadata (using RDF triples) and metadata scheme/vocabulary (using ontologies in RDFS/OWL). The Semantic Web also provides alignment of description languages/points of view. [Isaac 2005]

The required level of expressiveness will require an ontology with at least the full power of first-order logic for many tasks, though for some restricted tasks a description logic (such as the one used in the OWL semantic web ontology language) having an expressiveness somewhat less than first order, will be adequate. Semantic representation techniques are thus playing a key role in this area to facilitate answers to this demand. In this sense it is foreseeable that also small and medium cultural institutions will be soon reached by the semantic technologies.

Human languages are highly expressive, but are considered too ambiguous to guarantee an accurate automatic interpretation, given the current level of human language technology. To achieve perfect semantic interoperability, all communicating systems must use term (or symbol) definitions that are identical or can be accurately interconverted. Thus a common ontology is the ideal situation for semantic interoperability. Where that is impossible, lesser degrees of semantic interoperability may be achieved by techniques that automatically map the definitions used by one system to those of another.

How to achieve semantic interoperability for more than a few restricted scenarios is currently a matter of research and discussion. Some form of agreed common ontology, at least one that is sufficiently high-level to provide the defining concepts for more specialized ontologies, is believed by some to be essential. But there is as yet no single ontology accepted and used by more than a small number of leading-edge research groups.

Whether use of a single high-level ontology can be avoided by sophisticated mapping techniques among independently developed ontologies is under investigation. No one upper ontology has yet gained widespread acceptance as a de facto standard. Different organizations are attempting to define standards for specific domains. For example, WordNet[31] is a semantic lexicon for the English language.

Below, we list three research project that investigate the issue of semantic interoperability in the cultural heritage domain; by:

- Adopting a reference model
- Executing ontology mappings
- Creating metadata crosswalks

This list is not aimed to be comprehensive, but rather illustrate the different research approached

[31] http://wordnet.princeton.edu/

Research approach 1: eCHASE and the CIDOC CRM

In the eCHASE project, the CIDOC Conceptual Reference Model is employed (a description can be found above). In particular the recent CRM Core proposal is being used as the common model for different multimedia collections.[32] CIDOC CRM has been in development over the last ten years by the museum documentation standards group CIDOC and is in the process of ISO standardisation. CIDOC CRM is becoming increasingly used in the cultural heritage domain. It is capable of modelling the complex objects and relations within its scope, and can be extended to cover many specializations. The eCHASE project is using CIDOC CRM as the common metadata schema to cover the different metadata repositories from their partners' collections. [Sinclair, 2005]

By mapping the metadata which exists in each collection to a common ontology, interoperability is achieved across diverse collections. This allows not only the unified access sought by users but also introduces new capabilities due to the preservation of the rich interrelationships between information.

Research approach 2: STITCH project and the Multimedia Annotation on the Semantic Web Task Force

The prime research objective of STITCH[33] (***Semantic Interoperability To access Cultural Heritage***) is to develop theory, methods and tools that allow metadata interoperability through semantic links between the vocabularies. This research challenge is similar to what is called the "ontology mapping" problem in Semantic Web research.

To create the semantic links between the different resources, the project turns to the existing research work in ontology mapping. Several authors have proposed mapping relations for use in semantic linking. These include equality, equivalence, subclass, instance and domain specific relations. The project will uses proposals as a starting point and extends/revises this set of mapping relations. Research on identification of links will first focus on baseline methods for manual specification of links such as developed within the MACS project. This will be supplemented with techniques from ontology learning targeted at finding such links automatically.

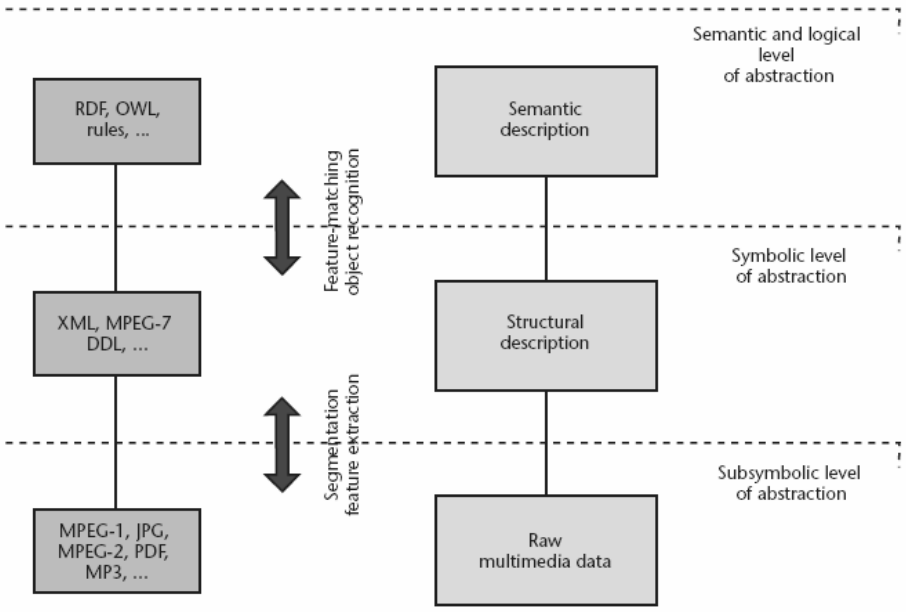
The state-of-the-art techniques are not full proof, so some form of human validation of the links will need to take place. This is not a big hurdle, as semantic links between vocabularies are a one-time thing. Another technique to consider is the generalization of existing annotations to semantic vocabulary links. For example, if according to a particular annotation the artist of a particular painting belongs to a certain art school, we may hypothesize that this link also exists for other works of the same artist.

The stack of RDF-based languages and technologies provided by the World Wide Web Consortium community is well suited to the formal, semantic descriptions of the terms in a multimedia document's annotation. However, because they lack the structural advantages of the XML-based approach and the work on multimedia document annotation already done within the framework of other standards, a combination of the existing standards seems to be the most promising path for multimedia document description in the near future. Therefore, The World Wide Web Consortium (W3C) has started a ***Multimedia Annotation on the Semantic Web Task Force***[34]

[32] http://eprints.ecs.soton.ac.uk/11567/01/echase.pdf

[33] http://www.cs.vu.nl/STITCH/

[34] http://www.w3.org/2001/sw/BestPractices/MM/



RDF = Resource Description Framework, OWL = Web Ontology Language, XML = Extensible Markup Language, and MPEG-7DDL = MPEG-7 Description Definition Language.[1]

The scope of this task force is illustrated by the image above. It shows the different levels of multimedia information and the type of annotation provided for each level. The subsymbolic abstraction level covers the raw multimedia information represented in well-known formats for video, image, audio, text,metadata, and so forth. These are typically binary formats, optimized for compression and streaming delivery. They are not necessarily well suited for further processing that uses, for example, the internal structure or other specific features of the media stream.

To address this issue, we can introduce a symbolic abstraction level, like the middle layer in the figure which provides this information. This is the MPEG-7 approach, which lets us use feature detectors' output, (multicue) segmentation algorithms, and so on to provide a structural layer on top of the binary media stream. Information on this level is typically serialized in XML. The standards that have been proposed and partly used in the literature for the representation of multimedia document descriptions (Dublin Core, MPEG-7, MPEG-21, Visual Resource Association [VRA], International Press Telecomunications Council [IPTC], and so on) mainly operate in this middle layer.

The problem with this structural approach is that the semantics of the information encoded in XML are only specified within each standard's framework (using that standard's structure and terminology). For example, if we use the MPEG-7 standard, then it is hard to reuse this data in environments that aren't based on MPEG-7 or to integrate non-MPEG metadata in an MPEG-7 application. This conflicts with the interoperability that is so crucial to Web-based applications.

To address this, we could simply replace the middle layer with another open one that has formal, machine-processable semantics by using a more appropriate, semantically enriched language like the Resource Description Framework (RDF). However, this would not take advantage of existing XML-based metadata, and more importantly, it ignores the advantages of an XML-based structural layer. Rather than changing the middle layer, a possible solution is to add a third layer (the logical abstraction level) that provides the semantics for the middle layer, actually defining mappings between the structured information sources and the domain's formal knowledge representation. An example of this is the Web Ontology Language (OWL). In this layer, we can make the implicit knowledge of the

multimedia document description explicit and reason with it—for example, to derive new knowledge not explicitly present in the middle layer. [Stamou, 2006]

Research approach 3: Getty Crosswalks

The Gerry Research Institute has produced charts that map several important metadata standards to one another, showing where they intersect and how their coverage differs[35]. Each of these standards can be said to represent a different "point of view" while Categories for the Description of Works of Art provides broad, encompassing guidelines for the information elements needed to describe an art object from a scholarly or research point of view, Object ID codifies the minimum set of data elements needed to protect or recover an object from theft and illicit traffic. The CIMI schema defines data elements for detailed museum information. The FDA guidelines focus on architectural documents, while the VRA Core Categories describe both the original work of art or architecture and its visual surrogate (the CDWA also includes data elements for visual surrogates; while VRA focuses on the surrogate, CDWA provides much richer, more detailed information for the original work). USMARC is a time-tested metadata standard used in the library world, while the Dublin Core metadata element set seeks to provide basic information elements to improve indexing and retrieval of resources on the Web.

Other cultural heritage metadata standards that are not included here are the AMICO (Art Museum Image Consortium) data dictionary, SPECTRUM, a standard developed for museums in the UK; the CIDOC Guidelines for Museum Object Information; and the International Council of Museums AFRICOM data standard, all of which map to Categories for the Description of Works of Art.

References

Isaac, Antoine. (2005). Accessing Cultural Heritage Collections Using Semantic Web Techniques. DE Conferentie, 2005. http://www.few.vu.nl/~aisaac/papers/Isaac-Talk-DE05.pdf

Sanderson, R., Young, J., & LeVan, R. (2005). SRW/U With OAI: Expected and Unexpected Synergies D-Lib Magazine 11(2)(Feb.2005)(http://www.dlib.org/dlib/february05/sanderson/02sanderson.html)

Sinclair, P., Lewis, P., Martinez, K., Addis, M., Prideaux, D., Fina, D. and Da Bormida, G. (2005) eCHASE: Sustainable Exploitation of Electronic Cultural Heritage (Poster). In Proceedings of 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, IEE Savoy Place.

Stamou, Giorgos, Jacco van Ossenbruggen, Jeff Pan and Guss Schreiber. (2006). Multimedia annotations on the semantic web. IEEE Multimedia, 13(1):86--90, January-March 2006.

[35] http://www.getty.edu/research/conducting_research/standards/intrometadata/crosswalks.html

Annex to Chapter 2: Technology Adaptation Assessment DigiCULT[36]

[The diagram was published in the end of 2004]

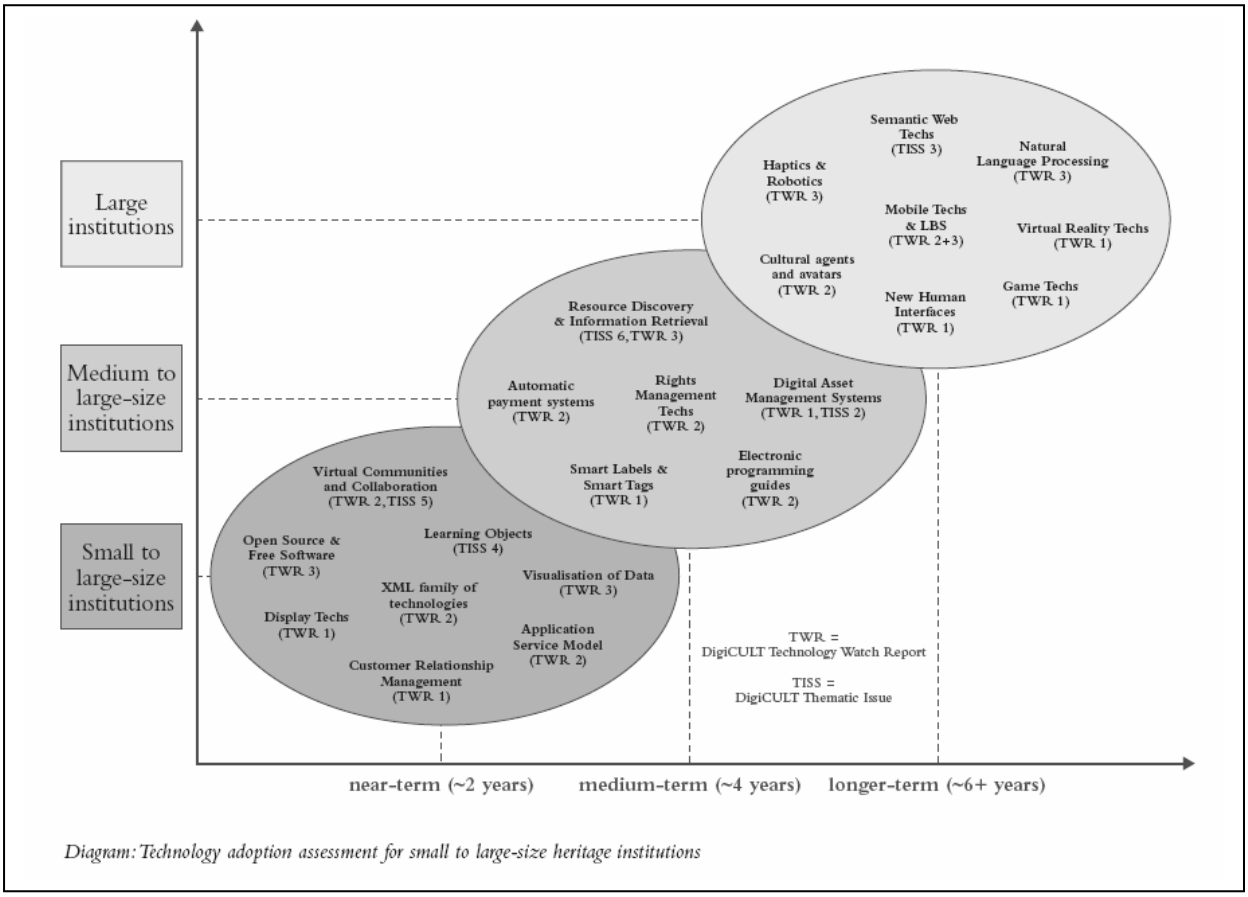


Diagram: Technology adoption assessment for small to large-size heritage institutions

What the diagram illustrates is the expectation that, over the next six years, only the large cultural heritage 'players' will adopt the latest group of technologies: Virtual Reality, cultural agents and avatars, new user interfaces (e.g. multimodal), games (e.g. multi-player online environments), haptics & robotics, mobile & location-based services, natural language processing, and Semantic Web technologies. These technologies will largely remain beyond the reach of small and medium-sized institutions. The initial investment for developing and implementing such state-of-the-art applications plus costs of running the application on a regular basis – the total cost of ownership (TCO) – are likely to be prohibitive for most institutions.

There may be scope for simple, low-cost Web based applications of games and virtual reality, but these are unlikely to become strong and longer-term attractions. As the diagram illustrates, small and medium-sized institutions will have to follow other strategies to attract on-site and online visitors such as virtual community projects, for example in regional history. Regarding management systems for digital assets, rights/licensing and payments, smaller institutions themselves will not find a business case as they do not, for example, hold an appropriate volume of marketable collection objects. However, such technologies may become relevant if smaller collections are digitized in the framework of a national or larger regional initiative, and the digital assets, rights and related transactions are then managed by digital heritage service centres. Thereby, collection metadata of smaller institutions could also be included in resource discovery networks, and some of their resources (e.g. photographs, postcards) may form highly valuable parts of Learning Objects in cultural and social history.

[36] http://www.digicult.info/downloads/dc_thematic_issue7.pdf

3. Vertical /Focussed Search Engines

by Carl Ibbotson with contributions from Marco Spadoni, Sam Minelli and Carol Peters

A search engine can simply be defined as a tool designed to retrieve information stored in some system. In the last decade or so, the web search engine has become of particular relevance and prominence, even an individual with the most modest of personal computer skills will be familiar with the search engines provided by Google[37] or Yahoo![38] These search engines allow users to request content from the World Wide Web that meets specific criteria by supplying a set of search terms, usually in the form of words or phrases. In this section, we briefly survey current search engine technology with particular focus on the areas of main interest to MultiMatch: domain-specific or vertical engines, engines specialised for multimedia and multilingual search and retrieval. We also give particular examples on the basis of the partners' own direct experience.

3.1 Generic Search Engines

All the major, current generic web search engines operate in a similar manner. General, broad-based engines aim to index as much of the World Wide Web as possible. They first crawl the web using automated software that follows every page link it finds. They then index and optimize this data into a database, and finally allow users of the search engine to submit queries to this optimized data.

A search results page is then returned to the user; this normally includes a list of web pages with titles, a link to the page and a short description showing where the keywords have matched the content. The popularity of Google's clean, unobtrusive interface and results page has influenced the design of other search engine interfaces, many of which look very similar.

3.1.1 Web Crawling

Due to the immense size of the World Wide Web, and limitations on both bandwidth and CPU time, crawling strategies become important. It has been noted that no search engine indexes more than 16% of the web[39] so choosing which pages to crawl, and when to crawl them are key decisions for a crawler.

Crawlers need to build a metric of importance for prioritizing pages on the Web. How this is done varies between providers. Often, crawling and indexing techniques and system architectures are guarded secrets, but all search engines employ some of the same basic methods. The importance of a page is a function of its perceived quality, and its popularity. Usually measured by how often the page is linked-to from other pages.

Due to the high rate of change of the Web, it is also crucial for a web crawler to sensibly determine how often to crawl a particular web resource. Typically, a crawler will employ a proportional update policy, meaning that pages that have previously demonstrated a high rate of change are generally crawled more often than pages that have shown a lower rate of change.

Large search engines such as Google, Yahoo! or MSN Live[40] have many thousands of machines positioned throughout the world that repeatedly crawl specific areas of the web, constantly providing new data to be indexed and stored. Web crawlers consume a huge amount of infrastructure and bandwidth, and are obviously expensive to run[41].

3.1.2 Indexing

Once web data has been crawled, it needs to be indexed. Different search engines do this in many different ways. Google, for example, indexes the entire page, or sometimes part of it, and often stores

[37] http://www.google.com
[38] http://uk.yahoo.com/
[39] http://www.nature.com/nature/journal/v400/n6740/abs/400107a0_fs.html
[40] http://www.live.com/
[41] http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1196112&isnumber=26907

additional meta-data about the page, such as titles and headings. AltaVista's indexing strategy involves storing every text word of the page being indexed.

How data is indexed is crucial. One of the most important elements of a search engine is the quality and relevance of the results it returns. When a user enters some search terms, the engine refers to its index of data to provide a result set. There will often be millions, maybe billions, of indexed pages containing the search terms. Returning the most useful and relevant pages to the user is often how search engines are evaluated, and each search engine provider handles ranking the result set in many different ways. Google uses its patented PageRank algorithm[42] to determine the relative importance of a particular document. It works by assigning a numerical weighting to every page it crawls, determined by how often the page is linked-to from other pages. From Google's own website:

> *PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."*

Google's PageRank algorithm, and their extensive infrastructure means their web search engine generates high quality, well-targeted search results, enabling them to gain huge popularity amongst Web users[43]. Accuracy and quality of results appears to be the quality that users value most in a search engine[44], and Google users believe Google has the most relevant results[45].

Other well-documented ranking algorithms such as Hilltop[46] and TrustRank[47] work on related principles. Hilltop gives additional ranking weight to 'expert' sites, those that are built around an individual topic, and therefore gives weight to pages that are linked to from this site. TrustRank gives additional ranking weight to 'trusted' sites, which are selected by hand. Ask.com uses an algorithm based on HITS, which presumes that a good hub is a document that points to many others, and a good authority is a document that many documents point to[48]. Hubs and authorities exhibit a *mutually reinforcing relationship*: a better hub points to many good authorities, and a better authority is pointed to by many good hubs.

Many other search engines have implemented their own page ranking systems, however the workings of such algorithms are often held as company-proprietary secrets to prevent misuse and copying.

3.1.3 Searching

Once data has been indexed, it can be searched by passing keyword searches to it. Traditionally this has involved simple keyword searches, which are directly matched up to indexed pages and meta-data. AltaVista was the first search engine to allow more advanced queries by allowing the user to use quotation marks to search for phrases, or mark some keywords as mandatory.

Ask.com was an attempt to allow the user to build queries, posed in the form of a natural language question. Ask.com has often being criticised for generating low accuracy search results when compared to other leading search engines with more sophisticated page ranking methodologies, and its popularity has wavered in recent years.

For particularly common user search terms, search engines do not build the result-set afresh each time. Instead the search engine builds the result set once, and periodically refreshes it.

---

[42] http://www.google.com/technology/
[43] http://searchenginewatch.com/showPage.html?page=2156451
[44] http://www.seobook.com/archives/001316.shtml
[45] http://www.internetretailer.com/article.asp?id=16570
[46] http://pagerank.suchmaschinen-doktor.de/hilltop.html
[47] http://pagerank.suchmaschinen-doktor.de/trustrank.html
[48] http://www2002.org/CDROM/refereed/643/node1.html

Additionally, most major search engines now offer their services though localised search engines For instance, on the Canada specific version of Google when a user searches for anything, the results will be of web sites with .ca domain extension

3.2 Vertical/ Focussed Search Engines

Vertical Search Engines work in a manner similar to the more broad-based search engines (such as Google and Yahoo!), however vertical search engine crawlers focus on highly refined pages and databases on the Web, and their indexes therefore contain more comprehensive information about specific topics in comparison to broad-based search engines.

Users of vertical search engines are often concerned only with results from a very specific niche (such as a medical database, or a job vacancy database), and are often unconcerned with the avalanche of data that accompanies a search performed on a broad-based search engine. For example, a Web user interested in buying a car would find far more relevant information from a niche search engine, such as Edmunds[49] than on google.com.

One of the problems of the more traditional, broad-based search engine is that the World Wide Web is growing at such an enormous rate, and pages are being updated so frequently that current search engine technology is struggling to continue to provide relevant, up-to-date result sets.

Additionally, a large part of the web remains impossible to index. The 'Deep Web' is a term, which describes sections of the Web that are not part of the 'surface web', and are therefore not able to be indexed. For example, dynamically generated web pages which act as search portals to specialised databases, or pages that are only accessible through dynamically generated links are considered to be in the 'Deep Web'. Because search engines can never link to these pages, they will never appear in search result sets. It is estimated that the Deep Web is several magnitudes larger than the surface web[50].

Vertical/Focussed search engines try hard to access the deep web by crawling it by subject category. Since traditional engines have difficulty crawling and indexing deep web pages and their content, deep web search engines like Alacra[51] (a business information search engine) create specialty engines by topic to search the deep web. Because these engines are narrow in their data focus, they are built to access specified deep web content by topic. These engines can search dynamic or password protected databases that are otherwise closed to search engines.

3.3 Media Targeted Search Engines

Using text-based search engines to retrieve multimedia content has been simple: Meta-data, or 'tags' are assigned to pieces of multimedia, allowing them to become searchable using standard techniques. For example, youtube.com allows users to upload their videos to the Web and share them with anyone. Before a user uploads their video, they would tag it with appropriate meta-data; for example if they upload a video clip of a boxing match, they may tag it with the words 'boxing', 'fight', 'punch', or whatever other words they considered relevant to the clip. Search engines would search only the Meta data, and treat it as simple text.

There are many web-based search multimedia search engines that serve multimedia content in these ways, Flickr.com, BBC Audio Search, WIND and Google Video are some examples.

IBM's Marvel[52], a image and video search engine, works on a similar principle, but takes it a step further. It has the ability to analyze multimedia content and automatically generate meta-data for that content by comparing it to a library of semantic models.

[49] http://www.edmunds.com
[50] http://www.press.umich.edu/jep/07-01/bergman.html
[51] http://www.alacra.com/
[52] http://domino.research.ibm.com/comm/research_projects.nsf/pages/marvel.index.html

3.3.1 Multimedia Search Engines

Under the heading of multimedia search engines, one should distinguish between search engines that retrieve multimedia data and those which accept multimedia queries. The first category describes engines that would return documents or pointers on documents of heterogeneous types understanding that the combination of their composing streams is an answer to the query (of any type). The second category is concerned with the form and formulation of the query. It may be interesting to formulate the query using different media. For example, this person (picture) saying something like this (audio and/or text).

While the distinction is interesting, search engines available in practice are of lower complexity. As mentioned above (Section 3.2) many search engines are focused on a single type of media and accept queries specific to that type. Queries are generally formulated using text. Text is not only the simplest media to manipulate and understand unambiguously, it is also the most accessible. A video search engine based on the query-by-example paradigm requires examples to be exhibited. These are not always easily accessible.

A number of search engines may still fall into our first category. These are generally information repositories where a navigation process has been enabled. This includes for example IMDB, the Internet Movie Database. Querying IMDB, one retrieves textual information (e.g. movie synopsis), video excerpts and summaries (trailers), pictures (making of) and structured information (actors, scenes, judgements). From there, Yahoo! Movies, and the INA TV archive can also be put into this category.

Most of the above relies either on structured manually created data (IMDB) or automatically inter-related data (Yahoo! Movies). Links are created over metadata, generally composed of text.

Looking at content-based search engines, all contributions essentially remain in the academic community as prototypes and applications rarely truly meets the general public. When doing so, functionalities are reduced and not engaged into a business process involving risk. This is the case for http://www.MyHeritage.com where one may find look-alike face picture of celebrities (“Find the Celebrity in You™”) or Retrievr (http://labs.systemone.at/retrievr/) which allows to query-by sketch in the Flickr image collection.

Looking at academic prototypes, we may non-exhaustively list Gift, Vicode and Web<img>Seek (Univ. of Geneva), Muvis (TUT), Ikona (INRIA), WebSeek (Columbia Univ.), MediaMill (Univ. Amsterdam), Fischlar (DCU), Informedia (CMU), or MARVEL (IBM). The list may be extended by citing almost all participants of the TRECVid benchmark (http://www-nlpir.nist.gov/projects/trecvid) who did develop their own multimodal retrieval systems. These search engines use truly multimodal content-based information to achieve the search process. All are based on low level signal processing (image/audio), language processing, machine learning and data-mining to infer semantic content (both from documents and queries), annotate documents and organize multimedia collections into comprehensive information structures. It is worth noting that the vast majority of these systems take benefit from user feedback and interactions to enhance their performance. However, their performance remains below large public needs (see the last TRECVid Evaluation: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html). Moreover, the “intelligent” strategies involved generally prevent such systems from very large-scale (i.e. planet-scale) applications.

WIND/Libero Image/Multimedia Search Engine

A commercial example of an Italian text-based multimedia search engine is provided by the Libero portal which offers users the possibility of searching among a fair amount of images, MP3 files and videos gathered from the Italian web.

The harvesting strategy is very simple: the engine scans the complete data base of html pages used to build indexes for Italian text search, and extracts links apparently pointing to images, MP3 files or videos. Candidate objects are then fetched, scanned for known magic numbers to make sure they really represent the kind of object the referrer declared, and digests are computed on their contents to help avoid duplicates.

Objects that pass the test are included in the database. Images and videos are further processed to extract fixed-size thumbnails (for videos only the few KB needed to extract some frames are really fetched from the net), and then thrown away to avoid copyright issues.

Indexing considers only textual information associated with every object:

- URL and title of the referring page
- Text surrounding the link
- "File name" of the URL representing the object (the complete path is frequently used for injecting spam)
- Contents of the ALT attribute (where applicable)
- Internal tags (author, title) for MP3 files

User queries are term based. Ranking of results takes into account matches in all indexed fields. In the case of MP3 files, the user may ask to sort results by date, in order to get information on the freshest addition to the web.

An image search service is provided by Google, and is very likely based on the same technology, i.e. search in indexed text "surrounding" the pictures and users perform term-based queries. Of course the coverage is much broader.

Technologies deployed for MultiMatch could improve significantly both engines by allowing content-based retrieval and clustering of results.

Alinari Search Engine

Alinari is currently developing intelligent search features in their site query functionality such as concept suggestion (similar to Google's 'perhaps you were looking for…') and keyword gender independence (male/female) and singular plural independence connected to RSS features: the user sets actively his personal preferences in a tutored context.

Table 3.1: Search features and market availability.

Alinari Search features		
Textual query	Keyword based query	Free input keyword
		Selection from a predefined list
		Selection from a predefined thesaurus
		Selection from a high level ontology
	Natural language based query	Annotation based query: exact term query
		Semantic based query
Visual query	Visual similitude	MPEG7 low level descriptors (see SCHEMA project)
		Statistical analysis
	Visual semantic query	Object detection (see MultiMatch project)
		Environment recognition
		Person detection
		Face recognition
		Mood detection
		Place recognition
		Historical period recognition
Human memory based search	Textual	Dictionary based suggestion (see Google-suggest: "perhaps you were looking for…")
		Semantic Textual suggestions
	Visual	Similitude visual suggestions
		Semantic visual suggestions

3.3.2 Future of Multimedia Searching

Several new types of multimedia search engine are beginning to surface. These are engines that actually search the content of multimedia files, rather than just the meta-data associated with it.

Podzinger[53] is a search engine for podcasts. It allows users to enter text based search terms, and then returns a list of podcasts containing those words. It works by using speech to text technology to convert the audio podcast into a text stream. This text stream can then be indexed and searched in a manner similar to standard search engine techniques. It is therefore possible to locate a podcast based on any single spoken word from the podcast, rather than just a limited set of Meta data tags associated with it.

Blinx has used a similar approach[54], but rather than searching podcasts, Blinx attempts to transcribe and search web-based TV channels. Its effectiveness appears questionable at the moment.

Retrievr[55] is a novel image search engine that features an interface allowing the user to sketch simple pictures, or upload images of their own. These are then matched against Flickr's database of images, and, in theory, similar images are displayed to the user. Retrievr's results would appear to be a little flaky at this stage in development.

Other similar types of multimedia search engine include tv-eyes[56] and singing- fish[57]. A discussion of multimedia and multilingual search interfaces is given in Chapter 7.

3.4 Multilingual Search Engines

Most of the search engines mentioned so far search by simply matching up input search words to indexed meta-data. Searching for "cat" for example will only ever match up exactly to the indexed phrase "cat". Most search engines would then prioritize their results to the locale of the user, but this is not a true multilingual search.

Yahoo!

Yahoo!France and Yahoo!Germany now provide a basic multilingual search functionality. You just have to activate the "Recherche multilingue" or "Suche Translator" option. Enter your query in your language and the search results will include not just the web pages written in your language, but also web pages written in other languages (French, English, German, Italian and Spanish). This functionality is currently available in a beta (testing) version and is not particularly intuitive to use; it is also not clear how the results are ranked and no option is provided for specifying in which languages the search should be performed. This functionality is of course highly relevant to MultiMatch and we will continue to monitor the developments in this service and investigate any changes and improvements.

Fotolia.com

At present, there are no major commercial search engines employing sophisticated cross-language retrieval functionalities. Fotolia.com is a stock photography database that offers the possibility of multilingual search. Using technology from Ultralingua, a company involved in producing translation software, the site search engine enables the retrieval of images whose metadata may be in a foreign language. Visitors enter a query in their native language, which is automatically translated and matched to image metadata in all languages. Therefore relevant results can be obtained, regardless of the language of the metadata (http://blog.fotolia.com/us/innovation/ultralingua.html). While this method is promising, when tested in practice it does not function perfectly. For example, typing in the

[53] http://www.podzinger.com/
[54] http://www.blinkx.tv/
[55] http://labs.systemone.at/retrievr
[56] http://www.tveyes.com/
[57] http://search.singingfish.com/

Italian query of "casa" results in a different (and smaller) set of results than typing in the English query "house." In theory, these two sets of results should be identical. This suggests that the area of cross-language search and information retrieval is still a domain in which further improvement can be made.

Quaero

Also of interest to MultiMatch is the activity of the Quaero project. Quaero was announced by Jacques Chirac during the French-German ministerial conference of Reims in April 2005, and was scheduled to be officially launched in early 2006 by the Agence de l'innovation industrielle. Close to 90 million Euros (110 million USD) from the governments of France and Germany will go towards development of Quaero. Quaero is mainly meant for multimedia search. The search engine will use techniques for recognizing, transcribing, indexing, and automatic translation of audiovisual documents and it will operate in several languages. There is also mention of automatic recognition and indexing of images.

According to some of the initial publicity and press releases Quaero will allow users to search using a "query image", not just a group of keywords. In a process known as "image mining", software that recognises shapes and colours will be used to look for and retrieve still images and video clips that contain images similar to the query image. (The software is supplied by LTU Technologies.) A technique called "keyword propagation" will be used so that when Quaero finds a descriptionless image which contains elements of or completely matches a properly labelled image, it will append the description from the labelled image to the unlabelled one. This will ensure faster searches and a definite enrichment of the web, also linguistically, as the primary interface and query terms will be in French and German.

However, despite the initial clamour, so far, the project is at a standstill; work is expected to begin in early 2007. MultiMatch is closely monitoring the developments of Quaero; we have already had some discussions with people involved and intend to invite representatives to our first workshops with the aim of promoting discussion and collaboration.

Research Prototypes

The development of multilingual search systems is still very much a research question and, as can be seen from the above, so far there has not been a lot of transfer of the research results into the application or commercial domains. An important source of literature with respect to the most recent research trends in this area is the website of the Cross Language Evaluation Forum (see www.clef-campaign.org). All the research institutions involved in MultiMatch are active collaborators in the CLEF activity.

Additional discussion of multilingual search systems can be found in Section 7.2.

3.5 Domain Targeted Search Engines

The aim of the service (http://arianna.libero.it/news/) is to collect, from a set of Internet newspapers and web magazines, all the published articles and to show them to the final end-user, grouping them either by category (Politics, Economics, Sports, etc.) or by “event”, i.e. grouping all the articles from different sources that are related with the same piece of news (including follow-ups).

The service is split in two main blocks:

- Data Management Service: an environment whose purpose is acquisition and management of news sources and retrieval and processing of articles. The environment can be thought of as a Web Service to which the data and their attributes are requested;
- Data Deployment Application: an environment whose purpose is querying the Data Management Service, and returning data to the final customer. The environment can be thought of as a Web application.

The most important stages of the pipeline constituting the DMS are:

- The Spider module, that repeatedly visits a list of news websites, several times a day, only retrieving relevant sections;

- The Extraction module, which is in charge of identifying and extracting interesting data (title, body, data, links to pictures) from unstructured pages. Identification is achieved by means of two orthogonal techniques:
    - Manually crafted per-site sets of regular expressions, built and validated through a web-based user interface, and applied at run-time;
    - Exploitation of anchor patterns in hub pages to address relevant data in the pointed leaf-pages (articles);
- The Categorization module which, after performing language normalization through a Natural Language Processing engine (tailored for the Italian language), associates each article with a category by means of self-updating Bayesian classifiers, initially trained on well known news sources;
- The Clustering module, in charge of grouping different articles dealing with the same event. This stage exploits the query-by-similarity functionality of the underlying full-text retrieval engine;
- The Indexer and Query Manager modules, build space- and time-efficient indexes and answering user queries.

End-users can make use of service data through the DDA environment in the following ways:

- By browsing static pages containing the most relevant news (in the service Home Page) or containing all the articles grouped by category (the articles Directory);
- By executing a standard, keyword-based query;
- By browsing a pictorial representation of the graph of the news, where nodes are entities (most frequently cited peoples, institutions, companies, cities etc.) and arcs are relationships witnessed by news articles mentioning (at least) two entities at the same time. A user click on a node redraws the graph centred around the selected entity, while a click on an arc returns all articles underlying the relationship between two entities.

When submitting queries, users can choose to sort returned articles either by reverse publication date, or by relevance. The relevance of an article is a function of

- Its affinity to the user query (standard keyword based scoring in title and body),
- Absolute score of the cluster to which the article belongs (a function of the number of articles in the cluster and spread of the cluster),
- Absolute score of the article (a function of the estimated precision of the categorization and importance of the site hosting the article)

A service very similar to Libero WebNews is Google News (e.g. http://news.google.it/). The features of the two services are very similar. However, whilst Libero WebNews currently provide the News Alert functionality only via RSS-feed (and not also via Email as Google News does), it is currently providing news from about 1180 news-sites, with respect to 250 sites claimed by Google News Italy.

3.6 Conclusions

Traditional search engines such as Google and Yahoo! are facing greater challenges as the World Wide Web grows faster than their indexing technology can keep up and popularity of the more focussed search engines is rising.

The publicly available multimedia search engines, which are of particular relevance to the MultiMatch project, currently offer varied levels of results. Podzinger's audio transcriptions appear to be of very high quality. Blinx's appear less so, and it's difficult to imagine using Retrievr in any practical way.

Most of the other Multimedia search engines still rely only on meta-data. IBM's Marvel intelligently generates its own meta-data after analyzing the media, but other search engines rely on a manual tagging process.

