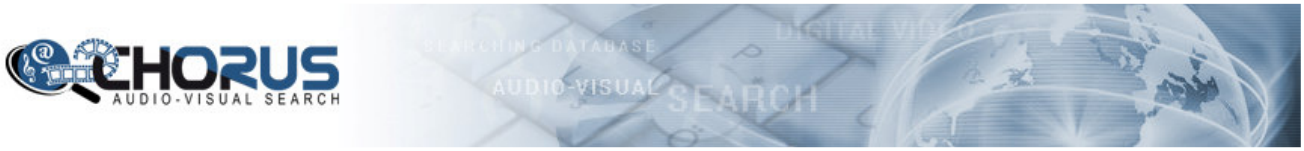


D2.1
State of the Art on Multimedia Search Engines

Deliverable Type *	:	PU
Nature of Deliverable **	:	R
Version	**:**	**Released**
Created	**:**	**23-11-2007**
Contributing Workpackages	:	WP2
Editor	**:**	Nozha Boujemaa
Contributors/Author(s)	:	Nozha Boujemaa, Ramon Compano, Christoph Doch, Joost Geurts, Yiannis Kampatsiaris, Jussi Karlgren, Paul King, Joachim Koehler, Jean-Yves Le Moine, Robert Ortgies, Jean-Charles Point, Boris Rotenberg, Asa Rudstrom, Nicu Sebe.

** **Deliverable type:** PU = Public, RE = Restricted to a group of the specified Consortium, PP = Restricted to other program participants (including Commission Services), CO= Confidential, only for members of the CHORUS Consortium (including the Commission Services)*
*** **Nature of Deliverable**: P= Prototype, R= Report, S= Specification, T= Tool, O = Other.*
***Version**: Preliminary, Draft 1, Draft 2,..., Released*

Abstract:

Based on the information provided by European projects and national initiatives related to multimedia search as well as domains experts that participated in the CHORUS Think-thanks and workshops, this document reports on the state of the art related to multimedia content search from, a technical, and socio-economic perspective.

The technical perspective includes an up to date view on content based indexing and retrieval technologies, multimedia search in the context of mobile devices and peer-to-peer networks, and an overview of current evaluation and benchmark inititiatives to measure the performance of multimedia search engines.

From a socio-economic perspective we inventorize the impact and legal consequences of these technical advances and point out future directions of research.

Keyword List: multimedia, search, content based indexing, benchmarking, mobility, peer to peer, use cases, socio-economic aspects, legal aspects

JCP-Consult	JCP	F
Institut National de Recherche en Informatique et Automatique	INRIA	F
Institut für Rundfunktechnik GmbH	IRT GmbH	D
Swedish Institute of Computer Science AB	SICS	SE
Joint Research Centre	JRC	B
Universiteit van Amsterdam	UVA	NL
Centre for Research and Technology - Hellas	CERTH	GR
Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e. V.	FHG/IAIS	D
Thomson R&D France	THO	F
France Telecom	FT	F
Circom Regional	CR	B
Exalead S. A.	Exalead	F
Fast Search & Transfer ASA	FAST	NO
Philips Electronics Nederland B.V.	PHILIPS	NL

Table of Contents

1. INTRODUCTION

The Chorus WP2 is dedicated to "Multi-disciplinary Analysis and Roadmap". The work is organized within thematic working groups dedicated to technical and non technical issues related to multimedia search engines. We list below the WGs topics and leaders:

- WG1: Audio-visual content indexing and retrieval technologies - Nicu Sebe (UvA) and Joachim Kohler (FhG)
- WG2: Evaluation, benchmarking and standards - Nozha Boujemaa and Joost Geurts (INRIA)
- WG3: Mobility, P2P, Heterogeneity - Jean-Yves le Moine (JCP)
- WG4: Socio-economic and legal aspects - Ramon Campano and Boris Rotenberg (IPTS)
- WG5: User interaction and group behavior Jussi Karlgren Jussi Karlgren and Åsa Rudstrom (SICS)
- WG6: Use-Cases and New services – Yiannis Kompatsiaris, Paul King (CERTH-ITI) and Christoph Dosch, Robert Ortgies (IRT)

The objective of this first deliverable is to establish the State of the Art regarding the critical issues identified through the WGs. We target to have a better view on the ongoing efforts mainly in the call6 European projects and (when possible) within the national initiatives. This information is needed before going a head in the Chorus roadmap activity and production. It is indeed necessary to have the clearest picture of the existing know-how and the existing problems as well to identify the bottlenecks. This first year effort will allow Chorus partners making the gap analysis between the expected new services and the necessary technological and non technological (socio-economic and legal aspects) evolution or mutation to make it possible. Of course, for the new services prospective, the WP2 will benefit from the feedback and the input of the Think-Tank participants and meeting (WP3 activity).

This document is organized as follows: In section 2 we set the scene of "multimedia search engines", which includes the users point of view, role and interaction (based on input from WG5), and existing uses-cases and services (input from WG6). In section 3, the state of the art from a technologinal point of view is produced inlcuding existing efforts within EC projects and NI[1] (input from WG1). Section 4 is dedicated to benchmarking and evaluation issues (input from WG2). In section 5, P2P and mobile search are investigated (input from WG3). Section 6 is dedicated to economical and social aspects of search and section 7 is targeting legal aspects. Both of these latter sections represent the input from WG4.

[1] National Intiatives

2. USER INTERACTION

Research in information access has heretofore mostly addressed the needs and necessities of topical text retrieval: that of focused search to find some known item or some topical information among a large collection of texts. Systems for text access have been traditionally evaluated in laboratory experiments to assess how well they meet the needs of topical retrieval - this has been done through the careful construction of test sets of simulated user needs and documents likely to meet those simulated needs. Many of the design principles for how users can be expected to act, how their actions can be simulated in laboratory testing, and how systems can be evaluated and designed on basis of user preferences will become obsolete or less pertinent once we move from mono-modal, mono-medial text documents to multi-medial information. This overview chapter will point out some of those trends and how the challenge they pose is met in today's designs information access; in future design cycles new efforts must be made, and the working groups and think tank processes of CHORUS will be contributing to the formation of such efforts.

2.1. Summary

Projects in the area of multimedia information access need to be vectored towards applications, needs, and requirements found or foreseen among users today and tomorrow. These requirements need to be formulated and based on studies and analyses made on data gathered from laboratory studies, observation studies, questionnaires and so forth. The evaluation of the scientific and other hypotheses can then be made with respect to the requirements as they are formulated.

The process of gathering user requirements, formulating needs and functionality, and evaluating hypotheses is complex and requires considerable methodological competence. The methodological tools and processes in question are research objects in their own right, and while the insight that users must be consulted in some form is fairly easy to come by, translating that insight to action and to adequate practice and craftsmanship is non-trivial: it is not to be expected that every multimedia information access research and development project will be able to provide competence in user studies. Examples of the intensive effort is given in the next chapter, which gives an outline of the effort of the current CHORUS projects as regards use case formulation: the projects have put considerable effort into anchoring their activities in a context of usage and use, but the concertation of these efforts and generalisation from results requires further analysis effort, since the common targets and goals are less defined than they might be.

To this end, a common framework of operationalised scenarios can be provided to future research endeavours, especially commission funded projects with similar targets and objectives – the projects will be able to relate their work to given cases, and if the cases are found to be constraining or ill suited to the research or development at hand, new cases can be defined using the previous ones as a model. This guarantees a higher level of compatibility between projects, and saves effort on the part of all.

As an added benefit this will function as a benchmarking on a high level of abstraction, allowing concertation meetings to be productive in terms of inter-project comparison, and as a method for funding agencies to channel research efforts to common goals.

One of the goals of the working groups on user interaction and use case formulation, as well as the goal of the think-tank activity currently under way, is to formulate common use cases and attendant scenarios for the consideration of future projects, calls, evaluation activities and the like.

2.2. User involvement in the system development process

More and more voices are raised to stress the importance of involving the user in the design and development of new services and products. The European commission also takes this stance. The

CHORUS Practitioner day in Amsterdam on July 11 2007[2] is one example when the user perspective was put forward at several occasions by representatives from the commission. Roberto Cencioni (head of Unit Knowledge and Interactive Content Technologies) discussed "Intelligent Content in FP7: Progress and Prospects". In his presentation, Dr. Cencioni suggested that the overall approach to research on intelligent content and semantics should be "centred around users, data and flows – a compelling 'use case' is as important as the underlying research". Also, in his presentation on "EU Research to master networked media future", Luis Rodríguez-Rosello (head of Unit Networked Media Systems) stated that future infrastructures, among other things, will need to "be user-centric, pervasive, ubiquitous and highly dynamic"; and that an important part of the on-going media revolution is that media becomes user centric and social.

There is no chance of providing new and more innovative services unless we look to the user for inspiration and understanding of needs. This viewpoint is shared by developers, researchers, commercial parties, and commission representatives alike.

2.2.1. User centered design

Modern systems development includes users in many parts of the development process, with a varying degree of involvement. Examples range from extended ethnographical studies of user behaviour to analyses of logged user behaviour in existing systems.

User centered design refers to design of system functionality starting from the user's perspective. Europe, in particular the Scandinavian region, has a long tradition of working with users to ensure that the systems produced are indeed suited to user needs and thus will be taken into use. User involvement reaches far beyond the design of interface components although a high degree of usability (Nielsen 1994) should always be strived for. What is important instead is to solve the right problem, i.e. to understand what the user needs the system to do.

2.2.2. The HCI perspective

Human computer interaction, HCI, is the research field concerned with the design of the borderland where humans and computational systems meet. Research in this area has resulted in a good understanding of the traditional windows-icons-menus-pointing way of interacting with computers. Human abilities and limitations have been taken into account from physiological, ergonomical and psychological standpoints, resulting in design and usability principles and guidelines for surface as well as structural aspects of human-computer interaction.[3]

Following the evolution where almost any person with any background and schooling has come to be a potential computer user, there is a growing need for expanding the view of HCI from the single human sitting alone at her desk interacting with some application on her stationary computer.

First and foremost, this user is a human and cannot be handled as part of the machinery. Humans do not act according to a set of rules as computers do. Within Computer Supported Cooperative Work (CSCW) it was early recognised that the actual work practices employed by people are very different from the formal procedures describing their work, and indeed, from peoples' own view of what it is that they do. For example, Bowers et al. (1996) found that problems with introducing new technology into a print shop were due to major differences between formal and actual procedures used for scheduling work. People engage in much more purposeful and artful ways of dealing with the complexities of real life. In order to design the right system functionality, the system designers need to take the actual working practices into account; and the experts on this topic are those

[2] From presentations at the Chorus Practitioner day at CVIR 2007 in Amsterdam, July 11 2007. See the Chorus website for further information on this event.

[3] There are numerous textbooks and references on these terms. For short descriptions of each term, I suggest turning to the socially constructed Wikipedia dictionary at http://wikipedia.org

performing the work, i.e. the future system users. Europe, and in particular Scandinavia, has a long tradition of involving users in the development process.

Another factor that has had a large impact on HCI is that today's computers and other devices to a large extent are mobile. One implication of mobility is of course that people may take their computers with them and leave their desks. Thus, computational systems can no longer be designed for a known and controlled environment – they may be used in any physical or social scenario. The borderlines between work and the rest of people's life become blurred. Moreover, the computer in the box on the desk can no longer be taken for granted. It may be replaced with one or several small devices with different interaction capabilities.

In addition to physically leaving the desktop, computational systems have long also been expanding their social context from the single, isolated user to the networked and networking user. In the last twenty years much research effort has therefore been spent on understanding and operationalising aspects of the context in which computational systems are used. These considerations and many other call for a different view on human-computer interaction that will take into account the versatility of human life. Studies of actual behaviour are necessary, triggering the introduction of ethnographic methods borrowed from the social sciences.

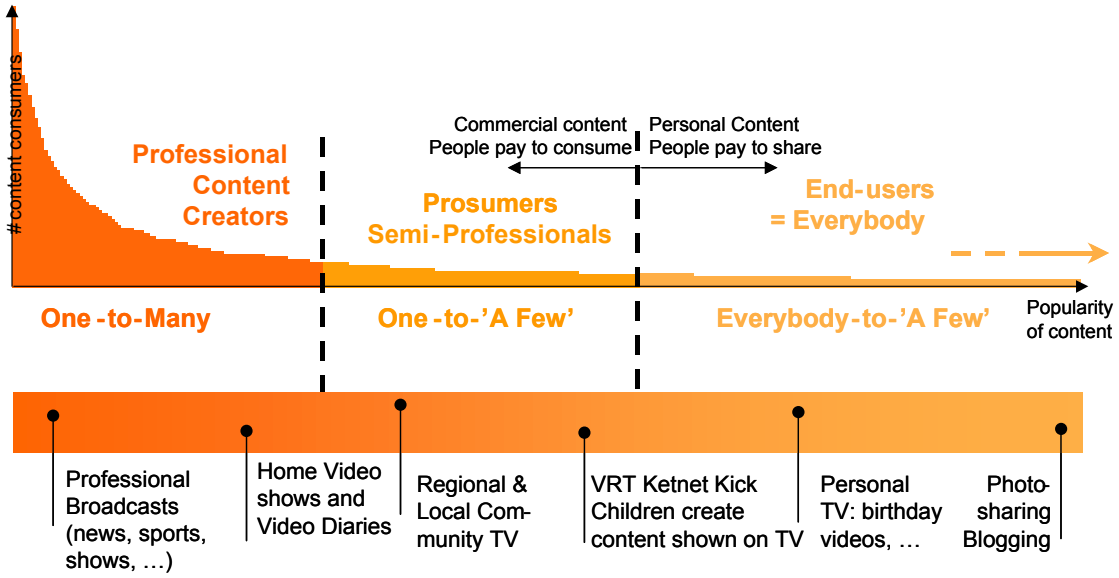
2.3. User generated content

The recent emergence of systems that allow user generated content stresses user involvement even more. This is particularly important in the area of multimedia search, since the content provided typically is multimedial. In addition to providing the actual content, users also provide structure to this content, i.e. the folksonomy of index terms emerging in Flickr. The appearance of user provided content (often referred to by the catchphrase "Web 2.0"), and the necessity of automatic analysis thereof (sometimes referred to as "Web 3.0") is one of the trends identified at the recent CHORUS workshop on National Initiatives [CHORUS Deliverable 4.3, November 2007].

Users are thus no longer restricted to being consumers of content. Quoting from a white paper produced by the "User Centric Media Cluster of FP6 projects", "[…] society is shifting from mainstream markets to individual and fragmented tastes where citizens evolve from a passive media consumer of mainstream content towards an active role in the media chain (see figure below). "[4]

A particular issue with user generated content is that not only do users create their own material from scratch, such as home video clips; they also use, re-mix and edit existing audiovisual material, treating the internet as a gigantic database of content. This poses new demands on multimedia search algorithms, both to provide support for users to generate such content, to find pertinent material to sample, clip, and combine - but also to rights holders who wish to track usage and modification of their materials in new and unexpected contexts. Designing tools for this sort of retrieval - beyond the most immediate ad-hoc services - will require new insights in user action, and these insights are not obviously capturable within a text retrieval frame-work, where e.g. the concepts of sampling and recombination have less application to user action, and where content analysis is on an entirely different level of complexity.

[4] "User Centric Media White Paper", created by "User Centric Media Cluster of FP6 projects". Coordinated by Networked Media Unit of the DG Information Society and Media of the European Commission. To be available from the website. http://www.



2.4. User centered design of multi-medial information access services

In a rapidly evolving situation such as is the case for the field of multi-medial information access services, any well and detailed specification of usage is likely to go stale and break down before the life cycle of the system is at end. Instead, we envision that a typical development project will use a modified rapid-prototyping inspired design model, consisting of fast and frequent iterations between low fidelity designs and concepts on the one hand and use case verification, user interviews, and feedback on the other, where the expert informants used as a basis for the usual rapid prototype evaluation are replaced with use cases.

The use cases are used for the informed design of interaction points – first specifications of what tasks the system will be required to fulfil. The user centred cyclic procedure will then collect experience on the adequacy of rapid prototype design sketches and refine the design for a concrete and deployable tool.

Previous research on textual retrieval systems could base its efforts on understood and under-specified notions of usage, based on topical retrieval of text; whatever the usage scenario, an underlying topical text retrieval engine could be taken as granted. Moving from text to image, video and other forms of potentially non-topical information sources will invalidate both the concrete feature extraction and content analysis done by the term occurrence statistics components of the retrieval systems as well as their target metrics: what are users looking for in video retrieval, and how do we envision they do so? Future research efforts in multi-medial retrieval must extend the target notion of topical relevance to cover other types of usage, and formulate use cases to cover new types of user action.

2.4.1. Beyond "relevance" as a target notion - How can we formalise our understanding of what users are up to?

The concept of *relevance* lies at the convergence of understanding users, information needs, items of information, and interaction. It ties together all proposed and current research projects in context sensitive information access. Relevance – the momentary quality of an information item that makes it valuable enough to view, read, or access in any way – is a function of task, document characteristics, user preferences and background, situation, tool, temporal constraints, and untold other factors.

In contrast, “Relevance”, as it is understood in evaluating information retrieval systems today is based on the everyday notion, but formalized further to be an effective tool for focused research.

Much of the success of information retrieval as a research field is owed to this formalization. But today, the strict, abstract, and formalizable relevance of the past decades is becoming something of a bottleneck.

"Relevance" does not take user satisfaction, quality of the information item, or reliability of source or channel into account. It is unclear how it could be generalized to the process of retrieving other than factual accounts. It is binary, where the intuitive and everyday understanding of relevance quite naturally is a gliding judgment. It does not does not take sequence of presentation into account - after seeing some information item, others may immediately lose relevance. And most importantly, it is completely abstracted away from every conceivable context one might care to investigate. This includes the various types of contexts the information item, the reader, the originator, and the session may be engaged in. (See e.g. Mizzaro, 1997 and 1998, for an overview of how relevance can be deconstructed.)

Trying to extend the scope of an information retrieval system so that it is more task-effective, more personalized, or more enjoyable will practically always carry an attendant cost in terms of lowered formal precision and recall as measured by relevance judgments. This cost is not necessarily one that will be noticed, and most likely does not even mirror a deterioration in real terms – it may quite often be an artefact of the measurement metric itself. Instead of being the intellectually satisfying measure which ties together the disparate and vague notions of user satisfaction, pertinence to task, and system performance, it gets in the way of delivering all three. Extending the notion of relevance so that it does not lose its attractive formalizable qualities but still takes context into account is not a straightforward task, and certainly has been attempted in various research endeavours with the text retrieval field in the past.

Extending information access beyond that of single-user single-session retrieval of factual items for professional use from text repositories, we find that multi-media, multi-user interaction, groupware, context-aware systems, user-generated content, entertainment use cases and various other features that broaden the interaction to need a new target notion, beyond that of relevance.

The notion of pertinence, user satisfaction, and context-sensitive relevance will occupy such a central position as to make it completely crucial for some extension to be agreed upon in the field, if the benefits of topical relevance to text retrieval can be emulated. If the concept of relevance is deconstructed, and information access systems made to model both reader and originator, we will better be able to satisfy the needs of information seekers, both professional and incidental.

The MIRA research project at Glasgow (cf. Mira research theme manifesto) note in their manifesto that quantitative evaluation and measurements from traditional information retrieval research do not transfer readily to new and emerging applications, such as multimedia technology. Qualitative evaluation of the new application in terms of user needs, goals, satisfaction has not been attempted yet (at time of writing). This is changing. An example of going beyond topical relevance for understanding user preferences was given by the CLEF interactive track (iCLEF) in year 2006. Previous iCLEF experiments have investigated the problems of foreign-language text retrieval and question answering, but moved to investigating image retrieval in many languages, with target notions such as "satisfaction" or "confidence" (Karlgren, Clough, and Gonzalo; 2006). Similarly, user satisfaction is a target notion (of several) in the Open Video Library project (Marchionini, 2006).

2.4.2. Use cases as a vehicle

Evaluation across projects, systems, and programs will be considerably simplified through cross-program formulation of use cases. Use cases are informally held descriptions of how a system is intended to be used or how it might be used. It is formulated as a goal oriented set of interactions between external actors, primarily users, and the system: it answers question such as "Who does what?" and "For what purpose?" (Jacobson et al., 1992; Cockburn, 2002).

Use cases track the requirements which are necessary to address in the development phase, and leave under-specified what needs to be left unattended, without bias to technical solutions. Most importantly, the use case should describe the user on an appropriate level of detail, take its point of departure from the goal of the user, and should describe what sequence of actions meets that goal.
Use cases should not address technology directly: the interaction should be described without dealing with system internals and do not need to specify platform or hardware. A scenario, describing system use, is an instance of a use case; use cases should attempt to generalise from the specifics of a scenario.

From a project point of view, use cases, formulated in the beginning steps of the project, help focus project attention on pertinent goals, and help prune project effort to avoid following paths of investigation which may be interesting but do not further those goals. This is a project-internal function, and is the most obvious benefit of putting effort in the formulation of use cases with attendant scenarios. But use cases have multiple functions in a project. In addition to helping project management by providing a challenge for the project personnel use cases are a convenient way of informing outside partners and others of project goals, objectives, and ambition.

From an external point of view – and this is what concerns us for the purposes of this report – use cases are a useful tool for formulating success criteria and benchmarking for e.g. funding agencies or peer review; they can, if well-formulated and accepted by the research community, serve to define a research path, to instantiate and operationalise research issues which otherwise might be left unanswered or unnoticed.