There are very few true multilingual search engines to compare. Fotolia.com appears to be one of the few, but its results appear inaccurate. Using the same search terms in different languages should

produce the same results, but searching for “cat” in different languages produces completely different result sets with little in common with each other.

The current state of both multimedia and multingual search still seems immature. Most multimedia searches rely on manually generated meta-data, and those which don’t have demonstrated a level of ineffectiveness. The very few multilingual services available are limited in effectiveness and not particularly user friendly. MultiMatch will be monitoring closely future developments in both areas.

4. Classification and Information Extraction

by Neil Ireson

Classification (also known as Categorisation) and Information Extraction are part of the Knowledge Discovery (KD) process, which attempts to find "interesting" patterns in data, i.e. those which reveal some underlying meaning (semantics). The KD process incorporates a number of other sub-processes including: Information Retrieval, Topic-tracking, Summarisation and Visualisation. KD was initially the focus of Data Mining research, where the data referred to that found in databases or spreadsheets, more recently, with the increase in computational resources and the availability of a mass of electronic media, the KD process encompasses a wider array of less structured media types, such as text, images, audio and video.

The Classification process allocates an object to one or more categories (or classes). Generally an object is viewed as a member of the category to which it is allocated, however in "fuzzy" or "rough" classification systems an object can also be a partial member of a category. Categories are generally used to contain objects which share a set of properties or attributes. Thus the classification process can be used to filter objects so that when a given category is selected, only objects with the desired properties are viewed or received.

The classification of media objects, such as text, images and videos, is the concern of library classification systems which organise the objects according to some predefined subject structure. For example, the most widely used library classification (taxonomic) systems, at least in the English speaking world, are the Library of Congress Classification and Dewey Decimal Classification systems. However the process of assigning (indexing) an object to a given category (or categories) in the classification is a laborious process involving careful consideration of the object's content. In addition such general classification schemes may not suit the requirements of the individual who wishes to identify and retrieve the classified objects. For specific domains or users alternative classification schemes may better suit their requirements and there may not be a ready mapping between the general and specific classification. Therefore research has focused on automatic approaches to facilitate the process of classification of objects according to their content.

Information extraction (IE) can be defined as the identification of specific instances of semantic elements (entities, events, relationships and their properties) within a given data object (i.e. a text or image). Thus IE can be viewed as the creation of an explicit structured representation (or metadata) from the information implicit in unstructured data. The IE task contrasts with the Information Retrieval (IR) as the result of IR is a sub-collection of objects, which are relevant to a given query; whilst the result of IE is a collection of facts extracted from the objects.

4.1 Pattern Recognition

Although there is a distinction between Classification and IE, IE can be considered as a classification process, the difference being that Classification is used to refer to the categorisation or labelling of an object as a whole, whilst IE refers to the categorisation, labelling or annotation of part of the object. In more general terms both Classification and IE can be considered as pattern recognition tasks; where a pattern is formed from features derived from an object. The recognition task maps (or classifies) a set of features onto a category, thus a media object (text, image or video), or part of that object, which exhibits a given pattern of features can be allocated to a semantic category, label or annotation. Categorisation, labelling and annotation can be considered to be synonymous processes, although annotation is generally seen as providing a more informative description than a simple label or category. Much research is devoted to the construction of automatic semantic annotation systems, due to the fact that manual annotation is a laborious task. This annotation task can be divided into three processes:

1. The processing of the media object to extract low-level feature descriptions.
2. Mapping between the low-level features and high-level of semantic concepts: the difference between these two descriptions of an object is referred to as the "Semantic Gap".
3. Understanding: moving from the annotation of a media object with a set of semantic concepts to a comprehension of the object as a whole (e.g. the narrative of the text or video, or the scene depicted by an image). Such the semantic interpretation may well depend upon the existence of (background) knowledge not contained within the media object.

The first process, feature extraction, is obviously dependant on the media type and will be discussed, below, in relation to each of the media types of interest to the MultiMatch Project (text, image and video. Pattern recognition is concerned with the second process, i.e. closing the Semantic Gap; the general (Machine Learning) approaches applied to the pattern recognition task will be discussed in the next section, with the specific applications in each of the sections on the media types. The third process, understanding, is beyond the scope of this document and the MultiMatch project.

4.2 Machine Learning

Most research into Classification and IE is concerned with the application of Machine Learning (ML) algorithms to the process of detecting classification patterns. The algorithms can be divided into three types, supervised, unsupervised and semi-supervised classification algorithms.

4.2.1 Supervised Classification

Supervised classification is based on the learning of a sequence of input/output pairs. It aims at producing the right result when it is given a new input. Supervised classification is achieved through the labelling of the data by a supervisor. When a new sample has to be added, it is labelled according to the already labelled data. The classification is based either on discrimination or on characterization. Discrimination consists in defining the frontiers between the already labelled data. New samples may then be added to the class they belong to by evaluating their position relatively to these frontiers. Characterization follows a different approach and intends to associate a set of invariants to each class. A new sample will belong to the class having the most similar properties. The following sections give a general introduction to the most widely used ML methods which have been employed in various Classification and IE tasks discussed below.

Decision Tree

The induction of decision trees was one on the original ML techniques developed and has been widely adopted due to its relatively simple implementation and transparency of the classification model. Most of the implementations are based around Quinlan's ID3 and C4.5 [Quinlan, 1993]. The algorithm iteratively partitions the example set according to the values of the most discriminative feature, i.e. the feature which provides the highest information gain.

Rule-based Models

Rule induction methods, unlike the global top-down approach of decision trees, develop a number of "if-then" type classifiers to cover the problem domain (represented by the training examples). These rules are not necessarily exhaustive (i.e. cover all the domain space) nor are they necessarily mutually exclusive (i.e. more than one rule can cover the same space). The "if" section of the rule determines the feature pattern, which constrains the rule coverage in the feature space; the "then" section determines the category to be associated with that part of the feature space. Rule induction algorithms attempt to create rules which are "consistent", i.e., do not cover any negative example and "complete", i.e. covers all positive examples. In practice the consistency and completeness constraints are relaxed to cope with uncertainty, imprecision and noise, in the problem domain and training examples. Rules are thus evaluated according to some measure based on their coverage and predictive accuracy, balancing the trade-off between generality (increased coverage) and accuracy (only covering positive examples).

To generate an individual rule most learner algorithms employ one of the following search strategies.

- Specialisation or top-down algorithms start from the most general rules and repeatedly specialise them imposing constraints in order to avoid covering negative examples.
- Generalisation or bottom-up algorithms start from the most specific rule that covers a given example; they then generalise the rule, relaxing its constraints to extend its coverage of without covering negative examples.

These learning strategies are attempting to generate rules which are either cases of Least General Generalisation or Most General Specialisation. There are other methods of rule induction such as using genetic algorithms [Holland, 1975] which cover the feature space then improve the rule set by combining "good" rules (using a crossover function) and performing local hill-climbing (using a mutation function).

One of the main attractions of rule-induction models is that (as with decision-trees) the model is human interpretable, i.e. that it is possible to determine the semantics behind the domain concept encapsulated by a rule.

Nearest-Neighbour

One of the simplest approaches to classification is to employ *nearest-neighbour classifiers*, also known as *memory-based* learning. The basic concept is to determine the distance between examples, thus an example with an unknown category can be assigned the category of its nearest neighbour, or more usually the most likely category given its K nearest neighbours. Obviously the complexity in the method is in determining distance function. The most straight-forward implementation use a standard Euclidian distance metric, however this assumes a very uniform problem space. More domain specific ML approaches can be applied to learning the appropriate feature weights or combinations. One of the principle difficulties with the application of nearest-neighbour learning is the prohibitive computational complexity when dealing with high dimensional feature spaces and large data sets. The Tilburg Memory-Based Learner (TiMBL), at http://ilk.uvt.nl/timbl/, provides the most widely used implementation of the approach.

Artificial Neural Networks (ANN)

ANN are based on an analogy to their biological counterpart, in the sense that they have simple processing nodes with a high degree of interconnection, processing involves the passing of simple scalar messages, learning occurs via the altering of weights which determine the interaction between nodes. The functioning of an ANN is determined by the topology of the network and learning algorithm applied to the adaptation of weights at the nodes. The most generally used topologies involve an input and output layer of nodes and one or more (hidden) layers. The topology of an ANN (the number of layers and nodes) determines its capacity, i.e. its ability to model a domain; however for complex domains the required capacity can cause difficulty in convergence of node weights.

Support Vector Machines (SVM)

SVM separate the problem domain space using hyperplanes; however one of the most appealing features of this approach is that as well as minimising the empirical error when dividing the example classes, the algorithm also positions the hyperplane such that it maximises the geometric margin between the proximate examples along the hyperplane. These examples are the support vectors; thus SVM are also known as maximum margin classifiers. An important development in SVM was to cope with non-linearity by employing a "kernel trick" [Boser, et al., 1992] which is used to transform the original feature space into a higher-dimensional space using a kernel function. Thus the hyperplane separation in the transformed space can represent a non-linear separation in the original space. The difficulty in implementation then becomes determining the appropriate kernel function for a given domain.

Hidden Markov Model (HMM)

A Markov Process is one in which a system stochastically changes from one state to another, in discrete steps. The change (transition) from the current state into the next state is dependent solely on the current state and not on any previous states. In a regular Markov model, the state is directly observable, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly observable, but variables influenced by the state are visible, thus the challenge is to determine the hidden parameters from the observable parameters. There are 3 canonical problems associated with HMMs:

1. To determine the probability of a particular state given the parameters of the model; solved by the forward-backward procedure.
2. To find the most likely sequence of hidden states that could have generated a given state given the parameters of the model; solved by the Viterbi algorithm [Viterbi, 1967]
3. To determine the parameters of the model (state transition probabilities), given a set of observed state sequences; solved by the Baum-Welch algorithm (a special case of the Expectation-Maximisation (EM) algorithm [Dempster et al., 1977]

One of the criticisms levelled at HMM is that in order to make the computations tractable an assumption of conditional independence between each discrete state has to be made (i.e. each state is independent of his ancestors and each observation depends only on current state). This may prove too restrictive for certain problem domains.

Maximum Entropy Model (MaxEnt)

Claude Shannon [Shannon, 1948] introduced the fundamental concept of entropy in information theory to measure the amount of uncertainty (or randomness) there is in a signal or event. MaxEnt modelling is used to determine the probability distribution which maximises the entropy given the known information (i.e. training examples). Applying MaxEnt involves constructing a stochastic model that accurately represents the behaviour of the "random" process by estimating the conditional probability that, given a context (set of features), the process will output a given result (category). The process involved in calculating the model is described in a number of relatively easily digestible tutorials [Berger, 1996; Ratnaparkhi, 1997].

The attractive feature of the MaxEnt model is that, given incomplete information is available (as is the case with IE tasks) inferences, derived from the probability distribution, are made solely on the available information.

Conditional Random Fields (CRF)

CRF can be viewed as a generalisation of the HMM and MaxEnt Model that aims to overcome the independence assumption drawbacks of HMM and the "Label Bias" problem exhibited by other Maximum Entropy Markov-based models. The Label Bias problem can be attributed to the local conditional modelling of each state, as states whose following-state distributions have low entropy will be preferred; despite these previous states possibly having no relation to the observations.

CRF is an undirected probabilistic graphical model where a node represents a discrete random variable, whose distribution is to be inferred, and an edge represents a dependency between the associated random variables. The distribution of each discrete random variable in the graph is conditioned on an input sequence provided by the feature space. A good introduction to CRF is provided by Hanna Wallach [2004].

Boosting

Boosting is a meta-learning approach for improving the accuracy of any given learning algorithm. The Boosting algorithm, which can be seen as a form of Probably Approximately Correct (PAC) learning [Schapire, 1990], iteratively combines (usually by using some majority voting method) weak classifiers (i.e. ones which are at least better than random) into a single accurate classifier. At each iteration the examples are weighted so that those incorrectly classified are "boosted" so that the new

weak classifiers focus on resolving the classification error. The most common boosting algorithm is AdaBoost [Freund and Schapire, 1999]

4.2.2 Unsupervised Classification (Clustering)

Perhaps the most problematic practical issue with using supervised classification systems is the need for a set of training examples; unsupervised classification systems remove the need for such a priori labelling of examples. An unsupervised classification process is only given a set of examples; these are then grouped (or clustered) according to the similarity and/or dissimilarity of their features. This process is also known as clustering and can be viewed as attempting to uncover the latent structure within a domain.

The principle issue in clustering is determining the appropriate distance metric to calculate the degree of similarity between two points in the feature space. Using the derived distance metric clustering generally involves minimising distances between examples within a cluster (intra-cluster variance) and maximising distance between examples in different clusters (inter-cluster variance). Clustering either exclusively allocates examples between clusters or examples can be partially or wholly members of one or more clusters, this is known as fuzzy clustering [Dunn, 1973]. One limitation of the most commonly used clustering algorithms is that either the number of clusters to be provided a priori, such as in the k-means algorithm [J. MacQueen, 1967] or the size of clusters has to be provided, as with QT (Quality Threshold) Clustering [Heyer et al, 1999].

One approach to removing the need for such a priori information is to use clustering techniques which place the clusters within a hierarchical structure. Hierarchical clustering can be either:

- Top-down beginning with a single cluster and splitting it to maximise some inter-cluster distance, and then continue splitting the clusters until there is one cluster per example.
- Bottom-up being with one cluster per example and combine the most similar cluster, and then continue to combine the most similar clusters until all examples are contained within a single cluster.

However such an approach is computationally expensive, especially when there is a large number of examples, n, to cluster as the complexity is in the order of $O(n^2)$.

A further issue with unsupervised learning is that although it does not require initial user input to create the classification; the output tends to require post classification operations in order to make the results meaningful, such as the allocation of labels or summaries, to the cluster, which is representative of their content.

4.2.3 Semi-supervised classification

Semi-supervised learning is a type of ML technique which makes use of a (typically small amount) of labelled data with a (typically large amount) of unlabelled data for training. Such methods use unlabeled data to either modify or give more weight to hypotheses deduced from the set of labelled data. Zhu [2005] provides a good review of the various approaches to semi-supervised learning.

Expectation-Maximisation (EM)

The goal of EM is to maximize the posterior probability of the model parameters (probability distribution means, standard deviations, and weights) given the data, in the presence of missing data, by applying the following process:

1. Initially estimate model parameters, generally based on some prior (domain) knowledge
2. a) Expectation (E) step: compute an expectation of the likelihood by including the missing (or latent) class variable as if it were observed.

   b) Maximisation (M) step: compute the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step.
3. Iterate step 2 by using the parameters calculated in the M step to initialise the E step and continue the process until a convergence threshold is satisfied.

The main concern when applying EM is avoiding convergence to local maxima. If the model convergences a local maximum, which is far from the global maximum, the use of unlabeled data is likely to have an adverse impact on learning. One proposed solution to alleviate this possibility is the selection of initial estimates using an active learning approach [Nigam, 2001].

Co-Training

In co-training [Blum and Mitchell, 1998], two classifiers are trained using disjoint features spaces. The features are divided into two class-conditionally independent sets, and a classifier is trained on the available labelled data, using each of the feature sets. Then those unlabelled examples for which one classifier is most confident in its prediction are labelled and added to the training set of the other classifier. The process is continued until some threshold level of accuracy on the training data is reached.

Expansion

Expansion is bootstrapping technique (i.e. one in which a process activates another process which serves the same purpose) which is related to query expansion from Information Retrieval, where terms are added to a query in an attempt to improve precision and recall. The process is initialised with a small set of labelled examples; from these, similar examples are found in the unlabelled data by expanding (relaxing) the feature values of the labelled examples. The similar examples are then assigned labels related to the associated labelled examples; these labels can be weighted according to the degree of similarity. The newly labelled data is then added to the training set and the process is repeated; with limits imposed on expansion to prevent making spurious inferences on examples too distant from the original labelled examples.

Active Learning

An active learning approach involves selecting the most appropriate sample of unlabelled data to label. The selection of the example can involve the use of a classifier to predict the labels on the unlabelled data to select the examples for which the classifier is most uncertain. Alternatively clustering techniques can be employed to select the most diverse set of examples. Unlike the other semi-supervised methods active learning then relies upon human intervention to label data, however the principle is to minimise the amount of data which needs labelling whilst maximising the quality of that data in term of building the classification model.

4.3 Text

4.3.1 Textual Data

Most of the research into Test Mining has come from the Natural Language Processing (NLP) domain which, for obvious reasons, has focused its attention on written text and transcribed speech. This is known as free or unstructured text, although there is, to a greater or lesser extent, a grammatical structure which can be exploited. Michelson and Knoblock [2005] have reported on some interesting work examining IE of unstructured and ungrammatical text.

Recently there has been more interest in the “mining” of semi-structured texts. In such texts the meaning is partially provided by the structure of the document in which the text appears. The documents may have titles, keywords or summaries and be divided into, possibly titled, sections. There might be internal or external (hypertext) references or text can be contained within tables. This type of document is exemplified by the HTML pages found on the internet, and the interest in being able to extract the information from the text on these pages is driven by the desire to exploit the potential of the billions of pages on the WWW.

Text Classification and IE systems generally presume that the input documents contain text from the domain of interest. However as well as the text providing a potential source of information to answer a given query, it may also contain noise the removal of which would improve the performance of the overall systems. This is often prevalent in web pages which may contain; adverts, menus, site-specific text and links, etc. which do not (directly) relate to the main content of the page. Being able to cleanly extract the relevant text has been highlighted as one of the key challenges for Web content mining [Liu and Chen-Chuan-Chang, 2004].

There are many factors which affect the interpretation of a piece of text, some of these are explicit and obvious such as its language (English, Russian, Japanese, etc.) or source (newspaper, journal article, audio transcript, web page, etc.). The text is also affected by the domain (art, sport, science, politics, etc.) to which it relates. The meaning will also be affected by the intention of the author; this may be to inform (news articles, user manuals, etc.), entertain (literature) or convince (argument, propaganda, marketing, etc.).

4.3.2 Text Analysis and Feature Extraction

The pre-processing of text to extract the relevant features is a necessary phase in all text mining techniques, to transform the text into a representation suitable for processing. Indeed there is often such a dependence on the application of specific pre-processing techniques that the distinction between the pre-processing and text mining technique is arbitrary.

Text Segmentation

The generic term "text segmentation" has analogies in analysis of other media type (i.e. images, video) in that it is a process which attempts to partition the data into coherent regions. For textual data, segmentation is used to refer to a number of different processes, the most basic being tokenisation where a text is partitioned into its atomic units; generally taken to be the word, term or token, although for certain applications (such as language or author identification) the text may be broken down to the character level. It is worth noting that although the process of tokenisation is considered to be a trivial task in Indo-European languages, the process is considerable more complex for Asian languages, such as Chinese, Japanese, Korean, Thai, Vietnamese, Mongolian, and Tibetan, where words cannot be fully identified by typographic features (e.g. spaces).

Similarly the text segmentation process of Sentence Boundary Detection is viewed as a trivial task in Indo-European languages; as boundaries are generally delineated using given characters, such as a full-stop or multiple newlines. The tokens and sentences derived from segmentation are used as input for further lexical and syntactic analysis (see below).

Another process in text segmentation relates to topic detection and tracking (TDT), this can be broadly divided into two forms; the detection of change-of-topic boundaries in a stream of text (such as speech transcripts or newswire feeds) and the partitioning of text into subtopics. Text classification, IE and indeed most other NLP techniques inherently rely on a notion of text documents, therefore the partitioning of a text stream into topic "documents" is a necessary precursor to the application of such techniques. Also the partitioning of long or complex documents into "sub-documents", each containing a coherent subtopic, can be of benefit to NLP techniques as it provides focused input and avoids information overload.

Research into TDT techniques can be divided into the generic machine learning areas of supervised and unsupervised learning. The performance of supervised learning techniques, as is generally the case with such approaches, is reliant on the amount and quality of learning material available, and tend to produce solutions which are not readily portable to other domains. Unsupervised techniques are more domain-independent, mainly relying on the concept of lexical coherence, i.e. topics can be differentiated by their distinct use of vocabulary. In addition to lexical coherence TDT techniques can also determine "cues" which mark the likely transition between topics.

Most of the work in this area has been based around the series of evaluation studies performed as part of the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program annually from 1998 to 2004 (see http://www.nist.gov/speech/tests/tdt/index.htm).

- *Semi-structured Documents*

The increasing use of the Internet as a means of communication has provided a large amount of machine readable XML/HTML documents which, as well as containing the text to communicate, contains structural information for the presentation of the text. This structural information can be used to segment the text into meaningful sub-sections [Luo et al., 2004]. This can be seen as an extension to the normal text segmentation process but with the use of HTML tags as "cues" for segment boundaries.

HTML documents, as well as providing additional information for segmentation, add a complexity over free text documents in that when the HTML is rendered the locality of text in the source HTML can be altered. As segmentation relies, to an extent, on the proximity of text to determine cohesion, the final presentation of the HTML must be considered. Thus can be done either by directly analysing the HTML code to extract its structure [Mukherjee et al., 2003], or by utilising the actual visual structure of the rendered HTML page [Kan, 2001; Yang, 2001; Gu et al., 2002].

Lexical Analysis

Lexical analysis provides an interpretation of the meaning behind individual words.

- *Part-Of-Speech (POS) Tagging*

POS tagging is the process of assigning grammatical classes to words in a sentence. The principal difficulty arises because some words can have multiple POS assignments depending upon their contextual use. Its importance stems from the fact that knowing the POS can be useful in subsequent text processing tasks; such as word-sense disambiguation and parsing.

- *Stemming and Lemmatization*

Both stemming and lemmatization attempt to find the base form of a given word (known as the "lexeme" for the word). Lemmatization is a more in-depth process which involves knowing the POS and may also require knowledge of the grammar. Stemming in contrast operates on a single word without knowledge of its context, and therefore cannot discriminate between words which have different meanings depending on POS. Therefore stemmers are less accurate than lemmatizers, they are however, easier to implement and faster. In most applications it is assumed that the use of a stemmer provides sufficient accuracy, however this may be more due to the fact that stemmers are available for a wide range of languages (see Snowball stemmer collection at http://www.snowball.tartarus.org/) and the difficulty in implementing a lemmatizer, rather than any strict empirical assessment of the cost/benefit of stemmers versus lemmatizers.

- *Word-Sense Disambiguation (WSD)*

WSD relates to the problem of "polysemy" where a word can have multiple meanings. For example, given the sentences, "the bank was breached by the water" and "she deposited her money in the bank", WSD determines whether "bank" refers to a river or financial bank. There are two main approaches to WSD; deep approaches and shallow approaches.

Deep approaches presume access to a comprehensive body of world knowledge. However these approaches are not very successful in practice, because of the difficulty in acquiring such knowledge in a computer-readable form (such as the Cyc project [Lenat, 1995], which is now OpenSource). Also there are many oddities introduced by the use of language, such as analogies and idioms, which may deliberately contradict the "proper" use.

Most WSD research focuses on shallow approaches which just consider a words context as defined by its surrounding words, i.e. river bank relates to water, fish, boats, etc. and financial bank relates to money, credit, manager, etc. These approaches define a window of N content words around each word to be disambiguated in the corpus, and statistically analysing those N surrounding words. Two shallow approaches used to train and then disambiguate are Naïve Bayes classifiers and decision lists. In recent research, kernel based methods such as support vector machines have shown superior performance in supervised learning. But over the last few years, there hasn't been any major improvement in performance of any of these methods.

It is instructive to compare the word sense disambiguation problem with the problem of part-of-speech tagging. Both involve disambiguating or tagging with words, be it with senses or parts of speech. However, algorithms used for one do not tend to work well for the other, mainly because the part of speech of a word is primarily determined by the immediately adjacent one to three words, whereas the sense of a word may be determined by words further away. The success rate for part-of-speech tagging algorithms is at present much higher than that for WSD, state-of-the art being around 95% accuracy or better, as compared to less than 75% accuracy in word sense disambiguation with supervised learning. These figures are typical for English, and may be very different from those for other languages.

- *Latent Semantic Indexing (LSI)*

The underlying idea behind LSI is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other [Landauer, 1998]. Thus LSI represents the meaning of a word as a kind of average of the meaning of all the passages in which it appears, and the meaning of a passage as a kind of average of the meaning of all the words it contains.

Syntactic Analysis

Syntactic Analysis is the study of the rules that govern how different words (categorised by their POS; nouns, adjectives, verbs, etc.) are combined into clauses, which, in turn, are combined into sentences. A sentence parsed in order to determine its grammatical structure with respect to a given formal grammar; this transforms input text into a data structure, usually a tree, which is suitable for further processing. Shallow parsing (or "chunking") is an analysis of a sentence which identifies the clauses (noun groups, verbs ...), but does not specify their internal structure, or their role in the main sentence. A frequent use of parsing in IE is to use the parse tree to extract the Subject-Verb-Object pattern from a sentence.

Use of Ontologies

The research on combining ontologies and IE involves both ontology building (generation and population) as an application of IE, and using ontologies to aid in the process of extracting information. In terms of aiding the IE process, given that a concept is present in a text, either because it has been annotated by a user or extracted by an IE system, ontologies can be used to provide "clues" to the other information which is likely to be in the text. Ontologies can also be used to disambiguate, as was mentioned above in WSD, for example given the text contains the word Paris, it is most likely to be a reference to the capital of France, unless the text also contains the geographical place name Texas in which case the ontological can be used to provide the information that Paris is a place in Texas, or if the page contains a Person who is a known celebrity then Paris is more likely to refer to "Paris Hilton", another celebrity. Of course the use of an ontology requires that an suitable and well-formed ontology exists and as was stated above, developing an ontology of reasonable size is an expensive task. However where such ontologies exist, such as in the biological domain, they have been found to be useful in providing information to text processing tasks [Honavar, 2001].

4.3.3 Text Classification (TC)

Text classification, that is the assignment of text documents to one or more categories based on their content, is an important component in many text analysis tasks such as; email "spam" filtering [Drucker et al., 1999], authorship attribution [Diederich et al., 2003], topic identification [Allan, et al., 1998] and (of specific interest to MultiMatch) Web page classification [Dumais and Chen, 2000]. However, much of the initial research into the use of ML for TC has been in the filtering of news stories, primarily because this was the first domain to provide a sizeable "Gold-Standard" corpus for training and evaluation of text classification systems [Lewis, 1997 and Lewis et al., 2004].

The automatic TC process involves: extracting the features from the text, selecting the most discriminating textual features (in its simplest form a set of keywords), allocating a weight to indicate the relative "importance" of the selected features in determining the semantics of the document (for supervised learning this is a measure of the degree to which a feature is indicative of a category) and define a similarity metric to determine the degree to which an object is assigned to a category (based on the combine the feature weights of an object). A good review of the ML approaches used for TC is provided by Sebastiani [Sebastiani, 1999 and 2002].

The feature extraction methods applied in TC tends to be relatively simplistic, in terms of applying the text analysis techniques described above. Textual documents are represented as a vector of terms (words) which are generally reduced to their lexeme (using stemming), and uninformative terms are removed using stop-word lists derived from large corpora (such as the Google stop-word list). However even such simple approaches are language specific. Attempts at applying state-of-the-art text analysis techniques (including parsing [Moschitti and Basili, 2004] and WSD [Kehagias et al., 2003])

have not shown substantial improvement in classification performance over the use of simpler representations.

Given a reasonably sized corpus the number of terms present in the vector representation of the text can be large (i.e. thousands of unique terms). For the application of ML techniques this can be problematic, thus dimensionality reduction (feature selection) methods are employed. The most commonly used approach for supervised learning systems is to select terms which are most indicative of a category; using measures such as Chi-square and Information Gain. An alternative is the use of Latent Semantic Indexing (LSI) to transform the original vector into a space with fewer dimensions [Liu, et al., 2004].

The weighting of the selected features (words or terms) to indicate their importance intuitively should be higher for those features that appear more often but are found in fewer documents. Thus the classic measure is given by the Term Frequency (TF) multiplied by the Inverse Document Frequency (IDF). The calculation of similarity between one document and another, or a document and a given category is determined by the co-occurrence of terms between the documents/category and the weight of those terms Salton and Buckley [1988] examine various approaches to term-weighting.

If sufficient training material is available for a given application domain then supervised ML techniques can be applied to feature selection and/or weighting, resulting in performance improvements over the use of the generic techniques described above. In TC a wide range of ML approaches have been applied including; nearest neighbour classifiers [Masand, et al., 1992], decision trees [Lam and Ho, 1998], Bayesian classifiers [McCallum and Nigam, 1998], Support Vector Machines [Joachims, 1998], rule learning algorithms [Cohen and Singer, 1996], neural networks [Li and Jain, 1998] and boosting [Schapire and Singer, 2000].

As has been stated for many applications a reasonable set of training data is too expensive to create so in order to overcome this document labelling bottleneck, semi-supervised methods have been applied [Nigam and Ghani, 2000; Nigam et al., 2000], however learning text classifiers from unlabelled data is still very much an active area of research.

The application of Text Clustering has tended to use the same basic techniques as text classification for feature extraction, selection, weighting and comparison. Although rather than measures being relative to the given categories, in clustering the measures relate to the categories constructed by the bottom-up or top-down clustering process. It is interesting to note that there has been some work which has shown that the addition of semantic information can aid the clustering process [Hotho, et al. 2003].

4.3.4 Information Extraction

Information Extraction from text, as a research field, has developed out of the more general field of Artificial Intelligence (AI) and more specifically from the area of knowledge representation. The mapping of natural language texts into more formal conceptual models originated with Roger Schank [1975] and Marvin Minsky's [1975] work in the 1970's. Schank's work formalised texts in terms of "scripts", where concepts within the text are interconnected by dependencies defined by a set of syntactic and semantic rules. Minksy developed a "frame" based representation where, each concept (entities, actions, events) is represented in a frame; the properties of the concept being represented as slots in the frame. The principal difference between to two forms of representation is that events and actions in scripts are ordered; i.e. represent procedural knowledge, whilst frames are linked into a tree or network structure, where a frame can be the value associated to another frame's slot. Such issues of knowledge representation are still important, and the goals of this original work are still fundamental to current research (i.e. the relationship between MUC "templates" and Minksy's frames). However recent IE research has become principally more concerned with the pragmatic process of acquisition rather than the representation of knowledge.

The overall IE process can be divided into a number of sub-tasks; named-entity recognition, coreference resolution, entity relation recognition and event recognition. The primary focus of research in IE has been the utilisation of Machine Learning (ML) techniques to aid in these tasks. The

following sections will outline the processes involved in each of the IE sub-tasks and discuss the key techniques applied to them.

IE Subtasks

- *Entity Extraction / Named Entity Recognition*

Entity Extraction or Named Entity Recognition (NER) is the identification of a term or phrase which refers to a specific entity. For example; a person or organization, place name, temporal expression, or certain types of numerical expression. Most of the research into IE has focused on the area of NER as it is the foundation of the other IE tasks; relation and event extraction.

The techniques employed in NER, to an extent, depend upon the entity to be extracted. Some entities, such as temporal expressions, have a relatively common representation and usage across domains. However other entities require more domain specific approaches, this is particularly true of Terminology Extraction, e.g. the extraction of protein or chemical names, which is an important sub-problem in NER. It is worth noting that the extraction of time expressions (TIMEX) is a significant area of NER research as the recognition of TIMEX is necessary for determining the temporal ordering, which is a fundamental task in event recognition. The work in this area has been stimulated by the availability of the 2004 ACE Temporal Expression Recognition and Normalization (TERN) corpus.

There are three basic approaches to identifying entities:

1. Gazetteers or Name lists

A look-up table which matches character strings with entities. Gazetteers work well for stable lists of names (such as days of the week, chemical elements, etc.) but are less useful where the list of names is constantly growing or changing. Even when the names are stable there is the problem of resolving ambiguous usage, for example Rose can be a flower, place name, persons name, colour, etc. There are however a growing number of useful resources being developed such as Getty Thesaurus of Geographic Names (TGN), which contains around 1.3 place names, and Union List of Artist Names (ULAN), which contains around 250000 artist names.

2. Orthography

Orthography NER considers the “internal” character pattern of an entity’s lexical representation. This works well for things like dates, phone numbers or postcodes which are readily recognized by their internal format (e.g., DD/MM/YY or chemical formulas). It is however not a technique generally applicable to the extraction of many entity types and thus is used in conjunction with the contextual pattern.

3. Contextual Patterns

Most of the work on NER has focused on the use of contextual patterns, where an entity is identified in the context of the surrounding terms. In the original MUC evaluations some of the best performing systems used hand-coded pattern rules using specific grammars (such as JAPE [Cunningham et al., 2002] which provide syntax for the creation of NER pattern rules. However the creation of such hand-coded rules requires a considerable amount of effort and, as with gazetteers, the performance of rules tends to be brittle when applied to domains with dynamical changing entity names or name usages. Therefore the majority of work has focused on alleviating the problems of determining contextual patterns for entity identification with the use of machine learning.

- *Coreference Recognition*

Coreference recognition finds multiple references to the same object within in a text. The coreferent objects can be expressed by; the same text, or in a modified version (i.e. James, Jamie, Dr J. Smith, etc.) or as pronouns and designators (“he treated the patient”, “The doctor called”). The references can occur both earlier (anaphoric references) or later (cataphoric references) in the text.

- *Entity Relation Extraction*

Relation Extraction identifies the occurrence, and type of relation between two entities, e.g. a person "is_located_at" a city, or gene "codes_for" a protein.

- *Event Extraction*

Event recognition extracts a collection of entities and relations which describe a single event. At the MUC conferences this task was referred to as template filling, while "Event Detection and Recognition" is the term adopted in the ACE program. The simplest approach is to assume that a given segment (sentence, passage or document) of text refers to a single event and fill the templates by combining entities and relations within that segment; resolving any of the co-reference between entities.

Supervised learning methods

The technology currently dominating IE is the supervised learning techniques. The basic approach is to formulate the IE problem as a pattern classification task; training the classification model on a set of pre-labelled positive and negative examples. The positive examples are provided by the labelled (or annotated) entities in the text, the negative examples are provided by the rest of the text. The ML systems can either develop models to identify entire entity in a text or to separately identify the positions defining the start and end of the entity. The pattern used to classify the examples is formed from the lexical, syntactic or semantic features derived from the text using the preprocessing techniques described above. In the training phase examples are extracted from a text by considering a window of features around the entities. ML algorithms are then employed to determine the patterns surrounding an entity which can be used in its identification. These patterns can then be applied to an un-annotated text to determine the likely placement of an entity; if start and end positions are identified then a process of pairing is used to resolve conflicting annotations. There has been a wide range of machine learning algorithms applied to the IE task; in the following sections we will discuss the key approaches.

Despite its general adoption for other tasks, decision tree induction has not been widely used in IE [Sekine, 1998] and [Karkaletsis, 2000], being two of the few examples) as it is less applicable to tasks, such as IE, where features are likely to have non-linear interactions, which adversely effects "greedy" induction processes, and possess a large number of values, which causes problems in determining the discriminative effect of features and limits the transparency of the final tree. Similarly nearest neighbour techniques have not been widely adopted, although Ahn recently examined their use in Event Extraction [Ahn, 2006]; however the work emphasised the approach to the modulisation of the task rather than extraction performance.

A good survey of the initial approaches to the use of rule-based induction for IE is provided by Muslea [1999]. Since then the two main applications of rule-learner to IE have been the LP2 generalisation technique [Ciravegna, 2001] and the uses of Inductive Logic Programming (ILP) [Aitken, 2002]. Simple rules have also been used for the "weak learners" in a boosting approach [Freitag and Kushmerick, 2000].

HMMs have been used widely in text analysis problems due to text, as an ordered sequence of tokens (or textual features), being readily formed as a Markov model. In IE, HMM have been used for the general NER task [Bikel et. al, 1997], as well as specific domains; in particular the biomedical domain [Leek, 1997; Shen, et al. 2003; Bunescu and Mooney, 2004]. In addition other probabilistic techniques have been applied to IE tasks; Maximum Entropy Model (ME) have been used for both Entity Recognition [Chieu and Hwee, 2002; Borthwick, 1998] and Coreference resolution [Kehler, 1997], and Andrew McCallum has championed the use of Conditional Random Fields (CRF) for NER [McCallum and Li, 2003; Sutton et al., 2006] and also for the extraction of information contained within web page tables [Pinto et al., 2003]. David Ahn has compared the use of CRF for TIMEX extraction [Ahn et al., 2005]; the work also applied MaxEnt to the normalisation of TIMEX statements.

A side from the attraction of using SVM due to their classification and generalisation capabilities, the use of kernel functions allows for a nature discrimination of graph representations as found in parse trees and structured (XML) documents. Therefore SVM have been used widely for the NER task [Isozaki and Kazawa, 2002; Finn and Kushmerick, 2004; Li et al., 2005; Iria, 2006], and specifically for TIMEX extraction [Hacioglu et al., 2005], as well as in coreference [Isozaki and Hirao, 2003] and relation extraction [Zalenko et al., 2003; Culotto and Sorensen, 2004].

Unsupervised/semi-supervised learning methods

Several approaches have applied clustering to IE where a word is characterised by its context and lexical features, for example NER [Lin and Pantel, 2000], relation extraction [Hasegawa, 2004], coreference resolution [Cardie and Wagstaff, 1999] noun phrase deal [Hasegawa et al., 2004] with Gooi and Allan [2004] extending the work to cross-document co-reference.

There are a number of approaches which have applied semi-supervised learning to the NER tasks. These employ bootstrapping techniques by initialising the algorithm with a set of optimised seed patterns which are used to extract a set of Named Entities, these are then marked-up in the unlabelled texts and new patterns are inferred and added to the set of initial patterns [Riloff and Jones, 1999; Collins and Singer, 1999; Etzioni et al., 2005; Nadeau et al., 2006]. Yangarber et al. [2000] use a similar approach, but perform the analysis at the pattern/document level to extract sentences rather than the Named-Entity/pattern level. A similar semi-supervised technique has also been used to extract relations [Brin, 1998].

Finn and Kushmerick [2004] compare a number of Active Learning approaches to IE, although the results are inconclusive a technique which selects documents most dissimilar to those in the labelled set and one which implements a co-train learning like approach improved over the baseline.

4.3.5 Evaluation

For an overview of the history and issues involved in evaluation of IE systems see Lavelli et al. [2004]. There have been a number of challenges which have provided both resources and incentive to stimulate research into Classification and IE.

Reuters: The initial Reuters corpus [Reuters-21578] was the main classification corpus for many years which was both positive in that it provided a means to compare techniques and negative in that it focussed research on a single domain. There is now a new corpus available (RCV1 [Reuters Corpus Volume 1]) which is much larger than the first.

Message Understanding Conference (MUC): was the main testing ground for IE approaches from its start in 1987 to its demise in 1998.

Automatic Content Extraction (ACE): (http://www.nist.gov/speech/tests/ace/) has replaced MUC and continues to organise various challenges for IE tasks.

Pascal Challenge: (http://tyne.shef.ac.uk/Pascal/) The Pascal Challenge on Evaluating Machine Learning for Information Extraction attempted to provide a level “playing-field” on which to assess relative approaches to ML for IE by providing a standard pre-processed corpus [Ireson et al., 2005].

4.3.6 Systems

There are many systems which provide varying degrees of text classification and IE functionality. The following list gives an indication of the most renowned systems which offer resources which are available for research purposes; there are also a number of commercial systems available (see Fan, et al. 2006 for an overview of these systems):

- Armadillo [Ciravegna et al., 2004]
- DIDEROT [Cowie et al., 1993]
- GATE [Cunningham et al., 2002]
- KIM [Popov et al, 2004]
- Know-It-All [Etzioni et al., 2004]

- LingPipe (http://www.alias-i.com/lingpipe/)
- Seeker/Semtag [Dill et al., 2003]
- Snowball and QXtract (http://snowball.cs.columbia.edu/)

4.4 Images

Image analysis is the quantitative or qualitative characterisation of two-dimensional (2D) or three-dimensional (3D) digital images to extract meaningful information. The characterisation of an image is based upon visual features which are extracted from that image, this can then be used to classify images with similar characteristics for applications such as content-based image retrieval (CBIR), which is also known as query by image content (QBIC). Applications may require the classification and retrieval of the entire image as a whole; however images may also be segmented into sub-regions which represent distinct objects within the image.

4.4.1 Feature Extraction

There are four main descriptors for the visual content of the image:

- Colour Features.
- Textural Features.
- Geometrical or Shape-based Features.
- Topological Features.

These features can either be global or local. Global image analysis considers the image as a whole, whilst local analysis first segments the image into several Regions Of Interest (ROI) then determines the properties and features of the ROI.

Colour Features

- *Colour Spaces*

A colour model is an abstract mathematical model describing the way colours can be represented as tuples of numbers, typically as three or four values or colour components. When this model is associated with a precise description of how the components are to be interpreted (viewing conditions, etc.), the resulting set of colours is called a colour space. The choice of a colour space depends on the information to be extracted or on the treatment to be applied.

- *Colour Histograms*

Colour histograms are used to encode the frequency distribution of pixel values either on a whole image or on some region of interest (ROI). Given a finite set of colours, it associates to each colour, its frequency in the image. It is invariant under any geometrical transformation (translation, rotation). When comparing two images or ROI using histograms it is necessary to compute the distance between both histograms using (dis)similarity measures such as Euclidean, χ-square, Kolmogorov-Smirnov and Kuiper distances [Brunelli, 2001]. Classical histograms and most of their derivatives do not take into account spatial distribution of pixels. Nevertheless Blob histograms [Qian, 2000] are able to differentiate pictures having the same colour pixel distribution but containing objects of different sizes. In order to reduce the histogram size, a few representative colours can be selected from the colour space, either using some generic heuristic or by analysing the image. This colour quantisation can be used as a basic descriptor of the image.

- *Colour Moments*

Colour moments have been shown to be both efficient and effective to represent the colour distribution of images [Stricker and Orengo, 1995]. They include the first order moment (mean), the second-order moment (variance) and the third order moment (skewness), thus an image can be described in only nine values (3 moments per colour component).

Textural Features

From a perceptual point of view, a texture may be defined by its ”coarseness”, ”repetitiveness”, “directionality” and “granularity”. However in terms of digital images, the texture of an image or

region is defined as a function of the spatial variation in pixel intensities (grey values) [Tuceryan and Jain, 1998]. The analysis of texture is used to determine regions of homogeneous texture, the boundaries between these regions can then be used to segment the image. Textural classification is also used to associate a region with a textural class (e.g. the material being represented (cotton, sand, etc), or a property of that material (smooth, coarse, etc).

The image analysis applied in the modelling of texture can be divided into three general methods:

- *Statistical Methods*

Statistical methods characterise image texture according to measures of the spatial distribution of grey values (e.g. moments of different orders, correlation functions, related covariance functions).

- *Structural Methods*

The structural methods of texture analysis assume that textures are composed of primitives (called texels). The texture is produced by the placement of these primitives according to certain placement rules. This class of algorithms, in general, is limited in power unless one is dealing with very regular textures. Structural texture analysis consists of two major steps: (a) extraction of the texture elements (texels), and (b) inference of the placement rule. A texture may then be characterized through properties of its texels (average intensity, area, perimeter, etc.) or the texel pattern as defined by the placement rules.

- *Model-based Methods*

Model based texture analysis methods study texture as a linear combination of a set of basis functions. The two main difficulties of such methods are first to find a suitable model to represent the texture (e.g. Fractal Model, Markov model, Fourier filter, Multi-channel Gabor filter, Wavelet transform) and then to compute the accurate parameters which capture the essential perceived characterization of the texture.

Geometrical or Shape-based Features

Using shape descriptors implies being able to extract accurate shapes from an image. Shape descriptors may be based on contour or edge detection together with statistical tools. Such methods are particularly suitable for simple images, which contain one shape easily distinguishable from the background. But better results may be obtained after a segmentation process, which is necessary when dealing when complex images.

Shapes can be described either by their contour or by the region they contain. Moreover they can be either seen from a global or from a local point of view. The former approach, which has been chosen for many shape descriptors, aims at capturing some overall property either of the shape itself (e.g.) or of its contour (e.g. Fourier descriptor). The latter approach is based on local observations on the region or more often on its contour (e.g. inflexion points). Global shape descriptors may be misled when occlusions occur whereas local ones are very sensitive to noise.

- *Region descriptors*

Simple geometrical attributes such as area, eccentricity, bounding box, elongation, convexity, compactness, and circular or elliptic variances are also often used to describe shapes. Although simple to compute, as they can be gathered in attributes vector that may be compared through the use of some accurate distance, their characterisation power is generally too weak to be used in isolation and they are often combined with more complex shape descriptors, such as those provided by geometrical moments.

- *Contour descriptors*

Fourier descriptors are one of the most popular tools to characterise and compare contours. A contour is first sampled into a given number of points. A shape signature function is then applied on the representative points of the contour (e.g. complex shape signature, distance to centroid, area, cumulative angular function, curvature). Such a function produces a set of values, which are encoded

through a Fourier transform and then normalized. Other methods include Autoregressive models and Wavelet transforms (particularly suitable for describing high curvature points).

Topological Features

Digital topology deals with properties and features of two-dimensional (2D) or three-dimensional (3D) digital images that correspond to topological properties (e.g., connectedness) or topological features (e.g., boundaries) of objects. Concepts and results of digital topology are used to specify and justify important (low-level) image analysis algorithms, including algorithms for thinning, border or surface tracing, counting of components or tunnels, or region-filling.

4.4.2 Image Segmentation

In order to analyse an image at the level of the objects it contains it is necessary to segment the image so that the image features can be related to the region representing the object. A segmentation process aims at accurately identifying the different areas of an image, either by computing an accurate partition of the image by detecting coherent regions or by detecting the boundaries between regions.

There are three broad approaches which are applied in ROI detection. Affine region detectors which detect regions covariant with a class of affine transformations; for a review of the various methods for detecting these regions see Mikolajczyk et al. 2006. The second approach is based on extracting a per pixel salience measure; after grouping pixels of similar saliency a hierarchical representation of salient regions may be obtained [Kadir et al., 2004; Rutishauser et al., 2004; and Walther et al., 2005]. Finally clustering can be applied to ROI as is usual with clustering it is possible to apply three basic methods; generating the clustering bottom-up (starting from a set of seed regions, combine the regions until some stop criteria are reached), top-down (by splitting the image into smaller regions) or a combination of both bottom-up and top-down (several clustering approaches are discussed in Llahi 2005]. The main difficulty in the application of such clustering methods is in deciding how to choose accurate criteria to characterize regions and determining a stopping condition for the algorithm.

4.4.3 Classification and IE

Image Classification and IE can be generally distinguished by processes which categorise the entire scene depicted in the image as oppose to those which categories a ROI or object within that image. Classification of images has been more widely examined due to the fact that image segmentation is not required and thus processes do not have to deal with segmentation inaccuracies, but mainly the difficultly in obtaining annotated images at the region or object level. Recently there has been a number of systems developed which aim to facilitate the process of image annotation [Halaschek, et al. 2005, Petridis et al. 2006, Chakravarthy et al. 2006], such systems are likely to stimulate more research into classification of images at the object level.

The image annotation process associates semantic descriptors, either keywords or ontological concepts, with some visual descriptors of the object contents. A variety of methodologies have been proposed for this process, the simplest approach is to merely consider the co-occurrences between semantic and visual descriptors [Mori 1999], however a number of ML techniques have also been applied to the task including; neural networks [Kosko 1992, Lin 1995, Stamou 2001, Tzouvaras 2003], genetic algorithms [Mitchell 1996], SVM [Vapnik 1995] and HMM [Rabiner 1986, Dugad 1996, Huang 1990].

4.4.4 Evaluation

ImageCLEF: (http://ir.shef.ac.uk/imageclef/) is the cross-language image retrieval track which is run as part of the Cross Language Evaluation Forum (CLEF) campaign.

4.5 Video

One of the features of video analysis is that it brings together a number of media types (image, audio and (via ASR) text) into a single connected setting. Thus video analysis has the opportunity of exploiting the data from these correlated, simultaneous channels, to extract information [Li et al., 2003; Huang et al.,1998 and Sundaram et al., 2000]. In addition there are other features which are specific to the media of video; those that involve the way in which the video frames are linked together using various editing effects (cut, fades, dissolves, etc.). The general video analysis process involves:

- Boundary detection: Segmenting the video stream into shots
- Key-frame extraction: Characterising the content of a shot/video
- Determining what objects are in the shot/video

The primary application of such a process is to allow the index of video in order to make it searchable, for content-based image retrieval systems; however the ultimate goal is to recognise the events portrayed and to understand the narrative of the video.

4.5.1 Feature Extraction

By analysing a video stream in terms of a structured sequence of shots, and then characterising the shots in terms of key-frames, the modelling of video content is reduced to extracting the content of structured still images. This means that the visual features extracted from video are mainly derived from the frame images, which where described above. In addition videos have the features which describe the motion of objects between frames, as well as features relating to the audio channel.

Boundary detection

The identification of the shot boundaries is a key essential step prior to performing shot-level feature extraction and any subsequent scene-level analysis. Shot transitions can be classified as of two types: abrupt transitions (cut) and gradual transitions (fade, wipe, dissolve, etc.). The approaches to detecting these shot transitions either make use of some statistical measure the change in frame features which indicate a transition (a review of several techniques is provided by Boreczky and Rowe [1996], and Dailianas et al. [1995] or use some form of Machine Learning (ML). In general visual features are used to identify the boundaries. However Huang et al [1998] and Sundaram et al [2000] both used a combination of video and audio; based on the idea that the audio should change as well as the video at the shot boundaries.

There are a number of ML approaches to Boundary Detection including nearest neighbour [Kender et al, 1998; Ren and Singh, 2004], neural nets [Ren and Singh, 2004], HMM for both shot boundary detection [Zhang et al. 2006] and higher level topic/story boundary detection [Phung et al. 2002; Chaisorn et al., 2003] and SVM [Feng et al.,2005].

Key-frame extraction

The usual approach to providing a higher level description for a video stream is to extract a set of key-frames which represent a summarisation of the content of the whole stream. The general technique employed is frame clustering [Yeung and Yeo, 1997; Zhuuang 1998; Mundir et al., 2005; Feng et al., 2005], each cluster being centred on a key-frame, thus the key-frames are maximally distinct from one another. The results of applying the clustering technique are dependent upon which features are used, the distance metric employed and the method for determining the number of key-frames (clusters) which sufficiently describe the video. Although clustering is the main key-frame extraction technique, other ML approaches have been applied to the problem, such as genetic algorithms [Avrithis et al., 1999].

Object extraction

The extraction of objects from video applies the techniques described above, for image object identification. As objects can be found in a number of sequential or disparate frames, they can also be used as features in key-frame extraction [Song and Fan, 2005; Lui and Fan, 2005]. Medioni et al. [2001] used object (car) detection with motion analysis to infer the event taking place in the video and thus the behaviour of the actors (drivers).

4.5.2 Classification and IE

As above the semantic classification of objects within a video relies mainly on the techniques applied to still images. However a number of approaches have been applied to the classification of whole videos according to global features using Decision Trees for educational videos [Phung et al., 2002] and news videos [Chaisorn et al., 2003] and more recently the use of SVM to filter videos which contain objectionable content [Jeong, et al., 2006].

Calic et al. [2005] present an interesting paper which discusses the specific issues that relate to the use of semantic information in video, and refers to a number of systems which uses some form of semantics for indexing, classification and retrieval.

4.5.3 Evaluation

TRECVid (http://www-nlpir.nist.gov/projects/trecvid/): The National Institute for Standards and Technology (NIST) have organised a challenge to evaluate video retrieval since 2001.

4.5.4 Systems

MediaMill (http://www.science.uva.nl/research/mediamill/index.php): a semantic video search engine.

aceMedia (http://www.acemedia.org/): knowledge and multimedia content technologies, which provides tools to automatically analyze content, generate metadata and annotation, and support intelligent content search and retrieval services.

References

Ahn, D. (2006). The stages of event extraction. Proceedings of the Workshop on Annotating and Reasoning about Time and Events, pages 1–8, Sydney, July 2006.

Ahn, D., Adafre, S. F. and de Rijke, M. (2005). Towards Task-Based Temporal Extraction and Recognition. Dagstuhl Seminar Proceedings of the Workshop on Annotating, Extracting and Reasoning about Time and Events

Aitken, J.S. (2002). Learning Information Extraction Rules: An Inductive Logic Programming approach. In van Harmelen, Prof Frank, Eds. Proceedings 15th European Conference on Artificial Intelligence, pages pp. 355-359, Lyon, France.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In Proceedings DARPA Broadcast News Transcription and Understanding Workshop (pp. 194–218). Lansdowne, VA: Morgan Kaufmann.

Avrithis, Yannis S., Doulamis, Anastasios D., Doulamis, Nikolaos D. & Kollias, Stefanos D. (1999). A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases Computer Vision and Image Understanding Vol. 75, Nos. 1/2, July/August, pp. 3–24, 1999

Beeferman, D., Berger, A., Lafferty, J. (1999). Statistical Models for Text Segmentation. Machine Learning

Berger, A. A Brief Maxent Tutorial http://www.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/tutorial.html

Bikel, D., Miller, S., Schwartz, R. and Weischedel, R. (1997). Nymble: a High-Performance Learning Name Finder ANLP 1997.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning theory (Madison, Wisconsin, United States, July 24 - 26, 1998). COLT' 98. ACM Press, New York, NY, 92-100. DOI= http://doi.acm.org/10.1145/279943.279962

Boresczky S. and Rowe, L.A. (1996). A comparison of video shot boundary detection techniques, Proc. SPIE 2664, 170-179, 1996

Borthwick, A., Sterling, J., Agichtein, E. and Grishman, R. (1998). Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition, WVLC 98.

Boser, B. E., Guyon, I. M. and Vapnik, V. N (1992). A training algorithm for optimal margin classifiers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144-152, Pittsburgh, PA, 1992. ACM Press.

Bunescu, R. and Mooney, R.J. (2004). Collective Information Extraction with Relational Markov Networks Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004), pp. 439-446, Barcelona, Spain, July 2004.

Calic, J., Campbell, N., Dasiopoulou S. and Kompatsiaris, Y. (2005). An Overview of Multimodal Video Representation for Semantic Analysis. European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies, EWIMT 2005, London, UK, November 30 - December 1, 2005

Cardie, C. and Wagstaff, K. (1999). Noun phrase coreference as clustering. In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 82-89.

Chaisorn, L., Koh, C., Zhao, Y., Xu, H., Chua, T.-S and Qi, T.(2003). Two-level multimodal framework for news story segmentation of large video corpus. 12th Text Retrieval Conference, Gaithersburg, MD, USA, 2003.

Chakravarthy, A., Ciravegna, F. and Lanfranchi, V. (2006). AKTiveMedia: Cross-media Document Annotation and Enrichment. In Poster Proceedings of the Fifteenth International Semantic Web Conference (ISWC2006).

Ciravegna, F. (2001). Adaptive Information Extraction from Text by Rule Induction and Generalisation, in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, August 2001.

Ciravegna, F., Chapman, S., Dingli, A., and Wilks, Y. (2004). Learning to Harvest Information for the Semantic Web. Proceedings of the 1st European Semantic Web Symposium, Heraklion, Greece, May 10-12, 2004

Cowie, J., Guthrie, L., Pustejovsky, J., Wakao, T., Wang, J., and Waterman, S. (1993) The Diderot Information Extraction System, to appear in Proc. First PA CLING Conference, Vancouver.

Culotta, A. and Sorensen, J. (2004). Dependency Tree Kernels for Relation Extraction. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.

Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002

Dailianas, A., Allen, R.B. and England, P. (1995). Comparison of automatic video segmentation algorithms, Proc. SPIE Photonics West, 2615, 2-16, 1995.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1):1--38, 1977.

Diederich, J., Kindermann, J., Leopold, E., and Paaß, G. (2003). Authorship attribution with support vector machines. Applied Intelligence, 19(1/2), 109–123.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A., and Zien, J. Y. (2003). SemTag and seeker: bootstrapping the semantic web via automated semantic annotation. In Proceedings of the 12th international Conference on World Wide Web (Budapest, Hungary, May 20 - 24, 2003). WWW '03. ACM Press, New York, NY, 178-186. DOI= http://doi.acm.org/10.1145/775152.775178

Drucker, H., Vapnik, V., and Wu, D. (1999). Support vector machines for spam categorization. IEEE Transactions on Neural Networks, 10(5), 1048–1054.

Dumais, S. T. and Chen, H. (2000). Hierarchical classification of Web content. In N. J. Belkin, P. Ingwersen, and M.-K. Leong (Eds.), Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval (pp. 256–263). Athens, GR: ACM Press, New York, US.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics 3: 32-57

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). Web-scale information extraction in knowitall: (preliminary results). In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM Press, New York, NY, 100-110. DOI= http://doi.acm.org/10.1145/988672.988687

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S. and Yates, A. (2005) Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artificial Intelligence, 165, pp. 91-134.

Fall, C. J., Törcsvári, A., Benzineb, K., and Karetka, G. (2003). Automated categorization in the International Patent Classification. SIGIR Forum, 37(1).

Fan, W., Wallace, L., Rich, S., and Zhang, Z. (2006). Tapping the power of text mining. Commun. ACM 49, 9 (Sep. 2006), 76-82. DOI= http://doi.acm.org/10.1145/1151030.1151032

Feng, H., Fang, W., Liu, S., and Fang, Y. (2005). A new general framework for shot boundary detection and key-frame extraction. In Proceedings of the 7th ACM SIGMM international Workshop on Multimedia information Retrieval (Hilton, Singapore, November 10 - 11, 2005). MIR '05. ACM Press, New York, NY, 121-126. DOI= http://doi.acm.org/10.1145/1101826.1101847

Finn, A. and Kushmerick, N. (2004). Multi-level Boundary Classification for Information Extraction. In Proceedings of the 15th European Conference on Machine Learning, Pisa, Italy.

Finn, A. and Kushmerick, N. (2003). Active learning selection strategies for information extraction. ECML-03 Workshop on Adaptive Text Extraction and Mining (Croatia)

Freitag, D. and Kushmerick, N. (2001). Boosted Wrapper Induction. AAAI 2000, 577-583

Freund, Y. and Schapire, R. E. (1999). A Short Introduction to Boosting: Introduction to Adaboost. Journal of Japanese Society for Artificial Intelligence, 14(5):771-780, September, 1999.

Giorgetti, D. and Sebastiani, F. (2003). Automating survey coding by multiclass text categorization techniques. Journal of the American Society for Information Science and Technology, 54(12), 1269–1277.

Gu, X.-D., Chen, J., Ma, W.-Y. and Chen, G.-L. (2002). Visual Based Content Understanding towards Web Adaptation, Proc. Adaptive Hypermedia and Adaptive Web-Based Systems, Malaga, Spain, 2002, pp. 164-173

Hacioglu, K., Chen, Y. and Douglas, B. (2005). Automatic Time Expression Labeling for English and Chinese Text. In Linguistics and Intelligent Text Processing, Volume 3406, 2005, 548-559

Halaschek-Wiener, C., Golbeck, J., Schain, A., Grove, M., Parsia, B. and Hendler, J. (2005) Photostuff - an image annotation tool for the semantic web. In 4th International Semantic Web Conference. 2005.

Hayes, P. J. and Weinstein, S. P. (1990). Construe/Tis: a system for content-based indexing of a database of news stories. In A. Rappaport and R. Smith (Eds.), Proceedings of IAAI-90, 2nd Conference on Innovative Applications of Artificial Intelligence (pp. 49–66).: AAAI Press, Menlo Park, US.

Heyer, L.J., Kruglyak, S. and Yooseph, S., (1999). Exploring Expression Data: Identification and Analysis of Coexpressed Genes, Genome Research 9:1106-1115

Holland, J. H. (1975), Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor

Honavar, V., Silvescu A., Reinoso-Castillo J., Caragea, D., Andorf, C. and Dobbs, D. (2001). Ontology-driven information extraction and knowledge acquisition from heterogeneous, distributed biological data sources, in: Proceedings of the IJCAI2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources, 2001.

Hotho, A., Staab, S., and Stumme, G. (2003). Wordnet improves Text Document Clustering. In Proc. of the Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference, Toronto, Canada, 2003.

Huang, J. Liu, Z. and Wang, Y. (1998). Integration of audio and visual information for content-based video segmentation. IEEE Int'l Conf. Image Processing (ICIP98), Special SessioSession on Content-Based Video Search and Retrieval. Oct. 1998. Chicago.

Ireson, N., Ciravegna, F., Califf, M. E., Freitag, D., Kushmerick, N. and Lavelli, A. (2005). Evaluating Machine Learning for Information Extraction, 22nd International Conference on Machine Learning (ICML 2005), Bonn, Germany, 7-11 August, 2005

Iria, J., Ireson, N. and Ciravegna, F. (2006) An Experimental Study on Boundary Classification Algorithms for Information Extraction using SVM

Jeong, C. Y., Han, S. W., and Nam, T. Y. (2006). Automatic Objectionable Video Classification System. Internet and Multimedia Systems and Applications 2006

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol (Eds.), Proceedings of ECML-98, 10th European Conference on Machine Learning (pp. 137–142). Chemnitz, DE: Springer Verlag, Heidelberg, DE. Published in the “Lecture Notes in Computer Science” series, number 1398.

Kan, M.-Y. (2001). Combining visual layout and lexical cohesion features for text segmentation. Columbia University Computer Science Technical Report, CUCS-002-01. 2001

Karkaletsis, V., Pailouras, G. and Spyropoulos, C. D. (2000). Learning decision trees for named-entity recognition and classification. In Proceedings of the ECAI Workshop on Machine Learning for Information Extraction, 2000

Kehagias, A., Petridis, V., Kaburlasos, V. G., and Fragkou, P. (2003). A comparison of word- and sense-based text categorization using several classification algorithms. Journal of Intelligent Information Systems, 21(3), 227–247.

Kender, J. R. and Yeo, B.-L. (1998). Video Scene Segmentation Via Continuous Video Coherence, Proc. CVPR ’98, pp 367-373, June 1998.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401–412.

Koster, C. H. and Seutter, M. (2003). Taming wild phrases. In F. Sebastiani (Ed.), Proceedings of ECIR-03, 25th European Conference on Information Retrieval (pp. 161–176). Pisa, IT: Springer Verlag.

Lam, W. and Ho, C. Y. (1998). Using a generalized instance set for automatic text categorization. In Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, pages 81–89, Melbourne, AU, 1998.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.

Lavelli, A., Califf, M. E., Ciravegna, F., Freitag, D., Giuliano, C., Kushmerick, N., and Romano, L. (2004). A Critical Survey of the Methodology for IE Evaluation. Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, May 26-28, 2004

Lenat, D. B. (1995). Cyc: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 38, no. 11 (November 1995).

Lewis, D. D. (1997). Reuters-21578 text Categorization test collection. Distribution 1.0. README file (version 1.2). Manuscript, September 26, 1997. http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. J. Mach. Learn. Res. 5 (Dec. 2004), 361-397.

Li, D., Dimitrova, N., Li, M., and Sethi, I. K. (2003). Multimedia content processing through cross-modal association. In MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia, pages 604–611, New York, NY, USA, 2003. ACM Press.

Li, Y. and Bontcheva, K. and Cunningham, H. (2005). SVM Based Learning System For Information Extraction. In: Proceedings of Sheffield Machine Learning Workshop. Lecture Notes in Computer Science. Springer Verlag.

Li, Y. H. and Jain, A. K. (1998). Classification of text documents. The Computer Journal, 41(8):537–546, 1998.

Liu, B. and Chen-Chuan-Chang, K. (2004). Editorial: special issue on web content mining. *SIGKDD Explor. Newsl.* 6, 2 (Dec. 2004), 1-4. DOI= http://doi.acm.org/10.1145/1046456.1046457

Liu, L. and Fan, G. (2005). Combined key-frame extraction and object-based video segmentation. IEEE transactions on circuits and systems for video technology. 2005, vol. 15, no7, pp. 869-884.

Liu, T., Chen, Z., Zhang, B., Ma, W., and Wu, G. (2004). Improving Text Classification using Local Latent Semantic Indexing. In Proceedings of the Fourth IEEE international Conference on Data Mining (Icdm'04) - Volume 00 (November 01 - 04, 2004). ICDM. IEEE Computer Society, Washington, DC, 162-169.

Luo, J., Shen, J. and Xie, C. (2004) Segmenting the Web Document with Document Object Model Services Computing, 2004 IEEE International Conference on (SCC'04), pp. 449-452

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

Maron, M. (1961). Automatic indexing: an experimental inquiry. Journal of the Association for Computing Machinery, 8(3), 404–417.

Masand, B., Linoff, G. and Waltz, D. (1992). Classifying news stories using memory-based reasoning. In Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval, pages 59–65, Kobenhavn, DK, 1992.

McCallum, A. and Li, W. (2003). Early Results for Named Entity Recognition with Conditional Random Fields, Fetures Induction and Web-Enhanced Lexicons, CoNLL 2003.

McCallum, A. and Nigam K. "A Comparison of Event Models for Naive Bayes Text Classification". In AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48. Technical Report WS-98-05. AAAI Press. 1998.

Medioni, G., Cohen, I., Bremond, F., Hongeng, S. and Nevatia, R. (2001) Event Detection and Analysis from Video Streams. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 8, pp. 873-889, Aug., 2001.

Michelson, M. and Knoblock, C. A. (2005). Semantic annotation of unstructured and ungrammatical text. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-2005).

Moschitti, A. and Basili, R. (2004). Complex linguistic features for text classification: A comprehensive study. In S. McDonald and J. Tait (Eds.), Proceedings of ECIR-04, 26th European Conference on Information

Retrieval Research (pp. 181–196). Sunderland, UK: Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2997.

Mukherjee, S., Yang, G., Tan, W. and Ramakrishnan, I. V. (2003). Automatic Discovery of Semantic Structures in HTML documents. International Conference on Document Analysis and Recognition (ICDAR). 2003

Mundir, P., Rao, Y. and Yesha, Y. (2005). Keyframe-based Video Summarization using Delaunay Clustering

Muslea, I. (1999). Extraction patterns for information extraction tasks: A survey. AAAI 1999 Workshop on Machine Learning for Information Extraction.

Nigam, K. (2001). Using Unlabeled Data to Improve Text Classification. Ph.D. Dissertation, Carnegie Mellon University.

Nigam, K. and Ghani, R. (2000). Analyzing the applicability and effectiveness of co-training. In A. Agah, J. Callan, and E. Rundensteiner (Eds.), Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management (pp. 86–93). McLean, US: ACM Press, New York, US.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2/3), 103–134.

Patwardhan, S. and Riloff, E. (2006). Learning Domain-Specific Information Extraction Patterns from the Web. http://www.cs.utah.edu/~sidd/papers/PatwardhanR06.pdf

Petridis, K., Anastasopoulos, D., Saathoff, C. , Timmermann, N., Kompatsiaris I. and Staab S.,

M-OntoMat-Annotizer: Image Annotation. Linking Ontologies and Multimedia Low-Level Features. Engineered Applications of Semantic Web Session (SWEA) at the 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES 2006), Bournemouth, U.K., 9-11 October 2006.

Phung, D. Q., Duong, T. V., Venkatesh, S., and Bui, H. H. 2005. Topic transition detection using hierarchical hidden Markov and semi-Markov models. In *Proceedings of the 13th Annual ACM international Conference on Multimedia* (Hilton, Singapore, November 06 - 11, 2005). MULTIMEDIA '05. ACM Press, New York, NY, 11-20. DOI= http://doi.acm.org/10.1145/1101149.1101153

Pinto, D., McCallum, A., Wei, X., and Croft, W. B. (2003). Table extraction using conditional random fields. In *Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval* (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM Press, New York, NY, 235-242. DOI= http://doi.acm.org/10.1145/860435.860479

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., and Kirilov, A. (2004). KIM – a semantic platform for information extraction and retrieval. Nat. Lang. Eng. 10, 3-4 (Sep. 2004), 375-392. DOI= http://dx.doi.org/10.1017/S135132490400347X

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kauffman.

Ratnaparkhi, Adwait, R. (1997). A Simple Introduction to Maximum Entropy Models for Natural Language Processing. IRCS Report 97--08, University of Pennsylvania, 3401 Walnut Street, Suite 400A, Philadelphia, PA, May 1997.

Ren, W. and Singh, S. (2004). Automatic Video Shot Boundary Detection Using Machine Learning. In Intelligent Data Engineering and Automated Learning – IDEAL 2004

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24, 5 (Aug. 1988), 513-523. DOI= http://dx.doi.org/10.1016/0306-4573(88)90021-0

Schapire, R. E. (1990). The strength of weak learnability. Machine Learning, 5(2):197--227, 1990.

Schapire, R. E. and Singer, Y. (2000). BoosTexter: a boosting-based system for text categorization. Machine Learning, 39(2/3), 135–168.

Schank, R.C. (1975). Conceptual Information Processing. New York: Elsevier.

Sebastiani, F. (1999a) Machine learning in automated text categorisation: A survey. Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, C.N.R., Pisa, IT, 1999.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 1–47.

Sekine, S., Grishman, R. and Shinnou, H. (1998) A Decision Tree Method for Finding and Classifying Names in Japanese Texts, WVLC 98.

Shannon, C. E. (1948). A Mathematical Theory of Communication, Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, 1948.

Shen, D., Zhang, J., Zhou, G., Su, J., and Tan, C. (2003). Effective adaptation of a Hidden Markov Model-based named entity recognizer for biomedical domain. In Proceedings of the ACL 2003 Workshop on Natural

Language Processing in Biomedicine - Volume 13 (Sapporo, Japan, July 11 - 11, 2003). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 49-56.

Song, X. and Fan, G. (2005). Joint Key-Frame Extraction and Object-Based Video Segmentation. wacv-motion, pp. 126-131, IEEE Workshop on Motion and Video Computing (WACV/MOTION'05) - Volume 2, 2005.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. Computational Linguistics, 26(4), 471–495.

Stricker, M. and Orengo, M. (1995). Similarity of color images, Proc. SPIE, vol. 2420, pp. 381–392, 1995

Sundaram, H. and Chang, S.-F. (2000). Determining Computable Scenes in Films and their Structures using Audio-Visual Memory Models. ACM Multimedia 2000, Oct 30 - Nov 3, Los Angeles, CA.

Sutton, C., McCallum, A. and Rohanimanesh, K. (2006). Dynamic Conditional Random Fields. Journal of Machine Learning Research (JMLR), Vol. 7, 2006.

Tuceryan, M. (1998). Textural Analysis. In The Handbook of Pattern Recognition and Computer Vision (2nd Edition), by C. H. Chen, L. F. Pau and P. S. P. Wang (eds.), pp. 207-248, World Scientific Publishing Co., 1998.

Tuceryan, M. and Jain, A. K. (1998). Texture Analysis. In The Handbook of Pattern Recognition and Computer Vision (2nd Edition), by C. H. Chen, L. F. Pau, P. S. P. Wang (eds.), pp. 207-248, World Scientific Publishing Co., 1998.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4), 315–346.

Viterbi, A J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13(2):260–269, April 1967. (The Viterbi decoding algorithm is described in section IV.)

Yang, Y. and Zhang, H. (2001). HTML Page Analysis Based on Visual Cues, Proc. of 6th International Conference on Document and Analysis, Seattle, USA, 2001

Yeung, M., and Yeo, B.-L. (1997). Video visualization for compact presentation and fast browsing of pictorial content. IEEE Trans. Circuits Syst. Video Technol. 7, 5 (Oct. 1997), 771–785

Wallach, H. M. (2004). Conditional Random Fields: An Introduction. Technical Report MS-CIS-04-21. Department of Computer and Information Science, University of Pennsylvania, 2004.

Zhang, W., Lin, J., Chen, X., Huang, Q. and Liu, Y. (2006). Video Shot Detection Using Hidden Markov Models with Complementary Features. First International Conference on Innovative Computing, Information and Control - Volume III (ICICIC'06), 593-596

Zhuang, Y., Rui, Y., Huang, T. S. and Mehrotra, S. (1998). Adaptive Key Frame Extraction Using Unsupervised Clustering

5. Multilingual/Multimedia Indexing

by Jaap Kamps

This chapter describes the state-of-the-art in the indexing of cultural heritage (CH) documents in various languages and of various media types. First, we discuss the special characteristics of cultural heritage documents. Second, we discuss the general approaches to indexing are currently being developed. Third, we detail for all the different media types the specific approaches available. Throughout the section, we'll indicate some of the open problems and challenges that are of most direct relevance to indexing cultural heritage documents as envisioned by the MultiMatch project.

5.1 Indexing Cultural Heritage Documents

As stated in Chapter 2[58], metadata plays a crucial role in providing access to cultural heritage. Cultural heritage institutions have invested enormous effort in gathering information about their precious objects, usually stored separately in library catalogue records, archival inventories, or museum registers. In non-digital collections, these descriptions of CH objects form the main access points for organizing, selecting and retrieving objects. For example, a controlled vocabulary which capture the topical subject of an object by a numerical code, such as DDC [2006] or UDC [2006], can provide subject access to CH objects even across language boundaries. In collections of digital CH objects, the combination of searching content as well as metadata provides powerful finding aids [Lesk, 2005]. However, combining different CH collections also implies combining different traditions of description, different controlled vocabularies, and different intended audiences in mind. Even when syntactically coded in a uniform format, such as [DCMI, 2006; RDF, 2006; OWL, 2006], the metadata will reflect the provenance of the particular object. Making sense of heterogeneous metadata is one of the greatest challenges for today's cultural heritages institutions.