A case in point is that of evaluation of multimedial retrieval systems. The target notion of relevance, with its companion evaluation metrics precision and recall, is ubiquitous in text retrieval evaluation, and is carefully designed to be neutral with respect to usage and differing user needs. This lack of explicit use cases has not hindered the evaluation of text retrieval systems to be a useful research vehicle to further the goals of system development. This has led to the systems most in use today being very efficient but also very similar. It has also led to a growing awareness of text search as a commodity: new services will be built on top of text search, not to replace it. The realisation that new services must be evaluated by their own criteria has led to the proposal and formulation of more carefully designed target notions with parameters able to model differing use cases (e.g. Järvelin and Kekäläinen, 2000) which allows the formal and quantitative evaluation of use cases, given their translation into target requirements. This development gives us the possibility of formal and rigorous evaluation even while aiming for different use and different services, retaining the best of both formal evaluation and tailoring requirements. An experimental framework, which invites the formulation of scenarios and tasks, has been proposed by e.g. Borlund (2003). Within the framework of a CHORUS working group, we will be able to discuss the user-centered approach to formulate common ground.

Evaluating multimedia search systems cannot be done directly within the text evaluation framework. Their character is different in important ways – images, video etc, do not wear their semantics on their sleeves in the way texts do, given the ease with which words can be extracted to become content cues: the target notion of relevance must be rethought to cope with a different operation framework. This affords the field the opportunity of rethinking which level of abstraction one might want to design for, and to formulate target notions for evaluation accordingly. It also is an opportunity for funding agencies to formulate more application oriented goals for the research community, and – in cooperation with the research community – to provide target notions for those goals.

2.5. Use Cases in Current Research Projects

In summary, involving users, through one of many instantiations of user-centered system design processes, is not only desirable, but essential for the provision of future valid and reliable results in research and development of competitive services in the arena of multimedial retrieval and access.

This can be accomplished either by research and development projects basing their work on an in-depth study of users, usage, and contexts - or by the informed selection (and possibly reformulation or modification) of some existing use case in the area. The formulation of such descriptions of pertinent usage factors allows some projects to concentrate on system-oriented research efforts, improving the working of their technology or algorithms. Other projects can provide the knowledge needed to design tools and services appropriately. Yet others to prove or disprove integrative efforts given by components developed by preceding projects. An important facet of such use-case based research is that use cases lay the table for designing appropriate evaluation schemes: without statement of what needs the effort is designed to address, evaluation risks not guaranteeing validity of results – with explicit formulation of needs, or reference to current practice in the field, this risk is neatly addressed.

In the following comprehensive report of use cases employed by the various projects under the CHORUS umbrella, we find variation as regards scope, abstraction, and technological boundedness of use cases. This variation is to be expected, given the differing points of departure of the various projects – but there are obvious similarities to work from as well.

2.6. Use-cases and Services

2.7. Overview: The CHORUS Analysis of Use Cases

Three specific CHORUS objectives are supported by this document:

- Integration and strengthening of the European Research Area by stimulating interaction and coordination on a EU level in the area of audio-visual search engines;
- Creation of a 'holistic', multi-disciplinary community of researchers sharing a common approach for the implementation and realization of audio-visual search engines in EU
- Identification of multi-technological topics of common interest, initiation of discussion on these topics, and development of shared views on how best to approach these technological issues.

The last objective is most relevant to the current deliverable. A Use Case scenario engenders a specific description of a problem to be solved. Research problems are understood and described in many different ways depending on the background, training and experience of the researcher. Without coordination among sibling research efforts, resources tend to be mis-allocated to efforts that have already been adequately investigated. In other words, the wheel is re-invented many times. With coordination among research partners within a common domain, prior solutions can be adopted and improved and freed resources can be brought to bear on new problems. This facilitates a quicker research and development cycle.

The goal of the Use Case summary is to ensure the following:

- Appropriate understanding (framing) of problems, methodologies, products, and users
- Standardization of identification and description of problems, methodologies and products
- Re-use of prior methodologies and products
- Identification of communities and industries that could benefit from the research

This enables CHORUS and the relevant community to identify research efforts that have already been adequately addressed using a given methodology so that prior solutions can be redeployed or further developed in an effective manner. Areas of interest that are not being sufficiently investigated will become more visible as well. Finally, a standardized framework for problem description and methodology deployment will help participating projects to compartmentalize and focus their research efforts effectively.

The two most important accomplishments of the current review are (1) the identification of a standard set of data dimensions for Use Cases, and (2) the development of a survey tool for standardizing Use Cases among projects. Although the summarizing data we present provide some insight into overall patterns among projects and national initiatives, it is rough. This review should be seen as a step towards understanding how to look at Use Cases in the future and how to collect better Use Case data.

2.8. Participating Projects

The following projects have been included in the current review: DIVAS, RUSHES, SAPIR, TRIPOD, VICTORY and VITALAS. Although we planned to include the VIDI-VIDEO, SEMEDIA and PHAROS projects when they are made available.

In addition, the following National Initiatives are included in this review: iAD, IM2, MultimediaN, MundoAV, QUAERO, and THESEUS. These initiatives consist of many projects below them which were not individually analyzed. Rather, overall project goals were reviewed for clues to the data dimensions identified from the initial project reviews mentioned above.

2.9. Dimensions of Data Analysis

Use Case scenario data has been parsed and analyzed along six dimensions. Values for each dimension have been categorized and normalized across projects. The result is a clean, standardized format for project descriptions. Dimensions are identified and defined below:

- Actions – The research goal.
- Corpora – The source of data used for analysis. Strictly speaking, all projects are working on Multimedia content. However, a distinction has been made between textual descriptions of multimedia (and their types) and the multimedia content itself (and its corresponding type).
- Methods – Standard methodologies needed to accomplishing the Action. These are the project requirements.
- Products – The deliverable that will result from the stated Action.
- Users – Specific users who could benefit from the stated Product.
- User Classes – Categories of users. This helps to identify industry sectors that could benefit from the research effort.

Multiple descriptive entries exist for each project if more than one value was found in the following dimensions: Action, Corpora, Method or Product.

2.10. Data Analysis

2.10.1. Actions

Summary

Stated project goals have been found to fall into the following standard set of Action categories. Occurrences for each Action across projects/national initiatives is indicated in the right column. In other words, *Retrieval (Browse)* was found to be an Action among three projects or national initiatives.

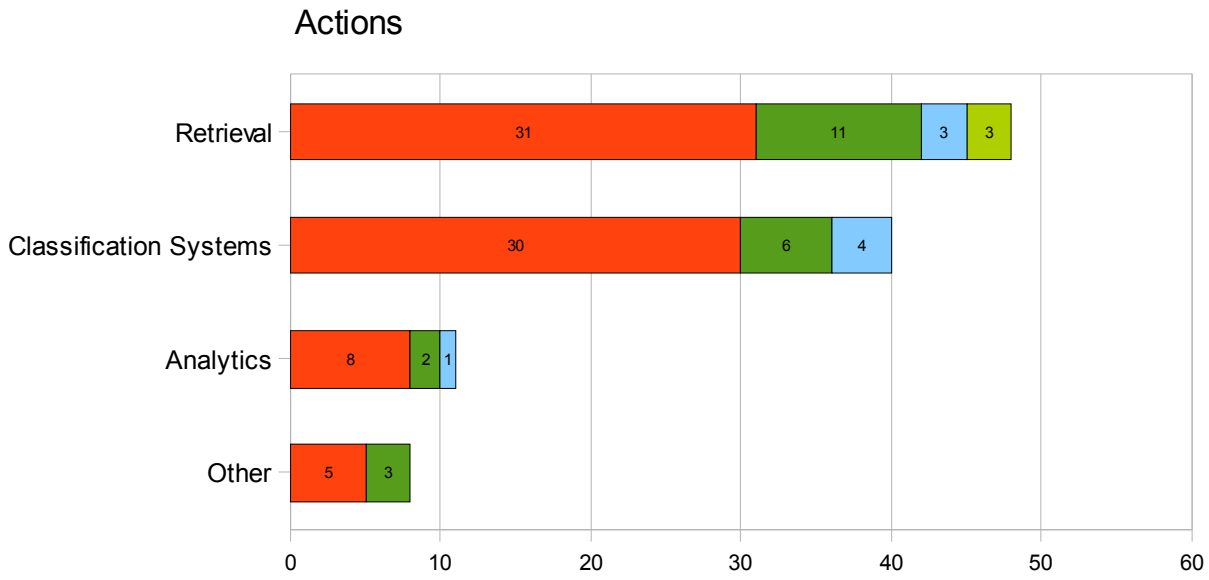
Categories and their corresponding subcategories listed below are sorted in descending order of occurrence. Occurrence numbers correspond to the numbers listed in the bar graph sections.

Retrieval (Search) is information retrieval characterized as the identification of a specific resource. *Retrieval (Browse)* is information retrieval characterized by exploratory behavior within a content collection for the purposes of discovery and research. Strictly speaking, *Content Delivery* is a type of Search (as opposed to Browse), since it aims to identify and deliver specific resources. However, we have maintained a separate sibling category for this Action in order to recognized the unique requirements of an overall system designed for the delivery of content. *Classification Systems* refers to the development of tools, and research into methodologies, for the purpose of enabling the classification (i.e., indexing) of artifacts or resources.

RETRIEVAL	**48**
Retrieval (General)	31
Retrieval (Search)	11
Retrieval (Browse)	3
Content Delivery	3
CLASSIFICATION SYSTEMS	**40**
Extraction/Indexing	30
Classification Systems (General)	6
Personalization	4
Analysis	**11**
Analysis (Multimedia)	8
Analysis (General)	2
Analysis (Text)	1
OTHER	**8**
Not Specified	5
Vague	3

Table 1: Action

The graph below conveys the information above in a more intuitive manner. Graph bars correspond to and are labeled by the major categories (i.e., Retrieval, Classification Systems, Analysis, Other), whereas subcategories (i.e., Content Delivery, Extraction/Indexing, Personalization) are represented as color-coded bar sections.



Analysis

As the graph above illustrates, the overwhelming majority of projects or national initiatives are either involved in *Retrieval* efforts (48 of 107, or 45%) or the development of *Classification*

Systems (40 of 107, or 37%). Most efforts within *Classification Systems* are focused on *Extraction/Indexing* (30 of 40, the first red section of the graph bar).

Unfortunately, Use Case data tended to be too general to determine with sufficient specificity the type of retrieval effort underway within most projects. Therefore, 31 of 48 are categorized as *Retrieval (General)*. However, we were able to determine that 11 of 48 projects are involved in *Search*, 3 of 48 are looking into *Browse* methodologies and another 3 of 48 are investigating *Content Delivery*.

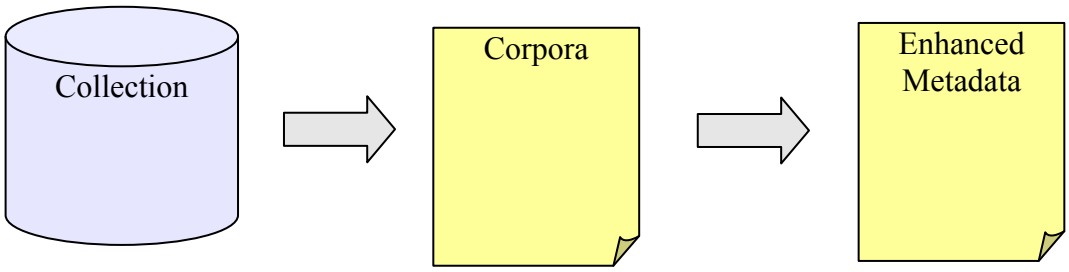
Surprisingly, only 11 projects or national initiatives have been found to be investigating *Analysis* techniques. *Multimedia Analysis* are being researched by a relatively meager 8 projects or national initiatives. A single project has stated that it is actively involved in researching *Text Analysis* techniques.

The small number of projects or national initiatives involved with *Multimedia Analysis* may be attributable to the erroneous classification of Actions for projects and national initiatives due to ambiguously defined goals within submitted Use Cases.

2.10.2. Corpora

Summary

Corpora does not necessarily describe or name the collection, resource or set of assets which are being investigated within a project or national initiative (i.e., multimedia news programs). Rather, a *Corpus* is defined here as the set of data that is used to produce a result stated by an *Action*. When dealing with a collection of audio and text, it generally represents an extraction of the collection or resource (viz., a product of transcription or annotation activities) and typically needs to be transformed in some additional way in order to semantically enhance (i.e., classify) the original set of resources. In other words, it is the immediate data precursor to the goal stated in the Action. However, this distinction does not always apply. In particular, when a project or national initiative is exploring Multimedia Analysis techniques to be applied directly to the collection, there will be no intermediate *Corpus*.

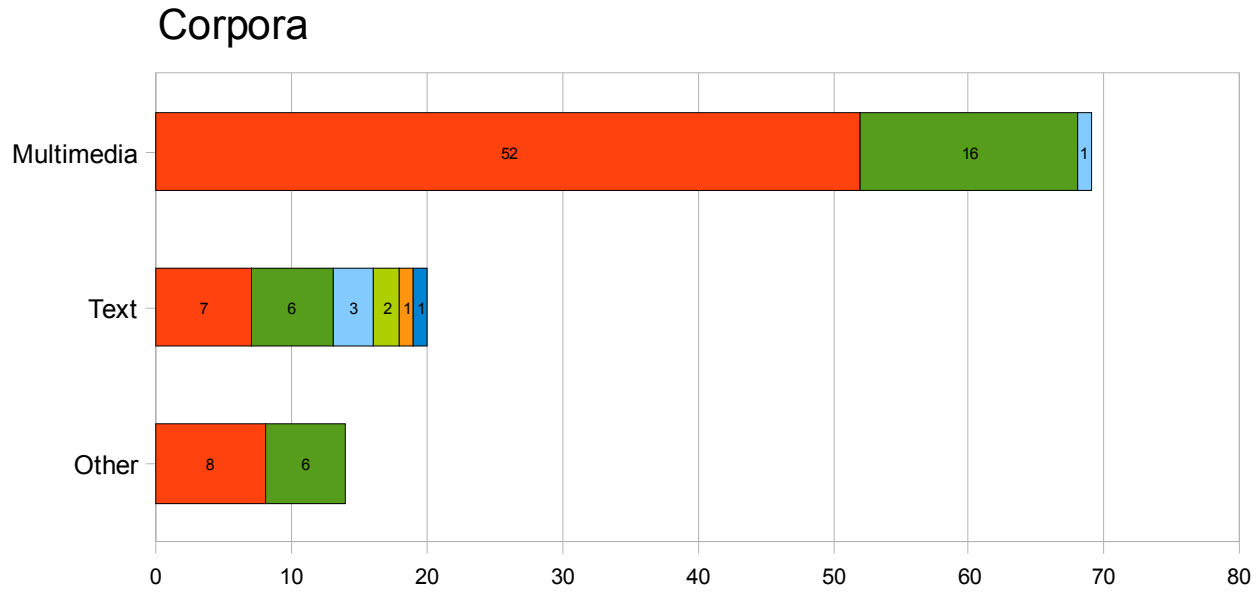


Analysis and normalization of the *Corpora* used by projects and national initiatives fall into the following set of categories. Subcategories under *Multimedia* typically describe the collection itself (as explained above). However, subcategories under *Text* point to an intermediate data precursor needed to achieve the result stated in the *Action*. Parenthetical qualifiers are added to text corpora in order to disambiguate the type of collection they represent (which, as stated above, is implicitly stated in multimedia corpus names).

MULTIMEDIA	69
Audiovisual	52
Image	16
Audio	1

TEXT	20
Annotations (Vague)	7
Controlled Metadata	6
Annotations (Video)	3
Annotations (Profiles)	2
Annotations (Image)	1
Concordances	1
OTHER	14
Unspecified	8
Vague	6

Table 2: Corpora



Analysis

By and large, *Multimedia* corpora represent the majority of corpora among research projects and national initiatives (69 of 103, or 67%). This means that most research efforts are focused on investigating algorithms that work directly on the collection items themselves in order to produce semantically enhanced access to the collection. This is in contrast to *Text* corpora (20 of 103, or 19%), which indicate that a research effort is investigating techniques that work with existing metadata (i.e., annotations) that describe the collection of interest in order to produce new semantically enhanced access possibilities into the collection.

It makes sense that a majority of efforts are focused on the direct enhancement of multimedia corpora since the projects and national initiatives under review are primarily concerned with multimedia content. Furthermore, there is much work to be done in the area of algorithm development and refinement for such things as appropriate temporal segmentation and feature detection.

However, it must be noted that enhanced semantic access into multimedia content can be accomplished in many ways. There is also much work to be done on developing methodologies to work with intermediate references to multimedia content (i.e., mapping multimedia features or keywords from annotations to concepts) as well as improving secondary retrieval factors, such as ontology development, systems design, standards development, human-computer interface design

and appropriate domain compartmentalization. It is important to note that people working on one of these two approaches must be cognizant of the capabilities of the other. In this way, unnecessary development efforts can be avoided.

2.10.3. Methodologies (Technical Requirements)

Summary

Analysis and normalization of problem descriptions from Use Cases yielded a standard set of methodologies that are generally recognized by the Library and Information Science (LIS) community. These methods can be thought of as requirements, since they represent a preferred approach to achieving the goal stated in the Action field. However, it should be noted that they are not closely related to typical Use Case requirements insofar as achieving the actual end-user goal, which should be abstracted away from technical considerations.

Query by Multimedia and *Extraction* are methods applied to multimedia corpora. *Query by Text* requires a text corpora, whereas *Semantic Classification* and *Statistical Classification* can be applied to either corpus type, depending on the specific method.

Controlled Metadata refers to an indexing vocabulary that has been formalized and adheres to some specification, such as a thesaurus or ontology. These languages minimally provide a means of controlling for synonymy, and typically provide hypernymy and meronymy functionality as well. *Controlled Metadata (Profile)* is a controlled indexing language that is applied, in particular, to users in order to provide customized access to a collection based on criteria such as user preferences and histories. *Query by Keyword* refers to an uncontrolled vocabulary.

Semantic Classification refers to techniques for mapping between semantic concepts and collection artifacts, such as transcribed audio, keywords from an annotation, or segmented video.

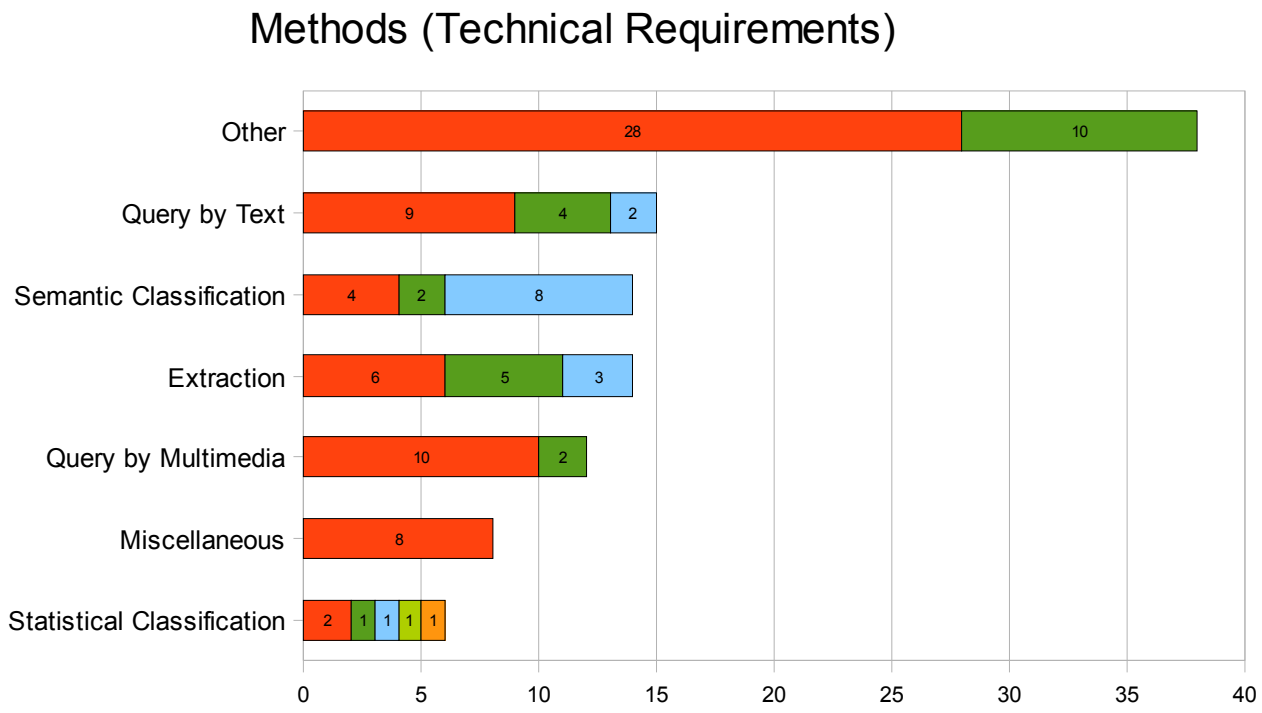
Query by Text and *Query by Multimedia* describe methods for implementing a retrieval system, as declared in the affiliated Action field.

The following table lists, in descending order of occurrence, the categories of methodologies used among projects and national initiatives to achieve goals stated in their corresponding Actions:

OTHER	38
Unspecified	28
Vague	10
QUERY BY TEXT	15
Query by Controlled Metadata	9
Query by Controlled Metadata (Profile)	4
Query by Keyword	2
SEMANTIC CLASSIFICATION	14
Audio-to-Concept	4
Visual-to-Concept	2
Text-to-Concept	0
Semantic Classification (General)	8
EXTRACTION	14
Feature Detection	6

Segmentation	5
Speech Recognition	3
MISCELLANEOUS	1
GUI Development	8
QUERY BY MULTIMEDIA	12
Query by Example	10
Query by Fingerprint	2
STATISTICAL CLASSIFICATION (Clustering)	6
Statistical Classification (General)	2
Relevancy Distance Metric	1
Relevance Feedback	1
Cross Modal Proximity	1
Machine Learning	1

Table 3: Methods (Requirements)



Analysis

Unfortunately, a large number of values for the *Methodology* dimension fall into the *Other* category (38%) due to insufficient information contained in the Use Cases. The next four categories are fairly evenly divided, with *Query by Text* (15%), *Semantic Classification* (14%), *Extraction* (14%), and *Query by Multimedia* (12%) making up a total 55% of methodologies. The miscellaneous category consisted of a single item called *GUI Development*.