It is an open problem how to provide uniform access to the myriads of formats in current combined collections, without the need for expensive manual or supervised revision of existing descriptions. There are two current approaches directly addressing this problem: The first approach is to treat the controlled vocabularies as a rigorous ontology, and attempt to define mappings between the different systems (e.g., [STITCH, 2006]). That is, the problem is now translated into a semantic interoperability or ontology mapping problem. The state-of-the-art techniques are far from full-proof, necessitating manual supervision [Handschuh and Staab, 2003]. Such effort is needed for each mapping covering a single pair of vocabularies. The viability of this approach depends on the number of different vocabularies involved, and on their rigorousness. The second approach is to treat the heritage descriptions as noisy and uncertain, and apply powerful methods from modern text retrieval (e.g., [MuSeUM, 2006]). Specifically, this approach makes very few assumptions on the presence or encoding of particular metadata, but exploits it whenever present. In essence, this is the famous ``dumb-down principle" [Weibel, 1995]: although metadata is based on a specific thesaurus or ontology, we can always fall back on the description of the terms in ordinary language. In theory, the second approach can be directly applied in the MultiMatch set-up where CH documents from many sources will be combined. It is an open question whether the approach is effective in practice considering the highly heterogeneous content ranging from authoritative information from CH sites to subjective views and opinions in personal Blogs. It is another open question whether the approach will scale to the volume of data envisioned in MultiMatch.

5.2 Indexing Approach

There are two basic approaches to indexing cultural heritage documents in various languages and of various media types. The first approach indexes all document sorts and media types separately, and

[58] In Chapter 2, we describe the metadata schemes typically adopted to describe digital objects. Chapter 3 discusses how Information Extraction techniques can be used to create explicit representations, i.e. metadata, from the information implicit in unstructured text. In this Chapter we examine the issues that have to be faced when applying or using manually or automatically assigned metadata for information access.

later integrates the results using distributed indexing techniques and fusion methods similar to those used in distributed IR [Callan et al., 1995]. The second approach is to define a single, complex document type definition that will form the basis for all material to be indexed: documents of various media types (text, audio, image, video, or mixed-content) and accompanying metadata. Despite much progress in searching by content in multimedia databases [Faloutsos, 1996] there is a clear trend toward the combination of various modalities [de Vries et al., 2000; Snoek and Worring, 2005]. Existing generic standards such as MPEG-7 (which is part of the XML family of languages) are able to cater for such a data model by incorporating multimedia content and metadata in a single semi-structured document.

Interestingly, researchers in IR are travelling down a similar path by integrating result ranking in the core of XML databases (e.g. [List et al., 2005]). Such systems radically depart from the standard "document as a bag-of-words" approaches, by preserving the document structure and using region algebras to score individual document components [Burkowski, 1992; Clarke et al., 1995]. The resulting database provides a general framework for complex object retrieval, allowing for a range of retrieval approaches without the need to re-index the collection. The most recent proposals allowing for complex retrieval models can be defined as logical queries on an XML database [Hiemstra and Michajlovic, 2005]. Currently available XML databases or retrieval systems such as the Cheshire [2006], MonetDB [2006], Lucene [2006], and MILOS [Amato et al., 2004] systems allow - to a greater or lesser extent- - this flexibility. It is an open question how to extend any of the existing systems to the specific demands of cultural heritage retrieval.

5.3 Indexing CH Media Types

5.3.1 Indexing Text

The state of the art indexing methods of cross-language information retrieval use dedicated tokenization methods [Hollink et al., 2004]. Some approaches consider various language-dependent morphological normalization techniques, such as lemmatization or stemming, and other approaches consider language-independent techniques, such as character n-gramming. Although approaches to the indexing of free-text are well studied, it is a major challenge how to preserve the document structure, if available, in the index, and how to ensure that the metadata associated with the documents is indexed in separate fields. As mentioned above, various metadata---both from the original CH documents as well as those automatically assigned by extraction and classification tools---are crucial for providing access to CH documents. It is an open question if, and how, special tokenization is required for the specific cultural heritage terminology, and certainly an issue that will be studied in MultiMatch.

5.3.2 Indexing Images

For indexing images, the state-of-the-art complex object database naturally supports indexing the binary image, features extracted from the image, and the metadata attributed to the images. Highly sophisticated methods have been developed for content-based image retrieval [Smeulders et al., 2000]. Examples are the extraction of salient features of images, such as low-level visual properties of texture, colour, and shape, or various multi-scale robust features. It is an open question whether specialized visual CH features can be fruitfully developed, for example for classifying different sorts of art objects. The output of visual feature extractors is typically stored in a dedicated indexing structure separated from the main index. Effective image retrieval methods still heavily rely on metadata, so all available textual information about the images will be carefully indexed. For the specific cultural heritage domain, it is also an open question how to integrate the specific metadata descriptions and textual context, with the content-based image features.

5.3.3 Indexing Speech and Audio

For indexing speech, effective systems do heavily rely on automatic speech recognition (ARS) techniques [Jelinek, 1997]. The quality of ARS for spoken English is well enough for effective text retrieval [Garofolo et al., 2000]. The situation for multilingual ASR is, however, quite different [Byrne et al., 2004]. Although some progress has been made, multilingual ASR systems are still in their infancy and developing such systems is clearly outside the scope of MultiMatch. Hence, MultiMatch

will most probably focus ASR for English, supplemented with existing transcriptions in other languages. In terms of the indexing techniques, due to the poor audio quality common in CH audio collections we have to rely on methods that are robust against noisy transcriptions. Fortunately, audio files in CH collections have typically rich metadata descriptions, and additionally or in lieu of transcriptions, can be indexed as a document surrogates. An alternative to generic ASR is to detect only a limited number of targeted CH concepts like names of artists in audio. This will greatly reduce the amount of training data needed to construct the appropriate language and acoustic models (and make it a realistic option for the generally scarce resources of CH institutes). Since this approach will only detect a limited number of concepts in the audio files, its index will serve as dedicated entry points to the stored audio files. Integrating these different sorts of transcripts and other document surrogates in a common index, whilst ensuring proper alignment with the audio file, is a major challenge. Little is known about the cost/benefit analysis: how much training data is needed to track a limited set of salient CH concepts, and what fraction of the search requests will these concepts cover?

5.3.4 Indexing Video

Traditional approach to indexing video would separate the audio and video streams, where the audio stream is indexed as text using automatic speech recognition techniques, and the video stream is - after shot boundary detection and key frame extraction - converted to content based image features. The integrated use of different sources is an emerging trend in video indexing research. This is a semantically informed multi-modal approach in which the visual, auditory and textual modalities are combined [Snoek and Worring, 2005]. First, a multi-modal approach to content segmentation is proposed; some of the content elements may be converted to text. Then, the different modalities are integrated to enhance the classification accuracy on semantic subtasks such as genre detection, logical units, and named events. Most successful integration methods are based on Hidden Markov Models or Bayesian networks. For the specific cultural heritage domain, it is an open question how to integrate the specific metadata descriptions and textual context, with the multi-modal video features.

5.4 Wrap Up

This concludes the overview of the state-of-the-art for indexing of cultural heritage documents in various languages and of various media types. We make three general observations. First, the state-of-the-art approaches have typically been applied in isolation, whereas the problems addressed by MultiMatch require them to be put coherently within a single unified framework. This integration of a whole range of approaches is in itself a major challenge. Second, the domain targeted by MultiMatch is substantively different in terms of its subject matter (i.e., the cultural heritage domain in a broad sense) and in terms of the nature of the document (i.e., with particular traditions of descriptions, heterogeneous document formats, and various media types). It is an open question to what extent the effectiveness of state-of- the-art approaches will carry over to this new domain. Third, although the existing approaches provide an excellent starting point for indexing within MultiMatch, there are quite a few open questions remaining that need to be addressed during the project. The progress made within the MultiMatch project will be carefully monitored in future revisions of this document.

References

Amato, G:, Gennaro, C., Rabitti, F., and Savino, P. (2004). Milos: A multimedia content management system for digital library applications. In Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL 2004, pages 14--25. Springer Berlin / Heidelberg, 2004.

Burkowski, F. J. (1992). Retrieval activities in a database consisting of heterogeneous collections of structured text. In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '92), pages 112--125, New York, NY, USA, 1992. ACM Press.

Callan, J. P., Lu, Z. and Croft, W. B. (1995). Searching distributed collections with inference networks. In SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pages 21--28. ACM Press, New York, 1995. Cheshire. Cheshire3 Information Retrieval Framework, 2006. http://cheshire3.sourceforge.net/.

Clarke, C. L. A., Cormack, G. V. and Burkowski, F. J. (1995). An algebra for structured text and a framework for its implementation. The Computer Journal, 38:43--56, 1995.

DCMI. Dublin Core Metadata Initiative, 2006. http://dublincore.org/.

DDC. Dewey decimal classification, 2006. http://www.oclc.org/dewey/.

de Vries, A. P., Windhouwer, M. Apers, P. M. G. Kersten, M. (2000). Information access in multimedia databases based on feature models. New Generation Computing, 18:323--339, 2000.

Faloutsos, C. (1996. Searching Multimedia Databases by Content. Kluwer Academic Publishers, 1996.

Garofolo, J. S., Auzanne, C. G. P. and Voorhees, E. M. (2000). The TREC spoken document retrieval track: A success story. In Proceedings of RIAO 2000: Content-Based Multimedia Information Access, pages 1--20, 2000.

Handschuh, S. and Staab, S. (2003). Annotation for the Semantic Web. IOS Press, Amsterdam, 2003.

Hiemstra, D. and Michajlovic, V. (2005). A database approach to information retrieval: The remarkable relationship between language models and region models. Technical Report 05-35, Centre for Telematics and Information Technology, 2005.

Hollink, V., Kamps, J., Monz, C. and de Rijke, M. (2004). Monolingual document retrieval for European languages. Information Retrieval, 7:33--52, 2004.

Jelinek, F. (1997). Statistical methods for speech recognition. MIT Press, Cambridge MA, 1997.

Lesk, M (2005). Understanding Digital Libraries. The Morgan Kaufmann series in multimedia information and systems. Morgan Kaufmann, San Francisco CA, second edition, 2005.

List, J., Mihajlovic, V., Ramirez, G., de Vries, A., Hiemstra, D., and Blok, H. E. (2005). TIJAH: Embracing IR methods in XML databases. Information Retrieval, 8:547--570, 2005.

Lucene. Open-source search software, 2006. http://lucene.apache.org/.

MonetDB. Open source high-performance database system, 2006. http://monetdb.cwi.nl/.

MuSeUM. Multiple-collection Searching Using Metadata, 2006. http://www.nwo.nl/catch/museum/.

OWL. Web Ontology Language, 2006. http://www.w3.org/2004/OWL/.

RDF. Resource Description Framework, 2006. http://www.w3.org/RDF/.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. and Jain, R. (2000). Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis Machine Intelligence, 22:1349--1380, 2000.

Snoek, C. G. M. and Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. Multimedia Tools and Applications, 25:5--35, 2005.

STITCH. SemanTic Interoperability To access Cultural Heritage, 2006. http://www.nwo.nl/catch/stitch/.

UDC. Universal decimal classification, 2006. http://www.udcc.org/.

Weibel, S. (1995). Metadata: The foundations of resource description. D-Lib Magazine, 1(7), 1995. http://www.dlib.org/dlib/july95/07/weibel.html.

6. Multilingual/Multimedia Information Retrieval

by Gareth J.F. Jones with contributions from Martha Larson and Stephane Marchand-Maillot

In common with many areas of language processing, the origins of information retrieval (IR) research are to be found in the exploration of techniques for electronic English language text archives. A number of successful models for information retrieval have been, and continue to be, developed with English language documents as their primary research focus.

However, English language document collections, and electronic text documents in any language, represent only a minority of the information sources that a user may wish to search to satisfy their information need. The need to expand the scope of IR research beyond English text has been recognised in the last 15 years. Increasing amounts of work have been conducted and reported which explore non-English IR, cross-language information retrieval (CLIR), multilingual information retrieval (MLIR) and multimedia information retrieval (MIR). This work has greatly increased understanding of the issues of multilingual and multimedia information retrieval and access. A range of techniques have been proposed, explored, evaluated and refined. However, the techniques are imperfect and many challenges remain to improve effectiveness and to extend the scope of retrieval tasks. This will require a deeper investigation of the issues and problems than has been carried out so far together with the introduction of novel techniques.

When efforts to expand the horizons of IR began it was not at all clear what retrieval methods should be adopted for these new tasks in order to achieve the greatest IR effectiveness. It was found that established IR methods transferred well to other languages, and linguistic media, speech and scanned text images. The reason for this result should perhaps not be too surprising given the rigor and care taken over the years to ground these models in sound theoretical analysis, and the extensive experimental evaluations that have characterized this work. Significant issues arise with respect to translation between search topics and documents for cross-language and multilingual information retrieval. For MIR, there are significant issues related to the definition of retrieval units, i.e. what should we look for in an image or video, and the accuracy with which features can be detected automatically once they have been defined.

This chapter continues in the next section with a brief review of the relevant details and indexing assumptions of text IR. Section 6.2 describes experimental work with non-English test collections, this is extended in Section 6.3 which gives results for cross-language and multilingual IR. Section 6.4 introduces multimedia IR and highlights some relevant experimental work. Finally, Section 6.5 draws conclusions from existing work and looks toward future applications and challenges.

6.1 Probabilistic Models and Feature Indexing

IR systems seek to satisfy a user's *information need*. Current IR systems attempt to do this by locating *relevant* documents from within which the user can extract the required information. Potentially relevant documents are selected and returned to the user based on a retrieval model taking the user's query as input. The retrieval model can make use of whatever information is made available about the documents from among which it is seeking to locate the relevant ones. Document information is most typically based on simple extracted attributes such as words present in a document, but may include phrases or other extracted features; additionally features may be annotated with functional details such as their part-of-speech or semantic details such as those representing a geographic place or a time. While such annotations are not generally used within retrieval models which are normally based on word-level features, they can be useful for document browsing interfaces using maps or timelines, or for more advanced retrieval applications such as question answering systems which usually include some degree of language processing to locate the answer to a user's questions from within the available documents.

Document retrieval models fall into two broad classes of Boolean and best-match, the latter being the dominant modality of searching in current IR research. Boolean retrieval uses search queries constructed using Boolean logic to select documents which match these criteria from the available collection; the documents are returned to the user unsorted. The user must then browse among the returned documents either randomly or using some potentially useful criteria such as the date of creation, author, or document source. The requirement for complex search queries and the absence of content-based ranking means that it is unattractive to the majority of users of search engines who lack the enterprise to construct complex queries and desire the simple way of determining which documents are most likely to be of interest to them provided by ranked best-match IR. Over the years, many best-match ranked retrieval models have been proposed and evaluated. The most popular models being: the vector-space approach [Salton & Buckley, 1988], the probabilistic model [Robertson & Sparck Jones, 1976], and more recent methods based on statistical language modelling [Ponte & Croft, 1998].

If we had a complete model of each document, describing all potentially important features, with a correspondingly detailed model of the information need expressed by the search request, we might expect perfect retrieval with all relevant documents having higher probabilities than non-relevant documents. Alas such document models do not currently exist, and indeed the expression of information need in the search request is often an insufficient or inaccurate expression of the user's information need. Due to these deficiencies, retrieved ranked document lists generally interleave relevant and non-relevant documents. The objective of research in ranked IR is to improve the reliability of these imperfect relevance probability estimates.

Every document can be assumed to be a unique event, and in general, we take it that the description of each document used for retrieval is similarly unique. A problem arises with this modelling assumption, since it is difficult to assign probabilities to unique events. A solution comes in the form of decomposing document descriptions into their non-unique components or attributes, whose association with relevance can be estimated. These attributes can be used in combination to synthesise a relevance probability estimate for each unique document. The derivation of the early form of this practical probabilistic model (the "binary independence model") is described in van Rijsbergen [1979], and the more recent extended form of the model (well known as the "Okapi BM25'" model) in Sparck Jones et al. [2000a]. In the BM25 model the likelihood of relevance for a document *j* is computed based on the sum of the *combined weights cw(i,j)* of the independent attributes *i* which occur in both the document and the current search request. *cw(i,j)* values are computed based on the classic IR attribute weighting features of across document collection frequency (the *collection frequency weight cfw(i)*) of attributes *i*, the *within document frequency* of an attribute *i* in the document *j*, and an adjustment of the weight to compensate for document length [Robertson & Walker, 1994].

In general for current IR systems, each document is modelled as a simple "bag-of-words" which lists the attributes occurring within the document and their frequency of occurrence. The degree of match between a document *j* and the search request is then simply computed as a matching score *ms(j)* of the sum of the weights of the attribute in common between the request and the document. A list of documents ranked by matching score is then returned to the users. Documents are thus represented within the IR system as (assumed) independent attributes. The models used for ranked retrieval tell us nothing about the language of these attributes or even the media of the documents. Of course, much of the experimental work that established the effectiveness of this model has been carried out using English text collections often taken from general news or agency sources, but in theory there should be no reason why they cannot be used effectively for other languages, media or data sources.

Several well established techniques are typically applied for automatic indexing of English language text documents. These include removal of frequent *stop words*, such as those in van Rijsbergen's list [van Rijsbergen 1979], *suffix stripping*, using a method such as the Porter algorithm [Porter 1980], standardisation of spelling, and conflation of synonyms. Whatever preprocessing is applied, the features used for retrieval are still independent attributes derived from the document. Combined with enhancements such as relevance feedback and pilot searching using large additional document

collections, these methods have shown effective retrieval in many evaluation tasks undertaken in the last 10 years or so.

The following sections look at the adaptations required for the application of IR methods to non-English documents, cross-language and multilingual information retrieval, and the effectiveness for multimedia information retrieval.

6.2 Non-English Information Retrieval

A key consideration when developing an IR system for a new language is the selection of the most suitable set of attributes to be used to index the documents. The lexical and structural differences between languages mean that the distributions of attributes within individual documents and across collections will vary between different languages. However, since the standard IR models make no explicit language dependent assumptions about these distributions, there is no reason to suppose that, with appropriately selected indexing units, they should not work effectively for any language.

From a linguistic perspective English actually provides a good starting point for the investigation of indexing methods and retrieval models. The basic word units of the language are easily identified, and the types and degrees of inflection of individual words are relatively simple compared to those of many other languages. There are of course many exceptions to these apparently simple rules of inflexion, and ongoing debate over the basic units of meaning, but generally these concerns can be safely ignored or handled by explicit exception lists for the purposes of IR indexing. Some other languages have similar properties to English while others introduce new issues which must be addressed for effective retrieval. This discussion outlines some of the features relating to indexing and retrieval of a range of representative languages.

From an IR perspective, languages such as French, Italian and Spanish can be addressed using adaptations of the techniques used for English. Thus for each language, we need to develop a suitable set of high frequency stop words that can be removed safely without affecting retrieval effectiveness, suffix stripping algorithms to conflate words to common stems, and appropriate synonym dictionaries [Wechsler, Sheridan, & Schäuble, 1997]. Standard IR methods using this approach have been shown to be effective in comparative evaluations of non-English IR tasks, for example within the Cross-Language Evaluation Forum (CLEF) workshop series [Savoy, 2004].

More complex issues are introduced by languages such as German and Dutch which are highly declensional with a rich system of inflections and cases [Braschler & Ripplinger, 2004]. In addition, in common with other Germanic languages, such as Swedish, and other languages such as Finnish, there is free compounding of words to express concepts developed from the component words. In these cases, although words are still the building blocks of the language, they are frequently combined into noun compounds without spaces. If one of these noun compounds appears in a search request and a document, there is a very good chance that this is a relevant document. However, the generative nature of the compounds means that often no match will be found for a search compound within the document set, even if the similar concepts are being described This can lead to many potentially relevant documents being missed, since they do not contain the compound in exactly the form used in the request. The general approach to this problem is to develop methods for compound splitting; these techniques may rely on the use of a compound dictionary or language specific rules for identifying word units within compounds, or a combination of both methods [Braschler & Ripplinger, 2004]. Of course, in addition to the decompounding of these concatenated words, indexing of these languages also benefits from the application of effective stemmers and removal of stop words.

Different issues arise in the case of east Asian languages such as Chinese and Japanese. The written form of these languages uses ideograms of Chinese origin. There are many thousands of these characters which usually have some meaning associated with them. Most words are formed by bringing two characters together. The meaning of the word is usually related to those of its constituent characters. Shorter words consisting of one character can express simple concepts and occasionally longer words more complex ones. While Chinese is restricted to a single character set, in the case of Japanese three additional character sets are in common usage: *hiragana* whose role is similar to

function words and verb suffices in English, *katakana* which are used to transliterate Western concepts, e.g. *computer* appears phonetically in Japanese katakana as *ko n pu ta*, and *romaji*, for Western characters sometimes used for numbers and proper nouns. The major concern when indexing languages of this type is the observation that there are no spaces between the words of each sentence. The text must thus be segmented into suitable representative units prior to indexing. Further since the ideogram character set is itself so rich, there is a question of what the best units for retrieval actually are.

A number of approaches have been explored for indexing these languages. The most basic method is simply to take each character as an indexing unit, a slightly more elaborate one is to use overlapping n-grams of characters of varying lengths, while the most complex strategy is to apply morphological analysis to identify the most likely word break points. A number of experiments using various Chinese and Japanese test collections exploring different approaches to segmentation have been carried out with inconclusive results, for example Huang & Robertson [1997] and Jones, Sakai, Kajiura, & Sumita [1998]. All the above approaches produce a good level of retrieval effectiveness.

6.3 Cross-Language and Multilingual Information Retrieval

Retrieval involving more than one language is broadly classified into two areas: cross-language information retrieval (CLIR), and multilingual information retrieval (MLIR). CLIR is concerned with the retrieval of documents in one language using search requests in another language, e.g. Dutch requests used to retrieve Italian documents. MLIR extends this to retrieval from a collection where documents are uniquely present in one language, but the collection overall covers documents in multiple languages, e.g. using an English request to retrieve from a collection with documents in English, Dutch, Spanish, and Italian. In practice, more complex situations are clearly possible. A single document may contain material in more than one language, and individual documents may be repeated in different languages within a collection. From these definitions it can be argued that CLIR is really a subset of MLIR. This section introduces research questions posed by CLIR and MLIR, and outlines the main solutions that have been proposed and explored to date.

6.3.1 Cross-Language Information Retrieval

The principal question that arises in the context of CLIR is: how should the language barrier between the search requests and documents be crossed? Should search requests be translated into the language of the documents, should the documents be translated into the language of the request, or both? Further, what is the best approach to carrying out this translation?

Request Translation vs. Document Translation

There are well rehearsed arguments for and against request or document translation, with the main issues relating to translation cost, at what stage it is carried out, its effectiveness for retrieval, the available translation and computational resources, and the storage implications.

Generally it is held that translating requests when they are entered will be fast enough, since they are likely to be short, not to interfere with interactive searching. Unfortunately, short requests often have minimal formal linguistic structure, and further because they are short, there is little information of the context in which the request words have been selected by the user. These factors mean that it will often be difficult to perform reliable deep linguistic analysis when attempting to perform translation of the request. One consequence of this is that it can be difficult to select the contextually appropriate translation of polysemous words. A further implication of attempting to translate short requests is that the mistranslation of individual words can have a significant impact on retrieval effectiveness. However, since the document collection to be searched will not have been translated, and is therefore accurate, redundancy effects are often found to help to ameliorate translation errors even for short requests. It is further frequently argued that, since deep linguistic analysis of a request may not be possible (or if possible may not be desirable, if it is likely to be unreliable), and since we are only seeking to transfer the words into another language, shallower translation methods may be better for request translation CLIR.

Consider now the alternative approach of document translation. Documents are generally much longer than search requests, and the content will often be linguistically well structured with large amounts of contextual information available. Thus translation of documents using formal linguistic analysis is potentially more accurate than it is for requests. This may not be the case for web content where content is often more informally structured without formal sentences. However, even in this case the amount of contextually related material in the document may assist in accurate translation. While they may generally be translated more accurately than short requests, translated documents will nevertheless contain a number of errors arising from incorrect analysis of the source text and limitations of the translation dictionaries. These errors will inevitably impact adversely on retrieval accuracy for CLIR. However, adopting document translation does mean that no translation has to take place when the search request is entered, so the retrieval stage itself is computationally faster and cheaper. Also, the search request is now accurate, with no possibility of translation error. A major disadvantage of document translation is the very high cost of translating all the documents. Although, since translation is done in advance of retrieval and only has to be done once, it can really be regarded as part of a very expensive indexing process. However, there are storage implications which arise from the need to maintain a separate search collection in each request language into which the documents are translated.

Experimentally both request and document translation have been shown to be effective, with at least one study showing that combining the retrieval output of both methods used independently can produce the best overall retrieval effectiveness [McCarley, 1999].

One way to address the problem of storage is to translate all documents into a single "pivot" language, most probably English, and then to translate the requests into this same language when they are entered. This has the disadvantage that since both the requests and documents are being translated, translation errors will be compounded with a consequential impact on retrieval effectiveness. Pivot languages can also be used when resources are not available to translate directly between the request and document languages [Gollins & Sanderson, 2001]. In this case they can be used for translation of both requests and the documents into the pivot language, or for sequential translation of either the requests or documents into the language of the other.

Translation Methods for CLIR

Another widely debated issue in CLIR is how the translation should be carried out. The issues here relate both to the actual best means of translation for CLIR, were a perfect translation resource to be available, and the most appropriate method, where technical and resource limitations mean that real translation systems are currently far from perfect. Broadly speaking the three translation strategies that have been explored for CLIR can be categorised as: dictionary-based, comparable corpora, and machine translation.

Most early work in CLIR advocated the use of bilingual dictionaries for topic translation, with a variety of elaborations to improve their effectiveness for this task [Hull & Grefenstette, 1996]. In its simplest form, this approach replaces each word in the search request with all possible translations of the word in the document language appearing in a bilingual dictionary. As well as including the appropriate translation, if it is available in the dictionary, this simple method often introduces many contextually inappropriate translations of this word. These incorrect translations have been shown to significantly degrade CLIR retrieval effectiveness relative to monolingual IR for the same set of requests and documents. It has been demonstrated that dictionary-based CLIR performance can be improved by using careful phrase translation and relevance feedback both prior to and after translation of the request [Ballesteros & Croft, 1998].

Given the problems with ambiguity arising from the use of bilingual dictionaries, and the gaps which occur with regard to their coverage of domain specific vocabulary items, alternative methods have been explored which align comparable corpora in the different languages [Sheridan & Ballerini, 1996]. Related terms appearing in this aligned content are used to translate requests in a context specific way. One of the problems with this strategy is that suitable related corpora are often not available for alignment. A widely explored way to overcome this problem is to use content from the

internet [Nie, Simard, Isabelle, & Durand, 1999]. In this approach, large numbers of web pages are collected and aligned, and then used for request translation. Nie et al. demonstrated that an improvement in retrieval effectiveness can be obtained by using the aligned web documents in combination with a bilingual dictionary.

Perhaps the most obvious solution to crossing the language barrier between requests and documents is to use a standard commercial machine translation system. Indeed for CLIR using document translation, machine translation would appear to be the only realistic option given the huge amount of ambiguity that the other translation methods would introduce. Certainly I'm not aware of work which attempts to translate whole document collections using a different method. The arguments in favour of machine translation for CLIR centre on the potential for accurate translation of the words, appearing in the request or the document, which can be achieved by bringing sophisticated translation resources to bear on the task. Current machine translation systems often produce rather unnatural prose output. However this is not a problem for CLIR where we are only interested in the reliable translation of words with good relevance selectivity. The arguments against machine translation for CLIR are based on the previously stated issues of poor linguistic structure in search requests, which can render them difficult for formal linguistic analysis using machine translation, with consequential translation failures and inappropriate translation of words. Dictionary limitations can also result in translation problems for both requests and documents. This latter issue is likely to pose particular challenges for domains and their associated specialist topics which will often be outside the general purpose vocabularies used for developing the standard versions of commercial translation systems. Specialised dictionaries can be available to adapt machine translation to specific domains, but these are only likely to be available commercially for domains where the financial returns are deemed likely to justify the significant investment required to develop them.

An experiment at Toshiba performed a comparative evaluation of progressively more sophisticated request translation strategies ranging from simple bilingual dictionary lookup, to part-of-speech tagging, sense disambiguation, and full machine translation for an English - Japanese CLIR task [Jones, Sakai, Collier, Kumano, & Sumita, 1999]. Perhaps surprisingly given the arguments against machine translation for CLIR, the best retrieval effectiveness was found using full machine translation. This result was observed for both natural language request statements, and requests modified to disrupt the linguistic structure by removing the function words prior to translation. More recent experiments have shown that a combination of machine translation and the BM25 ranked retrieval model combined with relevance feedback produces among the best reported effectiveness for the CLEF CLIR tasks [Jones & Lam-Adesina, 2001] [Lam-Adesina & Jones, 2003]. Analysis of the retrieval behaviour of individual requests showed that there is sensitivity to the failure to translate important words, usually previously unseen proper nouns. For example, failure to translate phonetic loan word proper nouns rendered in katakana in Japanese if they are not present in the translation dictionary significantly degrades retrieval effectiveness. This will often be a problem for bilingual dictionaries as well; although, the impact on retrieval performance may be masked by translation ambiguity issues. However comparable corpora should be able to capture these domain specific translations, as long as they include documents covering the appropriate related topics in their training set. It should be noted that in all cases the documents used in these experiments were taken from published news corpora, and the results may not extend to material that is not formally published and/or is outside the topics encountered in national and international news stories.

Many papers have been published describing CLIR results in more recent years. The references included here are generally those which first introduced or advocated a particular translation approach for CLIR, in each case subsequent work has often extended these methods. While machine translation shows good results when available, bilingual dictionaries and aligned corpora are an important translation resource for CLIR with language pairs for which well developed machine translation tools are not available, and most likely where structural and domain issues render machine translation systems less effective, although this latter points remains to be illustrated in practice. There are direct bilingual dictionaries available between most major languages pairs, and even for minority languages there are bilingual dictionaries to major languages such as English, while the expanding amounts of

electronic text available from many sources mean that corpus-based methods will become an increasingly important resource for translation in CLIR.

6.3.2 Multilingual Information Retrieval

In MLIR the IR system is expected to respond to a search request in one language by generating a ranked list of potentially relevant documents in multiple languages. Similar to CLIR, MLIR can be approached using either a request or document translation strategy. The challenges of MLIR include similar translation issues to CLIR; however it also introduces a significant new problem which arises because the documents in each language will often be in separate collections. In a practical system, document collections may be geographically distributed with no option to merge them into a single collection. However, even if the documents can be combined into a single physical collection, the fact that they are in different languages means that semantically related search terms cannot be conflated, and effectively the collection will still behave as separate, language specific, sub-collections. The major difficulty that arises for MLIR is how to take the separate outputs from searching individual collections and merge them into a single output list for delivery to the user, which reliably ranks relevant documents higher than non-relevant ones. For this reason, MLIR is often seen as being akin to monolingual distributed IR, where separate search collections are stored and searched independently for practical or commercial reasons; lists retrieved from the individual collections must then be merged to form a single ranked list output [Callan, 2000]. There are also potential issues or retrieval effectiveness arising from the separation of the overall ``virtual'' document collection into multiple smaller collections since the term weights may be less accurately estimated within the smaller collections.

In MLIR the merging problem arises since ranked lists from the separate collections will be generated using different indexing strategies, and, as discussed earlier, the features will have varied distributions for the individual languages. This means that the document matching scores from the retrieved ranked document lists will generally be incompatible. For example, documents retrieved from a collection with higher average matching scores will tend to be favoured in the merged list. Thus the list may be biased towards certain collections and hence languages, regardless of the actual relative likelihood of documents retrieved from these collections being relevant. If this problem is overcome, a further concern is that the matching score profiles of the lists may be different. Hence the lists cannot be merged in a simple reliable way. In general for distributed IR, difficulties of list merging vary depending on the number of differences between the IR systems used to compute the separate lists, and potentially the cooperation between the maintainers of the separate search engines. The separate retrieval engines may reliably make all statistics of their collections available to the merging algorithms, they may make some subset available (potentially of questionable reliability) or they may make no information available beyond identifying the documents and their retrieval rank [Callan, 2000]. The amount of information available from the separate collections affects the complexity of the merging strategy that can be adopted. If the separate retrieval systems use different retrieval ranking algorithms then the scores will clearly be incompatible, but even if an identical retrieval strategy is used for all the collections, the matching scores will be incompatible due to the different values used to estimate the term weights or other ranking parameters. In MLIR, these issues are compounded by problems arising from variations in the properties of the languages. For document translation MLIR, if the document index data are located physically together, the index files can be combined to form a single search collection. This removes the need for merging of separate lists. However, if the collections are distributed or query translation is being used, some method of merging must be adopted.

A variety of list merging algorithms of varying complexity have been proposed for distributed IR. A number of these have been applied for MLIR with varying degrees of success. The simplest approach involves ignoring the score incompatibility problem, and simply merging the ranked lists using their raw scores. More complex methods involve ranking the separate collections in terms of their estimated likelihood of containing relevant documents, combining these collection matching scores with the matching scores of individual documents to form a composite score, and using this combined score to generate the final merged document list. These methods have been shown to be effective for

monolingual distributed IR [Callan, 2000]. Unfortunately, they have not proved so successful for MLIR, where it has been difficult to improve performance beyond that achieved using the simplest methods [Lam-Adesina & Jones, 2003; Savoy, 2004].

In our experiments for the CLEF workshop MLIR task in 2003, we translated all the documents from their original languages of French, German and Spanish into English using machine translation. We then compared retrieval effectiveness of various list merging strategies with that for a single collection formed from the translated documents. Overall we found that the single collection method worked best indicating that all the merging strategies fell short of the performance that could potentially be achieved using these document sets [Lam-Adesina and Jones, 2003]. Once again our results showed that the BM25 Okapi probabilistic model produced among the best retrieval effectiveness for this task. Of course it will not always be possible to translate the entire retrieval collections and then combine them. More recent experiments using the CLEF 2003 MLIR tasks have shown that list merging can produce good retrieval results [Si and Callan, 2006]. However, merging remains an important ongoing concern for MLIR requiring further investigation.

6.3.3 Multilingual Web Retrieval

In recent years significant effort within the information retrieval research community has focused on the development of effective methods for retrieval of web content. This has gained momentum since the late 1990's, but is still a young area of research, and although many important results have already been attained, open problems remain that require further research, as is observed in Melucci and Hawking [2006].

Given its world-wide coverage, it is no surprise that the Web is inherently multilingual. Dominant world languages are all well-represented on the Web. Some multilingual Web content is created by translation between languages but, predominantly, documents appear in the languages they were originally authored in. The result is a heterogeneous body of information in which is content available in one or more languages with no guarantee that it will be duplicated in another language. The importance of developing approaches to improve access to multi-language Web collections has been recognized by the international research community, which has established exercises such as the Web track at CLEF, which promotes synchronization between researchers working in the area by developing systematic tasks, test-suites and evaluation of web content [Sigurbjörnsson et al. 2005; Balog et al. 2006].

Not only is the content of the Web multi-lingual, the users who wish to access this content are also polyglots [Sigurbjörnsson et al. 2005]. Especially in Europe, many users are able to make use of information presented to them in a range of languages. These users are quick to make use of the passive knowledge that they may have of a specific language, especially in cases when they realize that the information that they need is not available in another language. In addition, machine translation techniques offer a huge potential to support users in making use of information in languages that they do not understand at all.

Like classic information retrieval, web retrieval attempts to provide a user with information that satisfies an information need. However, many users undertake web search to find a particular URL or to perform a particular transaction rather than to find information [Broder 2002]. Also, frequently users like to browse in web collections, which means that web retrieval research must also focus on the question of providing information to a user who has no clearly formulated information need, but instead requires an overview of an area. One particularly challenging task is to provide web retrieval techniques that will support users who are browsing with the purpose of discovery of new subject areas that they were previously unaware of, or who are interested in finding new connections between topics that they are already familiar with.

Other differences between retrieval in digital libraries containing text documents and search in Web content concern the difference in the nature, structure and volume of information available on the Internet, as discussed by [Baeza-Yates and Ribeiro-Neto 1999]. On the Internet, data is changing

constantly. Because it is produced by a variety of sources, both professional and informal, Web data is fundamentally heterogeneous and its quality is variable. The amount of data available on the Internet is unrivalled in volume, constituting a particular challenge for Web search. Finally, data available on the Internet is distributed, meaning that before it can be indexed it must be gathered. Gathering of data requires a web crawler to discover and fetch web content so that it cab be indexed for searching. A challenge for the future is to design and implement web crawlers, whose efficiency stems from their intelligence, i.e. their ability to crawl only that material that will later be relevant to the information needs of the users of the search engine they were designed to feed. This issue is important for the harvesting of content for the MultiMatch search engine where crawled content should be drawn from the broad domain of cultural heritage.

Web retrieval can exploit normalization and term extraction techniques that have been developed for classic text retrieval, but also makes use of characteristics particular to Web content. Web retrieval makes critical use of the fact that web pages do not exist as isolated entities, but are connected to each other via hyperlinks. The most well known exploitation of this link structure is the PageRank algorithm which formed the starting point for the development of the Google search engine (see Chapter 3). A future direction for Web retrieval is to make full use of the structural information provided by the tree structure of XML documents and by the information contained in the XML tags. Fielded indexes that index path-tagged terms have demonstrated great potential and the future will surely see optimization of such techniques.

Alongside search engines which accept free text queries from users and deploy automatic methods to determine relevant websites, search engines based on hand crafted web-categories are also being developed. Such search engines supply users with high quality information, but suffer from the disadvantage that they do not provide wide coverage since the classification of sites into categories has to be done by hand and is very time consuming [Baeza-Yates and Ribeiro-Neto 1999]. A research direction for the future is to pursue approaches that will deliver the benefits of category-based search, but with reduction or near-elimination of human effort.

The Internet has witnessed the development of a profusion of search engines, each deploying its own crawler and its own search strategies. For this reason, the results delivered by one search engine have a great potential to complement the results returned by another. The bundling of search engine results is another important area of investigation for researchers involved with web retrieval.

The Internet is characterized by the existence of user communities which create content and interact with one another. The structure of these communities is an important source of information. Some communities engage in concerted effort to label web sites that are relevant to their interests with tags that will make them easily retrievable. Log files of user behaviour is another source of information. Patterns of previous searches can be used to refine future searches. For some types of searches, it is critical that a search engine returns reliable information to the user. Although every query deserves a reliable result, travel and medical queries can be particularly critical. For this reason, it is important to analyze the quality and the authority of web pages and for search engines to be aware that content providers may be trying to trick them into indexing pages that are not truly authoritative. (add probably several relevant citations which cover these points).

In sum, techniques required to tackle the challenge of web retrieval encompass, but extend approaches to text retrieval. Understanding how users formulate their information needs into queries for web search and exploitation of the particularities of web content are both necessary if web retrieval technology is to advance into the next generation. Web retrieval research stands to gain by embracing the multilingual nature of the Internet and leveraging complementary sources of information in multiple languages.

6.4 Multimedia Information Retrieval

The current expansion in archives of digital multimedia content is creating the need for tools to automatically search and retrieve material from these collections. Similar to the work on multilingual text documents, recent years have seen a rapid increase in research exploring Multimedia Information

Retrieval (MIR). Multimedia archives comprise material in one or more of audio or visual media, often accompanied by some form of manual electronic text annotation or metadata. Retrieval from these collections raises a number of issues with respect to both the indexing and retrieval processes. Multimedia content can be either static, in case of individual digitized images such as photographs or paintings, or temporal, comprising audio and/or video content. The static or temporal nature introduces various concerns with respect to the presentation to the user and browsing of retrieved content.

Indexing and retrieval methods for MIR depend on the media under consideration. Let us consider these in order of increasing complexity. Electronic text material available for MIR can either take the form of metadata or direct transcription of content. Metadata may describe the content in some way, e.g. the names or roles of the characters appearing in an image, or the events taking place in a video. Transcriptions of linguistic content may be generated manually or automatically. For example, the close captioning often broadcast with TV sources can be captured and used as a high quality transcription of the content for the purpose of retrieval and browsing.

Existing IR research has focussed very much on linguistic content, and so can in general be applied directly to manually annotated material associated with multimedia content. The usefulness of manually entered descriptive metadata will depend on the quality of the data, and its usefulness in addressing an individual need. Thus, while the visual content of an image may make it relevant to a particular request, if the descriptive metadata is not pertinent to the aspect of this item which makes it relevant, then the MIR system will fail to locate it. Therefore, the effectiveness of MIR will clearly be affected by the accuracy and richness of the annotation. Additionally, the complexity of the retrieval methods used for textual annotations may be influenced by their form; if the annotations are highly structured, this may be taken into account in the retrieval algorithms adopted.

Of more interest within recent and current research, is MIR based on automated annotation of the content. The following sections consider indexing and retrieval for first spoken documents, and then image and video data.

6.4.1 Spoken Document Retrieval

In many situations it is uneconomic or impractical to manually transcribe the spoken contents of multimedia documents, and thus transcriptions must be generated automatically using speech recognition technologies. Forming transcriptions in this way using current speech recognition tools has a number of limitations. The most significant issue is that, like machine translation systems used for CLIR, these tools make mistakes; incorrect words can be inserted into the transcription, correct words deleted, or one word incorrectly substituted for another one. These errors arise for a number of reasons relating to both the natural language data and the tools themselves. Speech recognition is inherently challenging for a number of reasons including the following: the speech may be poorly articulated, it may not follow expected linguistic patterns, it may be captured using poor quality equipment, there may be high levels of background or environmental noise, or there may be crosstalk where more than one speaker is talking at the same time. The accuracy of a speech recognition system is limited by the effectiveness of its acoustic models to accurately recognise the sound patterns of the current speaker, and of its language models to predict their use of word patterns. Current speech recognition transcription systems are also correctly described as “large vocabulary”, where only the words within a predefined vocabulary can be recognised correctly; other so called “out-of-vocabulary” words will be transcribed incorrectly by definition. In general, the overall accuracy of an automatically generated document transcript will depend on the extent to which the speech deviates from the trained parameters of the speech recognition system and the quality of the input speech signal.

The effect of recognition errors is to produce a “noisy” transcription which will have some similarities to the output of a machine translation system. The characteristics of the errors however are likely to be somewhat different. A machine translation system can determine its output, although it may experience problems with the naturalness of the word patterns generated, or be subject to limitations in the richness of the available vocabulary or linguistic structures. By contrast, a speech recognition system must do its best to transcribe the data presented to it. Automatic transcriptions often include

apparently random insertion and deletion errors. A potential problem for both machine translation and speech recognition though is how to appropriately handle input words outside their vocabulary.

Research into spoken document retrieval (SDR) began with a number of projects in the early 1990s. These examined various approaches to automatically indexing the spoken contents and were evaluated using locally developed test collections [Glavitsch & Schäuble, 1992; Jones, Foote, Sparck Jones, & Young, 1996]. When these projects started, the potential of IR techniques derived from experience with electronic text documents to transfer successfully to errorful spoken document index files was very much an open question.

It is a feature of speech recognition that the hardest words to recognise accurately are often short function words. Of course, these are generally not useful for retrieval, and hence SDR systems can still operate with good reliability in the presence of relatively high word recognition error rates. A further issue is that since important words within a document are often repeated, even if the word is recognised incorrectly when it occurs in one place, it may be correctly recognised elsewhere in the document. Whilst errors of this type will degrade the overall quality of term weights, the documents will still be retrieved. This distortion of term weights can result in some distortion of the ranked retrieval list, relative to that which would be achieved with a perfect document transcription, but overall high levels of retrieval effectiveness can still be achieved.

Interest in SDR increased significantly in the mid-1990's and a track was introduced at the annual TREC series in 1997. For the first time researchers were able to work with a common SDR test collection. The SDR track ran for 4 years, each conference increased the document collection size or the complexity of the retrieval task. During this time speech recognition technologies continued to advance. Using the best available transcription systems, achieving recognition average word errors rates of around 20% with a vocabulary of around 65,000 words, together with the BM25 model and retrieval enhancement techniques, such as relevance feedback and merging with in-domain large contemporaneous text collections, TREC SDR participants demonstrated similar overall retrieval effectiveness for manual and automatic document transcriptions [Johnson, Jourlin, Sparck Jones, & Woodland, 2001] [Garofolo, Auzanne, & Voorhees, 2000]. The success of the TREC SDR track indicated, at least for a task where the transcription system can be well trained for the domain of the document collection, in this case broadcast news, that SDR is effective using current speech recognition technologies.

More recently the Cross-Language Speech Retrieval (CL-SR) task at CLEF in 2005 and 2006 has explored speech retrieval for a more challenging document collection in a cross-language framework. Each document consists of multiple fields consisting of: an automatic transcription made with a large vocabulary automatic speech recognition system adapted to the domain of the data, a number of keywords assigned automatically based on these transcriptions, manual assigned keywords, a short manual summary of the document and a manually assigned list of proper nouns appearing in the actual audio of the document. This document set thus poses the challenges of SDR, but also the combination of multiple fields for effective retrieval. The optimal way of doing this is not obvious as explained in [Robertson et al, 2004]. Cross-language experiments carried out by the participants in the CLEF tasks show that speech retrieval behaves similarly to standard text retrieval in cross-language tasks; that is problems of translation between search requests and documents result in a reduction of retrieval effectiveness of between 10% and 20% [White et al, 2006].

6.4.2 Image and Video Retrieval

Whereas it is natural to use the same indexing units for spoken content and written linguistic content, the appropriate mechanism for indexing and retrieving from visual media is much less clear. Visual content can include natural scenes either in static images or moving video, as well as other image content, for example scanned or overlaid textual material.

Considering first the more straightforward case of textual content in images. The first stage in automatically indexing this material is to identify zones or regions in the image containing text. The text in these zones is then recognised using an optical character recognition (OCR) process. After this,

it can be indexed using a standard retrieval approach derived from experience with electronic text documents. Unfortunately, similar to speech recognition systems, OCR systems make mistakes; although the errors in this case are often of a different form. Instead of making whole word recognition errors, as is the case for speech recognition, OCR systems typically make errors in the recognition of individual characters. Each of these errors will usually introduce a new word into the indexing vocabulary of the collection. These words will not be useful indexing terms, since they will not match correctly with terms appearing in typed search requests, and they will also have disproportionately high collection frequency weights, since they are very rare within the document collection. A simple way to resolve this problem might be to attempt to correct automatically the spelling of these words using a dictionary. However, it is not always clear what the correct word should be. Indeed sometimes a word not present in the dictionary will actually have been correctly recognised by the OCR system, and attempting to correct OCR errors in this way may replace these accurately recognised words with incorrect words taken from the dictionary. As a consequence of this problem, "correcting" the OCR output with a dictionary may lead to a degrading of retrieval effectiveness. Another issue, similar to spoken document recognition, is that the accuracy of the output of an OCR system will be related to the difficulty of the recognition task. OCR accuracy will depend on the quality of the printing, the fonts used, and the contrast between the print and the paper. For example, modern laser printed output with a simple font is easier to recognise than older mechanically printed documents for which the paper may be yellowing with age. Significantly more difficult to recognise accurately is handwritten text, for which accuracy will obviously depend on how clearly it has been written, as well as the other factors affecting printed text [Rath, Manmatha, & Lavrenk, 2004]. Interestingly, while relevance feedback has been shown to be very effective for SDR [Johnson et al., 2001], the differences in error types encountered between OCR and speech generated transcripts, mean that it does not transfer to scanned text documents in a simple way and correction techniques must be applied to make it effective for this task [Lam-Adesina & Jones, 2006].

A much less well defined task is the retrieval of multimedia documents based on non-linguistic visual content. When examining a visual scene, we might want to identify any number of different features. For example, we may wish to recognise the individuals appearing in the image, the place where the scene is taking place, the objects in the picture, or perhaps the events being depicted. Identifying these features is very difficult. Indeed doing this in a robust way outside a very narrow pre-defined domain is currently not possible. Much visual media can be interpreted in a seemingly unlimited, often subjective, number of ways. This type of intelligent analysis will be beyond analysis of visual features alone, often requiring knowledge outside that available in the visual content itself. Of course, texts can frequently be interpreted in many ways as well, but for retrieval purposes, word level indexing has generally been shown to be effective without needing to determine any particular interpretation of the text. In the case of images, not only are attempts at recognising features unreliable, there is no obvious parallel means of selecting indexing units for open domain retrieval. Current video media retrieval systems either focus on very narrow domains, for example identifying pictures of predefined named individuals, or seek to index images using low-level features, such as colour or texture. Indexing images using such low-level features is perhaps comparable to identifying the letters in a text document without determining what the words are. A detailed summary of much of the work carried out in developing image and video retrieval technologies is described in [Smeulders et al, 2000], Much research is currently devoted to the segmentation of images into meaningful regions or to detect objects without extensive training to identify specific object types.

The difficulty in indexing images and of specifying search queries for them means that retrieval of visual media inherently requires more user interaction than text retrieval. For MIR systems, a user will typically initiate a search either using a text request which will locate some potentially relevant images or video based on their textual annotation, or they will select a sample image and request the retrieval system to "find me more like this", in response to which the system returns images with similar colour and texture profiles to those of the example. The user is then able to provide feedback on the images retrieved using this initial query, after which further searches are carried out, with feedback after each one, until the user's information need has been satisfied. Such "more like this" searches are typically based on generic MPEG-7 low-level image features of: global colour, regional colour, texture and

edges within the image. Some current research is extending this to explored interactive use of objects to enable users to select a combination of standard image features and detected objects in building more complex queries for feedback [Sav et al, 2006].

A significant challenge for MIR is the combination of the visual features with the textual metadata to provide an overall search output. Simple approaches to this are based on a simple data fusion strategy of forming separate ranked lists for each feature and then adding them in a weighted scalar sum. This is a simple strategy but can be effective, although it is important to assign the correct weights to each feature list. This also true of data fusion for text only retrieval, but is probably more crucial for MIR where the importance of individual features will be quite different for individual queries. For example, for one query colour of the query image may be important in finding relevant documents, whereas for another query a combination of metadata text and image texture may be important. A method to automatically select query dependent optimal features weights is introduced in [Wilkins, Ferguson and Smeaton, 2006].

While the above late fusion mechanism proves effective, it is important to define early approaches whereby the relevant features are combined at an early stage, thus enabling truly multimodal query. Important shortcomings however are the heterogeneity and normalisation of the features to combine. [Bruno *et al*, 2006] propose a distance-based learning strategy to combine multimodal feature at query time. Features are homogenized by considering relative distances rather than absolute values. A new representation space is thus created by an appropriate choice of pivot-like points. Efficient non-linear learning techniques (SVM, KFD) may then operate interactively within such a feature space based on user feedback to isolate portions of population relevant to the query.

The discussion so far really assumes that retrieval is of images with may be annotated with textual metadata. For video retrieval some additional processing and modelling is often required. Video is typically composed of events or scenes which are composed of a sequence of camera shots. Standard video processing typically first locates the shot boundaries, points at which the camera changes. Some camera changes are easy to locate others, such as gradual fades, can be problematic. Once shots have been identified, the next stage in video processing is typically to identify a single representative frame or "keyframe" for the shot. Retrieval for the shot then proceeds exactly as for static image retrieval using the keyframe. This of course assumes that a keyframe can be located which sufficiently represents the shot, such that it contains features which represent aspects of the shot that are going to appear in query images for which the shot is relevant. For some shots temporal features of the shot may be important in describing it, and in order to use this the temporal aspect of the image must be captured in some way.

Video shots are editing entities that may not be fully appropriate for video retrieval. A concept more advanced and probably more suited than the shot for searching is that of the *story*, somewhat close to the textual *topical* segmentation. In that case, the partition must be done according to semantic criteria gathered from a multimodal inspection of the streams (see e.g. Janvier *et al* [2005]). Semantic units are then said to be more appropriate for gathering relevance feedback than simple shots. The challenge here is to form a proper characterisation of the temporal evolution of the semantic information from multimodal features.

Since 2001 the TRECVID workshop has provided standard document collections for researchers to explore indexing and retrieval tasks for video data [Smeaton, Kraaij, & Over, 2004]. Tasks undertaken in TRECVID include: automated shot boundary detection, story boundary detection, visual feature recognition, locating named individuals or events in video, and interactive searching of a video archive. TRECVID is proving instructive in the development and evaluation of MIR technologies, but perhaps the clearest message so far is the large amount of work that remains to be done to achieve mature MIR systems.

6.4.3 Hybrid Searching for Multi-field Documents

The foregoing discussion has assumed that searching is based on a simple best-match ranked retrieval strategy. However, as has been mentioned a number of times documents are often accompanied by a

range of metadata fields, such as date of creation, author, publisher or publication venue. A common approach to exploiting this data in the search process is simply to fold it into the main document text field and use the attributes as search features. However, they can often be used instead, or additionally, as constraints on the search. For example to retrieve documents only published by a certain source or written by a named author within a specified time frame. Where the user has the requisite knowledge to impose these constraints limiting the document search space in this way can have significant benefits in terms of retrieval precision and efficiency of browsing. This can be particularly useful in multimedia environments where interactive constraints, particularly in audio browsing, mean that reducing the amount of material that must be explored in particularly useful [Brown et al, 1996].

6.5 Concluding Thoughts and Future Challenges

This chapter has demonstrated how fundamental work on English language text information retrieval has been successfully applied for multilingual and multimedia documents. For text retrieval in a new language it has been illustrated that the need is for the selection of appropriate indexing units and development of automatic indexing methods, including morphological processing, stop word lists, and suffix stripping algorithms. Research issues for CLIR relate primarily to translation methods to cross the language barrier between search requests and documents. In MultiMatch, we will advance automated translation in the CH area by using parallel corpora such as bilingual or multilingual metadata to automatically construct domain-specific dictionaries. These dictionaries will then be incorporated into a translation system with MT modules in order to translate search requests and CH documents.

For MLIR issues of translation are compounded with the need for effective merging of the document lists retrieved from different language collections. MultiMatch will investigate various techniques to find the optimal merging strategy for the CH domain and the multilingual indexes with which we will work. Speech and scanned text document retrieval have been shown to be remarkably robust to indexing errors in automatic recognition of their content. Research will be undertaken in MultiMatch to ascertain the most appropriate means of handling Cultural Heritage documents of these types. The ongoing issues of defining and recognising visual indexing features continue to be the focus of much research in visual media retrieval. However, there is already research underway exploring the use of the alternative language modelling approach to IR in visual retrieval [Westerveld & de Vries, 2004].

Solution of the problems of multilingual and multimedia information retrieval explored in this chapter does not represent the end of the story for research into information access technologies for this data. Research interest continues to evolve to embrace more challenging tasks. For example, work is currently being established in the areas of retrieval from multilingual collections of image and video archives, retrieval from multilingual web collections, and question-answering methods for multilingual and multimedia data.

References

Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier. (1999). Modern Information Retrieval. Addison Wesley. 1999.

Ballesteros, L., & Croft, W. B. (1998). Resolving Ambiguity for Cross-Language Retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 64-71, Melbourne, ACM.

Balog, K, Azzopardi, L., Kamps, J. & de Rijke, M. (2006). Overview of WebCLEF 2006. In Carol Peters, editor, Working Notes for the CLEF 2006 Workshop,2006. http://www.clef-campaign.org/

Braschler, M., & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval? Information Retrieval, 7(3-4), 291-316, Kluwer.

Broder, Andrei. A Taxomony of Web Search. (2002). SIGIR Forum, 36(2) 2002

Brown, M.G.,Foote, J.T.,Jones,G.J.F., Sparck Jones,K. and Young, S.J. (1996) Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval, Proceedings of ACM International Conference on Multimedia, Boston, U.S.A., pp307-316, ACM.

Bruno E., Moënne-Loccoz N., and Marchand-Maillet, S. (2006) Asymmetric learning and dissimilarity spaces for content-based retrieval. In CIVR, pp 330-339.

Callan, J. (2000). Distributed Information Retrieval. In W. B. Croft, editor, Advances in Information Retrieval, pp. 127-150. Kluwer.

Chakrabarti, Soumen. (2003). Mining the web: Discovering knowledge from hypertext data. Morgan Kaufmann. 2003.

Garofolo, J. S., Auzanne, C. G. P., & Voorhees, E. M. (2000). The TREC Spoken Document Retrieval Track: A Success Story. In Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access, pp. 1-20, Paris.

Gey, F., Kando, N. & Peters, C. (2002). Cross language information retrieval: a research roadmap. SIGIR Forum, 36(2) 72-80.2002.

Glavitsch, U., & Schäuble. P. (1992). A System for Retrieving Speech Documents. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 168-176. ACM.

Gollins, T., & Sanderson, M. (2001). Improving Cross Language Retrieval with Triangulated Translation, In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Rretrieval, pp 90-95, New Orleans, ACM.

Grossman, David and Frieder, Ophir. Information Retrieval: Algorithms and Heuristics. Springer. 2004.

Huang, X., & Robertson, S. E. (1997). Application of Probabilistic Methods to Chinese Text Retrieval. Journal of Documentation, 53(1), 74-79.

Hull, D. A., & Grefenstette. G. (1996). Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 49-57, Zürich, ACM.

Janvier B., Bruno E., Marchand Maillet S., and Pun T. (2005). A contextual model for semantic video structuring. In 13th European Signal Processing Conference, EUSIPCO'05, Antalya, Turkey.

Johnson, S. E., Jourlin, P., Sparck Jones, K., & Woodland, P. C. (2001). Spoken Document Retrieval for TREC-9 at Cambridge University. In E. M. Voorhees and D. K. Harman, editors, Proceedings of the Ninth Text REtrieval Conference (TREC-9), pp. 117-126. NIST.

Jones, G. J. F., Foote, J. T., Sparck Jones, K., & Young, S. J. (1996). Retrieving Spoken Documents by Combining Multiple Index Sources. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 30-38, Zürich,. ACM.

Jones, G. J. F., Sakai, T., Kajiura, M., & Sumita, K. (1998). Experiments in Japanese Text Retrieval and Routing using the NEAT System. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 197-205, Melbourne, ACM.

Jones, G. J. F., Sakai, T., Collier, N. H., Kumano, A., & Sumita, K. (1999). A Comparison of Query Translation Methods for English-Japanese Cross-Language Information Retrieval. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 269-270, San Francisco, ACM.

Jones, G. J. F., & Lam-Adesina, A. M. (2001). Exeter at CLEF 2001: Experiments with Machine Translation for bilingual retrieval. In Proceedings of the CLEF 2001: Workshop on Cross-Language Information Retrieval and Evaluation, pp. 59-77, Darmstadt, Springer Verlag.

Lam-Adesina, A. M., & Jones, G. J. F. (2003). Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual and Bilingual and Multilingual Retrieval. In Proceedings of the CLEF 2003: Workshop on Cross-Language Information Retrieval and Evaluation, Trondheim, Springer.

Lam-Adesina, A.M. and Jones, G.J.F., (2006) Using String Comparison in Contact for Improved Relevance feedback in Different Text Media, In Proceedings of the 13th Symposium on String Processing and Information retrieval (SPIRE 2006), Glasgow, Scotland, pp229-241, Springer

McCarley, J. S. (1999). Should we Translate the Documents or the Queries in Cross-language Information Retrieval. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 99), pp. 208-214, University of Maryland, MD, ACL.

Melucci, M. & Hawking, D. (2006). Introduction. A perspective on Web Information Retrieval. Information Retrieval Vol 9. 119-122. 2006

Nie, J.-Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In Proceedings of the 22nd AAnnual International ACM SIGI Conference on Research and Development in Information Retrieval, pp. 74-81, San Francisco, ACM.

Ponte, J. M., & Croft, W. B. (1998). A Language Modelling Approach to Information Retrieval. In Proceedings of the 21st Annual International ACM SIGIR International Conference on Research and Development in Information Retrieval, pp275-281, Melbourne, ACM.

Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14, 130-137.

Rath, T., Manmatha, R., & Lavrenko, V. (2004). A Search Engine for Historical Manuscript Images. In Proceedings of the 27th Annual International ACM SIGIR International Conference on Research and Development in Information Retrieval, pp369-376, Sheffield, ACM.

Robertson, S. E. (1977). The Probability Ranking Principle in IR. Journal of Documentation, 33, 294-304.

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science, 27, 129-146.

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 232-241, Dublin, ACM.

Robertson, S. E., Walker, S. & Beaulieu, M. M. (1999). Okapi at TREC-7: automatic ad hoc, filtering, vls and interactive track. In E. Voorhees and D. K. Harman, editors, Proceedings of the Seventh Text REtrieval Conference (TREC-7), pp. 253-264. NIST.

Robertson, S.E., Zaragoza, H., and Taylor, M (2004) Simple BM25 Extension to Multiple Weighted Fields, Proceedings of the 13th ACM International Conference on Information and Knowledge Management, pages 42-49, ACM.

Sakai, T., Koyama, M., Kumano, A., & Manabe, T. (2004). Toshiba BRIDJE at NTCIR-4 CLIR: Monolingual/Bilingual IR and Flexible Feedback. In Proceedings of NTCIR-4.

Salton, G, & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24, 513-523, Elsiver.

Sav, S., Jones, G.J.F. Lee, H., O'Connor, N.E., and Smeaton, A.F., (2006t) Interactive Experiments in Object-Based Retrieval, In Proceedings of the 5th International Conference on Image and Video Retrieval (CIVR 2006), Tempe, AZ, U.S.A., pp,1-10, Springer.

Savoy, J. (2004). Combining Multiple Strategies for Effective Monolingual and Cross-Language Retrieval. Information Retrieval, 7(1-2), 121-148, Kluwer.

Sheridan, P. & Ballerini, J. P. (1996). Experiments in Multilingual Information Retrieval using the SPIDER system. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 58-65, Zürich, ACM.

Si, L. and Callan, J. (2005.) CLEF 2005: Multilingual retrieval by combining multiple multilingual ranked lists." In Sixth Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Vienna, Austria.

Sigurbjörnsson, B., Kamps, J. & de Rijke, M. (2006). Overview of WebCLEF 2005. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, Henning Müller, and Maarten de Rijke, editors, Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language

Evaluation Forum (CLEF 2005), volume 4022 of Lecture Notes in Computer Science, pages 810-824. Springer Verlag, Heidelberg, 2006.

Smeaton, A. F., Kraaij, W., & Over, P. (2004). The TREC Video Retrieval Evaluation (TRECVID);' A Case Study and Status Report. In Proceedings of RIAO 2004 – Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, pp. 25-37, Avignon.

Smeulders, A.W., Worring, M., Santini, S., Gupta, A.,and Jain, R. (2000) Content-Based Image Retrieval at the End of the Early Years, IEEE Trans. Pattern Analysis and Machine Intelligence, 22(12):1349--1380, IEEE.

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000a). A probabilistic model of information retrieval: development and comparative experiments: Part 1. Information Processing and Management, 36(6), 779-808, Elisver.

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000b). A probabilistic model of information retrieval: development and comparative experiments: Part 2. Information Processing and Management, 36(6), 809-840, Elisver.

van Rijsbergen, C. J. (1979). Information Retrieval. (2nd edition) Butterworths.

Wechsler, M., Sheridan, P., & Schäuble, P. (1997). Experiments in Multilingual Information Retrieval using the SPIDER System. In Proceedings of the 5th RIAO Conference, Computer-Assisted Information Searching on the Internet, Montreal.

Westerveld, T. & de Vries, A. P. (2004). Multimedia Retrieval Using Multiple Examples. In Proceeding of the Third International Conference on Imag and Video Retrieval, pp. 344-352, Spinger.

White, R. W., Oard, D. W., Jones, G. J. F., Soergel, D., and Huang,X (2006): Overview of the CLEF-2005 Cross-Language Speech Retrieval Track, Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation, Vienna, Austria, pp. 744-759, Springer.

Wilkins, P., Ferguson, P. & Smeaton, A.F. (2006) Using Score Distributions for Querytime Fusion in Multimedia Retrieval}, Proceedings of MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval}, Santa Barbara, CA, ACM.

7. User Interaction & Interface Design.

by Paul Clough with contributions from Jennifer Marlow and James Carmichael

"Each new piece of information [users] encounter gives them new ideas and directions to follow, and, consequently, a new conception of the query."

Bates' berrypicking model for information-seeking [Bates, 1989].

The interface acts as the intermediary between users of information retrieval (IR) systems and the search system. In designing an interface for an IR system, the goal is to enable users to satisfy an information need without the assistance of a human intermediary [Brajnik et al., 1996]. A well-designed interface should assist users in clarifying their information needs, and subsequently help them formulate suitable queries and understand the results [Hearst, 1999; Shneiderman, 1997]. More recently, attention has been paid to human-computer interaction in information retrieval and interface design has been driven by the needs of end users, their information-seeking behaviour and psychological aspects of the users (see, e.g. Ingwersen & Järvelin 2005; Marchionini, 1992; Bates, 1989].

Belkin [2003] points out certain important aspects of functionality in information system design and in particular identifies two issues required to support users with information seeking tasks: (1) designing systems that support a variety of interactions and (2) personalizing the support for user interaction. The former suggests that systems should be designed with a holistic view of information seeking, e.g. adding a workspace to store items *between* individual searches and providing multiple functionalities. The latter recognises that aspects of search, such as a preferred ranking of documents, can be inferred from prior interactions between the user and the system.

Current interface design is linked strongly with research in Interactive Information Retrieval (IIR) that provides the necessary theories and frameworks for modelling user behaviour. Although a little dated in terms of describing the current state of the art, Hearst [1999] still provides an excellent general overview of user interfaces and interaction design for information retrieval systems.

7.1 Information Seeking and General Search Interfaces

With regards to search engine interfaces, it has been said that "Nearly every Web search engine offers users the identical search experience, regardless of the task they are trying to accomplish" [Rose, 2006: 797]. In order to create a more tailored and flexible search experience, users' needs and goals should be taken into consideration, in order to determine not only what users are searching for, but also why they are searching [Rose & Levinson, 2004].

People have different information needs and they make use of various information seeking strategies to solve those problems. For example, Broder [2002] analysed a large collection of queries from a search engine log and found at least three types of information need: *navigational* (find the URL of a specific web site, e.g. "BBC"), *informational* (find some information) and *transactional* (find a structured service to initiate further interaction). Rose and Levinson [2004] refined this work to create a hierarchy of users' goals. Henniger and Belkin [1996] describe analysing the process of satisfying information needs as a decision-making problem in which users learn and refine their needs as they interact with a repository.

Analysing the behaviour of users as they search for information provides informative and valuable insight into user interface design. For example, Gremett [2006] showed how an analysis of users shopping on Amazon.com revealed that in practice users would commonly mix searching and browsing while buying online products. Marchionini [1995] calls this a *mixed behaviour* strategy of information seeking in which a user searches for information by both navigational browsing and searching a site via some explicit search tool such as a search box.

In modern IR research, more emphasis is being placed on constructing models of the search process which go beyond a simplistic view of search as a one-shot matching function between the user's query

and collection of documents. Bates [1989] describes search as an interactive process that evolves in response to the information found: results from a search are not just documents, but also the knowledge accumulated along the way. Bates identifies different strategies that people follow during search including following relationships between documents (e.g. hyperlinks) or browsing over the structure of a collection. She suggests that IR interfaces would be more useful if these search strategies were supported at a higher level. Therefore, both search and browse functionalities should be present and tightly integrated, in order not to interrupt a user's exploration [Beale, 2006; Hearst et al., 2002].

Rose [2006] suggests there are three general areas in which knowledge of information seeking behaviour could inform the design of the user interface for Web search: (1) the goal of the user when conducting a search, (2) the cultural and situational relevance, and (3) the iterative nature of the search task itself. Recognising that users perform different tasks and understanding the user's goals would enable appropriate support mechanisms to be included in the interface design.

Users have different information needs, e.g. getting a specific piece of information, getting an answer to a question, getting advice and exploring a general topic. Modelling user's behaviour would enable provision of the most suitable support rather than creating a one-fits-all interface.

Recognising the search context is also important as the same query may have different meanings in different cultures or sub-communities (e.g. a user searching with the query "Madonna and baby" could have in mind the pop star if a music fan, or the painting if an art historian). Different results may also be relevant to the same user at different times. Interfaces offering localisation (e.g. ranking documents with country-specific URLs higher) could help support this.

Bates [1989] suggests that search is best modelled as an iterative process and that retrieval forms part of a dialogue between the user and system to gradually refine the results. Interface support for iteration could include relevance feedback in image retrieval, or lists of related query terms for query expansion or reformulation in text searching. Rose summarises by suggesting that user interfaces should provide different interfaces or forms of interaction to meet users' search goals, allow the user to select appropriate contexts for the search (e.g. language, search options, preferences), and support the iterative nature of the search task by inviting iteration and exploration.

Hearst [1999] notes that often when searching or browsing, individuals may become distracted and temporarily follow alternate paths. For this reason, it is recommended to provide ways of recording past queries and offering a means of storing intermediate results throughout the search. This also helps to reduce short-term memory load [Shneiderman et al., 1997, in Hearst et al., 2002].

White et al. [2006] also advocate the development of systems to support users who are engaged in exploratory search activities (i.e. those without a pre-defined or specific search task). Henninger and Belkin [1996] review current systems in terms of the key interface and interaction techniques such as querying, browsing and relevance feedback (to support the iterative refinement of the user's information need). They also advocate the use of task modelling and interaction modelling as key strategies to improve the design of retrieval systems.

Hearst et al. [2002] cite common search problems such as receiving empty results sets or disorganised result lists, and having difficulty forming special-syntax (Boolean) queries. Therefore, useful means of combating these problems can include providing suggestions for improving the query (if no results have been returned,) showing keywords in context, and giving brief search hints.

Regarding principles for future design interfaces, Rose [2006] advocates making different interfaces available to match different search goals. Another area to investigate is how to improve the browsing process, particularly because the common practice of displaying category lists takes up large amounts of space and often requires a user to guess which category heading will contain the related information of interest [Hearst, 1999].

Although related to Web search, the suggestions from Rose [2006] match existing best practices in designing interfaces to support information seeking. Resnick and Vaughn [2006] describe a set of best practices developed to assist in the design of search interfaces, these design principles are organised

into five domains: the corpus, search algorithms, user and task context, the search interface and mobility. Best practices include the use of faceted metadata [Hearst et al., 2002] within a controlled corpus, the use of spell-checking during user input, hybrid navigational support through combined search and browse, the use of past queries to frame the search context, the provision of a large query box (also confirmed by Belkin et al [2000] for more expressive queries), the organization of a large set of search results into categories, showing the keywords in context in search results and designing alternate versions of content specifically for mobile and handheld devices.

In summary, the emphasis on modern search engine interface design is on understanding and modelling the user's needs, identifying functionalities to support those needs and implementing systems which support the dynamic nature of the user's tasks and searching activities. These issues will be taken into consideration while designing the interfaces for MultiMatch.

7.2 Multilingual Information Access (MLIA)

There are multiple sides to providing multilingual information access (MLIA) and supporting interaction with users. These can range from adapting existing information for use by local communities to providing cross-language search. Current research is focused on aspects such as the design and usability of websites [Del Galdo & Nielsen, 1996; Yunker, 2003] and the provision of multilingual search functionalities [Oard, 1997].

7.2.1 Localisation (and Multilingual Interfaces)

On the Internet, adapting websites to meet the linguistic and cultural needs of the local communities they target is referred to as *globalisation*. The different versions are known as *localised* websites and often require specific design considerations (W3C, 2003; Eurescom, 2000; Del Galdo & Nielsen, 1996; De Troyer & Casteleyn, 2004]. These might include: identifying which languages a website should be translated into, an awareness of cultural issues (e.g. the use of specific terminology or offensive references), the availability of resources (e.g. manpower, translation tools), technical and maintenance issues, how to measure success and issues surrounding design. The W3C [2003] differentiate between *international* and *multilingual* websites: the former being defined as a website which is intended for an international audience while the latter is a website which uses more than one language. According to this definition, a multilingual site is also concerned with regional and cultural differences in addition to language. International sites are often multilingual, e.g. a global company with information presented in different languages.

Multilingual versions of a website (or search engine) may also exhibit different degrees of parallelism, ranging from a collection of monolingual sites at one extreme to a completely parallel site with identical structure, navigation and content at the other [Eurescom, 2000]. Typically a trade-off must be made between the cost and effort involved in creating such a site and its benefit. Further issues to consider include:

(i) the use of static versus dynamic content and whether off-line processing can be used to generate multilingual content,

(ii) query translation, in particular the advantages/disadvantages of using automatic as opposed to manual translation techniques. For example, digital libraries traditionally provide multilingual support via the use of multilingual thesauri such as Eurovoc[59], but this prohibits the use of free-text search and thereby limits interactivity.

7.2.2 Cross-Language Information Retrieval (CLIR)

An area of multilingual retrieval is Cross-Language Information Retrieval (CLIR) in which documents in different languages are searched by queries, also in different languages. This involves translating the query (in the *source language*) into the language of the document collection (*target language*), the documents into the query language or translating both queries and documents into a common language. Three major approaches for CLIR have emerged: (1) automatic machine translation where queries are

[59] http://europa.eu/eurovoc/ (Site Accessed: 04/10/06).

translated into the target language, (2) the use of machine readable bilingual dictionaries, and (3) the use of corpora to train or enable cross-language retrieval [Voorhees and Harman, 2000].

It is widely recognised that the design of an effective user interface is crucial for the successful implementation of any information system, particularly a search engine [Hearst, 1999; White and Ruthven, 2006]. Understanding the users, their searching behaviour, their needs, search tasks, situational context and their interaction strategies (among other factors) are all important elements of creating effective search applications (see, e.g. Ingwersen & Järvelin, 2005; Marchionini, 1992).