2.10.4. Products

Summary

Various product classes have been identified as relevant to the research efforts currently being reviewed and consideration should be given to whether a commercialization effort is appropriate following the completion of a project.

There are five major categories. The first category, *Retrieval Systems*, has three major subcategories. A *Social Sharing System* refers to a networked software tool that facilitates specialized data exchange within a customized environment. Two examples are a meeting browser or a system for sharing avatars among gamers within a gaming community. A *Targeted Delivery System* refers to content syndication.

Classified Content has six subcategories which describe two types of refined metadata: *Indexed Content* (which is further divided into the four collection domains of audiovisual, profiles, images and audio) and *Multimedia Segments*.

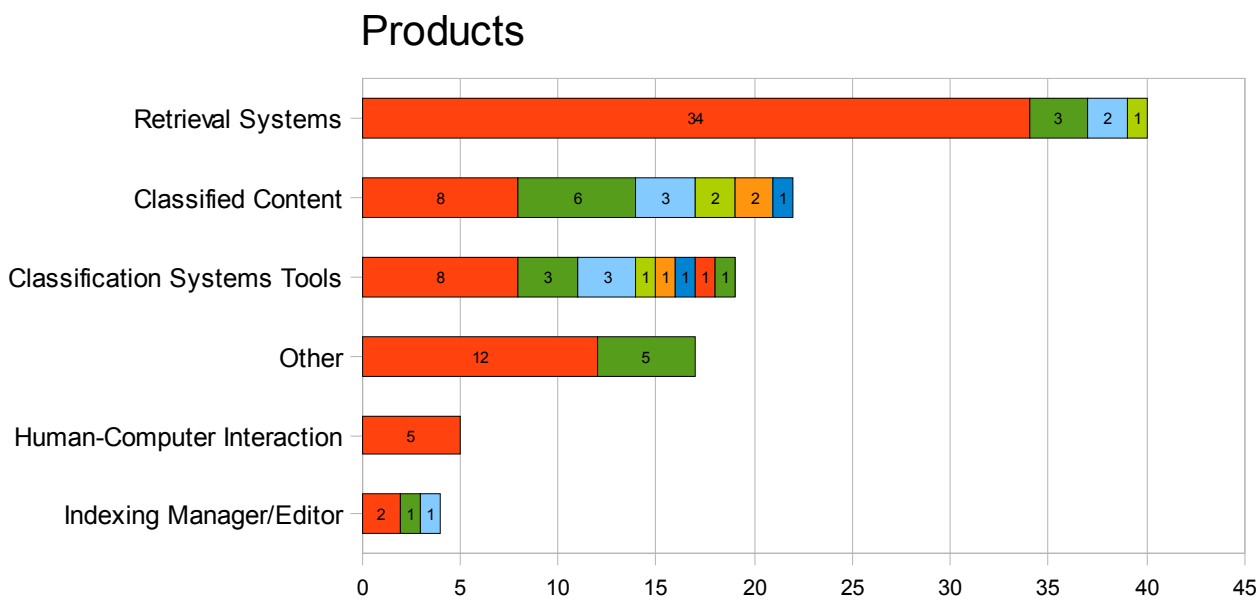
Classification System Tools refers to tools used to classify the content identified in the section above. There are three types of *Generators,* or automation tools. A *Taxonomy Generator* assists in the automatic creation of a controlled vocabulary for indexing artifacts within a given collection or knowledge domain. It usually starts with a set of resources and proceeds to the summarization and extraction of keywords from them. It then prunes and arranges these keywords into hierarchies with synonym references between conceptually similar keywords encoded.

An *Index Generator,* on the other hand, applies a controlled vocabulary (such as one that may be produced by a *Taxonomy Generator*) to a large collection of resources in order to describe them with indexing terms. *Concordance Generators* are simply tools that create lists of the major words within a collection of resources. These lists can take many forms. For example, "Named Entities" is a term used by commercial classification vendors that describes a list of proper nouns for a given domain. In the same way that controlled vocabularies (expressed, in particular, as ontologies) are undergoing vigorous development today, concordances should receive the same attention. A good classification ontology should, in addition to expressing an indexing vocabulary, contain references to various concordances, such as Named Entities. Concordances, used as an indexing aid, can greatly enhance classification efforts.

Human-Computer Interaction has been maintained as a major category even though it only has one subcategory beneath it. This is because we expect more effort to be devoted to this area of research in the future.

Indexing Manager/Editors are in contrast to the *Classification System Tools* described above. They are differentiated by the fact that they primarily rely on manual intervention and personal expertise to assist in the creation of classification tools (namely, controlled vocabularies). *Digital Asset Manager* has perhaps been misplaced under the *Indexing Manager/Editor* category since classification of assets plays only one part in the overall goal of these systems. However, the science domain for the Use Cases reviewed in this paper mainly covers semantic technologies. Therefore, the classification capabilities of a *Digital Asset Manager* was given the most weight when it was decided to place it in the current category.

RETRIEVAL SYSTEMS	40
Retrieval Systems (General)	34
Targeted Delivery System	3
Social Sharing System	2
Recommender System	1
CLASSIFIED CONTENT	22
Indexed Content (Audiovisual)	8
Multimedia Segments	6
Indexed Content (Profiles)	3
Indexed Content (Images)	2
Indexed Content (Audio)	2
Indexed Content (Vague)	1
CLASSIFICATION SYSTEM TOOLS	19
Index Generator (General)	8
Index Generator (Audio)	3
Taxonomy Generator	3
Ontology/Taxonomy Manager/Editor	1
Controlled Vocabulary Development	1
Standards Development	1
Clustering Algorithm	1
Concordance Generator	1
OTHER	17
Unspecified	12
Vague	5
HUMAN-COMPUTER INTERACTION	5
GUI	5
INDEXING MANAGER/EDITOR	4
Indexing Manager/Editor (General)	2
Profile Manager/Editor	1
Digital Asset Manager	1



Analysis

Naturally, the majority of research efforts seem to be focused on areas that can benefit *Retrieval Systems*. However, since retrieval is the general, overall goal of all the science reviewed in the projects and national initiatives for this report (whether it is considered *Classified Content, Classification Systems Tools,* or *Indexing Manager/Editor*), this should be considered a sort of generic placeholder. Indeed, the largest group of products within this category (34 of 40, or 85%) are described as *Retrieval Systems (General)*. Such a large number indicates that the projects and national initiatives may not have provided adequate information within their Use Cases in order to ascertain with more specificity what aspect of retrieval they were focusing on. However, the other three products under this general category are informative, specific and useful.

Classified Content is only useful insofar as the various projects donate their catalogs, indexes and/or collections (i.e., of segmented multimedia) to a beneficiary. There is a lot of classified content being generated by the various projects and national initiatives, and efforts should be undertaken to ensure that this knowledge is not lost.

Classification Systems Tools is an exciting area to be able to contribute to within the European market. In America, there is a vigorous and lucrative market for classification tools, which command a large and growing slice of military and intelligence expenditures. These tools underlie the most advanced intelligence efforts within various sectors that Americans excel at, such as aerospace, intelligence analysis, financial management and analysis, and media management. They are the central nerve center of retrieval and represent the most advanced intelligence efforts in the world.

Technological spin-offs of classification tool research is important for European military and political sovereignty. As a result, there should be a large economic market for regional tools.

Index Generators make up a majority of efforts in this area (11 of 19, or 58%). Unfortunately, when it comes to significant contributions to *Controlled Vocabulary Development* and *Standards Development*, there seems to be very little activity. This is regrettable because these two subcategories are key to making all of the other technologies and tools work well. Without (1) well defined knowledge domains, (2) useful vocabularies, and (3) applicable standards, it is impossible to define good Use Cases or design useful research problems. Monumental effort can be expended in the incremental development of algorithms that can segment video and extract features, but if there is only a vague sense of what concepts they should be mapped to for a given collection (for a particular audience and for a specific purpose), these algorithms will never seem to work well.

Difficult and small advances in algorithm development can potentially be addressed and solved easily within if our *Classification Systems Tools* are appropriately designed and specified.

Human-Computer Interaction makes up a very small percentage (5 of 107, or 4.7%) of the overall effort within projects and national initiatives. Although this area may seem of only peripheral importance to multimedia retrieval research, it should not be overlooked. The overall goal of creating a semantically enabled knowledge network can only be achieved with significant progress in interface design. Apple, Inc. understands this well and has enjoyed growing commercial success and technological achievements since the release of OS X. The retrieval metaphors we choose to work with for dealing with large spaces of complex information should inform our research efforts and play an integral role in the Use Cases we design. What metaphors will we use? How does this effect vocabulary development? Different concepts may emerge as more or less important within a given knowledge domain depending on how we handle them at the user interface level. For example, if we index a large collection of media as "Archived", maybe we could design an interface that automatically filters a given collection so that non-archived content is the only thing we can see and browse by default. This means that the concept of "Archive" may play a different role to the end user than to the content provider and it should inform the design of our experiments. Being able to visualize the end result of a classification technology is key to designing systems that work adequately well for potential commercialization efforts.

There seems to only be a small amount of effort that could result in products within the *Indexing Manager/Editor* category. This is surprising and may be a result of the mis-categorization of project and national initiative efforts due to incomplete or poorly understood Use Case data. In any case, these products need to be tracked. At the very least, they can be re-used and improved by other researchers and commercialization potential should definitely be investigated and followed as the product evolves.

2.10.5. Users

Summary

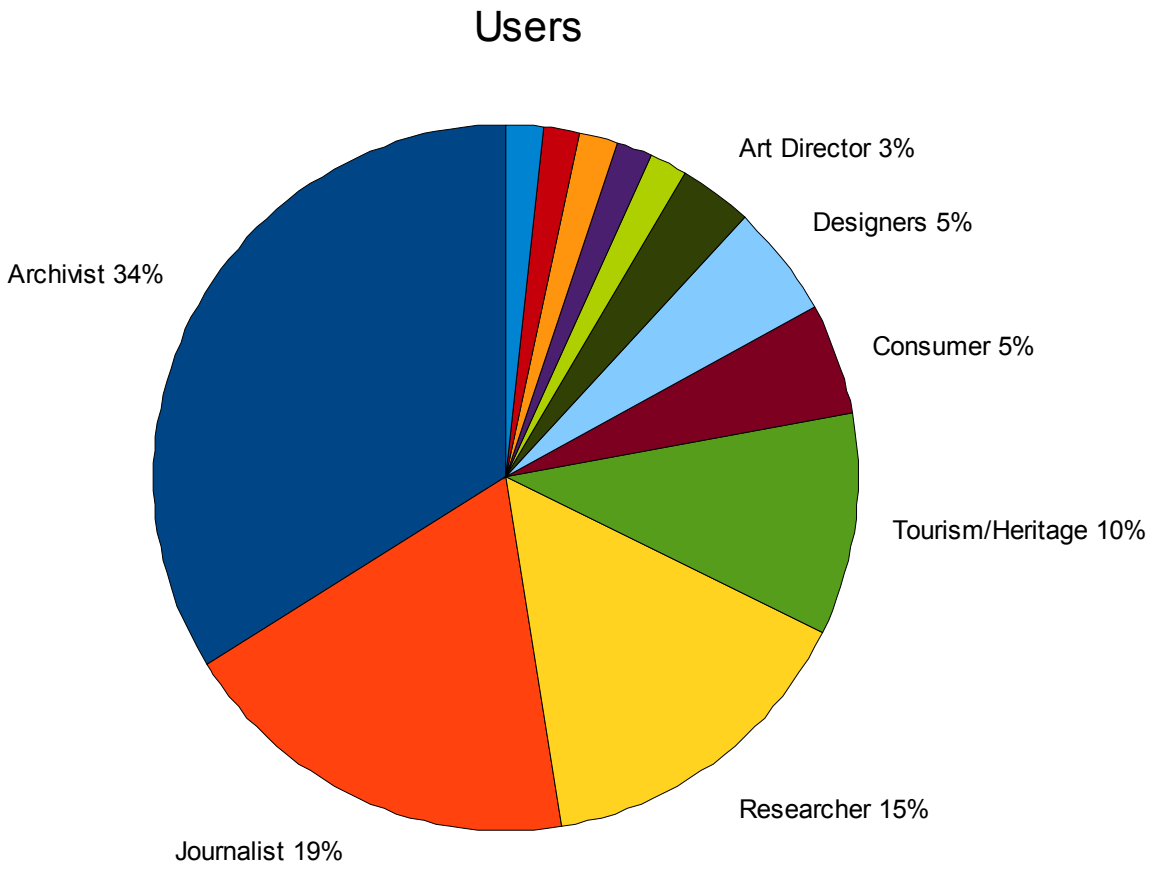
Some Use Cases provided information about the specific audience(s) that the various projects and national initiatives had in mind when they designed their research proposals. The audience is the heart of a Use Case; it is the first variable to be identified when approaching a retrieval problem. Who is retrieving the information? The next question is Why? For what purpose(s)? The answers to these two fundamental questions should form a sort of research mantra that informs every step of the inquiry process.

Without knowing who an audience is and why they are interested in some given content, we neglect to define essential parameters for our research effort. This effectively renders any problem intractable.

Knowing Users conveys another essential piece of information. All User categories can be mapped to an industry sector. This is important because it tells us who might be interested in potential commercialization efforts.

Most User names are self-explanatory. It should be noted that the differentiating factor between *Tourism/Heritage* and *Travel* is that the first refers to an industry and, as such, is a content provider. *Travel* refers to an end user, such as a vacationer.
The list of users below contain all that were mentioned in the submitted Use Cases.

Archivist	20
Journalist	11
Researcher	9
Tourism/Heritage	6
Consumer	3
Designers	3
Art Director	2
Decision Makers	1
Automotive	1
Travel	1
Maintenance/Installation/Support Personnel	1
Open Gaming Communities	1



Analysis

Predictably, the largest identified users are Archivists (20 of 59, or 34%) followed by Journalists (11 of 59, 19%). Together, they make up more than half the users that research efforts were designed to address. Researchers were identified in another 15% of Use Cases. Tourism/Heritage comprises 10% and so seems to be an important parameter for many of the research efforts in CHORUS.

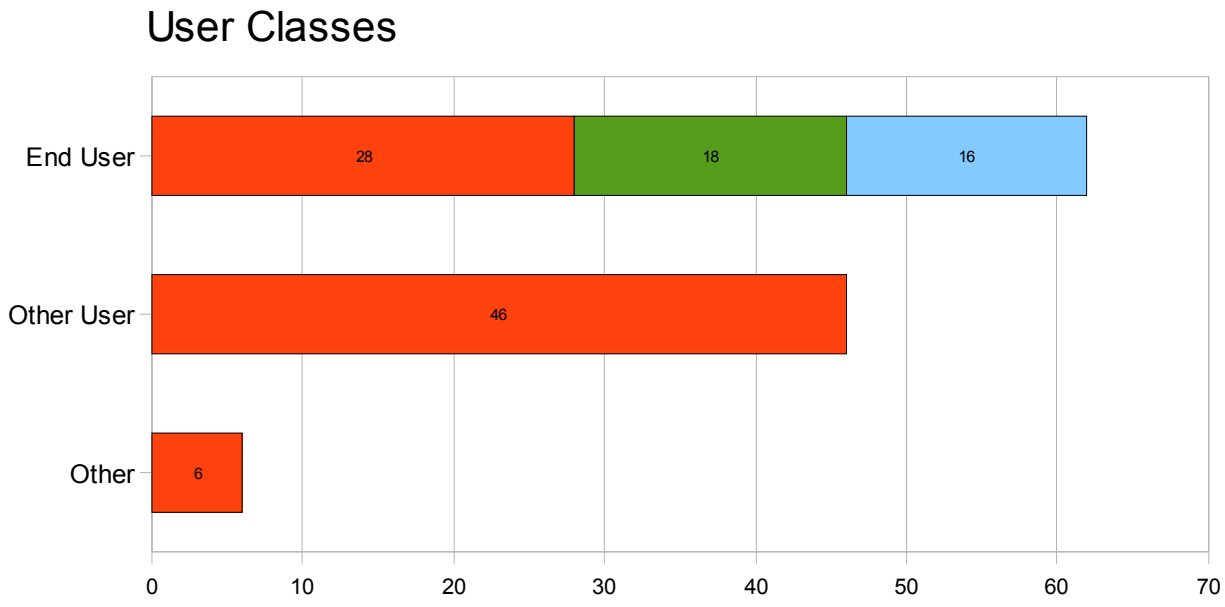
Surprisingly, Consumers are identified in only 5% of Use Cases. This could be an important audience. Such a low number may simply indicate mis-categorization of Use Case data due to insufficient, vague or ambiguous information. Incidence rates for this User among projects and national initiatives should be followed to ensure that consumers are receiving adequate attention.

2.10.6. User Classes

Summary

Users were further classified in broad categories in order to capture a more general trend in how research efforts are being framed. *User Classes*

END USERS	62
End User (Vague)	28
End User (Professional)	18
End User (Simple)	16
OTHER USERS	46
Content Provider	46
OTHER	6
Unspecified	6



Analysis

End Users comprise the majority type of user (*User Class*) in the submitted Use Cases (62 of 114, or 54%), whereas a very important class of users, *Content Providers,* were just under a majority (46 or 114, or 40%). This is satisfactory, but a bit surprising since successful retrieval depends primarily

on the action of *Content Providers* (viz., how well they index and manage their content). Comprising such a relatively small percentage (compared to *End Users*) may be a result of erroneous categorization of ambiguous data in submitted Use Cases.
In fact, 1 out of 4 (28 of 114) identified User Classes were inadequately defined and have been classified as *End User (Vague).*

2.11. Conclusion and Future Prospects

In the preceding comprehensive report of use cases employed by the various projects under the CHORUS as well as in the national initiatives covered at the Geneva workshop on National Initiatives, we find considerable variation in the formulation of the use cases. We also find clear patterns and a common perspectives. Variation is to be expected, since there has been no concertation of use case formulation effort; the similarities are heartening and give purchase for future concertation efforts.

During the course of CHORUS our objective is to provide target dimensions for the formulation of use cases for the commission to consider in future calls and for projects to use for concerted and fruitful benchmarking and evaluation efforts. This process is under way, both within the activities of the think tank meetings organised by CHORUS WP3, and within the working groups organised within CHORUS WP2, of which this text is the first deliverable.

The second deliverable will provide more concrete analyses for future efforts, taking the situation as described here as a starting point. For the second deliverable, CHORUS will further refine the analyses of current efforts, and with the help of those projects with more explicit user-oriented perspectives aim to build a more accurate and comprehensive snapshot of the overall field of research. In addition, the national initiatives, which currently are analysed on a programme level rather than a project level will have more fine-grained data to contribute. A rigorous collection of data dimension values will aid project leaders gain a clearer view of the problems they are attempting to solve as well as see how their research fits in with the efforts of all other CHORUS partners.

The second deliverable will also incorporate information from industrial partners, as provided in the think-tank processes. This will further validate the analyses made by research projects, and allow for the informed formulation of industrially as well as academically valid and useful prototypical challenge use cases, for future projects and funding cycles alike.

2.12. References

CHORUS Deliverable D 4.3, "*Agenda, viewgraphs and minutes of workshop 2 : National Initiatives on Multimedia Content Description and Retrieval", Geneva, October 10th, 2007"*.[Available at http://www.ist-chorus.org/geneva---october-10th-07.php]

Pia Borlund. (2003). The IIR Evaluation Model: a Framework for Evaluation of Interactive Information Retrieval Systems. In: *Information Research,* vol. 8, no. 3, paper no. 152. [Available at: http://informationr.net/ir/8-3/paper152.html]

J. Bowers, G. Button and W. Sharrock. (1995). Workflow from within and without: Technology and cooperative work on the print industry shop floor. Proceedings of ECSCW'95, 51-66. Kluwer.

Alan Cockburn. (2002). *Agile software development.* Addison-Wesley.

I. Jacobson, M. Christson, P. Jonsson and G. Overgaard. (1992). *Object-Oriented Software Engineering: A Use Case Driven Approach*, Addison-Wesley.

Kalervo Järvelin and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In: Belkin, N.J., Ingwersen, P. & Leong, M-K., eds. *Proceedings of the 23rd ACM Sigir Conference on Research and Development of Information Retrieval*, Athens, Greece, 2000. New York, N.Y.: ACM Press, pp. 41-48.

Jussi Karlgren, Julio Gonzalo, and Paul Clough. (2007). iCLEF2006 Overview: Searching the Flickr WWW Photo-Sharing Repository, Evaluation of Multilingual and Multi-modal Information Retrieval . *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, September 20-22, 2006, Revised Selected Papers, Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (Eds.), Vol. 4730, 2007, ISBN 978-3-540-74998-1, Softcover, pp.

Gary Marchionini. (2006). Human performance measures for video retrieval. In *Proceedings of the ACM Workshop on Multimedia Information Retrieval* (MIR2006), special session on Benchmarking Image and Video Retrieval Systems; Santa Barbara, CA, 2006.

Stefano Mizzaro. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810--832. John Wiley and Sons Inc., New York, NY. Republished in ”Historical Studies in Information Science.

Stefano Mizzaro. (1998). How many relevances in information retrieval? *Interacting With Computers*, 10(3):305--322. Elsevier: The Netherlands.

Evaluation frameworks for interactive multimedia information retrieval applications. MIRA theme statement (http://www.dcs.gla.uk/mira)

Jacob Nielsen. (1994). *Usability Engineering*, Morgan Kaufmann Publishers, ISBN 0-12-518406-9.