Providing effective access to multilingual document collections undoubtedly involves further challenges for the designers of interactive retrieval systems. In particular, deciding how best to support interaction within the search process can involve enabling: *query formulation* (e.g. offering the user additional query terms to refine their search such as synonyms), *query translation* (e.g. enabling the user to select from multiple query translations such as different word senses), *document selection* from search results (e.g. providing useable summaries for users to make informed decisions) and *document examination* (e.g. providing translated versions of documents for use by the end users). [Oard, 1997; He et al., 2006; Petrelli et al. 2006]

Practically, the interface may also enable users to indicate terms which should not be translated, identify phrases and signal out-of-vocabulary (OOV) terms (e.g. the CLARITY system [Petrelli et al., 2004; ibid. 2006]. Various studies analysing user interaction have highlighted the importance of the end user's multilingual ability. For example, Petrelli et al. [2002; ibid. 2006] consider users with competence in multiple languages (*polyglots*); whereas others such as Oard and Gonzalo [2002] and Ogden et al. [1999] consider users with no (or limited) knowledge of the target language (*monoglots*). This distinction between users alters the degree of multilingual support required in the search process (e.g. monoglot users may require the translation of retrieved documents or the back-translation of translated query terms).

The study of interactivity in CLIR ranges from studying aspects of the search process such as document selection [Oard et al., 2004; Resnik, 1997], query translation [Wang and Oard, 2001], presentation of search results [Ogden et al., 1999; Petrelli and Clough, 2005]; to the entire search process (e.g. Petrelli et al., 2002; Petrelli et al., 2005; Ogden et al., 1999; Ogden and Davies, 2000; Capstick et al., 2000; Peňas et al., 2001]. Example cross-language search systems (and interfaces) include the following: Keizai, ARCTOS, MULINEX, WTB, MIRACLE and CLARITY.

The Keizai system[60] [Ogden et al., 1999] uses a combination of automatic and user-assisted methods to build and refine cross-language queries. Queries composed of terms in multiple languages can be constructed. The user selects terms to be used in the search from a list of all possible senses of all possible translations. The result is displayed in the source language as a list of one-line summaries plus colour-coded keywords (the original word in Korean or Japanese is displayed in brackets). The Keizai system searches the Web to find documents in Japanese or Korean to answer a question in English. If the user decides to examine a document, they are able to translate the text into English using a link to an on-line MT system (Babelfish). In ARCTOS[61] [Ogden & Davis, 2000], each search term issued by the user is translated and boxed with the group of similar forms. Users can deselect translations, add new forms, or type new translations before the query is actually issued. Documents retrieved (in English, German, French and Italian) are displayed in a manner similar to Keizai.

MULINEX[62] [Capstick et al., 2000] allows users to choose the type of interface to work with: to either see all the translated query terms before proceeding with the search, or to completely hide the translation step. In Keizai and ARCTOS, when the query translation is shown, the user can edit the list and decide which terms will be included and which will not. MULINEX is multi-language (German, English, and French) and a separate column of translations is provided for each language. It also

[60] http://kythera.nmsu.edu:8099 (Site Accessed: 3/10/06).

[61] http://crl.nmsu.edu/~ogden/i-clir/cltr-interactive/arctos/page1.html (Site Accessed: 04/10/06).

[62] http://mulinex.dfki.de/demo.html (Site Accessed: 04/10/06).

suggests a list of additional terms the user might decide to include in the query. The retrieved documents are displayed as a list; for each document a set of category words in the user language and a summary in the document language are displayed. The user can click for a summary or the full-text translation in another language.

WTB (Web site Term Browser; [Peñas, Gonzalo, & Verdejo, 2001]) shows the terms generated during the query-expansion step grouped as families of terms, e.g. synonyms, hyponyms and hypernyms. Search results are presented as a cluster of documents grouped by relevant phrases. The system makes use of phrasal information to process queries and suggest relevant topics. By clicking on a line the user can explore the set of homogeneous documents represented by their title and an extensive set of relevant terms.

Figure 7.1: CLARITY user interface for CLIR

MIRACLE [Dorr et al., 2003; He et al., 2003] is a user-assisted CLIR system that groups translations for each query term in a tab and allows the user to view synonyms and examples of use. The list of terms actually used in the query is displayed below, followed by the list of retrieved documents for which the first two lines of machine-translated text are displayed. MIRACLE was designed with two aspects in mind: (1) a clear exposure to the user of the interaction design and (2) immediate feedback in response to user actions. Participation in the Cross Language Evaluation Forum (CLEF) interactive track (iCLEF) track has shown some interesting search behaviours from users such as adopting terms from relevant documents during query refinement (thereby confirming the need for document translations and consistency of translation resources used) and different strategies for query formulation [He et al., 2006].

CLARITY [Petrelli et al, 2005] has two interfaces: one to allow the users to modify the translation (*supervised mode*) and another interface (*delegated mode*). Using the delegated mode, the user simply enters the query, clicks the "Search" button and the results are then displayed. There is no user intervention during the query translation process. To modify the query, the user must re-enter it in the box. This system translates the queries into English, Finnish and Swedish. Figure 7.1 shows an example of the CLARITY interface (an English query searching Finnish documents).

Perhaps some of the most significant research undertaken to study the interaction with cross-language retrieval systems has been within iCLEF [Gonzalo & Oard, 2002].In 2000 iCLEF showed that users were able to determine the topic of retrieved documents, that they could often formulate effective queries (2002 and 2003), that users could find answers to factual questions (2004), find historical

images (2005), and most recently that users are able to perform multilingual searches using Flickr, the online photo management tool (2006).

7.2.3 Implementation of Multilingual Information Access

Table 7.1: Functionality offered by various online museums and art galleries [Marlow, 2006]

	Welcome pgs for f.l. (if more than one page available)	Multilingual. search of site (can locate material written in other languages)	CLIR (query translation)	Controlled vocabulary	Free-text Search	Easy to switch languages	Easy to return to original language
Tate Online		●	○	◉	◉	○	○
British Museum	●	●	○	◉	◉	●	○
National Gallery		○	○	◉	◉	○	●
V&A Museum		●	○	○	◉	●	●
Natl. Portrait Gallery		●	○	◉	◉	○	●
Louvre	●	●	◉ Lafayette database ○ Atlas database	● - Kaleidoscope	●	●	●
Guggenheim Bilbao	●	○	○	○	○	○	○
van Gogh Museum	●	●	○	●	○	◉	●
Rijks-museum		○	◉	●	●	○	○
Centre Pompidou			●	◉	◉	●	
MoMA		●	○	○	◉		
Met New York		●	○	○	◉	●	●
Guggenheim New York		○	○	◉	◉		
24 Hr Museum	◉	◉	○	○	●	○	●
Easyart.com	●	○	●	●	●	○	

● - multilingual offering ◉ - only in main language ○ - not offered

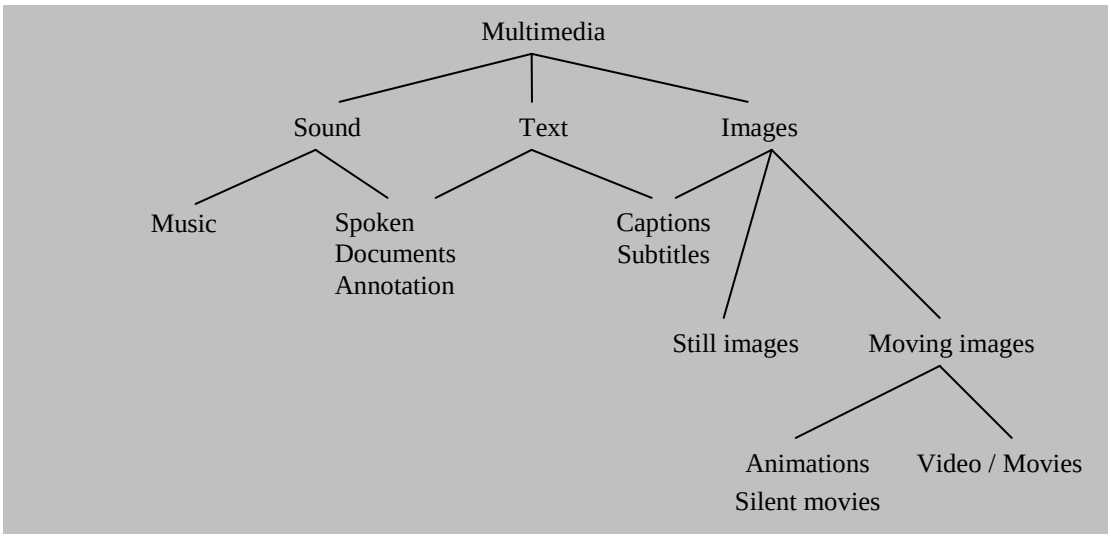
The Minerva survey [2006] examined the types of monolingual search functionalities provided by 671 European cultural and museum websites. Overall, it was reported that 51% of sites used no search tool at all, 24% offered free text indexing and 14% provided controlled vocabularies (some sites offered both). However, it is unclear how many of these search tools were available in more than one language. Marlow [2006] reviewed the functionality of a number of online museums and art galleries (shown in Table 7.1).

It is noteworthy that very few of the sites listed in Table 7.1 actually offer cross-language search functionality to users. This is typical of what generally obtains for most Internet search engines which tend to lack multilingual search facilities. The majority of cross-language research remains in the theoretical domain and has not often been implemented or made accessible to the end user [Peters and Sheridan, 2001]. Perhaps this is surprising given the motivation for multilingual search in [Oard, 1997], but Evans [2006] indicates that factors such as the limited effectiveness of translation, the lack of real-world user need for this kind of functionality, the complexity in effectively providing multilingual interaction and the additional cognitive burden pressed upon the user are all limiting factors.

7.3 Multimedia Information Access

Multimedia information retrieval (MIR) systems are designed to enable the searching of data in various modalities such as text, image, video and sound. Chu [2006] defines a taxonomy of multimedia information as shown in Figure 7.2, highlighting that multimedia information can be a combination of any single media. There are multiple ways of accessing visual objects (image and video) depending upon the information associated with the object: either information *about* the object (*metadata*) or information contained *within* the object (*audiovisual features*).

Figure 7.2: A taxonomy of multimedia information [Chu, 2006].



Images and video objects exhibit similar visual properties, the main difference being the additional spatio-temporal aspects of video [Gupta & Jain, 1997]. There is currently much research on combining both visual features and metadata as complementary evidence for both image and video retrieval and this is seen as one of the main research areas in current image retrieval research [Enser, 2004]. Further areas of research include both technical issues and establishing the requirements of users for multimedia information access. Ultimately the design of the interface and provision of functionality will depend on the needs of the end users, the indexing methods in use and available audiovisual data. We will start by discussing access to visual information (still images in section 4.1 and moving images or video in section 4.2). In section 4.3 we discuss access to audio information.

7.3.1 Still Image Retrieval

As with the design of any information system, an important part of the process is to establish what type of users will be using the system and their associated needs. For example, in describing image retrieval, Goodrum [2000] suggests that user interfaces must be influenced by considering the users' needs and their typical search tasks.

To date, most of the research and development in image retrieval has focused on providing functionality rather than giving sufficient attention to the needs of the end user [Eakins et al., 2004]. This has resulted in the design of interfaces which are inadequate (or unusable) for end users [Venters et al., 1997]. For example, a large body of research has grown up around developing algorithms to facilitate content-based retrieval (e.g. Smeulders et al., [2000]. However, studies of user needs have shown users' needs to be both linguistically and visually-orientated [Enser, 1995]. In practice, however, investigations (in particular domains) have shown that provision of text-based access is not just preferable but vital to many end users [Eakins et al., 2004; Markkula & Sormunen, 2000]. There are two main strategies for image retrieval:

(1) **description-based** (includes *text-based* and *concept-based*), which uses assigned free-text or terms from a controlled vocabulary (see, e.g. [Goodrum, 2000; Gupta & Jain, 1997; Rui et al., 1997; Smeulders et al., 2000; Veltkamp & Tanase, 2000].

(2) **content-based**, which makes use of low-level features derived from the visual content of an image Content-based retrieval [Smeulders et al., 2000] relies on indexing images by low-level attributes such as colour, shape and texture.

Since digitized images purely consist of arrays of pixel intensities with no inherent meaning, one of the key issues with CBIR and other image processing is to extract useful information from the raw data [Eakins and Graham, 1999]. By studying users' image retrieval requirements and the types of attributes images may exhibit, Eakins [1998] proposed a 3-level framework for image retrieval, classifying image queries by increasing complexity:

- **Level 1** comprises retrieval by primitive features such as colour, texture, shape or the spatial location of image elements. This level of retrieval uses features which directly extract from the images themselves, without the need to refer to any external knowledge base.
- **Level 2** comprises retrieval by derived features, involving some degree of logical inference about the identity of the objects depicted in the image. This requires reference some outside knowledge but in practice this level of query is very generally encountered (e.g. retrieval of objects of a given class such as "pictures of a passenger train on a bridge"; retrieval of individual objects or persons such as "pictures of Tony Blair" or "pictures of Nelson's Column").
- **Level 3** comprises retrieval by abstract attributes. This involves a large amount of high-level reasoning about the meaning and purpose of the objects depicted in the images. This level of query often requires some sophistication of the searcher and the reasoning judgment is often subjective. It would also require retrieval technique of level 2 to get the semantic meaning of various objects. For example, the retrieval of named events or types of activity "pictures of English folk dancing"; or retrieval of pictures with emotional or symbolic significance "pictures depicting *death*."

Description-Based Image Retrieval

Traditionally, the main approach for accessing images was based on formulating and serving text-based queries. Many of the early image retrieval systems were concept- (or text-) based utilising bespoke indexing schemes [Rasmussen, 1998] and overlapped substantially with the areas of databases and information science. Still images have unique meaning and properties that provide the basis for retrieval by users. For example, on considering the meaning of pictorial images, Panofsky [1955] categorised fine art images based on the "who, what, where and when" search paradigm and by the modes: iconography (specific requests), pre-iconography (general requests), and iconology (abstract images). Iconography describes a picture's actual subject matter (the what); iconology describes its deeper artistic or religious meaning (the why). Other authors such as Eakins and Graham [1999] have also discussed the categorisation of image attributes into various levels or strata. Pictures can therefore be described by their physical attributes (e.g. a picture of a dodo) and/or attributes of their subject (e.g. a picture of an extinct bird).

The main approach for accessing images is based on formulating and serving text-based queries which match between a user's query and image description. Rasmussen [1997] refers to descriptions of subject attributes as concept-based and Goodrum [2000] refers to descriptions based on texts associated with the images (e.g. captions, web pages) as text-based. There are many instances when images are associated with some kind of text semantically related to the image (e.g. metadata or captions); examples include collections such as historic or stock-photographic archives, medical databases, art/history collections, personal photographs (e.g. Flickr.com) and the Web (e.g. Yahoo! Images and AllTheWeb.com). Other attributes typically associated with an image which can be searched include date, time and information derived from the photographic equipment itself (e.g. the Exif[63] data provided by modern digital cameras).

Retrieval of images based on descriptions is typically through keywords (mostly derived from textual information accompanying an image) and controlled vocabularies associated with subject attributes. Searching with free-text (most keyword searches enable users to perform free-text search) or controlled vocabularies has shown to be an effective method of searching image repositories [Markkula & Sormunen, 2000; Rorvig: 1988].

Often, manually assigning indexing terms is a difficult task. The main problem is that the intrinsic meaning of an image is difficult to interpret and express in written form [Jorgensen, 1998]. In addition, assigning keywords to images is a very subjective task and suffers from low index term agreement across indexers and between indexers and user queries [Enser and McGregor, 1993]. Further, the amount (and availability) of visual material is growing at an astronomical rate and manual annotation is therefore impossible and in cases such as personal image collections, people often do not bother to annotate images. This has led to the popularity of approaches based on the automatic assignment of textual attributes [Turner, 1994]

Controlled vocabularies for text-based indexing can be found in the literature which describes the concepts of using certain established thesauri to describe image, e.g. Art & Architecture Thesaurus (AAT) [Petersen & Barnett, 1994]; Thesaurus for Graphic Materials [Parker, 1987] and ICONCLASS. They have applied existing cataloguing systems like Dewey Decimal System to describe images. See [Rasmussen, 1997] for further details of controlled vocabularies. An interesting extension of a controlled vocabulary is the *visual thesauri* which uses visual surrogates to represent concepts in addition to verbal descriptions (see, e.g. [Mostafa, 1994; Rasmussen, 1997]. This offers potentially interesting ways of using a controlled vocabulary (e.g. using the visual surrogates in a query-by-visual-example paradigm and using the pictures to create a language-independent representation of the controlled vocabulary). A summary of text-based retrieval products can be found in [Eakins et al, 1999], and previous research and prototypes described in [Clough and Sanderson, 2006].

[63] Exchangeable image file format is a specification for the image file format used by digital cameras: http://en.wikipedia.org/wiki/EXIF (site accessed: 13/11/06).

Content -Based Image Retrieval (CBIR)

In the early 1990s, because of the emergence of large-scale image collections and the aforementioned difficulties with manually indexing images, the development of content-based image retrieval (CBIR) was proposed by information researchers and scientists [Rui et al, 1999]. Content-based retrieval is implemented by automatically processing image attributes which are specified in user's queries. Typical image attributes include colour, shape, texture and spatial layout, all features which can be extracted using low-level feature extraction.

Retrieval based on colour similarity is often achieved by using a *colour histogram* for each image that identifies the distribution of colour pixels in an image. Image retrieval based on texture similarity is not regarded as very useful. However, the ability to match on texture similarity is often used most successfully when distinguishing between areas with similar colour in an image, e.g. between sky and sea [Eakins, 2000]. Queries by shapes are often achieved by selecting an example image provided by the system or by asking the user to sketch a rough shape. The primary mechanisms used for shape retrieval include "identification of features such as lines, boundaries, aspect ratio, circularity, and region and edge detection." [Goodrum, 2000]

Gudivada and Raghavan [1995] regard image retrieval at levels 2 and 3 of Eakin's framework as *semantic* image retrieval because they involve the addition of semantic information (typically by people). Most current CBIR techniques are designed for primitive levels (level 1), while some have attempted to tackle level 2 retrieval. However, this poses two non-trivial problems. The first is scene recognition: it is important to identify the type of scene presented in an image since this constitutes an important filter that can offer critical clues helping to recognise specific objects in an image. Object recognition is in itself a challenging problem in the area of computer vision. For example, Forsyth et al [1997] developed a technique for recognising naked people within images.

A number of general-purpose CBIR systems are commercially available on the Internet and most of these image retrieval systems support one or more of the following options: random browsing of images from the database, search by visual example, search by sketch, search by text and navigation with customised image categories [Chang et al, 1998]. Example content-based systems (both academic and commercial) include Virage's VIR Image Engine (VIR), Query By Visual Content (QBIC), VisualSEEk and Exacalibur's Image RetrievalWare. Web-based systems include WebSEEK, Informedia, Photobook and Alta Vista Photofinder. A full review of CBIR systems can be found in Veltkamp & Tanase [2000]. Most commercial and academic CBIR systems tend to offer either query-by-example functionality or support for user-input visual exemplars (e.g. colour).

One of the most cited examples of a commercial CBIR system is IBM's Query By Image Content or QBIC [Flickner et al., 1995]. It offers retrieval by combination of colour, texture or shape. Image queries can be formulated by selecting colour from a palette, sketching a rough shape of desired image, or specifying an example query image. The system extracts and stores the colour, shape and texture features from each image in its database, calculates similarity between query and stored images then displays the most similar image as thumbnails. In the cultural heritage domain, it can be used for colour and layout search in the State Hermitage Museum digital collection.[64]

WebSeek[65] [Smith et al, 1997], which was developed by Columbia University, is another content-based image retrieval system making keyword and colour based queries through a catalogue of images collected from the Web. The system allows the user to submit a query by choosing a subject from the available catalogue or entering a text topic. The results of the query may be used for another colour query in the whole catalogue or for sorting the results by decreasing colour similarity to the selected image. In addition, WebSeek allows the user to directly define a colour histogram's attributes in order to better refine the image search criteria.

[64] http://www.hermitagemuseum.org/fcgibin/db2www/qbicSearch.mac/qbic?selLang=English

[65] http://persia.ee.columbia.edu:8008

WebSeer [Swain et al., 1996] was developed by the department of computer science at the University of Chicago as an experimental system. Besides some common characteristics such as specifying image dimensions, file size, image type and submitting keywords describing the contents of the desired images, the system was also able to detect human faces based on a neural network. If the user is looking for people, he/she must indicate the number of faces as well as the size of the portrait. Face detection is believed to meet the needs of the Level 2 user.

Most existing CBIR systems retrieve images by image appearance, using automatic extraction and a comparison of image features such as colour, texture, shape and spatial layout. This well meets Level 1 of user's image query needs. However, for Level 2 and Level 3, evidence suggests that such a facility is actually of limited use in meeting image users' real needs [Eakins et al., 2004]. First, it is impossible to start a search if no suitable query image can be found or the user has no idea about what the image should look like, e.g. searching for a rare unseen animal. Second, users may find it difficult to manipulate search parameters such as the relative importance of colour, shape or texture because such visual features are not as intuitive as text [Eakins et al., 2004].

A large number of CBIR systems take sophisticated algorithms; however, it is not clear whether they can really address user needs. As a result, to narrow this semantic gap, a powerful and user-friendly query interface is needed where users can interact with systems by providing his or her evaluation or preference of a current retrieval result to the IR system [Rui et al, 1999].

Combining approaches

Combining both description and content-based approaches is likely to be more effective than any single method alone. Eakins and Graham [1999] comment that the use of keywords and image features in combination is desirable. This coincides with best practice in designing interactive retrieval systems which suggest that a variety of interaction approaches should be offered to meet the varying needs of users and their work tasks. Chu [2001] provides examples of research from the content-based community which has combined the two approaches. The current challenge is how best to integrate functionality to provide natural access for users to both low-level primitive features and high-level semantics. Systems such as WebSEEK [Chang et al.,. 1997] have shown the benefits of combining approaches (e.g. allowing users to initiate a search based on keywords or selecting terms from a controlled vocabulary, and then using content-based approaches during refinement or to provide a "more like this" function).

User interfaces and interaction

Interaction with image retrieval systems is similar to any other retrieval system and includes: query formulation, query reformulation/modification (e.g. through relevance feedback), browsing-searching and results presentation (in context). Typically in image retrieval systems, the user interface consists of a query formulation part and results presentation part [Veltkamp & Tanase, 2000:1]. Users can select images from the index (or database) by browsing one-by-one, or specify an image (or set of images) through the use of keywords, by using visual properties of an image (e.g. colour, texture etc.), or providing a visual exemplar (e.g. an example image or a sketch).

Various studies have been undertaken to establish what people search for in multimedia collections, e.g. newspaper image archives, picture archives and museums (Enser [1995]; Enser & McGregor [1992]; Armitage & Enser [1997]). Enser and McGregor [1992] categorised queries made to a large picture archive into those which could be satisfied by a picture of a unique person, object or event (e.g. Kenilworth Castle, Sergei Prokofiev, HMS Volunteer, Alan Turing), and those which could not (e.g. classroom scenes, Clyde cruisers, shopping arcades, air raids). These categories, unique and non-unique, were also subject to query refinement in terms of time, action, event or technical specification. For example a non-unique query such as "carnival" could be modified to create "the Rio Carnival, 1996" (unique), refined by location and time.

A recent study by Eakins et al. [2004] identified user needs within a framework based on a taxonomy of image content (i.e. classifying images from a low-level representation to high-level semantics) and how professionals search for and use image data (e.g. for illustration, learning, information processing

and generating ideas). Their findings reinforced previous studies (e.g. [Enser, 1995; Markkula & Sormunen, 2000]) whereby participants were primarily interested in concept-based retrieval rather than content-based. They also found the preferred method of querying was to type search terms rather than select from a hierarchy of terms or query by example. The use of text-based retrieval, however, presupposes that images are associated with textual metadata. In many scenarios this is a valid assumption, e.g. in stock photographic collections, on the Web and historical or cultural heritage archives. However, this is not always the case (e.g. for personal photographic collections).

Researchers have also considered the user's searching behaviour in image retrieval. For example, Cox et al. [1996] define at least three classes of image search: (1) target search – users find specific target images (e.g. art historian finding a specific painting), (2) category search – users seek one or more images from general categories (e.g. "sunsets" or "pictures of the Eiffel Tower"), and (3) open-ended browsing – users have a vague idea of their search needs and may change their mind repeatedly throughout the search. This last category includes exploratory tasks where users have no specific goal (e.g. browsing through a database for fun).

Two fundamental methods for accessing information include search and browse. Search consists of typing keywords; browse is more likely once an initial starting point is found Browsing support is often structured such that content is categorised into predetermined classes or a hierarchy (e.g. subject classification) into which users can further explore and navigate. However, this is typically useful only if it matches the user's expectations because it imposes a single view on a collection (alternatives are multiple alternative hierarchies, e.g. faceted metadata). Accessing information through browsing has demonstrated to be very effective in the domain of image retrieval (see, e.g. [Chang et al., 2004; Shen, 2003; Combs & Bederson, 1999]). When image browsing is combined with text searching, users are able to select their most preferred interaction mode and move between the two in a fluid way (see, e.g. [Hearst, 2002; Yee, 2003; Combs & Bederson, 1999]).

One of the biggest problems with retrieving visual information is the "semantic gap" between the low-levelled data representation (e.g. pixel light intensity values) and high-level–needs/concepts that the user desires [Enser and Sandom, 2003]. As Urban and Jose [2005] state, "the images' low-level feature representation does not reflect the high-level concepts the user has in mind." The problem of the semantic gap for information retrieval is that the meaning of an image can only be defined in context. The use of relevance feedback and browsing-searching techniques can assist with formulating the user's query and narrow the semantic gap (i.e. help the user to specify the query).

Query Specification

Queries to CBIR systems are most often expressed as visual exemplars (Query-By-Visual-Example or QBVE) or specifying image attributes such as colour (e.g. picking the desired colour from a palette). QBVE can be performed by supplying an example image being sought (either from within or outside the indexed collection of images), or sketching the desired shape of an example image (e.g. QBIC offers this [Flickner et al., 1995] and RetrieveR[66], a sketch interface to Flickr]. Eakins and Graham [1999) point out those content-based approaches based on colour, texture and shape are capable of delivering useful results, but in practice some of the features are far more useful than others (e.g. colour and texture retrieval often gives better results than shape matching). The advantages of this form of querying are its simplicity for novice users and ease of expressing more "visual" queries in domains where visual attributes are important (e.g. fine-art painting [Lombardi et al., 2004]).

However, this approach has some disadvantages. For example, the success of sketched queries may depend on the user's artistic abilities. Additionally, supplying a single example image may prove quite successful when searching for a single relevant image but will probably be less successful for retrieval of groups of images related to a category. Matching variants of a supplied image can be difficult (e.g. images distorted by rotation, skew and occlusion). A further problem is the semantic gap. Gupta and Jain [1997] state that query specification for visual information should not be limited to query-by-

[66] http://labs.systemone.at/retrievr/ (site accessed: 13/11/06).

example or the specification of visual properties of images and suggest nine further properties of a query language including: spatial arrangement, temporal arrangement and feature-space manipulation.

Most systems enable the user to evaluate or provide his or her preference of a current retrieval result to a CBIR system (*relevance feedback*) as a way of refining the query. This can be through specifying positive or negative examples, and Rui & Huang [1999] suggest that this can be used to narrow the semantic gap. Rui et al. [1998] suggest that systems involving CBIR must research into *where* in the interaction cycle users would want such support. However, CBIR systems are still not widely used by the general public after more than a decade of research effort. Urban and Jose [2005] suggest this is due to the continuing problem of the semantic gap and the fact that most current interfaces do not provide sufficient querying facilities and appropriate presentation of results.

Browsing

Many efforts have been undertaken to generate effective image indexing systems (e.g. ICONCLASS[67], the Getty Art and Architecture Thesaurus or AAT[68] and WordNet[69]) and these semantic classification systems are often used to complement search and provide browsing functionality. A study of interaction with WebSEEk found that users' preferred method of browsing was through theme-based navigation – rather than browsing through pages of image thumbnails – and preferred querying methods based on some specific subject matter rather than free-text search or advanced visual searches. The use of hierarchical structures for categorising and organising images not only facilitates browsing, but also helps to provide a context for the search results (e.g. users can browse through results in broader or narrower categories). There are several problems with using a controlled vocabulary, however, including the assignment of terms, the ambiguity of categories, and the user's unfamiliarity of subject categories used in the classification scheme (Getty photographic images).

One approach to render QBE more attractive is to use information derived from text associated with the image itself. For example, Yee et al. [2003] describe Flamenco, a text-based image retrieval system in which users are able to drill down results along conceptual dimensions provided by hierarchically faceted metadata. Categories are automatically derived from WordNet synsets based on texts associated with the images, but assignment of those categories to the images is then manual. This interface provides effective search and browse of images and supports exploratory search tasks. A further approach is to allow the users to generate their own taxonomies in the form of *folksonomies*. The online photo management tool, Flickr, allows this form of collaborative annotation through users assigning tags (keywords) to images. These then enable users to navigate to images with the same tags and a clustering of tags helps to organise images and facilitate browsing.

Results presentation/visualisation

Finding appropriate results that correspond to the user's searching and browsing requirements is the first task a system must achieve; however, an equally important consideration involves determining how best to present said results in an accessible and user-friendly manner. For example, Hearst [1999] recommends providing users with information about:

- How retrieved documents are related to the query
- How the retrieved documents relate to each other, and
- How the documents relate to the collection as a whole

Currently, the widely-used standard for displaying image results is to show a two-dimensional grid of thumbnails [Karadkar et al., 2006; Rodden et al., 2001; Combs & Bederson, 1999]. However, this is not necessarily an ideal approach.

[67] http://www.iconclass.nl/ (site accessed: 13/11/06).

[68] http://www.getty.edu/research/conducting_research/vocabularies/aat/ (site accessed: 13/11/06).

[69] http://wordnet.princeton.edu/ (site accessed: 13/11/06).

Chang & Leggett [2003] outline three main problems with current interfaces for searching and viewing image collections. First, querying by metadata is ambiguous and often does not accurately portray relations between image elements. Secondly, browsing is often time-consuming (involving a great deal of pointing and clicking) and not adaptive to users' needs. Finally, scrolling through many thumbnails is tedious, and if all results do not fit on one page, it is difficult to obtain a comprehensive view or understanding of the entire result set. Janecek & Pu [2004] note that since it is increasingly difficult to display all information in the limited space of one screen, there is often a balance that must be struck between showing a small amount of detailed information and providing a large amount of more abstract information.

Jörgensen & Jörgensen's [2005] study of image professionals revealed that 85.6% of the searches involved the browsing of results, implying that this behaviour is important in making an image selection. Therefore, developing a more effective way of enabling this to be done is the subject of much research. To combat some of the problems stated above, alternative approaches to visualising results displays have been explored.

With regards to the problem of having to scan a large set of results for relevant or related images, Liu et al [2004] developed a similarity-based results presentation that was meant to graphically depict the closeness of relationships between images, based on "regions of interest" within the images. The items were then arranged in a way so that closely related pictures were situated near and overlapped each other. To facilitate viewing, the user could control the overlapping ratio using a slider. Results of initial experimentation indicated that this approach helped to improve users' experience browsing results and sped up the search process.

Janecek & Pu [2004] advocate the use of semantic "fisheye" views to enable focusing in on relevant parts of a wide set of results. This type of visualisation helps users to examine local details while still maintaining a view of the broader context [Liu et al., 2004]. Moving the mouse over a particular element of the results display automatically brings it into greater focus. Thus, that which the user deems to be more interesting or important is emphasised, while the less important information remains in the background. The metrics used to determine "importance" are flexible and can thus be adjusted to enable a variety of search strategies.

Visualisation displays can also encourage query refinement in a variety of ways. For example, users can be given the opportunity to see a range of related items in order to decide if one of them fits their needs more closely. This can be particularly useful in the case where a query has multiple meanings (i.e. the word "Pluto" can refer to the astronomical entity or to the Disney character.) In this case, a clustering method could be helpful.

For image retrieval, clustering methods have been used to organize search results by grouping the top *n* ranked images into similar and dissimilar classes. Typically this is based on visual similarity and the cluster closest to the query or a representative image from each cluster can then be used to present the user with very different images enabling more effective user feedback. For example, Park et al. [2005] took the top 120 images and clustered these using hierarchical agglomerative clustering methods (HACM). Clusters are then ranked based on the distance of the cluster from the query. The effect is to group together visually similar images in the results. However, Rodden et al. [2003] performed usability studies to determine whether organization by visual similarity is actually useful. Interestingly, their results suggest that images organized by category/subject labels were more understandable to users that those grouped by visual features.

Other approaches have combined both visual and textual information to cluster sets of images into multiple topics. For example, Cai et al. [2004] use visual, textual and link information to cluster Web image search results into different types of semantic clusters. Barnard and Forsyth [2001] organize image collections using a statistical model which incorporates both semantic information extracted from associated text and visual data derived from image processing. During a training phase, they train a generative hierarchical model to learn semantic relationships between low-level visual features and words. The resulting hierarchical model associates segments of an image (known as *blobs*) with words and clusters these into groups which can then be used to browse the image collection.

As another form of clustering, Clough et al. [2005] propose automatically generating a set of conceptual hierarchies based on metadata, and then classifying representative images into the relevant place in the hierarchy. The result combines text and visual data and is essentially a hierarchical browsing facility with associated images displayed to illustrate and clarify the terms.

Visualising a collection overview can be slightly different from visualising results of a targeted search because rather than trying to locate a specific item, often the goal is to get a general understanding of a collection's underlying theme. To facilitate this, Chang & Leggett [2003] propose a streaming collage approach, whereby a collage of the collection's holdings is gradually and dynamically built over time, with similar items placed near one another in a way that highlights commonalities, links, and relationships. (see Figure 7.3). Another suggestion related to the browsing interface is to employ a zoomable image browser (Figure 7.4) as a way of maximising use of the screen space [Combs & Bederson, 1999].

Figure 7.3: Streaming Collage interface [Chang & Leggett, 2003]

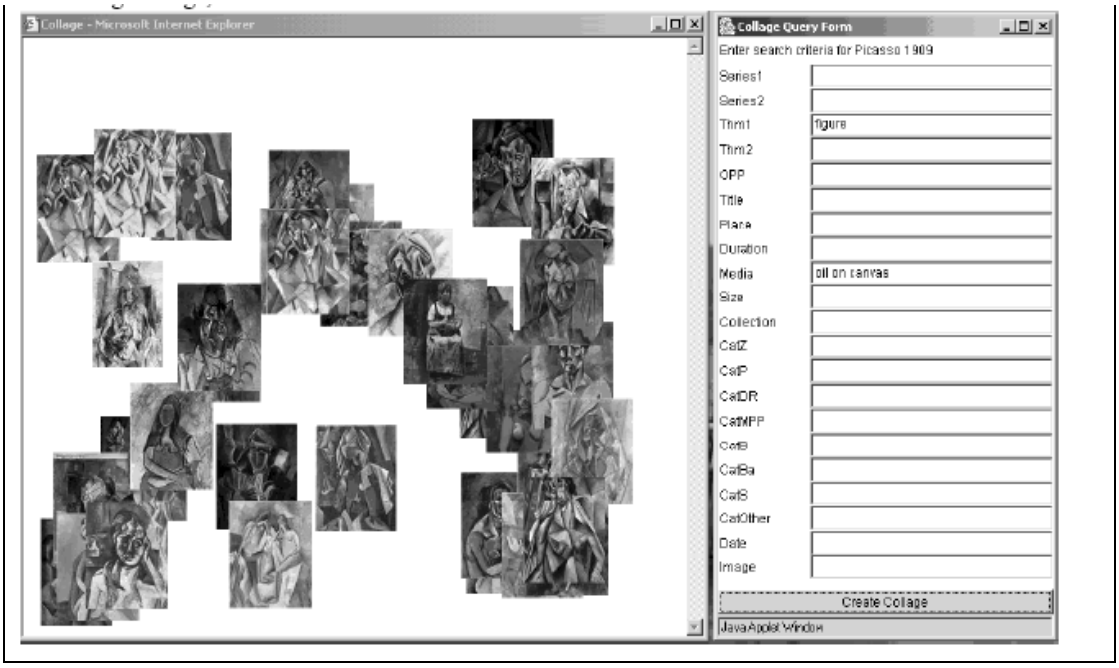
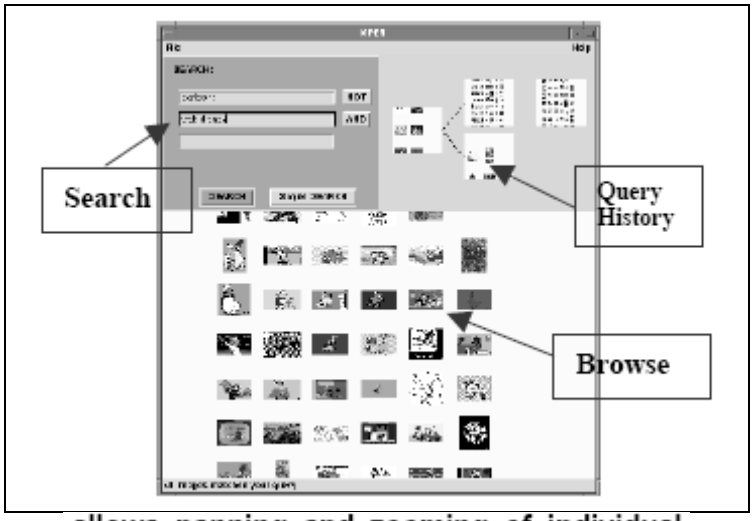


Figure 7.4: Zoomable image browser prototype [Combs & Bederson, 1999]



When retrieval is conducted across media, it is not clear how the results should be displayed. A single list of interleaved or fused heterogeneous multimedia objects to be explored in sequence may not be the best solution. Different metaphors and layouts have been proposed but limitedly to a single media (i.e. newspaper-like layout for text [Golovchinsky, 1997]; comic book [Boreczky, 2000] and storyboard [Christel, 2002] for video; or picture album for images [Kyu, 2004]. Karadkar et al. (2006] investigate various combinations of spatial and temporal layouts and their constraints on context during the design of an interface for a video and image retrieval system.

In summary, current image retrieval systems offer much functionality, some of which is not necessarily useful to users. It is important to study users, ascertain their needs, and determine their tasks to develop effective user interfaces. Rather than try and meet the needs of all users, it is important to provide functionality to meet specific user classes. For example, Jörgensen & Jörgensen's [2005] study of image professionals noted that these individuals had slightly different behaviours than more general users; these included a reliance on more descriptive and thematic queries than unique term searches.

Goodrum [2000] suggests that research is required that examines interface support for browsing, query formulation and iterative searching. Lee et al. [1994] emphasise that research must be undertaken to establish where in the interaction cycle CBIR would best be suited. Chang et al. [1997] have found with WebSEEk that users prefer to navigate through a clearly defined semantic structure organised in a hierarchical form (especially true for searching large repositories). After users have narrowed down results, the use of content-based methods can then be used to effectively organise, browse and view the content space.

7.4 Video Retrieval Interfaces

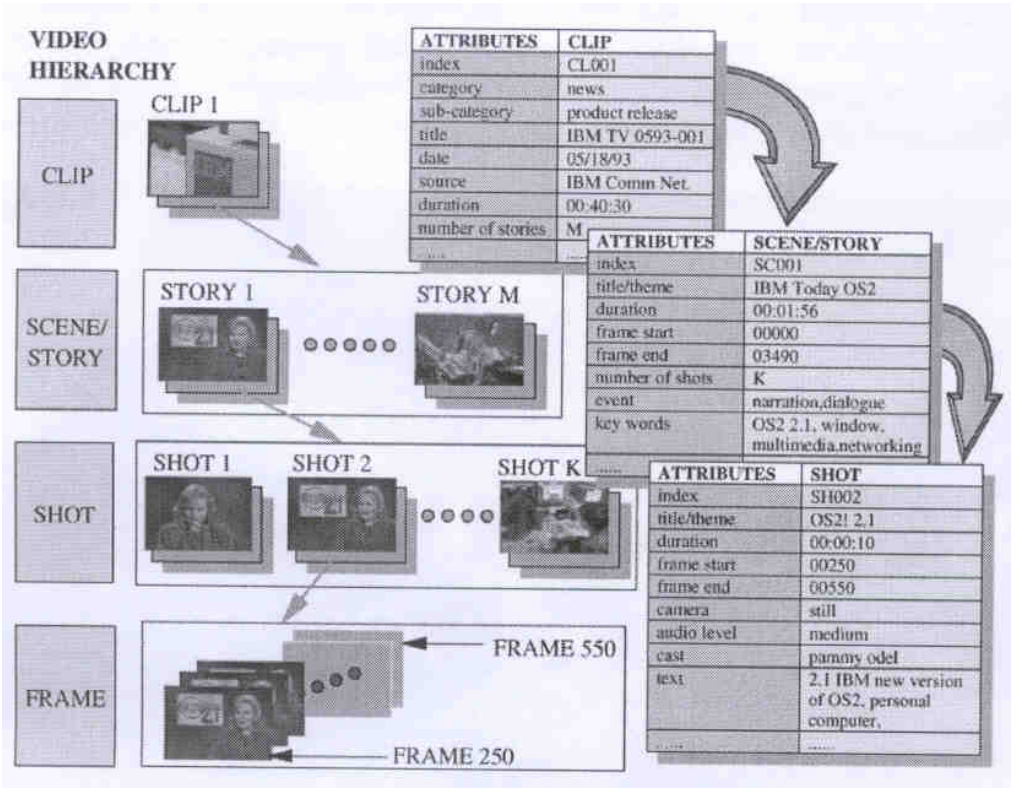
The process of searching, retrieving, and visualising videos differs from that of images due to the nature and format of video as a medium. For example, video is inherently multimodal and can contain visual, auditory, and textual elements [Snoek & Worring, 2005]. In addition, video is time-based and as a result, searching through clips to locate some information of interest can potentially be a tedious and lengthy process [van Houten et al., 2004]. Therefore, a video retrieval interface should make it easy for users to browse and/or search for relevant material in an efficient way.

7.4.1 Video indexing

Before videos can be searched, browsed, or manipulated, they must be indexed in some way [Snoek & Worring, 2005]. There are a variety of ways in which this can be done. One approach is to break a video clip down into its individual components and index these.

The atomic unit of the video *clip* is the *frame* (the equivalent of one exposure on a celluloid film track). A video *shot* is defined as the sequence of frames captured during a single "start recording" and "stop recording" camera operation. A *scene* is a sequential collection of shots unified by a common event or locale. A video clip is normally composed of a collection of scenes. There are several scene combination possibilities, one of which is the *dialog*, defined as a series of alternating shots depicting some form of communication between two or more entities (e.g. the "cut-away" shots switching between the in-studio news anchor and the on-location news reporter). Most video document indexing techniques exploit this inherent frame→shot→scene→clip hierarchical structure to automatically segment the video document into more manageable chunks. Yeo and Yeung [1997] schematically illustrate this hierarchy:

Figure 7.5: Video Decomposition Hierarchy (taken from Yeo & Young [1997])



Smeaton [2000] advises that manual annotation / mark-up of video should be kept to a minimum, with preference given to automatic techniques which yield consistent (even if occasionally incorrect or unexpected) results. Typical automatic *shot boundary detection* and *scene change detection* techniques[70] attempt video clip segmentation via the use of *scene transition graphs*, inter-frame and inter-shot colour histogram comparisons, and motion detection algorithms.

[70] Shot boundary and scene boundary detection techniques are quite similar, the principal difference being that the latter boundary detection technique works at a higher hierarchical level.

7.4.2 User Actions

Once video content is indexed, it is important to consider the ways in which users may wish to interact with it. Lee & Smeaton [2002] define the following potential user actions that a video library interface should support:

- browsing and selecting video programmes from a collection
- content querying of a video programme
- browsing the content of a video programme
- watching a video programme (all or part of one)
- re-querying the video digital library or within a programme

With regards to browsing or searching, Lee & Smeaton [2002] mention that searching is often done based on querying video metadata (i.e. the title, date, or description of a clip.) Smeaton [2002] explains that this can take the form of matching a query against some unit of information which can be as broad as a whole video or limited to some subset therein.

However, van Houten et al. [2004] assert that browsing is a more natural behaviour in the context of videos, because it can sometimes be difficult to articulate or find what one is looking for when using a keyword search. Yang & Marchionini [2005:1] agree that browsing is easier and faster for users, stating that "video information needs are sometimes hard to express in words, but are easily clarified when the picture/video clips are seen." Additionally, it is often the case that initial browsing often leads to the formulation of more specific search criteria.

Once an individual has located a video of interest, content browsing can occur in the form of allowing him/her to fast-forward and rewind through the clip, although alternative approaches do exist. Actual playback is often the final step and many interfaces support this by providing video player software (such as RealPlayer) to display the content. However, re-querying is also important to consider, as often a user will need to continue to interact with the system as his/her goals and information needs evolve [Lee & Smeaton, 2002).

7.4.3 Surrogates

After segmentation, the information extracted from the video clip must be displayed in a manner which is readily accessible and easily interpreted by the viewer. There are several approaches that can be taken when displaying the results of a video search. However, in general, some sort of surrogate must be presented. Yang et al. [2003:3] define a video surrogate as "a compact representation of the original video that shares major attributes with the object it represents." They go on to mention that the goal of a surrogate is to act as a summary and to enable the user to get the gist of the video's content. A successful surrogate allows the user to make accurate judgements about the relevance of a video without having to watch the entire clip.

There are a variety of surrogates that can be used, according to Yang et al [2003]:

- text surrogates (bibliographic information/metadata)
- still image surrogates (keyframes)
- moving image surrogates (sped-up versions of the video)
- audio surrogates (extracted audio information from the video)
- multimodal surrogates (a combination of video, audio, and text)

What many researchers seem to agree upon is that since humans process visual images more quickly than text and have an accurate recognition memory for pictures, providing easy visual access to video information is desirable [van Houten et al., 2004; Yang et al., 2003]. Christel et al. [2002] add that the combination of both textual captions and visual summaries is better than using textual summaries alone. Text can be extracted from associated closed-caption information (if available) or obtained using speech recognition programs.

In terms of the layout and presentation of video surrogates, Lee & Smeaton [2002: 11] propose that "keyframe-based browsing is similar to the now de-facto standard feature of 'thumbnail browsing' in image retrieval interfaces…" Keyframes are selected frames from a video displayed either as an individual image or as a temporally-ordered sequence of images. The idea behind choosing which keyframes to display is that they are the most representative of the overall video content. Christel et al. [2002] used synchronization metadata and inverse document frequency metrics to find the highest-scoring shot for an individual query. However, this is not always an easy task.

Keyframes may be chosen either manually or automatically (if automatically, they can be selected at regular time intervals in the video, or they can be taken from a certain place in each scene with the help of boundary detection methods.) Regardless, there is also the question of how many keyframes to display. Lee & Smeaton [2002: 14] mention that there is no easy answer to this question: "it will not be possible to say which level of granularity is best for every situation as one user in one situation will have different needs from another user."

7.4.4 Visualisation layouts

There are several ways in which the surrogates can be laid out within an interface. Many approaches have made reference to Shneiderman's [1998] mantra of "Overview first, zoom and details on demand" as a guiding principle. Some common approaches are [Lee & Smeaton, 2002]:

- Storyboards (a series of small keyframes displayed spatially on the screen in chronological order)
- Slideshows (the keyframes are displayed one at a time in a slideshow. The transition from one keyframe to the next can either occur automatically or can be controlled by the user.)
- Hierarchically arranged browsers (in which keyframes can be viewed by drilling down—best for structured programmes such as the news.)

However, these are not the only options. Other approaches to visualisation design will now be described. As previously mentioned, the most common display paradigm is the 2D story-board style grid layout. Since it is usually not feasible to display every frame in a shot, most video information visualisation techniques attempt to identify the frame within a shot- or scene- sequence which typifies the content of said sequence. This most typical frame is then displayed on the grid and configured to support some form of interactive playback – clicking on this representative frame will result in the playing of some or all of the frames within the same shot or sequence.

These representative interactive frames are usually arranged on a 2D grid in a sequential and/or hierarchical fashion. Figure 7.6 shows a typical frame sequence displayed in a strictly hierarchy-flattened sequential fashion, while Figure 7.7 depicts a similar 2D grid but with a frame→shot→scene hierarchy. It is to be note that Figure 7.7's three-tier display reflects this 3-level frame→shot→scene hierarchy with the top level corresponding to the scene and the 2 lower levels corresponding to the shot and frame collections respectively. Selecting a typical frame from the uppermost level (i.e. scene level) for playback will have a "drill-down" effect, i.e. the displays in the two lower levels will be updated to show:

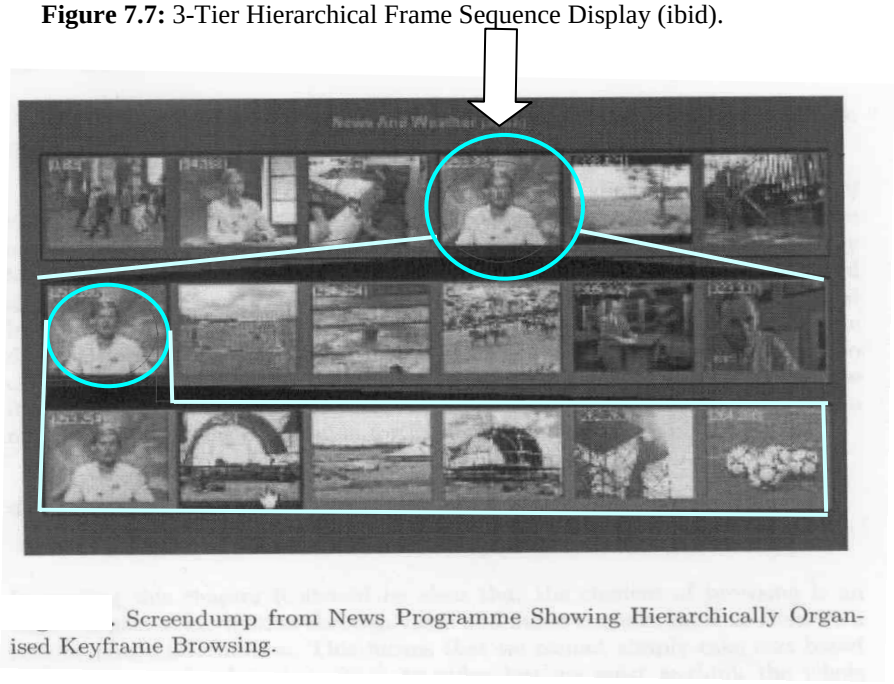
- the most-typical frames from all the shots at level 2
- representative frames amongst all the single frames at level 3[71]

[71] Of course, this hierarchy could extend one level higher with the uppermost level displaying a series of separate video clips, the second level would then display a series of scenes from any video clip which has been selected at the top level, the third level would then display a series of shots.

Figure 7.6: Hierarchy-flattened Frame Sequence Display [Yeo & Young, 1997]

. Screendump from Físchlár Showing Keyframes Taken from the Middle of a 30 minute News and Weather Broadcast.

Figure 7.7: 3-Tier Hierarchical Frame Sequence Display (ibid).



. Screendump from News Programme Showing Hierarchically Organised Keyframe Browsing.

Boreczky et al. [2000] have implemented a refinement of the 2D grid layout, where representative frames considered to be of greatest relevance are presented on bigger panels, in a fashion similar to that employed in comic books where climatic scenes in the narrative are given more space on the page. The frames are then slotted into position using a near-optimal "row block" packing algorithm, an example of which appears in Figure 7.8. Note that some panels (as is the case with panel 5 in this example) may be resized to better fit available space.

Yeo and Yeung [1997] also implement a (less sophisticated) variation of the comic book layout (Figure 7.5), but theirs does not incorporate the level of user interaction evidenced in the Boreczky model.

Figure 7.8: Boreczky's [2000] Comic Book Style Layout

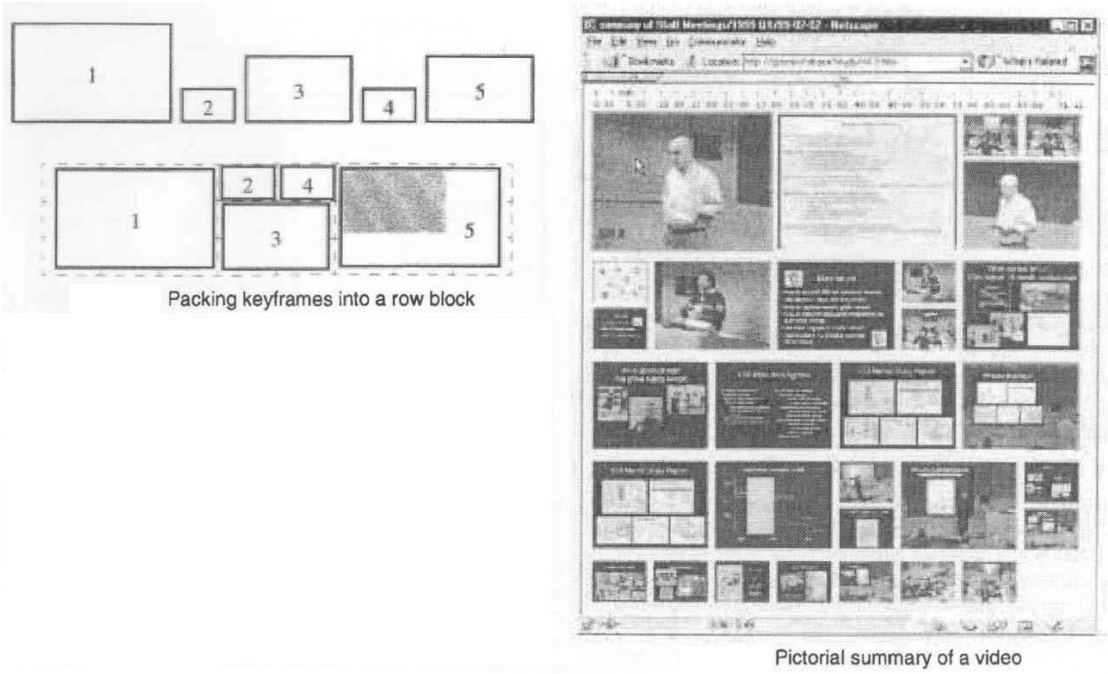


Figure 7.9: Examples of Boreczky's Playback Functionality

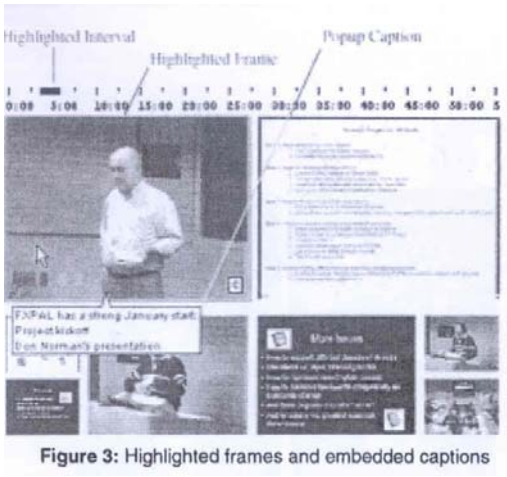


Figure 3: Highlighted frames and embedded captions

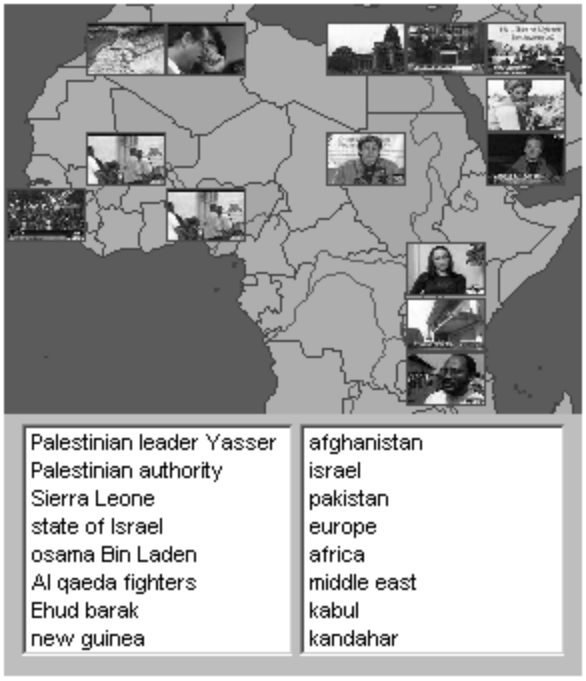
Figure 4: Playing the video

Smeaton [2002] reminds us that, in the case of video clip retrieval and indexing, it is important to use a variety of IR techniques which will be capable of processing all possible data types which may be embedded in the video object, these include:

- using OCR to decipher any text captions or titles (as is often the case with clips originating from news programs, documentaries, etc).
- Automatic speech recognition (ASR) and musical instrument recognition to process sound tracks. If full-blown ASR recognition returns low accuracy rates, Smeaton [2000] advocates a phone recognition approach, where the user's text-based request is decomposed into a string of phones and this phone string is compared to the phone sequences extracted from automatic phone recognition processing of the video clip's sound track. Sound tracks (and their associated video clips) with a high hit rate are deemed to be a good match and included in the list of returned documents.

Christel et al. [2002: 561] present the idea of visual collages as "new interactive tools facilitating efficient, intelligent browsing of video information by users as they follow their shifting information needs." A collage is a dynamic overview of video results where users can "drill down" or zoom in on areas that are of particular interest. More targeted browsing of videos by location can be done via a map collage interface (Figure 7.10), or by time via a timeline interface (Figure 7.11.) The collages also contain text that refers to the most frequently-occurring phrases in the videos (which are all news reports.) Overall, these collages incorporate automatically-generated data and the user's query context to create a dynamic and interactive way of exploring a large quantity of results.

Figure 7.10: Map collage interface [Christel et al., 2002].



Map collage, with common phrases and frequent locations for documents pertaining to Africa.

Figure 7.11: Timeline collage interface [Christel et al., 2002].

Collage generated from 20 video documents returned from "Dennis Tito" query.

Overall, general good practice to follow when designing a video retrieval interface is to support as many types of tasks and behaviours as possible, while making it easy to switch between different features [Lee & Smeaton, 2002]. Similarly, Smeaton [2002: 222] recommends providing "video navigation which seamlessly combines searching for objects, shots, or scenes, browsing and following hyperlinks between related video elements, and summarisation based on generated summaries or sets of keyframes" as the most efficient and useful way of enabling navigation through video libraries.

7.5 Audio Retrieval Interfaces

Indexing and retrieval of audio documents is, in principle, quite similar to its video counterpart with one significant exception: important progress has been made in developing methods for extracting semantic meaning from acoustic signals containing music and speech.

The three principal *music information retrieval* (MIR) techniques are *automatic musical instrument recognition*, *automatic music score transcription* and *automatic genre classification*. West and Cox [2005] have reported considerable success in classifying music recordings according to *genre* (e.g., musical style such as jazz, rock or classical, etc.). In terms of instrument recognition, Eggink and Brown [2004] achieved a recognition rate accuracy averaging 80% for certain types of instrument and given certain conditions[72].

Accuracy rates for *automatic speech recognition* (ASR) vary significantly depending on the constraints and scope of the task, ranging from in excess of 90% if the recogniser is small vocabulary and *speaker dependent* (i.e. trained on speech samples from the target speaker) to around 70% if the recogniser is *speaker independent*[73] and the number of word items to be recognised is quite large (e.g. in excess of 5,000). In the context of the speech indexing tasks to be attempted by the MultiMatch project, the most appropriate ASR system configuration would be speaker independent and large vocabulary. Furthermore, it would be necessary to devise some method of segmenting an audio clip or video clip sound track into thematically distinct units representing, for example, individual news stories or musical performances. These segmentation techniques are discussed in the following section.

7.5.1 Thematically indexing audio data

Thematic segmentation of speech and music has a well-established tradition with associated technologies being sufficiently mature as to permit commercial exploitation. A notable example of such technology is the THISL speech recognition and indexing system implemented by Renals et al. [2000] for the indexing of radio broadcasts from the United Kingdom's BBC news network. The segmentation methods employed by THISL are typical of most state of the art applications and consist of the following pattern recognition techniques:

- Detection of significant non-speech events: it is usually the case that individual news items will be separated by some type of non-speech event, usually in the form of a period of silence and/or a *station ident* – an ident being a short musical jingle or other distinctive audio event which is recognised as the acoustic equivalent of a company logo.
- Detection of a shift in term frequency: given that a news item normally has some unifying theme, it is quite likely that there will be some specific word or phrase which will be mentioned repeatedly for the duration of that news item but which will be mentioned less frequently – if at all – in subsequent or preceding news items.
- Detection of change in ambient noise quality: a sudden change in the loudness and quality of background noise is often an indicator of a change in physical location. This may in itself not

[72] The instrument recognition software application devised by Eggink and Brown proved more capable at recognising certain types of instrument (namely the wind instruments such as the flute). Furthermore, performance rates dropped if there were more than six other instruments being played simultaneously.

[73] In speaker independent ASR systems, the recogniser is trained on speech samples from a variety of individuals who typify the speaking style of the target population. Such training procedures will normally produce an ASR system which will work reasonably well for most but with an accuracy rate below that of a speaker dependent system customised for a specific individual.

indicate a boundary between two items, but when used in conjunction with the two techniques listed above, it can offer useful clues to facilitate segmentation.

Therefore, audio files can be indexed in a variety of ways, based on extracted metadata, acoustic indexing (i.e. using automatic speech recognition), or semantic indexing (based on topic or theme, as described above.)

7.5.2 Visualisation of audio search results

After audio files have been properly indexed to facilitate searching, the next issue involves determining how best to display the results of a search. Logan et al. [2004] describe two common ways for users to find an audio file: either by searching for keywords contained in the files' metadata or associated transcripts, or by conducting a "similarity search" for items that are related to a given file. In general, most searchable audio archives present results in a simple list of links to files, often ranked by supposed relevance [Van Thong et al., 2001; Foote, 1999].

A related consideration involves enabling a user to find the relevant content within a given media file (in case he or she only is interested in one section of a longer recording.) Foote [1999] mentions that the typical interface for audio playback and browsing is based on the tape recorder metaphor. In this presentation, the audio file is presented as a continuous stream which the user can navigate using play, stop, fast-forward, and rewind buttons.

However, this approach is fairly unsophisticated and current research has focused on optimising ways of letting users search for and browse audio content. Such research can take two different approaches, either focusing on improving the presentation and navigation of search results, or concentrating on novel ways of enabling navigation within a given file.

With regards to the first kind of approach, much of recent thinking focuses on presenting results "in a way that allows users to quickly identify the files that are really important for their particular information needs" [Hürst & Venkata, 2003]. Sometimes a brief amount of metadata relating to audio files (such as title, author, and file name) is displayed in search summaries, but this does not necessarily help a user to judge the file's relevance (or lack thereof.)

The SpeechBot project [Van Thong et al., 2001] attempted to address this problem by using speech recognition technology to automatically generate transcripts of audio files. Once the contents of an audio file have been transcribed, the retrieval task becomes essentially a text retrieval task: users enter search keywords and then can view the transcript to get an idea of the relevance of the result. Logan et al. [2004] mention that such an approach is advantageous because it is able to show the precise position of a word's occurrence within the file as a whole. However, the automatic nature of the transcription means that misrecognition of words or out-of-vocabulary terms can pose problems.

Although the SpeechBot transcriptions did contain such recognition errors, they were still deemed helpful in providing users with a general gist of the audio files' contents. They could then determine which files were worthy of further investigation based on these brief textual summaries. Overall, in the case of SpeechBot, it was determined that highlighting search query words in the transcription "was essential, and gave the user strong feedback on the relevance of the document even if the speech recognition output was sometimes hard to read and understand" [Van Thong et al., 2001: 12.]

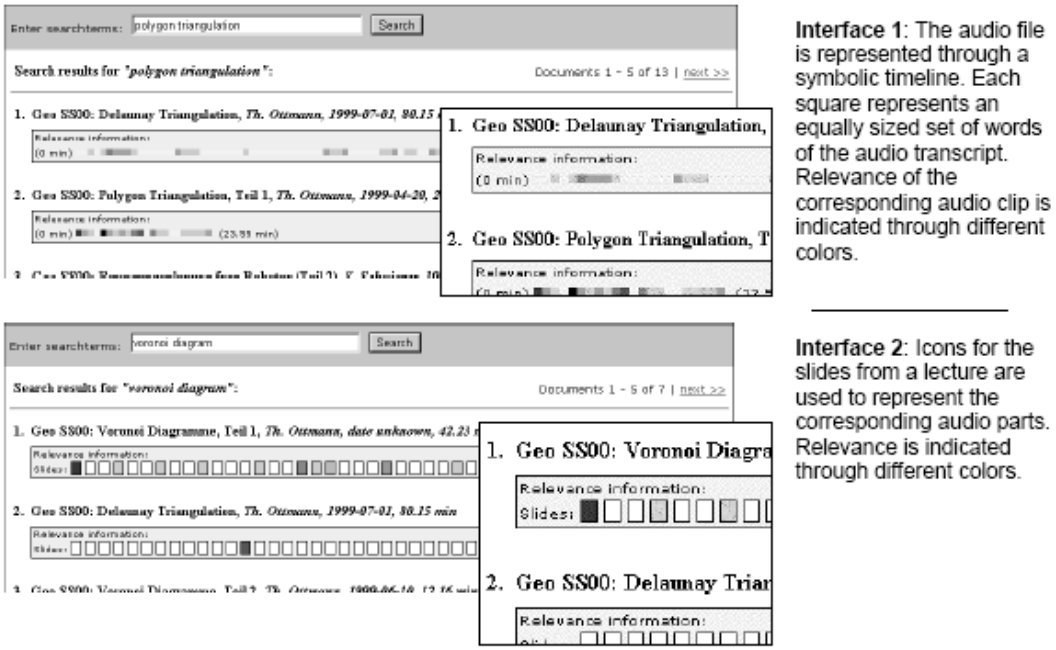
While SpeechBot is no longer publicly accessible on the Web, newer audio search sites devoted to podcast searching operate using a similar approach of automatic transcript generation. One such example is the PodZinger site (www.podzinger.com). This site uses a similar approach of presenting automatically generated transcripts that show the keywords in context.

The second type of approach to interaction with results involves navigating within a specific audio file. As discussed before, the "tape recorder" method is commonly used for this purpose but it often has drawbacks. First, it is time consuming to listen to a long audio file when only a small subsection contained somewhere within is of interest. Although many playback features offer some indication of a timeline (i.e. how much time has elapsed at a given point in the recording,) it can sometimes be difficult to go back and re-locate the exact position of a point of interest.

Again, new approaches have been explored in this area. Foote [1999] mentions a technology called SpeechSkimmer, which can compress audio recordings so that they can be played back at an accelerated but still comprehensible rate. Tucker & Whittaker [2006] tested different compression techniques in order to reduce the amount of time needed to listen to a file. Both excision (the removal of insignificant information) and compression (speeding-up) techniques were evaluated, and it was found that excision was generally more effective and better-liked by users than compression.

Even more useful than either of these methods, however, is the ability to skip directly to relevant portions of an audio recording. Hürst & Venkata [2003] explored ways of enabling this in the interface for a collection of archived lectures and presentations. They explored the idea of search using automatically generated transcripts but found in their case that these were not of high enough quality to be used even for gist or overall topic identification. As an alternative way of aiding visualisation, they designed a graphical timeline display with icons representing the subdivisions of the recording (in this case, each icon stood for one slide in a lecture.) The icons were then colour coded to show relevance to a search keyword: a darker coloured icon indicated higher relevance, suggesting that the keyword occurred most frequently in this section. Figure 7.12 displays two such timeline displays that were tested.

Figure 7.12: Graphical displays showing location of relevant words [Hürst & Venkata, 2003]

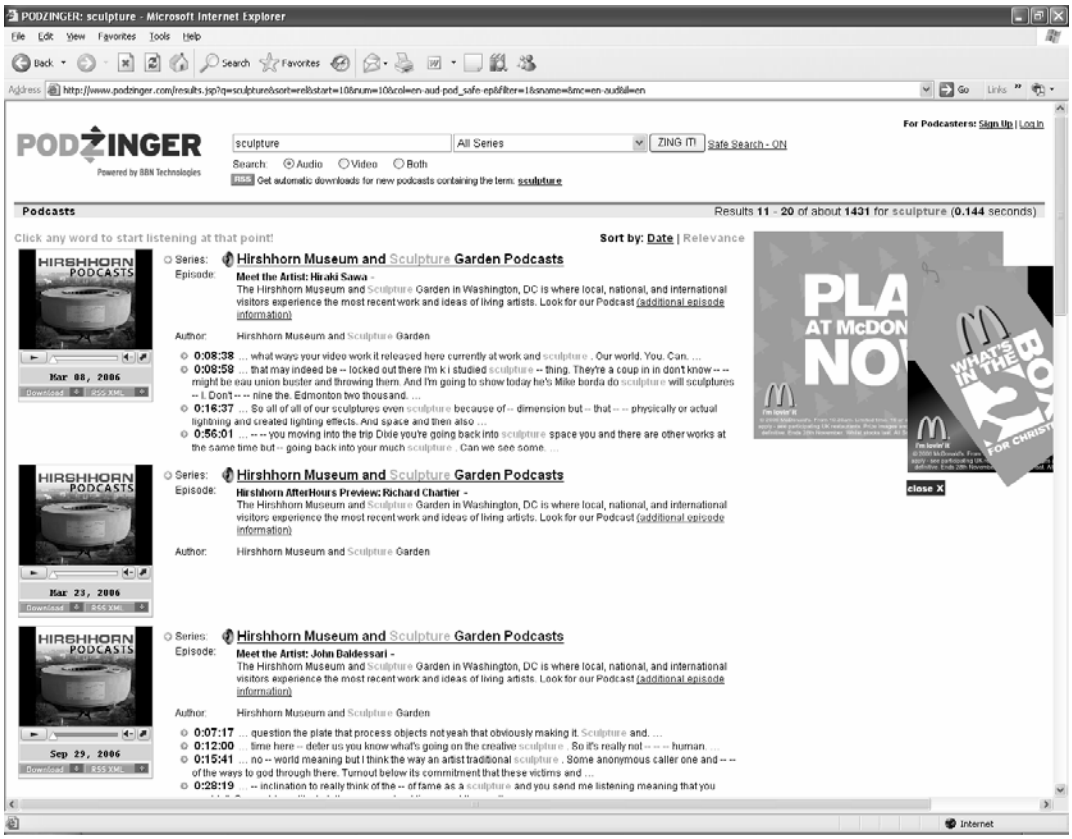


Interface 1: The audio file is represented through a symbolic timeline. Each square represents an equally sized set of words of the audio transcript. Relevance of the corresponding audio clip is indicated through different colors.

Interface 2: Icons for the slides from a lecture are used to represent the corresponding audio parts. Relevance is indicated through different colors.

The overall advantages of this design include giving easy access to the audio file at several intermediary points (often linked to a change in topic,) and visually displaying some indication of relevance. PodZinger also employs a means of quickly and easily locating the occurrence of keywords. Clicking on a term in the displayed transcription will automatically begin playing the audio file at the point where the word was mentioned (See Figure 7.13).

Figure 7.13: Sample results screen from PodZinger enabling the playing of the file at points where the keyword is mentioned.



In summary, it is not always easy to search audio files and display the results in a clear, informative format, but enabling users to get an overview of a file's content and its likely relevance is important. If they find a file that could be of interest, providing ways of quickly and efficiently browsing the content is also useful, particularly if the file is longer than a few minutes (or the length of the searcher's limits of patience.)

7.6 Example Multimedia Search Interfaces

Multimedia search engines can offer a variety of possible media formats to be searched. Based on a sample of 16 online multimedia search systems, Table 7.2 shows a breakdown of the number of combinations for each type (image, audio and video).

Table 7.2: Search Category Combinations Supported by Popular Internet IR Sites

Media types	Number of sites	Examples
Images Only	9	www.live.com www.clusty.com http://www.google.co.uk/imghp?hl=en&tab=wi&q=
Images, Audio, Video	4	www.alltheweb.com
Video Only	2	www.youtube.com
Audio & Video	1	www.singingfish.com

Table 7.2 summarises the main functionalities exhibited by the sample selected. Of the six sites that had content in more than one medium, only one of them (www.Singingfish.com) offered the possibility of searching several media types at once. For the rest, search had to be limited to a specific type (i.e. image OR audio OR video, but not a combination.) The results of the Singingfish search, however, are not separated by type.

Free text was the predominant means of searching. Only one site (www.YouTube.com) had the possibility of browsing by category. Most of the sites followed a similar layout and respected similar conventions. They were simple and based on the Google interface model. In terms of results presentation, again, clear conventions prevailed, with image results displayed in a grid and Audio/Video results shown as a list, often with a thumbnail and a brief description.

Table 7.3: Example online multimedia retrieval systems

Collection holdings	**Percentage**	**Example**
Images	86 %	See above
Audio	36 %	
Video	50 %	
Tabs for different media	60 % (3 of 5)	www.altavista.com
Searching functionalities		
Free text search	100 %	
Advanced search	53 %	
Search all types of media at once	20 % (1 of 5)	www.singingfish.com
Browsing functionalities		
Category list	14 %	www.youtube.com
Hierarchical browsing	0 %	
Tag cloud	7 %	www.youtube.com
Results		
Displayed in grid / rows	100 %	
Other display	7 %	www.live.com (infinite scroll bar)
Ability to refine search / change result layout	57 %	www.creative.gettyimages.com http://www.google.co.uk/imghp?hl=en&tab=wi&q=
Multimedia results segregated by type	40 % (2 of 5)	www.altavista.com
Recommendations / "more like this"	14 %	www.youtube.com
Clustering of results	6 %	www.clusty.com

7.7 Cultural Heritage Interfaces

Currently, most cultural heritage institutions have some sort of online presence in the form of a website. Museums and art galleries have homepages and sometimes specific archives or collections that are part of a larger body have web portals of their own. These websites often provide some degree of access to the associated institution's collection in a digitized format. The degree of material that is available and the sophistication of exploration of this content vary from site to site, depending on the resources available to the cultural heritage institution in question.