3. STATE OF THE ART IN AUDIO-VISUAL CONTENT INDEXING AND RETRIEVAL TECHNOLOGIES

3.1. Introduction

Multimedia information retrieval (MIR) is about the search for knowledge in all its forms, everywhere. Indeed, what good is all the knowledge in the world if it is not possible to find anything? This sentiment is mirrored as an ACM SIGMM grand challenge [Rowe and Jain 2005]: "make capturing, storing, finding, and using digital media an everyday occurrence in our computing environment."

Currently, the fundamental problem has been how to enable or improve multimedia retrieval using content-based methods. Content-based methods are necessary when text annotations are nonexistent or incomplete. Furthermore, content-based methods can potentially improve retrieval accuracy even when text annotations are present by giving additional insight into the media collections.

Our search for digital knowledge began several decades ago when the idea of digitizing media was commonplace, but when books were still the primary medium for storing knowledge. Before the field of multimedia information retrieval coalesced into a scientific community, there were many contributory advances from a wide set of established scientific fields. From a theoretical perspective, areas such as artificial intelligence, optimization theory, computational vision, and pattern recognition contributed significantly to the underlying mathematical foundation of MIR. Psychology and related areas such as aesthetics and ergonomics provided basic foundations for the interaction with the user. Furthermore, applications of pictorial search into a database of imagery already existed in niche forms such as face recognition, robotic guidance, and character recognition.

The earliest years of MIR were frequently based on computer vision (three excellent books: [Ballard and Brown 1982]; [Levine 1985]; [Haralick and Shapiro 1993]) algorithms focused on feature based similarity search over images, video, and audio. Influential and popular examples of these systems would be QBIC [Flickner, et al. 1995] and Virage [Bach, et al. 1996] circa mid 90s. Within a few years the basic concept of the similarity search was transferred to several Internet image search engines including Webseek [Smith and Chang 1997] and Webseer [Frankel, et al. 1996]. Significant effort was also placed into direct integration of the feature based similarity search into enterprise level databases such as Informix datablades, IBM DB2 Extenders, or Oracle Cartridges [Bliujute, et al. 1999; Egas, et al. 1999] towards making MIR more accessible to private industry.

In the area of video retrieval, the main focus in the mid 90s was toward robust shot boundary detection of which the most common approaches involved thresholding the distance between color histograms corresponding to two consecutive frames in a video [Flickner, et al. 1995]. Hanjalic, et al. [1997] proposed a method which overcame the problem of subjective user thresholds. Their approach was not dependent on any manual parameters. It gave a set of keyframes based on an objective model for the video information flow. Haas, et al. [1997] described a method to use the motion within the video to determine the shot boundary locations. Their method outperformed the histogram approaches of the period and also performed semantic classification of the video shots into categories such as zoom-in, zoom-out, pan, etc. A more recent practitioner's guide to video transition detection is given by Lienhart [2001].

Also in the area of speech and audio indexing many different algorithms and system are developed to structure and index audio content automatically. One of the first systems in the area of spoken document retrieval is the Thisl Broadcast News Retrieval system [Abberley, et al. 1997]. This systems apply a large vocabulary continuous speech (LVCSR) system to generate word transcription for broadcast data. The automatically transcribed word sequences are attached with

time codes for each word. The transformation from speech to text allows the usage of standard text retrieval mechanism. NIST has carried out several TREC Spoken Document Retrieval evaluations. In TREC-6 to TREC-9 from 1997 – 2000 the indexing task for broadcast news was made more challenging regarding the quality and amount of processed speech data. Other well performing systems for indexing broadcast news (BN) are the systems from LIMSI (Gauvain et al.), the HTK group of the University of Cambridge [Woodland, 1999] and the BBN system. One result of this research work was that the text retrieval performance was not affected by higher error rates for the BN task which varies between 15% and 30%. It has also been shown that the segmentation of complex speech recordings including many speaker changes, music and background noises decrease the system performance. Oard et al. startet 2002 the MALACH project. This system combines speech recognition technology and text retrieval algorithms to index multilingual speech recordings from an oral history archive.

In the area of music indexing and retrieval one of the first systems were developed by Foote et al. Based on low level features of audio processing which were mainly invented and standardized in MPEG-7 Audio several Audio-ID systems were developed by several groups. The Audio-ID technology generates a fingerprint of a segment of music and provides fast matching algorithms to find this fingerprint in a large pre-processed archive. Currently the focus of music retrieval has been changed to genre and mood classification.

Starting near the turn of the 21st century, researchers noticed that the feature based similarity search algorithms were not as intuitive or as user-friendly as they had expected. One could say that systems built by research scientists were essentially systems which could only be used effectively by scientists. The new direction was toward designing systems which would be user friendly and could bring the vast multimedia knowledge from libraries, databases, and collections to the world. To do this it was noted that the next evolution of systems would need to understand the semantics of a query, not simply the low level underlying computational features. This general problem was called “bridging the semantic gap”. From a pattern recognition perspective, this roughly meant translating the easily computable low level content-based media features to high level concepts or terms which would be intuitive to the user. Examples of bridging the semantic gap for the single concept of human faces were demonstrated by Rowley, et al. [1996] and Lew and Huijsmans [1996]. Perhaps the earliest pictorial content-based retrieval system which addressed the semantic gap problem in the query interface, indexing, and results was the ImageScape search engine [Lew 2000]. In this system, the user could make direct queries for multiple visual objects such as sky, trees, water, etc. using spatially positioned icons in a WWW index containing 10+ million images and videos using keyframes. The system used information theory to determine the best features for minimizing uncertainty in the classification.

At this point it is important to note that the feature based similarity search engines were useful in a variety of contexts [Smeulders, et al. 2000] such as searching trademark databases [Eakins, et al. 2003], finding video shots with similar visual content and motion or for DJs searching for music with similar rhythms [Foote 1999], automatic detection of pornographic content [Forsyth and Fleck 1999; Bosson, et al. 2002], and copyright infringement detection [Jaimes 2002, Joly 2003]. Intuitively, the most pertinent applications are those where the basic features such as color and texture in images and video; or dominant rhythm, melody, or frequency spectrum in audio [Foote 1999] are highly correlated to the search goals of the particular application.

3.2. Recent Work

In this section we discuss representative work [Dimitrova 2003; Lew 2006; Sebe, et al. 2003 (CIVR)] done in content-based multimedia retrieval in the recent years. The two fundamental necessities for a multimedia information retrieval system are (1) Searching for a particular media item such as a particular object or concept; and (2) Browsing and summarizing a media collection.

In searching for a particular media item, the current systems have significant limitations such as an inability to understand a wide user vocabulary, understand the user's satisfaction level, nor do there exist credible representative real world test sets for evaluation nor even benchmarking measures which are clearly correlated with user satisfaction. In general current systems have not yet had significant impact on society due to an inability to bridge the semantic gap between computers and humans.

Learning algorithms are interesting because they potentially allow the computer to understand the media collection on a semantic level. Furthermore, learning algorithms may be able to adapt and compensate for the noise and clutter in real world contexts. New features are pertinent in that they can potentially improve the detection and recognition process or be correlated with human perception. New media types address the changing nature of the media in the collections or databases. Some of the recent new media include 3D models (i.e. for virtual reality or games)).

For the most recent research, there currently are several conferences dedicated to the field of MIR such as the ACM SIGMM Workshop on Multimedia Information Retrieval (http://www.liacs.nl/~mir), the ACM International Conference on Image and Video Retrieval (http://www.civr.org), the International Conference on Music Information Retrieval (ISMIR), IEEE International Conference on Acoustics, Speech, and Signal Processing, and the INTERSPEECH conference. For a searchable MIR library, we suggest the community driven digital library at the Association for Multimedia Search and Retrieval (http://www.amsr.org). Additionally, the general multimedia conferences such as ACM Multimedia (http://www.sigmm.org) and the IEEE International Conference on Multimedia and Expo (ICME) typically have MIR related tracks.

3.2.1. Learning and Semantics

The potential for learning in multimedia retrieval is quite compelling toward bridging the semantic gap and the recent research literature has seen significant interest in applying classification and learning [Therrien 1989; Winston 1992; Haralick and Shapiro 1993] algorithms to MIR. The Karhunen-Loeve (KL) transform or principal components method [Therrien 1989] has the property of representational optimality for a linear description of the media. It is important to distinguish between representational optimality versus classification optimality. The ability to optimally represent a class does not necessarily lead to optimally classifying an instance of the class. An example of an improvement on the principal component approach was proposed by Capelli, et al. [2001] where they suggest a multispace KL for classification purposes. The multispace KL directly addresses the problem of when a class is represented by multiple clusters in feature space and can be used in most cases where the normal KL would be appropriate. Zhou and Huang [2001] compared discriminating transforms and SVM for image retrieval. They found that the biased discriminating transform (BDT) outperformed the SVM. Lew and Denteneer [2001] found that the optimal linear keys in the sense of minimizing the distance between two relevant images could be found directly from Fisher's Linear Discriminant. Liu, et al. [2003] find optimal linear subspaces by formulating the retrieval problem as optimization on a Grassman manifold. Balakrishnan, et al. [2005] propose a new representation based on biological vision which uses complementary subspaces. They compare their new representation with principal component analysis, the discrete cosine transform and the independent component transform.

Another approach toward learning semantics is to determine the associations behind features and the semantic descriptions. Djeraba [2002 and 2003] examines the problem of data mining and discovering hidden associations during image indexing and consider a visual dictionary which groups together similar colors and textures. A learning approach is explored by Krishnapuram, et al. [2004] in which they introduce a fuzzy graph matching algorithm. Greenspan, et al. [2004] performs clustering on space-time regions in feature space toward creating a piece-wise GMM framework which allows for the detection of video events.

3.2.1.1. Concept Detection in Complex Backgrounds

One of the most important challenges and perhaps the most difficult problem in semantic understanding of media is visual concept detection in the *presence of complex backgrounds*. Many researchers have looked at classifying whole images, but the granularity is often too coarse to be useful in real world applications. Its typically necessary to find the human in the picture, not simply global features. Another limiting case is where researchers have examined the problem of detecting visual concepts in laboratory conditions where the background is simple and therefore can be easily segmented. Thus, the challenge is to detect all of the semantic content within an image such as faces, trees, animals, etc. with emphasis on the presence of complex backgrounds.

In the mid 90s, there was a great deal of success in the special case of detecting the locations of human faces in grayscale images with complex backgrounds. Lew and Huijsmans [1996] used Shannon's information theory to minimize the uncertainty in the face detection process. Rowley, et al. [1996] applied several neural networks toward detecting faces. Both methods had the limitation of searching for whole faces which prompted later component based model approaches which combined separate detectors for the eyes and nose regions. For the case of near frontal face views in high quality photographs, the early systems generally performed near 95% accuracy with minimal false positives. Non-frontal views and low quality or older images from cultural heritage collections are still considered to be very difficult. An early example of designing a simple detector for city pictures was demonstrated by Vailaya, et al. [1998]. They used a nearest neighbor classifier in conjunction with edge histograms. In more recent work, Schneiderman and Kanade [2004] proposed a system for component based face detection using the statistics of parts. Chua, et al. [2002] used the gradient energy directly from the video representation to detect faces based on the high contrast areas such as the eyes, nose, and mouth. They also compared a rules based classifier with a neural network and found that the neural network gave superior accuracy. For a good overview, Yang, et al. [2002] did a comprehensive survey on the area of face detection.

Detecting a wider set of concepts other than human faces turned out to be fairly difficult. In the context of image search over the Internet, Lew [2000] showed a system for detecting sky, trees, mountains, grass, and faces in images with complex backgrounds. Fan, et al. [2004] used multi-level annotation of natural scenes using dominant image components and semantic concepts. Li and Wang [2003] used a statistical modeling approach toward converting images to keywords. Rautianinen, et al. [2001] used temporal gradients and audio analysis in video to detect semantic concepts.

In certain contexts, there may be several media type available which allows for multimodal analysis. Shen, et al. [2000] discussed a method for giving descriptions of WWW images by using lexical chain analysis of the nearby text on webpages. Benitez and Chang [2002] exploit WordNet to disambiguate descriptive words. They also found 3-15% improvement from combining pictorial search with text analysis. Amir, et al. [2004] proposed a framework for a multi-modal system for video event detection which combined speech recognition and annotated video. Dimitrova, et al. [2000] proposed a Hidden Markov Model based using text and faces for video classification. In the TRECVID [Smeaton and Over 2003] project, the current focus is on multiple domain concept detection for video retrieval.

3.2.1.2. Relevance Feedback

Beyond the one-shot queries in the early similarity based search systems, the next generation of systems attempted to integrate continuous feedback from the user toward learning more about the user query. The interactive process of asking the user a sequential set of questions after each round

of results was called *relevance feedback* due to the similarity with older pure text approaches. Relevance feedback can be considered a special case of *emergent semantics*. Other names have included query refinement, interactive search, and active learning from the computer vision literature.

The fundamental idea behind relevance feedback is to show the user a list of candidate images, ask the user to decide whether each image is relevant or irrelevant, and modify the parameter space, semantic space, feature space, or classification space to reflect the relevant and irrelevant examples. In the simplest relevance feedback method from Rocchio [Rocchio 1971], the idea is to move the query point toward the relevant examples and away from the irrelevant examples. In principle, one general view is to view relevance feedback as a particular type of pattern classification in which the positive and negative examples are found from the relevant and irrelevant labels, respectively.

Therefore, it is possible to apply any learning algorithm into the relevance feedback loop. One of the major problems in relevance feedback is how to address the small training set. A typical user may only want to label 50 images when the algorithm really needs 5000 examples instead. If we compare the simple Rocchio algorithm to more sophisticated learning algorithms such as neural networks, its clear that one reason the Rocchio algorithm is popular is that it requires very few examples. However, one challenging limitation of the Rocchio algorithm is that there is a single query point which would refer to a single cluster of results. In the discussion below we briefly describe some of the recent innovations in relevance feedback.

Chang, et al. [1998] proposed a framework which allows for interactive construction of a set of queries which detect visual concepts such as *sunsets.* Sclaroff, et al. [2001] describe the first WWW image search engine which focussed on relevance feedback based improvement of the results. In their initial system, where they used relevance feedback to guide the feature selection process, it was found that the positive examples were more important towards maximizing accuracy than the negative examples. Rui and Huang [2001] compare heuristic to optimization based parameter updating and find that the optimization based method achieves higher accuracy.

Chen, et al. [2001] described a one-class SVM method for updating the feedback space which shows substantially improved results over previous work. He, et al. [2002] use both short term and long term perspectives to infer a semantic space from user's relevance feedback for image retrieval. The short term perspective was found by marking the top 3 incorrect examples from the results as irrelevant and selecting at most 3 images as relevant examples from the current iteration. The long term perspective was found by updating the semantic space from the results of the short term perspective. Yin, et al. [2005] found that combining multiple relevance feedback strategies gives superior results as opposed to any single strategy. Tieu and Viola [2004] proposed a method for applying the AdaBoost learning algorithm and noted that it is quite suitable for relevance feedback due to the fact that AdaBoost works well with small training sets. Howe [2003] compares different strategies using AdaBoost. Dy, et al. [2003] use a two level approach via customized queries and introduce a new unsupervised learning method called feature subset selection using expectation-maximization clustering. Their method doubled the accuracy for the case of a set of lung images. Guo, et al. [2001] performed a comparison between AdaBoost and SVM and found that SVM gives superior retrieval results. Haas, et al. [2004] described a general paradigm which integrates external knowledge sources with a relevance feedback mechanism and demonstrated on real test sets that the external knowledge substantially improves the relevance of the results. Ferecatu [Ferecatu2005] proposed a hybrid visual and conceptual image representation within active relevance feedback context. A good overview can also be found from Muller, et al. [2000].

3.2.2. New Features & Similarity Measures

Research did not only proceed along the lines of improved search algorithms, but also toward creating new features and similarity measures based on color, texture, and shape. One of the recent

interesting additions to the set of features are from the MPEG-7 standard [Pereira and Koenen 2001]. The new color features [Lew 2001, Gevers2001] such as the NF, rgb, and m color spaces have specific benefits in areas such as lighting invariance, intuitiveness, and perceptual uniformity. A quantitative comparison of influential color models is performed in Sebe and Lew [2001].

In texture understanding, Ojala, et al. [1996] found that combining relatively simple texture histograms outperformed traditional texture models such as Gaussian or Markov features. Jafari-Khouzani and Soltanian-Zadeh [2005] proposed a new texture feature based on the Radon transform orientation which has the significant advantage of being rotationally invariant. Insight into the MPEG-7 texture descriptors has been given by Wu, et al. [2001].

Veltkamp and Hagedoorn [2001] describe the state of the art in shape matching from the perspective of computational geometry. Sebe and Lew [2002] evaluate a wide set of shape measures in the context of image retrieval. Srivastava, et al. [2005] describes some novel approaches to learning shape. Sebastian, et al. [2004] introduce the notion of shape recognition using shock graphs. Bartolini, et al. [2005] suggest using the Fourier phase and time warping distance.

Foote [2000] introduces a feature for audio based on local self-similarity. The important benefit of the feature is that it can be computed for any audio signal and works well on a wide variety of audio segmentation and retrieval applications. Bakker and Lew [2002] suggest several new audio features called the frequency spectrum differentials and the differential swap rate. They evaluate the new audio features in the context of automatic labeling the sample as either speech, music, piano, organ, guitar, automobile, explosion, or silence and achieve promising results.

Fauqueur et al. [Fauqueur2004] devise a new histogram based color descriptor that uses distributions of quantised colors, previously employed in global image feature techniques, in the local feature extraction case. Considering that description must be finer for regions than for images they propose region descriptor of fine color variability: the Adaptive Distribution of Color Shades (ADCS). They combine ADCS with an appropriate similarity measure to enable its use in indexing.

Equally important to novel features is the method to determine similarity between them. Jolion [2001] gives an excellent overview of the common similarity measures. Sebe, et al. [2000] discuss how to derive an optimal similarity measure given a training set. In particular they find that the sum of squared distance tends to be the worst similarity measure and that the Cauchy metric outperforms the commonly used distance measures. Jacobs, et al. [2000] investigates non-metric distances and evaluates their performance. Beretti, et al. [2001] proposes an algorithm which relies on graph matching for a similarity measure. Cooper, et al. [2005] suggest measuring image similarity using time and pictorial content.

In the last decades, a lot of research has been done on the matching of images and their structures [Schmid, et al. 2000, Mikolajczyk and Schmid 2004]. Although the approaches are very different, most methods use some kind of point selection from which descriptors are derived. Most of these approaches address the detection of points and regions that can be detected in an affine invariant way.

Lindeberg [1998] proposed an “interesting scale level” detector which is based on determining maxima over scale of a normalized blob measure. The Laplacian-of-Gaussian (LoG) function is used for building the scale space. Mikolajczyk and Schmid [2004] showed that this function is very suitable for automatic scale selection of structures. An efficient algorithm to be used in object recognition was proposed by Lowe [2004]. This algorithm constructs a scale space pyramid using difference-of-Gaussian (doG) filters. The doG can be used to obtain an efficient approximation of the LoG. From the local 3D maxima a robust descriptor is build for matching purposes. The disadvantage of using doG or LoG as feature detectors is that the repeatability is not optimal since they not only respond to blobs, but also to high gradients in one direction. Because of this, the localization of the features may not be very accurate.

An approach that intuitively arises from this observation is the separation of the feature detector and the scale selection. The commonly used Harris detector [Harris and Stephens 1988] is robust to noise and lighting variations, but only to a very limited extent to scale changes [Schmid, et al. 2000]. To deal with this Dufournoud, et al. [2000] proposed the scale adapted Harris operator. Given the scale adapted Harris operator, a scale space can be created. Local 3D maxima in this scale space can be taken as salient points but this scale adapted Harris operator rarely attains a maximum over scales. This results in very few points, which are not representative enough for the image. To address this problem, Mikolajczyk and Schmid [2004] proposed the Harris-Laplace detector that merges the scale-adapted Harris corner detector and the Laplacian based scale selection.

During the last years much of the research on scale invariance has been generalized to affine invariance. Affine invariance is defined here as invariance to non-uniform scaling in different directions. This allows for matching of descriptors under perspective transformations since a global perspective transformation can be locally approximated by an affine transformation [Tuytelaars and van Gool 2000]. The use of the second moment matrix (or autocorrelation matrix) of a point for affine normalization was explored by Lindeberg and Garding [1997]. A similar approach was used by Baumberg [2000] for feature matching.

All the above methods were designed to be used in the context of object-class recognition application. However, it was found that wavelet-based salient points [Tian, et al. 2001] outperform traditional interest operators such as corner detectors when they are applied to general content-based image retrieval. For a good overview, we refer the reader to Sebe, et al. [IVC 2003].

Some recent works focus on detecting more perceptible local structure. Szumilas et al. [Szumilas2007] extract feature centre locations at places where a symmetry measure is maximized. Next, boundary points along rays emanating from the centre are extracted. Boundary points are defined as edges or transitions between relatively different regions, and are extracted by hierarchical clustering of pixel feature values along the ray. Rebai et al. [Rebai2007] focus their interpretable interest points on radial symmetry centers detected by a Hough like strategy generalized to several tangential angles.

To eliminate the Out-Of-Vocabulary (OOV) problem in the area of spoken document retrieval subword units for indexing are introduced [Larson2007]. Based on phone or syllable transcriptions generated by an automatic speech recognition system fuzzy matching algorithms, like the Levenshtein based fuzzy search, arbitrary textual search query can be formulated. Here new indexing paradigm are required to provide a short reaction time during retrieval.

3.2.3. 3D Retrieval

In the early years of MIR, most research focussed on content-based image retrieval. Recently, there has been a surge of interest in a wide variety of media. An excellent example, “life records”, which encompasses simultaneously all types of media is being actively promoted by Bell [2004]. He is investigating the issues and challenges in processing life records - all the text, audio, video, and media related to a person's life.

Beyond text, audio, images, and video, there has been significant recent interest in new media such as 3D models. Assfalg, et al. [2004] discuss using *spin-images*, which essentially encode the density of mesh vertices projected onto a 2D space, resulting in a 2D histogram. It was found that they give an effective view-independent representation for searching through a database of cultural artifacts. Funkhouser, et al. [2003] develop a search engine for 3D models based on shape matching using spherical harmonics to compute discriminating similarity measures which are effective even in the presence of model degeneracies. An overview of how 3D models are used in content-based retrieval systems can be found in Tangelder and Veltkamp [2004].

3.2.4. Browsing and Summarization

There have been a wide variety of innovative ways of browsing and summarizing multimedia information. Spierenburg and Huijsmans [1997] proposed a method for converting an image database into a movie. The intuition was that one could cluster a sufficiently large image database so that visually similar images would be in the same cluster. After the cluster process, one can order the clusters by the inter-cluster similarity, arrange the images in sequential order and then convert to a video. This allows a user to have a gestalt understanding of a large image database in minutes.

Sundaram, et al. [2002] took a similar approach toward summarizing video. They introduced the idea of a video skim which is a shortened video composed of informative scenes from the original video. The fundamental idea is for the user to be able to receive an abstract of the story but in video format.

Snoek, et al. [2005] propose several methods for summarizing video such as grouping by categories and browsing by category and in time. Chiu, et al. [2005] created a system for texturing a 3D city with relevant frames from video shots. The user would then be able to fly through the 3D city and browse all of the videos in a directory. The most important frames would be located on the roofs of the buildings in the city so that a high altitude fly through would result in viewing a single frame per video.

Uchihashi, et al. [1999] suggested a method for converting a movie into a cartoon strip in the Manga style from Japan. This means altering the size and position of the relevant keyframes from the video based on their importance. Tian, et al. [2002] took the concept of variable size and positions of images to the next level by posing the problem as a general optimization criterion problem. What is the optimal arrangement of images on the screen so that the user can optimally browse an image database.

Liu, et al. [2004] address the problem of effective summarization of images from WWW image search engines. They compare a rank list summarization method to an image clustering scheme and find that their users find the clustering scheme allows them to explore the image results more naturally and effectively.

3.2.5. High Performance Indexing

In the early multimedia database systems, the multimedia items such as images or video were frequently simply files in a directory or entries in an SQL database table. From a computational efficiency perspective, both options exhibited poor performance because most filesystems use linear search within directories and most databases could only perform efficient operations on fixed size elements. Thus, as the size of the multimedia databases or collections grew from hundreds to thousands to millions of variable sized items, the computers could not respond in an acceptable time period.

Even as the typical SQL database systems began to implement higher performance table searches, the search keys had to be exact such as in text search. Audio, images, and video were stored as blobs which could not be indexed effectively. Therefore, researchers [Egas, et al. 1999; Lew 2000] turned to similarity based databases which used tree-based indexes to achieve logarithmic performance. Even in the case of multimedia oriented databases such as the Informix database, it was still necessary to create custom datablades to handle efficient similarity searching such as k-d trees [Egas, et al. 1999]. In general the k-d tree methods had linear worst case performance and logarithmic average case performance in the context of feature based similarity searches. A recent improvement to the k-d tree method is to integrate entropy based balancing [Scott and Shyu 2003].

Other data representations have also been suggested besides k-d trees. Ye and Xu [2003] show that vector quantization can be used effectively for searching large databases. Elkwae and Kabuka [2000] propose a 2-tier signature based method for indexing large image databases. Type 1 signatures represent the properties of the objects found in the images. Type 2 signatures capture the inter-object spatial positioning. Together these signatures allow them to achieve a 98% performance improvement. Shao, et al. [2003] use invariant features together with efficient indexing to achieve near real-time performance in the context of k nearest neighbor searching.

Other kinds of high performance indexing problems appear when searching peer to peer (P2P) networks due to the curse of dimensionality, the high communication overhead and that all searches within the network are based on nearest neighbor methods. Muller and Henrich [2003] suggest an effective P2P search algorithm based on compact peer data summaries. They show that their model allows peers to only communicate with a small sample and still retain high quality of results.

3.3. Summary of Multimedia Analysis in European Research

The goal of this section is to summarize the multimedia analysis research that takes place within several European projects and national initiatives. We explicitly mention the research partners and their contribution to different type of media analysis: (1) speech, music, and audio analysis; (2) image analysis; (3) 3D analysis in images and video; (3) video analysis; and (4) text and semantics. Please note that most of these research efforts do not restrict to a single media but they are rather addressing the multimedia problem and advocate the use of cross-media inference and analysis. We are also summarizing in the end of the section the main issues regarding the state of the art in analysis of different media focussing on the following issues: (1) objectives; (2) Approaches and technologies; (3) Systems; (4) Applications; and 95) challenges.

3.3.1. Multimedia Analysis in European Projects

The research topic audio-visual indexing and retrieval is in the main focus of the 9 funded IST projects of the strategic objective “Audio Visual Search Technologies”. Table 1 shows which partners in the nine projects work on indexing and retrieval technologies for the different media types. This information was collected from the different projects and was augmented by us in the cases when the information was not available or was incomplete.

The table shows that all types of media are well covered by the funded EU projects. In the IP projects (Vitalas) all media types are presented. 3D indexing and retrieval is the main focus of the Victory project while Rushes addresses also this subject. In these projects special 3D search engine technology will be developed. It is also obvious that research on video processing is a very active research area. Motivated by work in the context of TrecVid many research groups continue their research work to improve video retrieval performance and a good example here is Vidi-Video.

	Speech/Audio	Image	3D	Video	Text/Semantics
DIVAS	FhG IDMT Sail Labs			Elecard	
PHAROS	Univ. P. Fabra FhG IDMT Sail Labs	EPFL		EPFL Open Univ., UK	Web Models L3S Research
RUSHES	Brunel Univ.	Brunel Univ.	FhG HHI	Queen Mary Univ. Brunel Univ.	Queen Mary Univ. Brunel Univ.
SAPIR	IBM Univ. of Padova	CNR		Eurix	Xerox
SEMEDIA				Joaneum Research Fundacio Barcelona Univ. P. Fabra UPC Barcelona Digital Video Systems Univ. of Glasgow	
TRIPOD		Dublin City Univ.			Sheffield Univ.
VICTORY			Certh/ITI		
VIDI-VIDEO	INESC Lisboa	U. Surrey UvA ITI U. Florence		UvA ITI U. Florence	
VITALAS	FhG IAIS	INRIA Robotiker		INRIA CWI Certh/ITI	Univ. of Sunderland EADS

Table 1: Overview about the AV indexing activities in the 9 IST projects with information about the active partners

3.3.2. Multimedia Analysis in National Initiative

Many national projects do research in the area of audio-visual indexing and retrieval. Although the overall focus of the national projects differs the underlying technologies are quite similar. Table 2 presents a summary of the research activities in the national projects for the different types of media.

In all national projects a strong participation of industrial partners can be observed. The research activities are application driven with a clear market focus. In the German Theseus project tools for semantic knowledge engineering and future Web applications will be developed. The main objective of the French Quaero project is to provide applications for the multimedia business sector. MultimediaN shows already concrete results and demo applications for advanced multimedia search applications. IM2 is carrying out research in the area of meeting annotation which requires innovation in the area of multimedia indexing and communication modelling.

	Speech/Audio	Image	3D	Video	Text/Semantics
Quaero (French)	Limisi RWTH Aachen Univ. Karlsruhe VecSys IRCAM	INRIA Univ. J. Fourier Jouve		INRIA LTU Univ. J. Fourier	Jouve Limsi INRIA
Theseus (German)	FhG IAIS M2Any	FhG HHI FhG First Siemens CT	FhG HHI FhG IGD	FhG HHI Siemens	Univ. Karlsruhe FhG IAIS DFKI FZI
iAD (Norway)				Dublin Univ.	Fast
MultimediaN (Dutch)	U. Twente TU Delft	CWI U. Amsterdam		U. Amsterdam CWI TU Delft Philips	U. Twente
IM2 (Swiss)	IDIAP	EPFL IDIAP		U. Fribourg IDIAP	
Mundos (Spanish)				CineVideo20	

Table 2: Overview about the AV indexing activities in the national research projects with information about the active partners

3.3.3. State-of-the Art in European Research

State-of-the-Art: Speech Analysis

- Objectives
  - Automatic indexing of huge audio archives using speech technology
- Approaches/Technologies
  - Speech recognition: HMM based LVCSR systems, Spoken Document Retrieval, Subword indexing (SAPIR, VITALAS, PHAROS, Quaero, Theseus, MultimediaN, IM2)
  - Speech Segmentation: speaker clustering and recognition (DIVAS, VITALAS, Quaero, Theseus, MultimediaN, IM2)
  - Speech-to-video transcoding (DIVAS)

- Systems
  - IST AV-projects: IBM speech system (SAPIR), Audiomining System from Fraunhofer IAIS (VITALAS), Sail Labs Technolgoy (DIVAS), AudioSurf from Limsi & Vecsys (Quaero)
  - Others: BBN, HTK-Group Cambridge, LIMSI, RWTH Aachen, Nuance, etc.
- Applications
  - Indexing of broadcast news/archives (VITALAS, DIVAS, VIDIVIDEO, Quaero, Theseus, MultimediaN)
  - Podcast/Videocast search (Potzinger, Blinkx)
  - Audio archives (Parliament data, historical archives)
- Challenges
  - Variability of content (e.g. background noise)
  - domain dependency
  - scalability of subwords approaches
  - language dependency

State-of-the-Art: Music Analysis

- Objectives
  - Automatic indexing and classification of large music collections
- Approaches/Technologies
  - Music segmentation: Spectral Flatness (MPEG-7 Audio), Genetic Algorithms, Viterbi (DIVAS, PHAROS, Quaero, Theseus)
  - Music retrieval and Recommendation (SOMs) (SAPIR, Theseus)
- Systems
  - IST projects: Fraunhofer IDMT (DIVAS, PHAROS), M2Any (Theseus), IRCAM (Quaero)
  - Others: Barcelona Music & Audio Technologies, FhG AudioID, PlaySom (Univ. Vienna), SyncPlayer (Univ. Bonn), etc.
- Applications
  - Indexing of music collections
  - Query by humming
  - Audio-music identification
  - Recommendation engines

- Challenges
  - Genre Classification
  - Polyphonic instrument recognition
  - Affective analysis

State-of-the-Art: Image Analysis

- Objectives
  - Indexing and retrieval of images, object recognition
- Approaches/Technologies
  - Low level image processing (histograms, shapes, textures, MPEG7-visual, SIFT) (SAPIR, VIDIVIDEO, VITALAS, SMEDIA, TRIPOD, Rushes, Quaero, Theseus, MultimediaN, IM2)
  - Image similarity measurements (Rushes, VIDIVIDEO, VITALAS, Theseus, IM2)
  - Relevance Feedback (Rushes, SMEDIA, VITALAS), etc.
- Systems
  - Ist projects: INRIA (VITALAS), Univ. of Amsterdam & Univ. of Florence (VIDIVIDEO), etc.
  - Others: IBM (QBIC), Webseek, MPEG-7 search system (Univ. Munich), IKONA (INRIA), Riya, Nevenvision, etc.
- Applications
  - Content based retrieval in image collections
  - Object recognition
  - Face recognition (security, photo collections)
  - Automatic annotation of image collections with keywords and textual descriptions
- Challenges
  - Semantic gap
  - Image segmentation
  - Sensory gap

State-of-the-Art: Video Analysis

- Objectives
  - Automatic segmentation of videos, video retrieval, object recognition in videos

- Approaches/Technologies
  - Shot detection, keyframe generation (DIVAS, Rushes, SAPIR, VIDIVIDEO, VITALAS, Quaero, Theseus, MultimediaN, IM2)
  - Object tracking based on motion based features, closed captions recognition, etc. (Rushes, VIDIVIDEO, VITALAS, Quaero, Theseus, MultimediaN, IM2)
  - Object detection and recognition (ANN, Adaboost, SIFT) (VIDIVIDEO, VITALAS, SMEDIA, VITALAS, Quaero, Theseus, MultimediaN, IM2)
  - Video annotation and summarization (Rushes, SMEDIA, VITALAS, Quaero, Theseus, MultimediaN, IM2)
  - Metadata workflow management (SMEDIA, PHAROS Quaero, Theseus, MultimediaN)
  - Video event detection (SMEDIA, VITALAS, VIDIVIDEO)
- Systems
  - IST projects: Univ. Amsterdam & Univ. Florence (VIDIVIDEO), Joaneum Research (SMEDIA), CERTH/ITI (VICTORY), VITALAS (INA/INRIA, CERTH-ITI), Fraunhofer IAIS (Theseus),
  - Others: Virage, TrecVideo-particpants, Informedia, Univ. of Marburg, etc.
- Applications
  - Indexing of broadcast material, media observation
  - Indexing of videocast material,
  - Recommendation Engines
  - Video fingerprinting, logo detection, security, etc.
  - 3D video (Rushes, VICTORY, Theseus)
- Challenges
  - Detection of complex concepts
  - Segmentation into more semantic based units (i.e. complex scenes)
  - Thousands of different objects
  - Multimodality, fusion

State-of-the-Art: Text/Semantic Analysis

- Objectives
  - Automatic indexing and classification of text based documents

- Approaches/Technologies
  - SVM, PLSI, Named Entity Recognition (Rushes, SAPIR, VITALAS, Quaero, Theseus, iAD, MultimediaN)
  - Bayesian semantic reasoning (Rushes, Theseus)
  - Caption augmentation (TRIPOD)
- Systems
  - IST projects: EADS/Univ. of Sunderland text classification (VITALAS), Yahoo (SMEDIA), Univ. of Karlsruhe (Theseus), Empolis (Theseus)
  - Others (many): Recommind, ITxY, Xtramind (DFKI), Autonomy, Gate, etc.
- Applications
  - Classification of news and documents in companies
  - Email filtering
  - Text based search engines
  - Semantic analysis of multimedia (automatic) annotations
- Challenges
  - Semantics, Ontologies

3.4. Future Directions

Despite the considerable progress of academic research in multimedia information retrieval, there has been relatively little impact of audio-visual content indexing and retrieval research into commercial applications with some niche exceptions such as video segmentation. One example of an attempt to merge academic and commercial interests is Riya (www.riya.com). Their goal is to have a commercial product that uses the academic research in face detection and recognition and allows the users to search through their own photo collection or through the Internet for particular persons. Another example is the MagicVideo Browser (www.magicbot.com) which transfers research in video summarization to household desktop computers and has a plug-in architecture intended for easily adding new promising summarization methods as they appear in the research community. An interesting long-term initiative is the launching of Yahoo! Research Berkeley (research.yahoo.com/Berkeley), a research partnership between Yahoo! Inc. and UC Berkeley with the declared scope to explore and invent social media and mobile media technology and applications that will enable people to create, describe, find, share, and remix media on the web. Nevenvision (www.nevenvision.com) is developing technology for mobile phones that utilizes visual recognition algorithms for bringing in ambient finding technology. However, these efforts are just in their infancy and there is a need for avoiding a future where the multimedia information retrieval (MIR) community is isolated from real world interests. We believe that the MIR community has a golden opportunity to the growth of the multimedia search field that is commonly considered the next major frontier of search [Battelle 2005].

To assess research effectively in multimedia retrieval, task-related standardized databases on which different groups can apply their algorithms are needed. In text retrieval, it has been relatively straightforward to obtain large collections of old newspaper texts because the copyright owners do

not see the raw text being of much value, however image, video, and speech libraries do see great value in their collections and consequently are much more cautious in releasing their content. While it is not a research challenge, obtaining large multimedia collections for widespread evaluation benchmarking is a practical and important step that needs to be addressed. One possible solution is that task-related image and video databases with appropriate relevance judgments are included and made available to groups for research purposes as is it done with TRECVID. Useful video collections could include news video (in multiple languages), collections of personal videos, and possibly movie collections. Image collections would include image databases (maybe on specific topics) along with annotated text - the use of library image collections should also be explored. One critical point here is that sometimes the artificial collections like Corel might do more harm than good to the field by misleading people into believing that their techniques work, while they do not necessarily work with more general image collections.

Therefore, cooperation between private industry and academia is strongly encouraged and is currently taking place within the European projects and national initiatives mentioned before. The key point here is to focus on efforts which mutually benefit both industry and academia. As was noted earlier, it is of clear importance to keep in mind the needs of the users in retrieval system design and it is logical that industry can contribute substantially to our understanding of the end-user and also aid in realistic evaluation of research algorithms. Furthermore, by having closer communication with private industry we can potentially find out what parts of their systems need additional improvements toward increasing user satisfaction. In the example of Riya, they clearly need to perform object detection (faces) in complex backgrounds and then object recognition (who the face is). For the context of consumer digital photograph collections, the MIR community might attempt to create a solid test set which could be used to assess the efficacy of different algorithms in both detection and recognition in real world media.

To summarize the major research challenges listed in the previous section of particular importance to the audio-visual content indexing and retrieval research community are the following challenges: (1) Semantic search with emphasis on the detection of concepts in media with complex backgrounds; (2) Multi-modal analysis and retrieval algorithms especially towards exploiting the synergy between the various media including text and context information; (3) Experiential multimedia exploration systems toward allowing users to gain insight and explore media collections; (4) Interactive search, emergent semantics, or relevance feedback systems; and (5) Evaluation with emphasis on representative test sets and usage patterns.

3.5. References

D. Abberley, D. Kirby, S. Renals and T. Robinson, "The THISL broadcast news retrieval system ", Proc. of ESCA ETRW Workshop on Accessing Information in Spoken Audio, Cambridge (UK), April 1999

Amir, A., Basu, S., Iyengar, G., Lin, C.-Y., Naphade, M., Smith, J.R., Srinivasan S., and Tseng, B. 2004. A Multi-modal System for the Retrieval of Semantic Video Events. *CVIU 96(2)*, 216-236.

Assfalg, J., Del Bimbo, A., and Pala, P. 2004. Retrieval of 3D Objects by Visual Similarity. *ACM MIR,* 77-83.

Bach, J.R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R., AND Shu, C.F. 1996. Virage image search engine: An open framework for image management. In *SPIE: Storage and Retrieval for Still Image and Video Databases*, 76-87.

Balakrishnan, N., Hariharakrishnan, K., AND Schonfeld, D. 2005. A New Image Representation Algorithm Inspired by Image Submodality Models, Redundancy Reduction, and Learning in Biological Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27(9)*, 1367-1378.

Ballard, D.H. and Brown, C.M. 1982. *Computer Vision*. Prentice Hall, New Jersey, USA.

Bakker, E.M. AND Lew, M.S. 2002. Semantic Video Retrieval Using Audio Analysis. *In CIVR*, 262-270.

Bartolini, I., Ciaccia, P., AND Patella, M. 2005. WARP: Accurate Retrieval of Shapes Using Phase of Fourier Descriptors and Time Warping Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(1), 142-147.BATTELLE, J. 2005. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Portfolio Hardcover, USA.

Baumberg, A. 2000, Reliable feature matching across widely separated views, *CVPR*, 774–781.

BELL, G. 2004. A New Relevance for Multimedia When We Record Everything Personal. In *ACM Multimedia*.

Bentiez, A. B. AND Chang, S.-F. 2002. Semantic knowledge construction from annotated image collection. In *ICME*.

Beretti, S., Del Bimbo, A., AND Vicario, E. 2001. Efficient Matching and Indexing of Graph Models in Content-Based Retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence 23(10)*, 1089-1105.

Bliujute, R., Saltenis, S., Slivinskas, G., AND Jensen, C.S. 1999. Developing a DataBlade for a New Index. In *Proceedings of IEEE International Conference on Data Engineering*, 314-323.

Bosson, A., Cawley, G.C., Chan, Y., AND Harvey, R. 2002. Non-retrieval: Blocking Pornographic Images. *In CIVR*, 50-60.

Byrne W. et al. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, 12(4):420-435, July 2004

Cappelli, r., Maio, D., AND Maltoni. D. 2001. Multispace KL for Pattern Representation and Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23(9)*, 977-996.

Chang, S.-F., Chen, W., and Sundaram, H. 1998. Semantic visual templates: Linking visual features to semantics. In *ICIP*, 531–535.

Chen, Y., Zhou, X.S., AND Huang, T.S. 2001. One-class SVM for Learning in Image Retrieval, In *ICIP*, 815-818.

Chi, P., Girgensoh, A., Lertsithichai, S., Polak, W., AND Shipman, F. 2005. MediaMetro: browsing multimedia document collections with a 3D city metaphor. In *ACM Multimedia*, 213-214.

Chua, T.S., Zhao, Y., and Kankanhalli, M.S. 2002. Detection of human faces in a compressed domain for video stratification, *The Visual Computer 18(2)*, 121-133.

Cooper, M., Foote, J., Girgensohn, A., AND Wilcox, L. 2005. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications* 1(3). 269-288.

Dimitrova, N., Agnihotri, L., and Wei, G. 2000. Video Classification Based on HMM Using Text and Faces. *European Signal Processing Conference*.

Dimitrova, N., Zhang, H. J., Shahraray, B., Sezan, I., Huang, T., AND Zakhor, A. 2002. Applications of video-content analysis and retrieval. *IEEE Multimedia 9(3)*, 42-55.

Dimitrova, N. 2003. Multimedia Content Analysis: The Next Wave. In *CIVR*, 9-18.

Djeraba, C. 2002. Content-based Multimedia Indexing and Retrieval, *IEEE Multimedia 9*, 18-22.

Djeraba, C. 2003. Association and Content-Based Retrieval, *IEEE Transactions on Knowledge and Data Engineering 15(1)*, 118-135.