However, overall, a majority of these sites do have common features which include both search and browse functionalities at the very minimum. A summary of the relative proportions of functionalities taken from a sample of 56 cultural heritage sites is presented in Table 7.4.

Table 7.4: A summary of the functionality of selected multimedia search engines

Functionality	Percent	Example
Free text search	91 %	
Browse by category	71 %	www.archinform.net
Advanced search	70 %	
News/Calendar	61 %	www.tate.org.uk
Registration/login	45 %	
Multilingual	34 %	www.louvre.fr
Geographical search / Map	29 %	http://whc.unesco.org/en/map
Shopping	29 %	
Search within results / See "more like this"	29 %	www.fotolia.com
Ability to segregate multimedia results by type (if applicable)	29 %	www.archive.org
Feedback section	23 %	
Timeline / Search by time (12 sites total; 25% of these offer search by time only, 75% have a timeline (2 of the 8 were interactive)	21 %	www.birth-of-tv.org
View results in popup window	21 %	
Change results layout (order by..)	21 %	www.artandarchitecture.co.uk
Hierarchical browse	20 %	http://www.staffspasttrack.org.uk/
Sitemap	20 %	
Controlled vocabulary	9 %	www.tate.org.uk
Colour/layout search	7 %	www.hermitagemuseum.org
Query translation	5 %	www.fotolia.com
Multimedia results arranged by type	5 %	http://ec.europa.eu/avservices/home/index_en.cfm
Faceted browse	3%	http://orange.sims.berkeley.edu/cgi-bin/flamenco.cgi/famuseum/Flamenco
Allow user annotation	2%	BRICKS workspace

Overall, most of the sites surveyed offered basic, expected, useful ways of searching and browsing their collections but were not very interactive or advanced. As technological capabilities have improved, there has been an increasing realisation that the current functionalities for accessing cultural heritage information online can be enhanced and upgraded. For example, it has been argued that in

the area of humanities, a keyword-based search "is not sufficient because one is above all interested in *relations* e.g. between artists, their works, the friends, their studies, who they inspired, etc." [Benjamins et al., 2004: 433.]

Kravchyna [2004] surveyed five categories of users to assess their information needs when using museum websites. The categories included were (i) museum professionals, (ii) scholars/art historians, (iii) the general public, (iv) university students, and (v) high school teachers. Across all groups, primary purposes for using museum sites were to determine the main exhibits and activities of interest, to gain knowledge about museum collections, and to learn of any upcoming activities by consulting any available event calendars. Additional priorities that were unique to the scholar group were related to gathering information for research (i.e. looking for specific images or looking for textual information on a museum object.) Therefore, while some needs crossed group boundaries, there were also group-specific requirements.

Current research and projects are focusing on new ways to aggregate, search and display multimedia cultural heritage material originating from several different sources. A selection of these projects will now be discussed briefly.

7.7.1 Cultural Heritage Projects

There are a variety of projects, both past and present, focusing on some degree to the electronic cultural heritage of Europe. These include:

- The European Library project (www.theeuropeanlibrary.org)
  - focusing on searching the content of European national libraries
- MICHAELplus (www.michael-culture.org)
  - creating a multilingual, open source platform with a search engine able to retrieve objects from cultural heritage collections across Europe
- BRICKS (www.brickscommunity.org)
  - integrating existing digital resources into a shared and common digital library
- ECHO
  - making a web-based digital library service for the historical film collections of various European national audiovisual archives
- Birth of TV project (www.birth-of-tv.org)
  - (internet archive of films from the early days of European television)

These projects have used or plan to use a variety of methods to implement their creations; however, most of them rely on exploiting metadata, thesauri, and controlled vocabularies in one way or another. Another set of related projects (SCULPTEUR and its successor, eCHASE) adopt a more advanced, ontology-based system in order to describe complex relationships and enrich the searching or browsing process.

SCULPTEUR Project

The objective of the SCULPTEUR project was "to create a distributed multimedia digital library for storing, searching and retrieving of more diverse multimedia types, with significant support for 3D objects" (www.sculpteurweb.org). It particularly focuses on "new ways to create, search, navigate, access, repurpose and use multimedia content from multiple sources over the Web" [Addis et al., 2005: 1]. The project's main goal is finding new ways of searching and navigating online museum collections.

The SCULPTEUR functionalities include basic, common features such as free text search and controlled vocabulary. However, it also incorporates novel ways of searching by concept and content.

The concept search is based around the use of a common ontology (CIDOC CRM), which encourages interoperability. It is meant to serve as a unifying query interface for heterogeneous databases. The CIDOC-based structure enables one to visualise the ontology itself. In addition, the interface also

incorporates mSpace technology (for a sample, see http://beta.mspace.fm). MSpace facilitates the navigation of multidimensional spaces such as those provided by a given ontology; thus, it is essentially a form of faceted browsing.

With regards to searching by content, functionality provided allows users to find or compare objects based on colour, pattern, and shape. This can potentially simplify the search in various situations, depending on the searcher's objectives. Overall, it must be noted that these more advanced search features were not developed for use by the general public but rather for the interface's target audience (i.e. museum professionals or similar "power users") [Addis et al., 2005]. Other features of the SCULPTEUR interface include:

- A lightbox for storing search results
- Attribute map (graphical representation of metadata attributes)
- Results overview
- Query history

eCHASE Project

The eCHASE project draws on the past experiences of SCULPTEUR. Its objective is to create "a single, on-line site that provides a contextualized access point for the multimedia cultural content currently distributed across the museums, galleries, photo libraries and audiovisual archives of Europe" (www.echase.org).

Therefore, its mission is to link related content items from a variety of sources into a coherent whole, using aggregation and contextualization [Sinclair et al., 2005]. The eCHASE portal will focus in particular on content related to the cultural heritage of Central and Eastern European countries. Functionalities to be offered include:

- Searching and browsing of content (via text and context-based queries)
- A facility to collect and annotate objects (a lightbox)

Like SCULPTEUR, the eCHASE architecture will employ CIDOC CRM as a common metadata schema which is capable of describing complex relationships between the objects in the database. Once again, the mSpace system will be used for browsing, and as a result users will be able to navigate multi-dimensional spaces through interaction with the interface.

Other functionalities the project will provide include thesaurus navigation in the form of thesaurus trees or concept hierarchies, and a geographical gazetteer for visualizing place information. The former will present the structure of the data in a way that allows users to focus queries on a specific place in a specific country. The latter will utilize Google Map technology along with latitudinal and longitudinal data to present a zoomable map of the place in which a given object was created.

7.7.2 Typical Functionality

Timelines and Maps

Both SCULPTEUR and eCHASE are similar to MultiMatch in terms of their overall scope and goals. They share characteristics such as the use of the CIDOC ontology and have similar features (i.e. a lightbox, search and browse, etc.) However, some features proposed by MultiMatch go beyond the offerings of these similar projects. Differentiating features proposed by MultiMatch could include increased interactivity in browsing functionalities: for example, with the use of timelines and/or maps.

Bates, Wilde & Siegfried [1993] analysed humanities scholars' search strategies and noted that most online searches were based around subjects, as opposed to specific works or authors. Other popular search terms were related to geographical names, dates and historical periods.

According to Allen [2005: 260], while event-oriented timelines are commonly-used graphical devices, "surprisingly, only a few systems have employed *interactive* event-oriented timelines as a framework

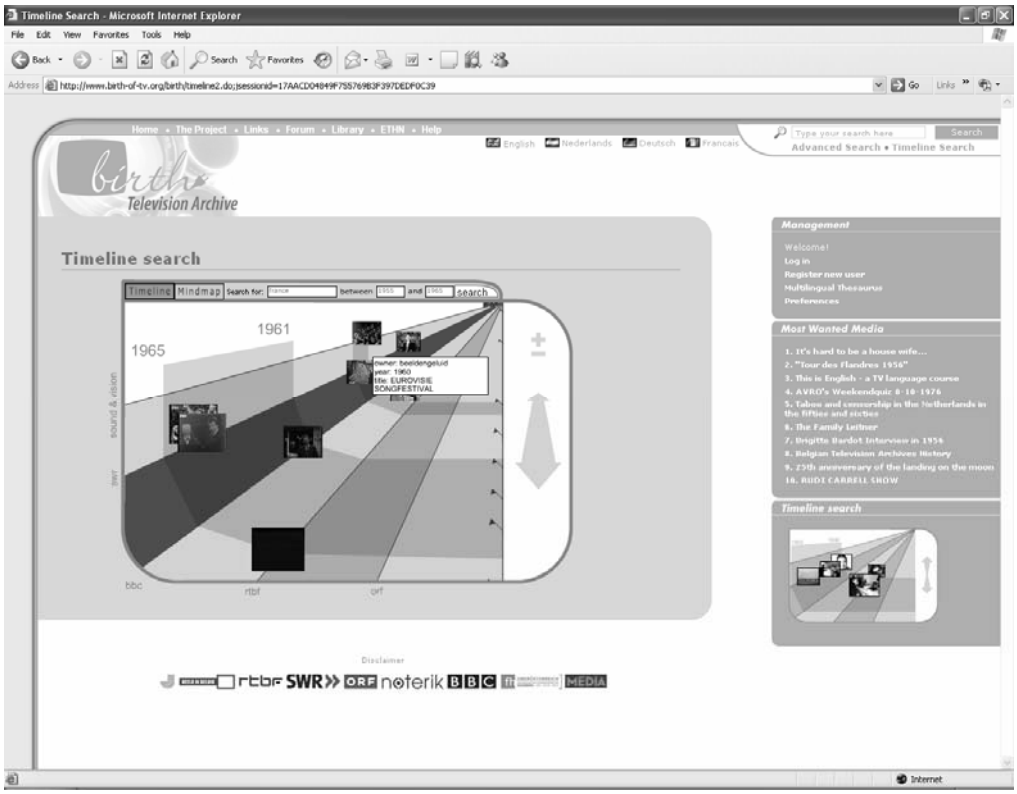
to support information access." The use of interactive timelines can be useful in the cultural heritage domain for several reasons.

First, investigations in this area often incorporate elements relating to place, time, topic, and creator, with a particular interest in change over time and relationships in context [Buckland & Lancaster, 2004]. Timelines can inform, show context, encapsulate ideas, and provide contextual links [Allen, 1995]. Secondly, a visual presentation is often easier to understand than a purely textual display [Shneiderman, 1998].

Examples of dynamic timelines (some of which are linked to maps) can be seen here. Some are related to cultural heritage and others are more history-oriented.

- www.ina.fr/fresque
  (Interactive, multimedia timeline of French radio and television history)
- http://digitalhistory.uh.edu/timeline/timelineO.cfm (Integrated map and timeline relating to American history)
- www.birth-of-tv.org/birth/timeline2.do (see Figure 7.14)
  (Birth of TV project's timeline of television history)
- http://ecai.org/Area/AreaTeamExamples/Korea/tm_korea.html
  (TimeMAP visualization of Korean history: integrated map and timeline)

Figure 7.14: Sample interactive timeline interface (Birth of TV project)



7.8 Discussion and New Directions

Current challenges in the area of online information retrieval include determining how best to classify, organise, and present objects of diverse origins and media types in a way that is intuitive and easy for the user to navigate. For example, although research indicates it may be beneficial, the use of a

faceted browsing feature has not been widely adopted by most websites. Additionally, relevance feedback is often unimodal and does not always help users to specify exactly what facets they would like to use to search for similar items (i.e. colour, subject, etc.) It can also be difficult to appropriately cluster results in the case where a query can have multiple meanings. With regards to cross language functionality in the form of query translation, again, this feature is not prevalent and when it is employed, it often does not function perfectly. Finally, there is the issue of the semantic gap query resolution as discussed previously.

These areas all represent potential opportunities for MultiMatch to experiment with new and potentially different means of improving the information-seeking experience. MultiMatch will exploit existing interfaces and incorporate ideas including the following:

- Faceted search and browse
- Multimodal search and reformulation (multimodal relevance feedback)
- Interactivity and exploration (variety of interaction methods)
    - Multiple ways to access the collection, i.e. multiple views (search/browse based on facets and time etc.)
    - Providing multimodal prompts, such as audiovisual surrogates (e.g. collection overviews), to assist the user in initiating searches and refining search parameters.
    - Use of workspaces, potentially to provide relevance feedback (dragging items into the workspace tells the system to "find more like this")
- Interaction and relevance feedback (through browsing)
    - Implementation of functionality to support the formulation of multimedia queries for complex needs, an example of this would be allowing the user to input both text (e.g. "van Gogh") along with an image selected from some pre-existing 'visual hints' gallery to assist in a query about the famous Dutch artist.
    - Implementation of some type of on-line storage facility (i.e. a "lightbox" analogous to the shopping cart on an e-commerce site) for the collection of relevant items along the way [Bates, 1989]. Items stored in such a shopping cart object could be retained or discarded as appropriate.
- Previews and overviews (dynamic queries)
    - Creating a more interactive search experience for the user
- Use of visual thesaurus to help bridge the semantic gap
    - also provides prototypical images for multimodal query expansion
- Use of multilingual thesaurus and facets
    - e.g. as implemented in the Birth of TV project
- Providing an adaptive, personalised interface
    - e.g. for images, relevance depends on work context, therefore rather than ranking images, we can create other displays and allow browsing

Overall, there are currently several related sites and projects with similar aims and functions to those of MultiMatch. MultiMatch must therefore find a way to distinguish itself from its "competitors." In one sense, it is unique in that it provides a set of characteristics (multilinguality and multimediality) that may exist elsewhere, but usually are not found together in this combination.

In the cultural heritage domain, people often use "creative and exploratory thought processes involved in translating conceptual ideas to visual instantiations" [Jörgensen & Jörgensen, 2002: 1357.] Given this, there are a number of areas in which MultiMatch can endeavour to improve upon current practices in terms of information seeking, retrieval, and presentation.

For example, most multimedia or cultural heritage sites follow fairly standard ways of presenting browse and search results, even though these may not be the most effective methods of doing so. Inspired by research on alternative means of visualizing search results, MultiMatch can consider adopting different and more interactive methods, including but not limited to clustered concept hierarchies, visual collages, fisheye views, or other methods beyond the standard thumbnail grid display.

Additionally, interactivity will be a main emphasis of the MultiMatch interface, since searching or browsing is often a fluid and evolving process in which users' needs and strategies may constantly change. How best to support these needs will be a major focus of MultiMatch which will draw on a user-centred approach to interface design that takes into account user input and requirements. Ways of facilitating interaction may include the development of features for storing items and searches, refining queries, giving relevance feedback, navigating between results, and exploring relationships between items on a variety of planes.

Given that the cultural heritage field is heavily based on themes and relationships between people, places, time periods, and media, it will be necessary to provide ways of describing and navigating said relationships, be this through a more advanced type of faceted browsing, using concept maps, or including interactive means of visualizing interactions or connections over time and geographical location (e.g., seeing when, where, and by whom artworks related to Shakespeare's "A Midsummer Night's Dream" were produced.)

The present state-of-the-art research provides a variety of technological or design concepts that enable new and innovative ways of interacting with virtual objects; however, many of these have yet to be implemented in practice on a wide scale. In theory, new ideas and concepts are meant to improve upon the weaknesses of current practice, but it is not always the case that these methods are appreciated by users. Therefore, by examining and testing a variety of approaches with potential user groups, MultiMatch can endeavour to build an interactive, innovative interface that is first and foremost successful at meeting its users' needs.

References

Addis et al. (2005). New Ways to Search, Navigate, and Use Multimedia Museum Collections over the Web. In J. Trant and D. Bearman (eds.). *Museums and the Web 2005:Proceedings*, Toronto: Archives & Museum Informatics, published March 31, 2005 at http://www.archimuse.com/mw2005/papers/addis/addis.html

Allen, R.B. (1995). Interactive Timelines as Information System Interfaces. Symposium on Digital Libraries, Japan, 175-180

ibid. (2005). A focus-context browser for multiple timelines. Proceedings of the 5th ACM-IEEE-CS joint conference on digital libraries, 260-61.

Armitage, L. & Enser, P. (1997). Analysis of user need in image archives. Journal of Information Science, 23(4), 287-299.

Bates, M.J. (1999). The design of browsing and berrypicking techniques for the online search interface. Online Review, 13, 407-424.

Bates, M.J., Wilde, D.N., & Siegfried, S. (1993). An analysis of search terminology used by humanities scholars: The Getty online searching project, report no. 1. The Library Quarterly, 63(1), 1-39.

Beale, R. (2006). Improving Internet interaction: From theory to practice. JASIST 57(6): 829-33.

Belkin, N. J. (2003). Interface techniques for making searching for information more effective. Retrieved October 4, 2004 from http://home.earthlink.net/~searchworkshop/docs/belkin-final.pdf

Benjamins, V.R., Contreras, J., Blázquez, M., Dodero, J.M., Garcia, A., Navas, E., Hernandez, F., & Wert, C. (2004). Cultural Heritage and the Semantic Web. In Proceedings of the First European Semantic Web Symposium, 433-44.

Bernard, K. and Forsyth, D. (2001) Learning the Semantics of Words and Pictures. In Proceedings of the Intentional Conference on Computer Vision, 2, pp. 408-415.

Boreczky, J., Girgensohn, A., Golovchinsky, G., Uchihashi, S. (2000). An Interactive Comic Book Presentation for Exploring Video. CHI Letters Vol. 2, No. 1

Brajnik, G., Mizzaro, S., & Tasso, C. (1996). Evaluating user interfaces to information retrieval systems: A case study on user support. Proceedings of 19th annual SIGIR Conference on research and development in information retrieval, 128-136.

Broder, A. (2002). A taxonomy of web search. ACM SIGIR Forum, 36(2), 3-10.

Buckland, M., & Lancaster, D.L. (2004). Combining Place, Time, and Topic: The Electronic Cultural Atlas Initiative. Digital Library Forum (D-Lib) Magazine, 10(5), 4.

Cai, D., He, Xiaofei., Li, Zhiwei., Ma, W-Y., and Wei, J-R. (2004) Hierarchical clustering of WWW image search results using visual, textual and link information. In: Proceedings of the 12th annual ACM international conference on Multimedia, 952-959.

Capstick, J., Diagne, A.K., Erbach, G. Uszkoreit, H., Leisenberg, A., & Leisenberg, M. (2000). A system for supporting cross-lingual information retrieval. Information Processing and Management , 36(2), 275-289

Chang S.F., Eleftheriadis, A., & McClintock, R. (1998). Next-generation content representation, creation, and searching for new-media applications in education. Proceedings of the IEEE, Vol.86(5), 884-904.

Chang, M. & Leggett, J. (2003). Collection understanding through streaming collage. In Proceedings of the Information Visualization Interfaces for Retrieval and Analysis (IVARA) Workshop, associated with the Joint Conference on Digital Libraries, Houston, Texas.

Chang, M., Leggett, J. J., Furuta, R., Kerne, A., Williams, J. P., Burns, S. A., and Bias, R. G. (2004). Collection understanding. In Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (Tuscon, AZ, USA, June 07 - 11, 2004). JCDL '04. ACM Press, New York, NY, 334-342.

Chang, S., J.R. Smith, M. Beigi & A. Benitez. (1997). Visual Information Retrieval from Large Distributed Online Repositories. Communications of the ACM, 63-71.

Christel, M.G., Hauptmann, A.G., Wactlar, H.D., Ng, T.D. (2002). Collages as dynamic summaries for news video. Proceedings of the tenth ACM international conference on Multimedia, 561-569.

Chu, H. (2001). Research in image indexing and retrieval as reflected in the literature. J. Am. Soc. Inf. Sci. Technol. 52 (12), 1011-1018.

Chu, H. (2006) Information Representation and Retrieval in the Digital Age, Information Today Inc (August 2003), ISBN: 1573871729.

Clough, P., Joho, H. & Sanderson, M. (2005). Automatically Organising Images using Concept Hierarchies. Workshop held at the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Workshop: Multimedia Information Retrieval, Salvador, Brazil.

Clough, P. and Sanderson, M. (2006). User Experiments with the Eurovision Cross-Language Image Retrieval System. In Journal of the American Society for Information Science and Technology (JASIST) Special Topic Section on Multilingual Information Systems, 57(5), 697 - 708.

Combs, T. T. A., & Bederson, B. B. (1999). Does Zooming Improve Image Browsing? In: Proceedings of DigitalLibrary (DL 99) New York: ACM, 130-137.

Cox, I. J., Miller, M. L., Omohundro, M., & Yianilos, P. N. (1996). Target Testing and the PicHunter Bayesian Multimedia Retrieval System. Proceedings of the Third Forum on Research and Technology Advances in Digital Library (ADL'96), pp. 66-75, Washington, D.C. USA. IEEE Computer Society Press.

Del Galdo, E.M., & Nielsen, J. (1996). International User Interfaces. New York: John Wiley & Sons.

De Troyer, O., & Casteleyn, S. (2004) Designing Localized Web Sites. In Proceedings of the 5th International Conference on Web Information Systems Engineering (WISE2004), 547–558.

Dorr, B., He, D., Luo, J., & Oard, D. (2003). iCLEF at Maryland: Translation selection and document selection. In C. Peters (Ed.), Working Notes for the CLEF 2003 Workshop.

Eakins, J. P. (1996). Automatic image content retrieval – are we getting anywhere? Proceedings of Third International Conference on Electronic Library and Visual Information Research (ELVIRA3), De Montfort University, Milton Keynes, pp 123-135.

Eakins, J.P. (1998). Techniques for image retrieval. Library and information briefings, 85, 1-15.

Eakins, J.P. (2000). Retrieval of Still Images by Content. Lectures on Information Retrieval, Springer Berlin/Heidelberg, pp. 111-138.

Eakins, J. & Graham, M. (1999). Content-based image retrieval: A report to the JISC Technology Applications Programme. Technical report, Institute for Image Data Research, University of Northumbria at Newcastle.

Eakins, J. Briggs, P. and Burford, B. (2004), Image Retrieval Interfaces: A User Perspective. Lecture Notes in Computer Science, Vol3115, pp. 628-637.

Eggink, J., Brown, G. (2004). Instrument Recognition in Accompanied Sonatas and Concertos. ASP, 217-220.

Enser, P. (1995). Pictorial Information Retrieval. Journal of Documentation, 51(2), 126-170.

Enser, Peter, Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science* 26(4), 2000, 199-210.

Enser, P. & McGregor, C. (1993), Analysis of visual information retrieval queries, British Library Research and Development Report, 6104.

Enser, Peter. & Sandom, Christine, Towards a comprehensive survey of the semantic gap in visual image retrieval. In: Bakker, E.M. et al. (eds.) *Image and video retrieval; Second International Conference, CIVR 2003*, Urbana-Champaign, IL, USA, July 24-25, 2003 Proceedings (Lecture Notes in Computer Science, Vol. 2728. Berlin: Springer-Verlag, 2003, 291-299.

Eurescom (2000). Multi-Lingual Web Sites: Best Practice Guidelines and Architecture (P923). Eurescom Project report, Available online: http://www.eurescom.de/Public/projectresults/P900-series/923d1.asp

Evans, D. (2006). From R&D to practice – challenges to multilingual information access in the real world. Presented at SIGIR Workshop on New Directions in Multilingual Information Access (MLIA), Seattle, Washington.

Flickner M, Sawhney H, Niblack W (1995). Query by image and video content: the QBIC system. IEEE Computer 28(9), pp. 23-32.

Foote, J. (1999). An overview of audio information retrieval. Multimedia Systems, 7: 2-10.

Forsyth, D. A. et al (1996). Finding pictures of objects in large collections of images, Proceedings International Workshop on Object Recognition, Cambridge, 1996.

Golovchinsky, G. (1997). Queries? Links? Is there a difference? Proceedings of CHI 1997, 407-414.

Goodrum, A. (2000). Image information retrieval: An overview of current research. Informing Science, 3(2), 63-66.

Gremett, P. (2006). Utilizing a user's context to improve search results. JASIST 57(6), 808-812.

Gudivada, V.N. & Raghavan, V.V. (1995). Content based image retrieval systems. Computer 28(9): 18-22.

Gupta, A. & Jain, R. (1997). Visual information retrieval. Communications of the ACM, 40(5), 70-79.

He, D., Wang, J., Oard, D., & Nossal, M. (2003). Comparing user-assisted and automatic query translation. In LNCS 2785, C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), Advances in cross-language information retrieval, 400-415.

He, D., & Oard, D. (2006). Studying the Use of Interactive Multilingual Information Retrieval. In New Directions of Multilingual Information Access, A workshop of Annual Conference of SIGIR 2006.

Hearst, M. (1999). "User Interfaces and Visualization". In: Baeza-Yates, R. & Ribeiro-Neto, B. (eds.), Modern Information Retrieval, 257-323. New York: ACM Press.

Hearst, M., et al. (2002). Finding the flow in web site search. Communications of the ACM, 45(9).

Henninger, S., & Belkin, N. (1996). Interface issues and interaction strategies for information retrieval systems. Conference companion on human factors in computing systems: Common ground, 352-353.

Hürst, W., & Venkata, L. (2003). Interface issues for accessing and skimming speech documents in context with recorded lectures and presentations. Proceedings of HCI International 2003, pp. 656-660.

Ingwersen, P. and Järvelin, K. (2005). The turn: integration of information seeking and retrieval in context. Dordrecht, The Netherlands: Springer.

Janecek, P., & Pu, P. (2004). Opportunistic search with semantic fisheye views. EFPL Technical Report: IC/2004/42.

Jörgensen, C., & Jörgensen, J. (2002). Image querying by image professionals. JASIST 56(12): 1346-59.

Karadkar, U., Nordt, M., Furuta, R., Lee, C., Quick, C. (2006). An exploration of space-time constraints on contextual information in image-based testing interfaces. ECDL 2006, LNCS 4172, 391-402.

Kravchyna, V. (2004). Information needs of museum visitors: Real and Virtual. PhD dissertation, University of North Texas.

Lee, H., & Smeaton, A. (2002). Designing the User Interface for the Físchlár Digital Video Library. Journal of Digital Information, 2(4).

Liu, H., Xie, X., Tang, X., Li, Z., & Ma, W. (2004). Effective browsing of web image search results. Proceedings of MIR '04, New York, 84-90.

Logan, B., Moreno, P., Van Thong, J.M., Marston, J., & MacCarthy, G.(2004). NewsTuner: A simple interface for searching and browsing radio archives. IEEE International Conference on Multimedia and Expo (ICME).

Lombardi, T., Cha, S., & Tappert, C. (2004). A graphical user interface for a fine-art painting image retrieval system. Proceedings of the 6th ACM SIGMM international workshop on multimedia information retrieval, 107-112.

Marchionini, G. (1992). Interfaces for end-user information seeking. Journal of the American Society for Information Science, 43(2):156-163.

Marchionini, G. (1995) Information Seeking in Electronic Environments, Cambridge University Press (May 26 1995), ISBN: 0521443725.

Markkula, M., & Sormunen, E. (2000). End-user searching challenges: Indexing practices in the Digital Newspaper photo archive. Information Retrieval, 1(4), 259-285.

Marlow, J. (2006). Designing a localisation strategy for Tate Online: requirements and recommendations. MSc dissertation for Master of Arts in Multilingual Information Management, University of Sheffield.

Minerva Project (2006). Multilingual Access to the digital European cultural heritage. http://www.mek.oszk.hu/minerva/survey/delir20060130. [Accessed 16 August 2006]

Mostafa, J. (1994). Digital image representation and access. In M.E. Williams (Ed.), Annual Review of Information Science and Technology, vol. 29.

Oard, D. W. (1997). Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries. D-Lib Magazine, December 1997.

Oard, D., & Gonzalo, J. (2002). The CLEF 2001 Interactive Track. Evaluation of Cross- Language Information Retrieval Systems, Springer-Verlag LNCS 2406.

Oard, D., Gonzalo, J., Sanderson, M., López-Ostenero, F., & Wang, J. (2004). Interactive Cross-Language Document Selection. Information Retrieval, Vol. 7 (1-2), 205-228.

Ogden, W., Cowie, J., Davis, M., Ludovik, E., Nirenburg, S., Molina-Salgado, H., et al. (1999). Keizai: An interactive cross-language text retrieval system. Paper presented at the Machine Translation Summit VII, Workshop on Machine Translation for Cross-Language Information Retrieval, Singapore, PRC.

Ogden, W.C., & Davis, M.W. (2000). Improving cross-language text retrieval with human interactions. Proceedings of the Hawaii International Conference on System Scuence (HICSS-33), Vol. 3.

Panofsky, E. (1955). Meaning in the Visual Arts: Papers in and on art history. Garden City, NY: Doubleday Anchor Press.

Park, G., Baek, Y., and Lee, H-K. (2005) Re-ranking algorithm using post-retrieval clustering for content-based image retrieval. Information Processing and Management, 41(2), 177-194.

Parker, E. B. (1987), LC Thesaurus for Graphic Materials: Topical Terms for Subject Access. Washington, D. C., Library of Congress.

Penas, A., Gonzalo, J., & Verdejo, F. (2001). Cross-language information access through phrase browsing. Paper presented at the 6th International Conference of Natural Language for Information Systems (NLDB'01), Madrid, Spain.

Peters, C. & Sheridan, P. (2001) Multilingual information access. In Lectures on information Retrieval, M. Agosti, F. Crestani, and G. Pasi, Eds. Springer Lecture Notes In Computer Science Series, vol. 1980. Springer-Verlag New York, New York, NY, 51-80.

Petersen P. and Barnett P. (1994). Guide to indexing and cataloguing with the Art & Architecture Thesaurus, The Getty Art History Information Program. OUP.

Petrelli, D., Hansen, P., Beaulieu, M., Sanderson, M. (2002). User requirement elicitation for Cross-Language Information Retrieval. New Review of Information Behaviour Research, 3, 17-35.

Petrelli, D., P. Hansen , M. Beaulieu, M. Sanderson, G. Demetriou, P. Herring. (2004). Observing Users - Designing Clarity: A Case study on the user-centred design of a cross-language retrieval system. JASIST, 55(10), 923-934.

Petrelli, D., Levin, S., Beaulieu, M., & Sanderson, M. (2006). Which User Interaction for Cross-Language Information Retrieval? Design Issues and Reflections. JASIST - special issue on Multilingual Information Systems. 57(5), 709-722.

Petrelli, D. and Clough, P. (2005) Concept Hierarchy across Languages in Text-Based Image Retrieval: A User Evaluation, In the working notes of the CLEF workshop, Vienna, Austria, 21-23 September 2005, online.

Rasmussen, E.M. (1997) Indexing Images, Annual Review of Information Science and Technology (ARIST), Vol. 32, pp. 169-196.

Renals, S., Abberley, D., Kirby, D., & Robinson, T. (2000). Indexing and Retrieving Broadcast News. Speech Communication, 32, 5-20.

Resnick, P. (1997). Evaluating Multilingual Gisting of Web Pages in Cross-Language Text and Speech Retrieval. AAAI Technical Report SS-97-05.

Resnick, M., & Vaughn, M. (2006). Best practices and future visions for search user interfaces. JASIST 57(6): 781-787.

Rodden, K., Basalaj, W., Sinclair, D., & Wood, K. (2001). Does organization by similarity assist image browsing? Proceedings of SIGCHI Conference on Human Factors in Computing, 190-197.

Rorvig, M. (1988). Psychometric measurement and information retrieval. In M.E. Williams (Ed.), Annual Review of Information Science and Technology (ARIST), 23, 157-189.

Rose, D.E. (2006). Reconciling information-seeking behaviour with search user interfaces for the Web. JASIST 57(6): 797-799.

Rose, D. and Levinson, D. (2004). Understanding User Goals in Web Search. In Proceedings of WWW 2004, New York, USA. ACM.

Rui, Y, Huang, T., Ortega, M., and Mehrota, S. (1998). Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval .IEEE Trans. on Circuits and Systems for Video Technology , Special Issue on Segmentation, Description, and Retrieval of Video Content, 8 (5), 644-655.

Rui, Y., & Huang, T. (1999). A novel relevance feedback technique in information retrieval. Proceedings of the 7th ACM international conference on Multimedia, 67-70.

Rui, Y., Huang, T., & Mehrotra, S. (1997). Content based image retrieval with relevance feedback in MARS. Proceedings of the International Conference on Image Processing, 815-818.

Shneiderman, M. (1998). Designing the User Interface: Strategies for Effective Human-Computer Interaction. Reading, MA: Addison Wesley.

Sinclair, P., et al. (2005). eCHASE: Exploiting Cultural Heritage using the Semantic Web. In Proceedings of the 4th International Semantic Web Conference, ISWC 2005, Galway.

Smeaton, A. (2000). Indexing, Browsing and Searching of Digital Video and Digital Audio Information. ESSIR 2000, LNCS 1980, 93-110.

Smeaton, A. (2002). Challenges for content-based navigation of digital video in the Físchlár Digital Library. LCNS 2383, 215-224.

Smeulders, A., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. Pattern analysis and Machine Intelligence, 22(12), 1349-1380.

Smith J. and Chang, S (1997). An Image and Video Search Engine for the World Wide Web. Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases V, 84-95.

Snoek, C., & Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. Multimedia Tools and Applications, 25, 5-35.

Swain, M., Frankel, C. and Athitsos, V. (1996). WebSeer: An image search engine for the World Wide Web. Technical Report TR-96-14, Department of Computer Science, University of Chicago.

Tucker, S., & Whittaker, S. (2006). Time is of the essence: An evaluation of temporal compression algorithms. Conference on Human Factors in Computing Systems (CHI), Montreal, Canada.

Turner, J. (1994). Determining the subject content of still and moving image documents for storage and retrieval: an experimental investigation. PhD thesis, University of Toronto.

Urban, J. and Jose, J.M. (2005). Exploring results organization for image searching. In Proceedings of INTERACT 2005: human-computer interaction, LNCS1973, v.3585, 958-961.

Van Houten, Y., Schuurman, J., & Verhagen, P. (2004). Video content foraging. CIVR Proceedings, pp. 15-23.

Van Thong, J., Moreno, P., Logan, B., Fidler, B., Maffey, K., & Moores, M. (2001). SPEECHBOT: An experimental speech-based search engine for multimedia content in the web Cambridge Research Laboratory Technical Report Series CRL 2001/06.

Veltkamp, R., & Tanase, M. (2000). Content-based image retrieval systems: A survey. Technical report UU-CS-2000-34, Department of Computer Science, Utrecht University.

Venters, C., Eakins, J. and Hartley, R. (1997). The user interface and content-based image retrieval systems, 19th Annual BCS-IRSG Colloquium on IR.

Voorhees, E.H. and Harman, D (2000). Overview of the sixth text retrieval conference (TREC-6). Information Processing and Management. 36. 1, pp. 3-35.

W3C (2003) W3C FAQ: International and Multilingual websites, http://www.w3.org/International/questions/qa-international-multilingual

West, K. and Cox, S.J., (2005). Finding an Optimal Segmentation for Audio Genre Classification. In Proc. 6th International Conference on Music Information Retrieval (ISMIR 2005), London.

White, R. W. and Ruthven, I. (2006). A study of interface support mechanisms for interactive information retrieval. JASIST 57(7), 933-948.

Yang, M., Wildemuth, B.M., Marchionini, G., Wilkens, T., Geisler, G., Hughes, A., Gruss, R., & Webster, C. (2003). Measuring user performance during interactions with digital video collections. ASIST 2003 Contributed Paper, 3-11.

Yang, M., & Marchionini, G. (2005). Deciphering visual gist and its implications for video retrieval and interface design. CHI '05 extended abstracts on Human factors in computing systems, 1877-1880.

Yeo, B., Yeung, M. (1997). Retrieving and Visualizing Video. Communications of the ACM, 40 (12), 43-52.

Yunker, J. (2003). Beyond borders - Web globalization strategies. Indianapolis, IN: New Riders Publishing.

Acknowledgments

The authors would like to thank all their colleagues in the MultiMatch project for many useful discussions and much input.

In addition, we must express our immense gratitude to the following external reviewers for having accepted to read through and comment earlier versions of this report:

- Antonella Fresa, Technical Coordinator of MICHAEL project, for the chapter on Cultural Heritage Technologies
- Nicholas Kushmerick, QL2 Software Inc., Seattle, USA (previously Dublin City University, Ireland), for the chapter on Information Extraction and Classification
- Arjen de Vries, Centrum voor Wiskunde en Informatica (CWI), The Netherlands, for the chapter on Multilingual/Multimedia Indexing.
- Douglas W. Oard, University of Maryland, for his suggestions for the chapter on Multilingual/Multimedia Indexing, some of which will be added in a future revision of this report.
- Jussi Karlgren, Swedish Institute of Computer Science (SICS), for the chapter on User Interaction and Interfaces

And finally Costantino Thanos, ISTI-CNR, for having the patience to read and provide valuable feedback with respect to the entire report.