Dufournaud, Y., Schmid, C., AND Horaud, R. 2000, Matching images with different resolutions, *CVPR*, 612–618.

Dy, J.G., Brodley, C.E., Kak, A., Broderick, L.S., AND Aisen, A.M. 2003. Unsupervised Feature Selection Applied to Content-Based Retrieval of Lung Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence 25(3)*, 373-378.

Eakins, J.P., Riley, K.J., AND Edwards, J.D. 2003. Shape Feature Matching for Trademark Image Retrieval. *CIVR,* 28-38.

Egas, R., Huijsmans, N., Lew, M.S., and Sebe, N. 1999. Adapting k-d Trees to Visual Retrieval. *In Proceedings of the International Conference on Visual Information Systems*, 533-540.

Eiter, T., and Libkin, L. 2005. *Database Theory*. Springer, London. 2005.

Elkwae, E.A. and Kabuka, M.R., 2000. Efficient content-based indexing of large image databases. *ACM Transactions on Information Systems 18(2)*, 171-210.

Fan, J., Gao, Y., and Luo, H. 2004. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *ACM Multimedia.* 540 –547.

Fauquer, J. and Boujemaa, N. 2004 Region-based image retrieval: Fast coarse segmentation and fine color description, *Journal of Visual Languages and Computing*, 15(1):69-95.

Ferecatu, M, Boujemaa, N., and Crucianu, M. 2005 *Hybrid visual and conceptual image representation within active relevance feedback context,* 7th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'05).

Flickner, M. Sawhney, H. Niblack, W. Ashley, J. Qian Huang Dom, B. Gorkani, M. Hafner, J. Lee, D. Petkovic, D. Steele, D. Yanker, P. 1995. Query by image and video content: the QBIC system, *IEEE Computer*, September, 23-32.

Foote, J. 1999. An Overview of Audio Information Retrieval. *ACM Multimedia Systems 7(1)*, 42-51.

Foote, J. 2000. Automatic audio segmentation using a measure of audio novelty. In *ICME*. 452–455.

Forsyth, D.A., AND Fleck, M.M. 1999. Automatic Detection of Human Nudes, *International Journal of Computer Vision 32(1)*, 63-77.

Frankel, C., Swain, M.J., and Athitsos, V. 1996. WebSeer: An Image Search Engine for the World Wide Web. *University of Chicago Technical Report 96-14*.

Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., AND Jacobs, D. 2003. A search engine for 3D models. *ACM Transactions on Graphics 22(1)*, 83-105.

Gauvain, J.L, Lamel, L., Adda, G. and Jardino, M. The LIMSI 1998 Hub-4E Transcription System, *Proc.* DARPA B*roadcastNews Workshop*, pp. 99-104, Herndon, VA, February, 1999.

Gevers, T. 2001. Color-based Retrieval. In *Principles of Visual Information Retrieval*, M.S. LEW, Ed. Springer-Verlag, London, 11-49.

Greenspan, H., Goldberger, J., AND Mayer, A. 2004. Probabilistic Space-Time Video Modeling via Piecewise GMM. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26(3)*, 384-396.

Guo, G., Zhang, H.J., and Li, S.Z. 2001. Boosting for Content-Based Audio Classification and Retrieval: An Evaluation, In *ICME*.

Haas, M., Lew, M.S. AND Huijsmans, D.P. 1997. A New Method for Key Frame based Video Content Representation. In *Image Databases and Multimedia Search*, A. SMEULDERS AND R. JAIN, Eds., World Scientific. 191-200.

Haas, M., Rijsdam, J. and Lew, M. 2004. Relevance feedback: perceptual learning and retrieval in bio-computing, photos, and video, In *ACM MIR,* 151-156.

Hanjalic, A., Lagendijk, R.L., and Biemond, J. 1997. A New Method for Key Frame based Video Content Representation. In *Image Databases and Multimedia Search*, A. Smeulders and R. Jain, Eds., World Scientific. 97-107.

Haralick, R.M. and Shapiro, L.G. 1993. *Computer and Robot Vision*. Addison-Wesley, New York, USA.

Harris, C. and Stephens, M. 1988, A combined corner and edge detector, *4th Alvey Vision Conference*, 147–151

He, X., Ma, W.-Y., King, O. Li, M., and Zhang, H. 2002. Learning and inferring a semantic space from user's relevance feedback for image retrieval. In *ACM Multimedia*. 343–347.

Howe, N. 2003. A Closer Look at Boosted Image Retrieval. *In CIVR*, 61-70.

Jacobs, D.W., Weinshall, D., AND Gdalyahu, Y. 2000. Classification with Nonmetric Distances: Image Retrieval and Class Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22(6)*, 583-600.

Jafari-Khouzani, K. AND Soltanian-Zadeh, H. 2005. Radon Transform Orientation Estimation for Rotation Invariant Texture Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27(6)*, 1004-1008.

Jaimes, A and Chang, S-F. 2002 Duplicate Detection in Consumer Photography and News Video, *ACM Int. Conf. on Multimedia*, 423-424.

Larson, M., Eickeler, S. and Köhler, J. Supporting Radio Archive Workflows with Vocabulary Independent Spoken Keyword Search. Proceedings of SIGIR 2007 Workshop Searching Spontaneous Conversational Speech. 2007

Jolion, J.M. 2001. Feature Similarity. In *Principles of Visual Information Retrieval*, M.S. LEW, Ed. Springer-Verlag, London, 122-162.

Joly, A., Buisson, O., AND Frelicot, C. Robust content-based copy detection in large reference database, *Int. Conf. on Image and Video Retrieval*, 2003

Krishnapuram, R., Medasani, S., Jung, S.H., Choi, Y.S., AND Balasubramaniam, R. 2004. Content-Based Image Retrieval Based on a Fuzzy Approach. *IEEE Transactions on Knowledge and Data Engineering 16(10)*, 1185-1199.

Levine, M. 1985. *Vision in Man and Machine*, Mcgraw Hill, Columbus.

Lew, M.S. AND Huijsmans, N. 1996. Information Theory and Face Detection. In *Proceedings of the International Conference on Pattern Recogntion*, 601-605.

Lew, M.S. 2000. Next Generation Web Searches for Visual Content. *IEEE Computer*, November, 46-53.

Lew, M.S. 2001. *Principles of Visual Information Retrieval*. Springer, London, UK.

Lew, M.S., Sebe, N., Djeraba, C., AND Jain, R. 2006 Multimedia Information Retrieval: State of the Art and Challenges, *ACM Transactions on Multimedia Computing, Communication, and Applications,* 2(1):1-19.

Lew, M.S. and Denteneer, D. 2001. Fisher Keys for Content Based Retrieval. *Image and Vision Computing 19*, 561-566.

Li, J. and Wang, J.Z. 2003. Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25(9)*, 1075-1088.

Lienhart, R. 2001. Reliable Transition Detection in Videos: A Survey and Practitioner's Guide. *International Journal of Image and Graphics 1(3)*, 469-486.

Lindeberg, T. 1998, Feature detection with automatic scale selection, *International Journal of Computer Vision*, *30(2)*:79–116

Lindeberg, , T. and Garding, J. 1997, Shape-adapted smoothing in estimation of the 3D shape cues from affine deformations of local 2D brightness structure, *Image and Vision Computing, 15(6)*:415–434,1997

Liu, B., Gupta, A., and Jain, R. 2005. MedSMan: A Streaming Data Management System over Live Multimedia, *ACM Multimedia,* 171-180.

Liu, H., Xie, X., Tang, X., Li, Z.W., Ma, W.Y. 2004. Effective browsing of web image search results. In ACM MIR, 84-90.

Liu, X., Srivastava, A., and Sun, D. 2003. Learning Optimal Representations for Image Retrieval Applications. *In CIVR*, 50-60.

Lowe, D. 2004, Distinctive image features from scale-invariant keypoints, *International Journal* of Computer Vision, 60(2), 91–110.

Mikolajczyk, K. and Schmid, C. 2004, Scale and affine invariant interest point detectors *International Journal of Computer Vision, 60(1)*, 63–86.

Muller, H., Muller, W., Marchand-Maillet, S., Pun, T., AND SQUIRE, D. 2000. Strategies for Positive and Negative Relevance Feedback in Image Retrieval. In ICPR, 1043-1046.

Müller, W. AND Henrich, A. 2003. Fast retrieval of high-dimensional feature vectors in P2P networks using compact peer data summaries. In *ACM MIR*, 79-86.

Ojala, T., Pietikainen, M., and Harwood, D. 1996. Comparative study of texture measures with classification based on feature distributions, Pattern Recognition 29(1), 51-59.

Pereira, F. and Koenen, R. 2001. MPEG-7: A Standard for Multimedia Content Description. *International Journal of Image and Graphics* 1(3), 527-546.

Rautiainen, M., Seppanen, T., Penttila, J., and Peltola, J. 2003. Detecting Semantic Concepts from Video Using Temporal Gradients and Audio Classification. In CIVR.

Rebai, A, Joly, A., and Boujemaa, N. 2007 Interpretability Based Interest Points Detection, *ACM International Conference on Image and Video Retrieval.*

Rocchio, 1971. Relevance Feedback in Information Retrieval. In *The Smart Retrieval System:* Experiments in Automatic Document Processing, G. Salton, Ed. Prentice Hall, Englewoods Cliffs.

Rowe, L.A. and Jain, R. 2005. ACM SIGMM retreat report on future directions in multimedia research. *ACM Transactions on Multimedia Computing, Communications, and Application 1(1),* 3-13.

Rowley, H., Baluja, S., and Kanade, K. 1996. Human Face Detection in Visual Scenes. *Advances in Neural Information Processing Systems 8,* 875-881.

Schmid, C., Mohr, R., Bauckage, C. 2000, Evaluation of interest point detectors, International Journal of Computer Vision, 37(2), 151–172

Schneiderman, H. AND Kanade, T. 2004. Object Detection Using the Statistics of Parts, International Journal of Computer Vision 56(3), 151-177.

Sclaroff, S., La Cascia, M., Sethi, S., and Taycher, L. 2001. Mix and Match Features in the ImageRover Search Engine. In *Principles of Visual Information Retrieval*, M.S. LEW, Ed. Springer-Verlag, London, 259-277.

Scott, G.J. AND Shyu, C.R. 2003. EBS k-d Tree: An Entropy Balanced Statistical k-d Tree for Image Databases with Ground-Truth Labels. *In CIVR*, 467-476.

Sebastian, T.B., Klein, P.N., AND Kimia, B.B. 2004. Recognition of Shapes by Editing Their Shock Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26(5)*, 550-571.

Sebe, N., Lew, M.S., AND Huijsmans, D.P. 2000. Toward Improved Ranking Metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10)*, 1132-1143.

Sebe, N., AND Lew, M.S. 2001. Color Based Retrieval, *Pattern Recognition Letters 22(2)*, 223-230.

Sebe, N., AND LEW, M.S. 2002. Robust Shape Matching. *In CIVR*, 17-28.

Sebe, N., TIAN, Q., LOUPIAS, E., LEW, M.S., AND HUANG, T.S. 2003. Evaluation of Salient Point Techniques. Image and Vision Computing 21(13-14), 1087-1095.

Sebe, N., LEW, M.S., ZHOU, X., AND HUANG, T.S. 2003. The State of the Art in Image and Video Retrieval. *In CIVR*.

Shao, H., Svoboda, T., Tuytekaars, T., and Van Gool, L. 2003. HPAT Indexing for Fast Object/Scene Recognition Based on Local Appearance. In CIVR, 71-80.

Shen, H. T., Ooi, B. C., and Tan, K. L. 2000. Giving meanings to www images. In *ACM Multimedia*, 39–48.

Smeaton, A.F. and Over, P. 2003. Benchmarking the Effectiveness of Information Retrieval Tasks on Digital Video. In CIVR, 10-27.

Smeulders, A., 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 and Machine Intelligence 22(12)*, 1349-1380.

Smith, J. R. and Chang, S.F. 1997. Visually Searching the Web for Content. *IEEE Multimedia* 4(3), 12-20.

Snoek, C.G.M., Worring, M., van Gemert, J., Geusebroek, J.M., Koelma, D., Nguyen, G.P., de Rooij, O., AND Seinstra, F. 2005. MediaMill: exploring news video archives based on learned semantics. In *ACM Multimedia*, 225-226.

Spierenburg, J.A. AND Huijsmans, D.P. 1997. VOICI: Video Overview for Image Cluster Indexing. In BMVC.

Srivastava, A., Joshi, S.H., Mio, W., AND Liu, X. 2005. Statistical Shape Analysis: Clustering, Learning, and Testing. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27(4)*, 590-602.

Sundaram, H., Xie, L., and Chang, S.F. 2002. A utility framework for the automatic generation of audio-visual skims. ACM Multimedia, 189-198.

Szumilas, L., Donner, R., Langs, G., and Hanbury, A. 2007 Local structure detection with orientation-invariant radial configuration, *CVPR*.

Tangelder, J. and Veltkamp, R.C. 2004. A survey of content based 3d shape retrieval methods*, In Proceedings of International Conference on Shape Modeling and Applications*, 157-166.

Tian, Q., Sebe, N., Lew, M.S., Loupias, E., and Huang, T.S. 2001. Image Retrieval using Wavelet-based Salient Points. *Journal of Electronic Imaging* 10(4), 835-849.

Tian, Q., Moghaddam, B., and Huang, T.S. 2002. Visualization, Estimation and User-Modeling. In CIVR, 7-16.

Tieu, K. and Viola, P. 2004. Boosting Image Retrieval, *International Journal of Computer Vision 56(1)*, 17-36.

Therrien, C.W. 1989. Decision, Estimation, and Classification, Wiley, New York, USA.

Tuytelaars, T. and Van Gool, L. 2000, Wide baseline stereo matching based on local affinely invariant regions, *British Machine Vision Conference*, 412–425.

Uchihashi, S., Foote, J., Girgensohn, A., AND Boreczky, J. 1999. Video Manga: generating semantically meaningful video summaries. In ACM Multimedia, 383-392.

Vailaya, A., Jain, A., and Zhang, H. 1998. On Image Classification: City vs Landscape. In Proceedings of Workshop on Content-based Access of Image and Video Libraries, 3-8.

Veltkamp, R.C. and Hagedoorn, M. 2001. State of the Art in Shape Matching. In *Principles of Visual Information Retrieval*, M.S. Lew, Ed. Springer-Verlag, London, 87-119.

Winston, P. 1992. Artificial Intelligence, Addison-Wesley, New York, USA.

Wu, P., Choi, Y., Ro., Y.M., and Won, C.S. 2001. MPEG-7 Texture Descriptors. *International Journal of Image and Graphics 1(3)*, 547-563.

Yang, M.H., Kriegman, D.J., AND Ahuja. N. 2002. Detecting Faces in Images: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1), 34-58.

Ye, H. and Xu, G. 2003. Fast Search in Large-Scale Image Database Using Vector Quantization. *CIVR*, 477-487.

Yin, P.Y., Bhanu, B., Chang, K.C., AND Dong, A. 2005. Integrating Relevance Feedback Techniques for Image Retrieval Using Reinforcement Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1536-1551.

Zhou, X.S. and Huang, T.S. 2001. Comparing discriminating transformations and SVM for learning during multimedia retrieval. In *ACM Multimedia,* 137-146.

Annex A: Overview of the 9 IST projects

The project-coordinators were asked to give feedback about their projects regarding the research activities in the area of multimedia indexing and retrieval. The filled in questionnaires sent by the 9 projects are included in this Appendix.

Overview of Divas

Project name	Divas
co-ordinator	Nikos Achilleopoulos, Archetypon S.A. Information Technologies
Budget in Mio. Euro	budget: 3,188
project start	1.1.2007
project duration (in month)	24
Objectives	
main objectives	Design and implement a multimedia search engine based on advanced direct video and audio search algorithms applied on encoded (compressed) content
objectives regarding AV search engine technology	Direct Search in Compressed Audio and Video
target / final product	DIVAS Algorithms, system level demonstrator (available over the web for user evaluation), studies and designs methodologies for application integration
internal user groups of the project results	ESCOM, BeTV
external user groups of the project results	Potential: AudioVisual Archive of any type
scenarios for deployment	ESCOM: indexing and search of audiovisual archive, BeTV: DRM monitoring
data sources	ESCOM and BeTV: Videofiles and Audiofiles
metadata inventories	Escom and BetV Metadata
Modalities (please give detailed answers in the additional sheets of this form)	
speech/audio indexing	yes
image indexing	no
video indexing	yes
text+semantics	no
multimodal fusion	no
retrieval models/techniques	no
official benchmarking/which one	
evaluation	
social networks	no
what standardization body are you addressing (if applicable)	MPEG-7
Please identify your use cases (if any)	
Are the details confidential?	no/partly
System development/integration	yes, planned for 2008
Distributed system	possible
p2p technology	no
mobile access	yes
DRM	yes

Overview of Rushes

Project name	**RUSHES**
co-ordinator	Fraunhofer HHI, Dr. Oliver Schreer
Budget in Mio. Euro	4,55 (2,67 funded)
project start	01.02.2007
project duration (in month)	30
Objectives	
main objectives	to design, implement, and validate a system for indexing, accessing and delivering raw, unedited audio-visual footage known in broadcasting industry as "rushes".
objectives regarding AV search engine technology	to provide services for querying audio-visual footage using keywords, semantics or actual footage examples.
target / final product	(1) to allow home users to have advanced search functionalities and low access latency when navigating rushes databases; (2) to allow professional users to conduct automatic content cataloguing and semantic based indexing to link raw content with metadata; (3) to illustrate the benefits of using semantic technologies in video annotations/indexing; (4) to summarise AV sequences using representative frames.
internal user groups of the project results	specific departments of RUSHES industrial partners such as ATC, FAST, ETB
external user groups of the project results	broadcasters, search engine development companies, other European projects and clusteringinitiatives (CHORUS)
scenarios for deployment	regular meeting scenario, movies, advertisements, TV news report
data sources	text, television and other resources, and radio and other audio resources.
metadata inventories	MPEG-7 XML
Modalities (please give detailed answers in the additional sheets of this form)	
speech/audio indexing	high- and low-level audio features, e.g. envelop, frequency, time, space, etc., will be extracted for a proper classification.
image indexing	high- and low-level video features, e.g. color, texture, action, space, etc., will be extracted for a proper classification.
video indexing	this consists of image and audio indexing. Video indexing cannot be accomplished unless the two components' indexing is performed. This involves alignment of visual and audio signals, interaction of two components, and other process. In terms of video summarisation/annotation, this can be performed using an attention model that considers human visual models for motion, audio, and event detection.
text+semantics	Graph Matching, Kernel analysis … can be used for similarity search.
multimodal fusion	Audio and visual observations can be fused in the domain of Bayesian network
retrieval models/techniques	probably Hidden Markov Model, or Mixture Gaussian Model
official benchmarking/which one	TRECVID
evaluation	the basic idea is the statistical analysis based on the test on benchmarking data
social networks	Britain's universities, some local companies such as BT, Microsoft, HP, IBM, Motorola.
what standardization body are you addressing (if applicable)?	MPEG/ITU-T, JPSearch, DVB, SMPTE, IPTC
Please identify your use cases (if any)	journalists working at broadcasters will use the RUSHES system for semi-automatic indexing and annotation as well as for retrieval of rushes material.
Are the details confidential?	yes

System development/integration	
Distributed system	yes
p2p technology	no issue in RUSHES
mobile access	no issue in RUSHES
DRM	no issue in RUSHES

Overview of Sapir

Project name	**SAPIR**
co-ordinator	Yosi Mass, IBM
Budget in Mio. Euro	4,5
project start	01.01.2007
project duration (in month)	30
Objectives	
main objectives	The broad scope of SAPIR is to develop theories and technologies for next-generation search techniques that would effectively and efficiently deliver relevant information in the presence of exponentially growing (i.e. dynamic) volumes of distributed multimedia data. Fundamental to our approach is the development of scalable solutions that address the requirements of future generations of massively distributed data produced in a variety of applications. The scale of the problem can be gauged from the fact that almost everything we see, read, hear, write and measure will soon be available to computerized information systems.
objectives regarding AV search engine technology	While structured search methods apply to attributed-type data that yield records that match the search query exactly, SAPIR offers a more modern approach to searching information through similarity searching which is used in content-based retrieval for queries involving complex data such as images, videos, speech, music and text. Similarity search is based on gradual rather than exact relevance using a distance metric that, together with the database, forms a mathematical metric space. The obvious advantage of similarity search is that the results can be ranked according to their estimated relevance. However, current similarity search structures, which are mostly centralized, reveal linear scalability in respect to the data search size, which is not sufficient for the expected data volume dimension of the problem. With the increasing diversity of digital data types covering practically all forms of fact representation, computerized data processing must provide adequate tools for similarity searching.
target / final product	Define APIs and show a prototype that can do feature extractions form the different medias and index and search large volumes using a P2P architecture.
internal user groups of the project results	SAPIR partners
external user groups of the project results	The APIs will be published and be available for external users. We will have to decide which components that are developed by SAPIR will be available also.
scenarios for deployment	We have worked on 5 possible scenarios for the technology - 1. Advanced home messaging 2. The music and text scenario 3. Tourist searching 4. Hollywood@home 5. The journalist's helpers
data sources	We may start by testing image + text + metadata on the Flickr image collection.
metadata inventories	From Flickr and automatically extracted
Modalities (please give detailed answers in the additional sheets of this form)	See next sheets
speech/audio indexing	
image indexing	
video indexing	

text+semantics	
multimodal fusion	
retrieval models/techniques	
official benchmarking/which one	
evaluation	
social networks	We currently work on definitions of Social Networks and how they can improve the search results. This is part of WP7
what standardization body are you addressing (if applicable)	MPEG-7, MPEG-21
Please identify your use cases (if any)	We work on 5 User scenarios as described above. Use cases can be derived from those scenarios.
Are the details confidential?	No
System development/integration	We defined indexing ans Search APIs. We currently work on first implementation of the Search APIs. We will upgrade the APIs as work progress and also add Content Managtement/Feature extraction APIs.
Distributed system	
p2p technology	The main objective of the project is a large scale search using P2P technology.
mobile access	Will be supported as part of a dedicated WorkPackage (WP7)
DRM	This will be developed as part of a dedicated WorkPackage (WP6).

Overview of Semedia

Project name	**SEMEDIA Search Environments for Media**
co-ordinator	Prof. Ricardo Baeza-Yates, Yahoo! Research
Budget in Mio. Euro	Funding: 2,73
project start	01.01.2007
project duration (in month)	30
Objectives	The overall objective of SEMEDIA is to develop a collection of audiovisual search tools that are heavily user driven, preserve metadata along the chain, are generic enough to be applicable to different fields (broadcasting production, cinema postproduction, social web). This will be achieved through five specific objectives:
main objectives	O1. **To develop techniques to extract metadata from 'essence'** in ways that allow the automatic inference of high-level structural information from the content of new, partly annotated media data produced in a range of professional and amateur contexts. O2. To create tools for navigating intelligently and searching efficiently in very large bodies of media in heterogeneous, distributed, networked data storage systems. O3. To design and evaluate efficient user interfaces that allow fast browsing. O4. To integrate the results in a series of prototypes for real production and postproduction environments, and evaluate them with real data sets, user groups and industry work flows. O5. To develop strategies for wide dissemination of the results and their incorporation into marketable products.
objectives regarding AV search engine technology	"
target / final product	Tools will be integrated into industrial partner's systems. An integrated demonstrator will also be produced.
internal user groups of the project results	Yes, industrial partners have formed internal user groups.
external user groups of the project results	Yes, an external user group has been organized.

scenarios for deployment	Yes, however, it is available to the Consortium only. In Month 12, user scenarios will be made available to the Public.
data sources	Yes, industrial partners (BBC, CCRTV-ASI, S&M, and Yahoo!) have made data available to the consortium partners.
metadata inventories	Yes, meta-data inventories related to the data sources are being build.
Modalities (please give detailed answers in the additional sheets of this form)	
speech/audio indexing	N/A
image indexing	Yes, to the extend that it contributes to the video indexing and retrieval task.
video indexing	Yes, this is the main focus of the SEMEDIA project
text+semantics	Yes, to the extend that it contributes to the video indexing and retrieval task.
multimodal fusion	Yes
retrieval models/techniques	Yes
official benchmarking/which one	Possibly an adaptation of TRECVID
evaluation	3 Types: 1. System perf. 2. Usability 3. Retrieval perf.
social networks	Flickr and online models based on video
what standardization body are you addressing (if applicable)	tools produced will use "standard" APIs, whenever possible, we will adopt existing standards. MPEG7 is the current candidate.
Please identify your use cases (if any)	initial user scenarios produced (consortium only). In month 12, revised scenarios will be produced and available publicly.
Are the details confidential?	in month 12, scenarios will be available publicly.
System development/integration	planned that tools will be integrated into industrial partners systems. Integrated demos will also be produced.
Distributed system	Yes, but it is not the main focus of the project.
p2p technology	No.
mobile access	No.
DRM	Yes, Digital Rights Management is a concern and is being addressed.

Overview of Tripod

Project name	**Tripod**
co-ordinator	University of Sheffield, Mark Sanderson
Budget in Mio. Euro	funding: 3.15
project start	01.01.2007
project duration (in month)	36
Objectives	
main objectives	The primary objective of Tripod is to revolutionise access to the enormous body of visual media. Applying an innovative multidisciplinary approach Tripod will utilise largely untapped but vast, accurate and regularly updated sources of semantic information to create ground breaking intuitive search services, enabling users to effortlessly and accurately gain access to the image they seek from this ever expanding resource.
objectives regarding AV search engine technology	Create image search facilities that serve broader user needs than current keyword or content-based approaches provide
target / final product	Package Tripod's tools as a suite of services to prepare Tripod for exploitation in a wide range of markets

internal user groups of the project results	Ordnance Survey, United Kingdom; Centrica, Italy; Geodan Holding BV, The Netherlands; Fratelli Alinari Istituto Edizioni Artistiche SpA, Italy; Tilde, Latvia
external user groups of the project results	Photographic agencies
scenarios for deployment	
data sources	Mapping data from OS & Geodan; photographs from Alinari & Tilde
metadata inventories	
Modalities (please give detailed answers in the additional sheets of this form)	
speech/audio indexing	
image indexing	
video indexing	
text+semantics	×
multimodal fusion	
retrieval models/techniques	
official benchmarking/which one	
evaluation	
social networks	
what standardization body are you addressing (if applicable)	
Please identify your use cases (if any)	
Are the details confidential?	
System development/integration	
Distributed system	
p2p technology	
mobile access	
DRM	

Overview of Victory

Project name	**VICTORY**
co-ordinator	Dr. Dimitrios Tzovaras
Budget in Mio. Euro	project budget: 3,869
project start	01.01.2007
project duration (in month)	30
Objectives	
main objectives	O1: The first objective of VICTORY is to develop the MultiPedia repository and the mechanisms to support its wide access by the community. The centralised MultiPedia repository will consist of only the 3D models that contain the global truth of the objects stored in the repository. The accompanying MultiPedia information (2D images, text, annotations, etc.) will be available on a peer-to-peer basis. Tools will be supported by the repository administration mechanism for population, management and reorganisation of the centralised content. The content will be adequately categorised in order to support special interest groups targeting mainly industrial applications (automotive, games, simulations, etc.). Also, the repository will act as the main access point for the P2P framework and thus it will support mechanisms for adding MultiPedia content from the peers connected each time to the VICTORY network.

	O2: The second objective of VICTORY is to develop novel 3D search and retrieval algorithms (see below) O3: The third objective of VICTORY is the development of novel search and retrieval framework that allows an easy integration of different search methodologies (see below). O4: The fourth objective of VICTORY is the development of a P2P scheme so as to utilise not only the distributed data storage, but also the computational power of each peer for the pre-processing, interpreting, indexing, searching, retrieving and representing of MultiPedia data. Through the VICTORY framework, users will be able to handle, share and retrieve 3D and audio-visual data among peers around the world. Moreover, every peer will be responsible for extracting and indexing the features of the shared 3D data, thus the efficient manipulation of the 3D data will be accomplished. The P2P-based middleware will provide the means (intelligence, semantics, and communications protocols) allowing the negotiation and determination of peer resources sharing. The key driver will be the user QoE realised as the combination of a multitude of Quality of Services (communications quality, processing speed, 3D content rendering quality, power consumption, etc) impacting the user experience.
objectives regarding AV search engine technology	O2: The second objective of VICTORY is to develop novel 3D search and retrieval algorithms which will be based on a) content, which will be extracted taking into account low-level geometric characteristics and b) context, which will be high-level features (semantic concepts) mapped to low-level features. In the existing 3D search and retrieval methods no semantic information (high-level features) is attached to the (low-level) geometric features of the 3D content, which would significantly improve the retrieved results. Therefore, the second objective of the proposed system is to introduce a solution so as to bridge the gap between low and high-level information through automated knowledge discovery and extraction mechanisms. High level features will be a) appropriate annotation options provided by the system or generated by the user dynamically (active learning) and b) relevance feedback where the user will mark which retrieved objects he thinks are relevant to the query (user's subjectivity). The strength of the VICTORY approach is the ability to translate both explicit and tacit knowledge of the user into semantic information by analysing user's explicit operations like manual annotation, query by example, feedback and intuitive interactions with the system like browsing or objects manipulations. This acquired knowledge will be exploited to automatically propagate annotations through the existing object database of each peer and to adapt the retrieval process to the user's subjectivity. The input of the system will consist of mixed-media (MultiPedia) queries such as text (annotation), 2D images (taken by the user's mobile device), sketches made by the user and 3D objects. Therefore, 2D/3D combined algorithms are going to be developed and integrated to the search engine. O3: For supporting sophisticated 3D content search and retrieval, a search framework is needed that allows for combining text-/metadata-based searching with 3D object searching. An ontology helps to cluster the 3D objects and to either • use the ontology as organizing principle to navigate through the objects, or • use the ontology to restrict/guide the search through the objects Thus, the third objective of VICTORY is the development of novel search and retrieval framework that allows an easy integration of different search methodologies. It will result in an integrated platform which allows processing and accessing data and knowledge by using ontology based management and retrieval mechanisms. The challenge within VICTORY means to bridge the gap between textual-/metadata oriented data respectively and to apply this really innovative

	technology to MultiPedia content, especially such as 3D-objects.
target / final product	3D search engine
internal user groups of the project results	Companies: EMPOLIS, HYPERTECH
external user groups of the project results	Automotive, aeronautic, game industries, all
scenarios for deployment	see use cases
data sources	internet, automotive industries
metadata inventories	
Modalities (please give detailed answers in the additional sheets of this form)	
speech/audio indexing	
image indexing	
video indexing	
text+semantics	
multimodal fusion	
retrieval models/techniques	CERTH/ITI algorithms (see www.victory-eu.org)
official benchmarking/which one	Princeton Shape Benchmark (see www.victory-eu.org)
evaluation	
social networks	
what standardization body are you addressing (if applicable)	MPEG-7
Please identify your use cases (if any)	
Are the details confidential?	YES
System development/integration	
Distributed system	YES
p2p technology	YES
mobile access	YES
DRM	YES

Overview of Vidi-Video

Project name	VIDI-Video
co-ordinator	Prof. A. Smeulders
Budget in Mio. Euro	3.6 Meuro
project start	2/1/2007
project duration (in month)	36
Objectives	
main objectives	boost the performance of video search by developing a 1000 element thesaurus for automatically detecting instances of semantic concepts in the audio-visual content
objectives regarding AV search engine technology	Semantic search using a large-scale learned vocabulary
target / final product	Semantic video search engine
internal user groups of the project results	Fondazione Rinascimento Digitale, Italy Beeld en Geluid, The Netherlands

external user groups of the project results	Potential: audiovisual archives in broadcasting, surveillance, conferencing, diaries and logging.
scenarios for deployment	Use within the processes of the archives involved. In later stage other parties.
data sources	Sound and Vision archives, archives of FDR and partners, TRECVID, Surveillance data.
metadata inventories	Existing annotations of the archives.
Modalities (please give detailed answers in the additional sheets of this form)	
speech/audio indexing	Yes
image indexing	No
video indexing	Yes
text+semantics	No
multimodal fusion	Yes
retrieval models/techniques	Yes
official benchmarking/which one	TRECVID, VOC
evaluation	Yes
social networks	Yes (for obtaining annotations)
what standardization body are you addressing (if applicable)	MPEG-7, RDFS/OWL
Please identify your use cases (if any)	
Are the details confidential?	Partly
System development/integration	Yes
Distributed system	Yes
p2p technology	No
mobile access	No
DRM	No

Overview of Vitalas

Project name	**VITALAS**
co-ordinator	INRIA, ERCIM
Budget in Mio. Euro	6 millions
project start	January 1st, 2007
project duration (in month)	36
Objectives	
main objectives	Use-case driven project that aims that aims to provide advanced solution for indexing, searching and accessing large scale digital audio-visual content.
objectives regarding AV search engine technology	Cross-media indexing and retrieval, interactivity and context adapting, scalability
target / final product	Pre-industrial prototype system dedicated to intelligent access services to multimedia professional archives
internal user groups of the project results	Audiovisual archives (INA) and broadcasters (IRT), Photo press agency BELGA
external user groups of the project results	Photo press agency AFP
scenarios for deployment	
data sources	INA, IRT, BELGA
metadata inventories	IPTC BELGA annotations, INA video archives annotations
Modalities (please give detailed answers in the additional sheets of this form)	

speech/audio indexing	yes
image indexing	yes
video indexing	yes
text+semantics	yes
multimodal fusion	yes
retrieval models/techniques	yes
official benchmarking/which one	Not yet. Probably TRECVID. Maybe ImageCLEF, ImagEval.
evaluation	Technical evaluation + end user tests
social networks	no
what standardization body are you addressing (if applicable)	Content representation (e.g. JPEG), query languages (e.g. Xquery), evaluation of multimedia retrieval systems (e.g. JPsearch)
Please identify your use cases (if any)	Automatic labelling of visual concepts in images, global navigation in a set of results, Interactive browsing of a search results, Search by concept, Face identification, Personalization, Search by example, Visual and audio categorization, Search by concept in video content.
Are the details confidential?	yes
System development/integration	3 prototype versions. V1 due to January 2008
Distributed system	Yes: Web services, distributed similarity search structures
p2p technology	No
mobile access	No
DRM	No

Modules on Speech/Audio Indexing and Retrieval

Project Divas

module/task	**Music Segmentation**
investigator/partner	Fraunhofer IDMT
applied algorithms/approaches	segmentation algorithm based on Foote's segmentation
pre-existing technology before project start	MP3, Music Segmentation, Speech Segmentation
research challenge/innovation/not addressed	improvement of the music segmentation algorithm; music segmentation directly from the compressed domain
type and amount of processed data	compressed audio data; more than 1000 pieces of music
success criteria, recognition/indexing rate	at the moment the recognition performance is about 70%
risk	
demo (available/planned/not forseen)	demo is available

module/task	**Speech Segmentation**
investigator/partner	SAIL LABS
applied algorithms/approaches	make models more robust by statistical training using compressed audio and application of transforms
pre-existing technology before project start	segmentation component using specifically trained phone-level models and a GMM/BIC-based approach for segmentation.
research challenge/innovation	keep aproximate same level of segmentation results in spite of compressed audio data and corresponding limitation of incorporated information
type and amount of processed data	
success criteria, recognition/indexing rate	see separate table of current non-compressed vs compressed data segment recogition results
risk	moderate
demo available	yes

Project Rushes

module/task	**Audio retrieval**
investigator/partner	Brunel University, UK
applied algorithms/approaches	possibly HMM with perception model (how human beings link speech with visual components)
pre-existing technology before project start	HMM implementation by others
research challenge/innovation/not addressed	feature extraction/selection, speech-to-video transmoding
type and amount of processed data	real audio data, at least 20 persons and each one > 10 minutes
success criteria, recognition/indexing rate	in the used database, hopefully > 85%
risk	consistency of recognition
demo (available/planned/not forseen)	will be available

Project Sapir

module/task	Speech
investigator/partner	IBM
applied algorithms/approaches	We use an Automatic Speech Recognition (ASR) system for transcribing speech data. The ASR generates lattices that can be considered as directed acyclic graphs. Each vertex in a lattice is associated with a timestamp and each edge (u,v) is labeled with a word or phone hypothesis and its prior probability, which is the probability of the signal delimited by the timestamps of the vertices u and v, given the hypothesis. The 1-best path transcript is obtained from the path containing the best hypotheses using dynamic programming techniques. For indexing and search purposes, it is often more convenient to use a compact representation of a word lattice, called word confusion network (WCN). Each edge is labeled with a word hypothesis and its posterior probability, i.e., the probability of the word given the signal. The main advantages of WCN are that it provides an alignment for all of the words in the lattice and also posterior probabilities. Note that the 1-best path can be directly extracted from the WCN.
pre-existing technology before project start	We have an ASR technology and we adapt it to represent the features in MPEG-7 and then use it for indexing and search in the SAPIR p2p architecture.
research challenge/innovation/not addressed	Write a UIMA Annotators that extract the features and represent them in MPEG-7. Index and search using a P2P architecture
type and amount of processed data	TBD
success criteria, recognition/indexing rate	TBD
risk	Efficiency dimension - SAPIR basic (features similarity search) performance can degrade for large volume of content and/or large number of peers, resulting in scalability issues. Effectiveness dimension - Feature search does not improve over text only search, resulting in little gain over existing approaches.
demo (available/planned/not forseen)	Planned

module/task	Music
investigator/partner	UPD - University of Padova
applied algorithms/approaches	Music ContentObjects can be instantiated in three main forms: digital audio recordings with possible compression, MIDI (Musical Instrument Digital Interface) files with temporal information, and digital scores. All the forms may be of interest for the final user, depending on the required audio quality, on the available bandwidth, on the usage, and on copyright restrictions. Many formats correspond to audio and score forms, yet for the aims of this project, only open formats will be addressed, such as MP3 and aiff for audio or Lilypond [11] and Guido [12] for scores. The first step in music processing will regard the automatic extraction of high level features, which are shared by all the forms. The main content descriptors, are the rhythm and the melody of the leading voice.
pre-existing technology before project start	UPD has technology for Music feature extraction
research challenge/innovation	Write a UIMA Annotators that extract the features and represent them in MPEG-7. Index and search using a P2P architecture
type and amount of processed data	TBD
success criteria, recognition/indexing rate	TBD
risk	Same as for speech
demo available	Planned

Project Vitalas

module/task	**Speech Mining and Segmentation**
investigator/partner	Fraunhofer IAIS
applied algorithms/approaches	The speech recordings from the content providers (e.g. INA) are segmented automatically in homogenous segments. Further, a speech/non-speech detection is performed. Here algorithms based on Gaussian Mixture Techniques are applied. The indexing of the speech files is performed by subword recognition on syllable level. Here Hidden-Markov-Models are used for the context-dependent modelling of the phones. For the subword retrieval process a dynamic time warping approach in combination with the Levensthein distance metric is applied to enable a fuzzy search to eliminate the Out-Of-Vocabulary problem.
pre-existing technology before project start	Speech recognition engine for the German language.
research challenge/innovation	Robust indexing on large scale corups
type and amount of processed data	Speech and video recordings from the INA archive (mainly in French)
success criteria, recognition/indexing rate	Tbd
risk	High
demo available	Segmentation and indexing demo for German is available

module/task	**Jingle detection**
investigator/partner	Fraunhofer IAIS
applied algorithms/approaches	Fingerprints extraction, combination of features, Gaussian filtering techniques
pre-existing technology before project start	Zero crossing rate and spectral flatness
research challenge/innovation	Solve fade-in/fade-out problems and signal overlap problems, scalability
type and amount of processed data	Audio-visual archives, 10 000 hours
success criteria, recognition/indexing rate	Currently being defined
risk	Medium
demo available	Segmentation and indexing demo for German is available

Project VidiVideo

module/task	**Audio Analysis**
investigator/partner	INESC
applied algorithms/approaches	Speech recognition, Machine learning, MEL features
pre-existing technology before project start	Speech recognition for Portugese, features for speech recognition, no use of audio events in search engines
research challenge/innovation/not addressed	Non-news data, integration of many different features, audio-visual integration in early stages,
type and amount of processed data	Broadcast TV, >500 hours
success criteria, recognition/indexing rate	Average Precision
risk	Methods don't generalize to the new domains
demo (available/planned/not forseen)	Integrated demo for whole of VidiVideo

Modules on Image Indexing and Retrieval

Project Rushes

module/task	Image retrieval
investigator/partner	Brunel University, UK
applied algorithms/approaches	HMM or Gaussian Mixture Model and perception model (how human beings search similarity, based on the selected visual features)
pre-existing technology before project start	HMM and Gaussian Mixture Model by others
research challenge/innovation	robust image retrieval
type and amount of processed data	real static images > 5000 frames
success criteria, recognition/indexing rate	in the used database hopefully > 85%
risk	consistency
demo available	will be available

Project Sapir

module/task	Image
investigator/partner	CNR
applied algorithms/approaches	Images will be indexed using the following five standard MPEG-7 visual descriptors - Scalable Color, Color Structure, Color Layout, Edge Histogram, Homogeneous Texture
pre-existing technology before project start	We use the MPEG-7 Reference software with some modifications
research challenge/innovation	Write a UIMA Annotators that extract the features and represent them in MPEG-7. Index and search using a P2P architecture
type and amount of processed data	> 20M images
success criteria, recognition/indexing rate	TBD
risk	Same as for Speech
demo available	Planned

Project Vitalas:

module/task	Similarity search in large datasets
investigator/partner	INRIA, INA
applied algorithms/approaches	Low level features extraction (global and local features), Probabilistic similarity search structure
pre-existing technology before project start	Low level global features, SIFT like local features, in-memory similarity search structures
research challenge/innovation	More interpretable visual features, similarity search in very large features datasets, more generic and efficient similarity search structures
type and amount of processed data	3 millions professional images (3 millions global features, up to 3 billions local features)
success criteria, recognition/indexing rate	Currently being defined
risk	Medium
demo available	Not yet

module/task	Objects and visual concepts recognition
investigator/partner	INRIA
applied algorithms/approaches	Low level visual features, machine learning, multiple instance learning
pre-existing technology before project start	Low level global features, SIFT like local features, SVM, boosting
research challenge/innovation	More interpretable and complementary low level features, Automatic relevant visual concepts selection, Large and relevant learning sets generation, Large sets of concepts
type and amount of processed data	Annotated professional images (3 millions)
success criteria, recognition/indexing rate	Currently being defined

risk	High
demo available	Not yet

module/task	Visualization maps
investigator/partner	INA, INRIA
applied algorithms/approaches	Graph based and diagram representations, semi-supervised clustering
pre-existing technology before project start	Proximity information maps
research challenge/innovation	Interactive feedback of the user, fusion of heterogeneous similarity measures
type and amount of processed data	Annotated professional images (10000)
success criteria, recognition/indexing rate	Currently being defined
risk	Medium
demo available	Not yet

Modules on 3D Indexing and Retrieval

Project Rushes

module/task	3D video scene description
investigator/partner	FhG/HHI, Germany
applied algorithms/approaches	camera motion and 3D scene structure clustering
pre-existing technology before project start	initial algorithm
research challenge/innovation	real life data
type and amount of processed data	1 hour rushes material from EiTB
success criteria, recognition/indexing rate	not yet defined
Risk	too inaccurate for real life data
demo available	no

Project Victory

module/task	3D Search engine
investigator/partner	CERTH/ITI
applied algorithms/approaches	see www.victory-eu.org
pre-existing technology before project start	www.3d-search.iti.gr
research challenge/innovation	all the techniques used are innovative
type and amount of processed data	thousands of 3D models
success criteria, recognition/indexing rate	retrieval accuracy>95%
Risk	
demo available	(see www.victory-eu.org)

Modules on Video Indexing and Retrieval

Project Divas

module/task	Video segmentation, indexing and search
investigator/partner	ELECARD
applied algorithms/approaches	"Scene change detection" segmentation algorithm, "Scene change" index search algorithm, "Brightness histogram (horizontal)" index creation and index comparing algorithm, "Key frame extraction" algorithm for index creation on compressed domain
pre-existing technology before project start	initial algorithms
research challenge/innovation	H.264 segmentation, indexing and search on compressed domain
type and amount of processed data	30 hours video from Escom, 30 hours video from BeTV

success criteria, recognition/indexing rate	Scene change detection segmentation algorithm - 90% accuracy
Risk	Some content (encoded with codecs other, than H.264 and MPEG-2) needs full decoding
demo available	not yet

Project Rushes

module/task	**Relevance feedback**
investigator/partner	Queen Mary University London, UK
applied algorithms/approaches	Support vector machines
pre-existing technology before project start	initial algorithm
research challenge/innovation	real life data
type and amount of processed data	about 158 hours news video from Trecvid 2006
success criteria, recognition/indexing rate	above 70% at the last iteration
Risk	amount and quality of data
demo available	yes

module/task	**AV information retrieval**
investigator/partner	Brunel University, UK
applied algorithms/approaches	HMM scheme
pre-existing technology before project start	wavelet implementation for feature selection/ranking
research challenge/innovation	real life data
type and amount of processed data	
success criteria, recognition/indexing rate	it should be more than 95%
Risk	inaccurate for real life data
demo available	yes

module/task	**Video annotation and summarisation**
investigator/partner	Brunel University, UK
applied algorithms/approaches	semantic feature based annotation and frame based summarisation
pre-existing technology before project start	IBM UIMA package
research challenge/innovation	comprehensive search
type and amount of processed data	Internet resources
success criteria, recognition/indexing rate	hopefully > 70%
Risk	diversity of information
demo available	will be available

Project Sapir

module/task	**Video segmentation**
investigator/partner	Eurix
applied algorithms/approaches	The video processing module segments a video into temporal units using different levels of granularity: keyframes, shots and clusters. Shots and clusters represent the first level of decomposition, while keyframes are used at the second level.
pre-existing technology before project start	initial algorithm
research challenge/innovation	Extract data and represent it in MPEG-7. Then use the MPEG-7 for indexing and retrieval.
type and amount of processed data	TBD
success criteria, recognition/indexing rate	TBD

Risk	Efficiency dimension - SAPIR basic (features similarity search) performance can degrade for large volume of content and/or large number of peers, resulting in scalability issues. Effectiveness dimension - Feature search does not improve over text only search, resulting in little gain over existing approaches.
demo available	No

Project Semedia

module/task	**Quick overview of media including sparsely annotated material**
investigator/partner	JRS
applied algorithms/approaches	new approaches for browsing & navigation within huge sparsely annotated material and classificators
pre-existing technology before project start	several low level analysis modules, framework for GUI development
research challenge/innovation	development of algorithms and GUIs for browsing & navigation including e.g. setting detection, finding of retakes and classificators
type and amount of processed data	Substantial subsets from BBC, CCRTV, S&M and flickr test data
success criteria, recognition/indexing rate	Application dependent, varying from high recall to high precision
Risk	minimal
demo available	planned to be integrated into the post-production demonstrator

module/task	**Low level indexing for efficient searches of A/V databases**
investigator/partner	FBM-UPF
applied algorithms/approaches	bag of visual words approach based on: local region detectors: Harris, Hession, MSER; Sift and GLOH descriptors; Aggregation of visual object representations
pre-existing technology before project start	N/A
research challenge/innovation	Combining content-based image retrieval with social media object annotations
type and amount of processed data	Millions of Flickr photos and high-quality video.
success criteria, recognition/indexing rate	Measured in term of recall/precision and accuracy
Risk	Scalability of the approach of Internet size (hundreds of millions of photos and or video)
demo available	Planned to be integrated into the second version of the web-based-communities demonstrator

Module/task	Efficient combination of metadata sources
investigator/partner	JRS
applied algorithms/approaches	development of methods and tools to efficiently combine metadata coming from different sources relating to the same essence; development of methods to ensure metadata consistency and content over the entire production workflow
pre-existing technology before project start	results from a diploma thesis we performed within this area
research challenge/innovation	successfully apply technologies from the semantic web area within multi media description formats such as MPEG-7; development of identity resolution (find out which annotations are the same) and find a upper hierarchy/ontolgy to describe the content neutral and coherent
type and amount of processed data	A substantial sub-set of Flickr photo annotations
success criteria, recognition/indexing rate	Application dependent, varying from high recall to high precision
Risk	minimal
demo available	planned to be integrated into the web-based-communites demonstrator and maybe within post-production demonstrator

Module/task	Data architectures and security in networked media environments
investigator/partner	UPC, DVS
applied algorithms/approaches	Fast metadata extraction from cluster filesystem storages, Caching algorithms
pre-existing technology before project start	DVS Spycer Content Management System, results from UPC caching research
research challenge/innovation	Efficient content management on cluster filesystem storages
type and amount of processed data	Media data from broadcast and postproduction, some TB
success criteria, recognition/indexing rate	Better scalability, higher throughput
Risk	Minimal
demo available	Planned

Module/task	Media mining techniques
investigator/partner	UG
applied algorithms/approaches	affect -based models for mining event patterns in football video data sets
pre-existing technology before project start	
research challenge/innovation	event detection by analysing audio, video ad textual streams
type and amount of processed data	World CUP Football data set
success criteria, recognition/indexing rate	Application dependent, varying from high recall to high precision
Risk	minimal
demo available	Planned

Module/task	Interface design for context-aware adaptive search, browsing and annotation
investigator/partner	FBM-UPF
applied algorithms/approaches	New algorithms for semantic clustering, surrogate formation, layout management, and new approaches for direct interaction, minimalistic design
pre-existing technology before project start	Calm technolgy approach, inteface ecology, information visualization techniques for large information spaces, latent semantic analysis, statistical models of user interaction
research challenge/innovation	Increasing contact with media spaces, designing for a prolonged exploration, building an immersive experience
type and amount of processed data	videos collected from social sites along with their contextual information, news articles and RSS feeds. In total 5000 videos and 10000 articles
success criteria, recognition/indexing rate	Prolonged immersive exploration of information spaces, social

	interaction, intuitive affordances of interaction mechanism

Module/task	Prototypes in Media Postproduction Environments
investigator/partner	All partners
applied algorithms/approaches	Selection of approaches developed above
pre-existing technology before project start	S&M Cakes production management system and DVS Spycer content management system
research challenge/innovation	Use of selected approaches in real-world postproduction content management tools
type and amount of processed data	Dataset from S&M postproduction, probably a few TB
success criteria, recognition/indexing rate	Usefulness of the integrated research approaches, User satisfaction
Risk	Integration problems
demo available	Planned
Risk	Minimal
demo available	series of demos are planned starting from the first of december

Module/task	**Feedback-Only Search**
investigator/partner	FBM-UPF
applied algorithms/approaches	Development of specialized algorithms for feedback-intesive situations. Comparison to standard statistical classifiers.
pre-existing technology before project start	Many standard statistical classifiers (e.g., SVM)
research challenge/innovation	Identification of feedback-intensive situations and performance comparison of statistical classification techniques to retrieval and feedback functions. Analysis of advantages of specialized algorithms compared to standard classifiers.
type and amount of processed data	Substantial subsets from BBC, CCRTV, S&M, and Y!I test data
success criteria, recognition/indexing rate	Achieve a better understanding of techniques for feedback-intensive situations. Success of new algorithms will be measure by high accuracy in classifying.

Module/task	**Prototypes in Broadcast Media Environments**
investigator/partner	All partners
applied algorithms/approaches	Selection of approaches developed above
pre-existing technology before project start	CCRTV-ASI Digition Suite for professional asset management
research challenge/innovation	Use of selected approaches in real-world broadcast content management tools
type and amount of processed data	A sub-set of the CCRTV online and archieve media files, probably a few TB
success criteria, recognition/indexing rate	Usefulness of the integrated research approaches, User satisfaction
Risk	Integration problems
demo available	Planned
Risk	Minimal
demo available	Planned

Module/task	**Integrated retrieval and mining models**
investigator/partner	UG
applied algorithms/approaches	event mining models and new retrieval models
pre-existing technology before project start	event detection algorithms
research challenge/innovation	Integration of retrieval model with mining data set
type and amount of processed data	TREC VID data set, world cup data set
success criteria, recognition/indexing rate	precision, recall
Risk	minimal
demo available	planned

Module/task	**Prototypes for media access, search and retrieval in web-based communities**
investigator/partner	All partners
applied algorithms/approaches	Selection of approaches developed above
pre-existing technology before project start	Yahoo! Web servers
research challenge/innovation	Use of selected approaches in real-world web community environments
type and amount of processed data	Sub-set of media files from Yahoo! Communities, probably a few TB
success criteria, recognition/indexing rate	Positive user feedback
Risk	Integration problems, lack of acceptance by users
demo available	Planned

Project VidiVideo

Module/task	**Visual analysis**
investigator/partner	UvA, CVC
applied algorithms/approaches	Keypoints, color spaces, machine learning, motion pattern analysis
pre-existing technology before project start	Various feature detection methods, SVM based learning of concepts
research challenge/innovation	Complete invariant feature sets, Motion features,
type and amount of processed data	Broadcast TV, >500 hours
success criteria, recognition/indexing rate	Average Precision
Risk	Ambition of 1000 usable detectors too high.
demo available	Yes

Module/task	**Learning**
investigator/partner	Surrey, UvA
applied algorithms/approaches	Machine learning
pre-existing technology before project start	mostly SVM based classifiers
research challenge/innovation	Integrated multi-media features, fusion low-high level semantics, class specific detectors
type and amount of processed data	Broadcast TV, >500 hours
success criteria, recognition/indexing rate	Average Precision
Risk	inbalance in training/testing set
demo available	No

Project Vitalas

module/task	**Rigid local entities retrieval**
investigator/partner	INA, INRIA
applied algorithms/approaches	Low level local features extraction, similarity search structure, tracking and spatio-temporal fusion
pre-existing technology before project start	SIFT like local features, common similarity search structures
research challenge/innovation	More discrimant local features, Large video datasets, spatio-temporal fusion
type and amount of processed data	10000 hours of video (INA)
success criteria, recognition/indexing rate	Currently being defined
risk	Medium
demo available	Not yet

module/task	**Large set of cross-media concepts extraction**
investigator/partner	CERTH-ITI, UoS, CWI
applied algorithms/approaches	Low level features, Hierarchy of classifiers, machine learning
pre-existing technology before project start	Low level features, SVM, mediamill
research challenge/innovation	Cross-media fusion, Large hierarchy of classifiers, several thousands of concepts
type and amount of processed data	1000 hours of video (INA)
success criteria, recognition/indexing rate	Currently being defined
risk	High
demo available	Not yet

Modules on Text Indexing and Retrieval

Project Rushes

module/task	Semantic reasoning
investigator/partner	Queen Mary University London, UK
applied algorithms/approaches	Bayesian networks
pre-existing technology before project start	Initial algorithm
research challenge/innovation	Availability of semantic features
type and amount of processed data	semantic annotation of 10 concepts in 12000 images
success criteria, recognition/indexing rate	Improved accuracy compared with initial annotation
risk	Availability and accuracy of semantic features
demo available	yes

module/task	Text semantic retrieval
investigator/partner	Brunel University, UK
applied algorithms/approaches	biologically driven segmentation/clustering->feature extraction/selection->Support Vector Machine for classification
pre-existing technology before project start	some segmentation implementation, e.g. kernel based.
research challenge/innovation	segmentation and features to be selected for similarity search
type and amount of processed data	on-line documents
success criteria, recognition/indexing rate	> 90% (possibly)
risk	unknown
demo available	yes

Project Sapir

module/task	Text
investigator/partner	Xerox
applied algorithms/approaches	Four kinds of information will be generated by text processing: 1. word-level indexing information: information about word occurrences in the text; used for keyword searching 2. named entity information: annotation of names of people, places, dates; used for searches with semantic constraints 3. extracted facts: structured information induced from text; used for searches with semantic constraints 4. summary: a selection of important sentences that allows a user to determine quickly whether a ContentObject is relevant
pre-existing technology before project start	Some text analytics tool from Xerox
research challenge/innovation	Write a UIMA Annotators that extract the features and represent them in MPEG-7. Index and search using a P2P architecture
type and amount of processed data	TBD
success criteria, recognition/indexing rate	TBD
risk	Same as for Speech
demo available	Planned

Project Tripod

module/task	**Caption augmentation for images with existing captions**
investigator/partner	Tripod partners
applied algorithms/approaches	Expanding captions with words from Web pages and from map data
pre-existing technology before project start	Very little currently being done
research challenge/innovation	Making the approach work well
type and amount of processed data	Thousands of images
success criteria, recognition/indexing rate	Acceptance of image captions by photolibraries
risk	Medium
demo available	Not yet

Project Vitalas

module/task	**Text search module**
investigator/partner	EADS, CWI
applied algorithms/approaches	Vectorial approaches
pre-existing technology before project start	TF/IDF, Inverted lists
research challenge/innovation	Large scale
type and amount of processed data	Annotations (manually and automatically generated) of audio-visual and photo angency archives (10000 hours of video, 3 millions images)
success criteria, recognition/indexing rate	Currently being defined
risk	Low
demo available	Not yet

module/task	**Word sense disambiguation**
investigator/partner	University of Sunderland
applied algorithms/approaches	Statistic methods
pre-existing technology before project start	EuroWordNet
research challenge/innovation	Large scale
type and amount of processed data	Annotations (manually and automatically generated) of audio-visual and photo angency archives (10000 hours of video, 3 millions images)
success criteria, recognition/indexing rate	Currently being defined
risk	Low
demo available	Not yet

Annex B: Overview of the national research projects

Project	Quaero
Budget	• €100m for >5 years and more than 20 partners • Granted by French 'Agence de L'innovation Industrielle' • State aid to be authorised by DG Competition of European Commission
Duration	>5 years
Country	France with the participation of German partners
Partners	***Private companies*** : Thomson, France Telecom, Jouve, Exalead, Bertin Technologies, LTU Technologies, Vecsys, Synapse Development ***Public research labs*** : LIMSI-CNRS, RWTH-Aachen, Karlsruhe University, INRIA, LIG-UJF, IRCAM, ENST-GET, IRIT, INIST-CNRS, MIG-INRA, LIPN ***Public institutions*** : INA, BNF, LNE, DGA Some contacts have been established with other European potential participants
Main Objectives and challenges	Develop demonstrators or applications corresponding to identified use cases in the domain of access and manipulation of multimedia and multilingual content • Search, navigate, distribute, produce Develop the corresponding enabling technologies for multilingual and multimodal content processing
Main applications and use cases	1. Consumer Multimedia Search Engine 2. Multimedia Search Services to enrich European portals 3. Personalised Video on interactive consumer networked devices Anytime and Anywhere 4. Recondition the Audiovisual Cultural Heritage 5. Professional Digital Media Asset Management for Broadcasting Industry 6. Platform for Text and Image Annotation
Research and Technologies	• Search and extraction infrastructure • Content processing infrastructure • Document capture and processing • Speech recognition • Translation • Musical analysis • Object recognition in images and video • Face detection and recognition • Video segmentation and structure analysis • Object tracking and event recognition in videos • Man machine interaction • Security
Benchmarking of project results	Evaluation is the founding principle of Quaero's technological research and development organisation. Evaluation will be used as a tool for facilitating and structuring technology transfer between research organisations and leaders of use cases. Periodic evaluation campaigns shall be conducted within the program to assess global progress in each of the technology areas addressed in the program. These evaluation campaign shall be build on the most advanced procedures developed and organized by national or international bodies and programs such as NIST, CLEF, Technolangue, Technovision…

Project	Theseus
	http://www.bmwi.de/BMWi/Navigation/Technologie-und-Innovation/Informationsgesellschaft/multimedia,did=184810.html http://theseus-programm.de
Budget	Overall volume: 200 Mio. Euro (Funding: 90 Mio. Euro)
Duration	5 years
Country	Germany
Partners	***Industry:*** Empolis/Bertelsmann (co-ordinator), SAP, Siemens, Deutsche Thomson, Lycos, Morsophy, m2any, Intelligent Views, Ontoprise ***Research and public organisations:*** Fraunhofer Gesellschaft zur Förderung der angewandten Forschung (FhG), Institut für Rundfunktechnik (IRT), Deutsche Nationalbibliothek (DNB), Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Forschungszentrum Informatik (FZI), VDMA-Verband, Gesellschaft für Forschung und Innovation (VFI), universities (Karlsruhe, München, Darmstadt, Dresden, Konstanz, Erlangen)
Main Objectives and challenges	The main objective is to generate innovation in the area of semantic technologies to strengthen the role of the German IT industry and to establish new services in this area. The technologies are mainly for new internet based applications and services.

Main applications and use cases	There are several applications foreseen. They are realized in sub projects (calls "use cases"): • Alexandria: semantic internet platform to process and organize user generated content, semantic internet search platform • Contentus: Processing of cultural audio visual content of the German National Library • Medico: semantic image technology for Clinical Decision Support and Computer Aided Diagnosis. • ORDO: automatic semantic processing of huge text and audio visual corpora, semantic search tools • Processus: development of knowledge intensive tools to optimize generic production workflow • Texo: semantic based interconnection between service provider and service users
Research and Technologies	• Image and video processing • 3D analysis • Ontology • User interaction and semantic modelling • Machine learning • Digital rights management
Benchmarking of project results	In the Core Technology part of the project one work package is dealing with benchmarking of the other technology and research work. For the benchmarking the Fraunhofer IDMT is responsible

Project	**iAD – information access disruptions**
Budget	Ca. €30m
Duration	8 years, start in 2007
Country	Norway
Partners	• Fast Search & Transfer (Host) • Accenture • Schibsted • Cornell University • AIC Dublin (DCU, UCD) • NTNU Trondheim • University of Tromsø • University of Oslo • Norwegian School of Management
Main Objectives and challenges	• Core research for next generation precision, analytics and scale in information access • Build international networks to identify and execute on global disruption opportunities enabled by emerging services in the information age
Main applications and use cases	
Research and Technologies	**Schema agnostic indexing services** • Schema-agnostic end2end design • Consolidation of query model **Processing high-speed data streams** • Capturing & extracting knowledge from data streams: • Pervasive sensor networks, RFID readers, multimedia feeds, … **Scalable infrastructure for push and pull based computing** • Robust principles and services for next generation infrastructure for distributed information access **Extreme precision and recommendation in multimedia access** • Extreme precision solutions for access to multimedia content • Social networks with recommender functions **Understanding and managing the disruptive potential of iAD** • Analyze business and societal impact • Assess disruptive potential
Benchmarking of project results	

Project	MultimediaN http://www.multimedian.nl/en/multimedian.php
Budget	30 MEuro
Duration	Phase 1: 2002 – 2004 Phase 2: 2004 – 2009
Country	Netherland
Partners	• Center for Math and Computer Science • Philips Research • Technical University Delft • Telematica Institute • TNO • University of Amsterdam • University of Twente + 39 affiliated business partners
Main Objectives and challenges	MultimediaN is a public-private partnership focusing on science and technology of multimedia interaction & search engines. MultimediaN contributes to the solution of four fundamental problems: 1. The accessibility of much multimedia content is low. 2. The information is fragmented: sound can't be matched to text, text can't be matched to speech. 3. A lot of information contributes to the 'information overload' that is characteristic of today's society. 4. Multimedia information is often badly organized as a result of legacy systems, self-created standards and heterogeneity in terminologies.
Main applications and use cases	MultimediaN is divided in fundamental, integration, and application projects. The fundamental projects (Learning Features, Multimodal Interaction, and Ambient Multimedia Databases) create knowledge that is new on a world level. The integration projects (Semantic Multimedia Access, Professional Dashboard, and Video At Your Fingers) develop knowledge in which existing video-, audio- and speech technology are combined. The application projects (E-Culture and Personal Information Services) are pilots, which create application knowledge in an application context. • Learning Features • Multimodal Interaction • Ambient Multimedia Databases • Semantic Multimedia Access • Professional's Dashboard • Video At Your Fingertips • E-Culture (N9C) • PERsonal Information Services
Research and Technologies	MultimediaN covers the following research topics: • Image, picture, video processing and indexing • Audio and speech recognition and indexing • Textual processing • Knowledge modelling, mining • System engineering (databases, standards)
Benchmarking of project results	The modules are evaluated in several international benchmarking initiatives. For video indexing a special track of TRECVidio was established in which data from MultimediaN was used for evaluation.

Project	**Interactive Multimodal Information Management (IM2)**
Budget	Phase 1: • SNFS funding: 15’349’000.- CHF • Self & third-party funding: 19’655’000.- CHF Phase 2: • NSF funding: 14’000’000 • Self & third-party funding: 14’000’000.- CHF
Duration	3 x 4 years (4 phases), project start: January 2002
Country	Switzerland
Partners	IDIAP Research Institute, Martigny (co-ordinator) Partners: EPFL, Univ. Geneva, Univ. Fribourg, ETHZ, and Univ. Bern
Main Objectives and challenges	IM2 has the objective to develop advanced methods for indexing multimedia content and to provide advanced multimodal human computer interfaces. Therefore investigations in the area of human-human communication are carried out.
Main applications	The application scenario so far is the indexing and modelling of face-to-face meetings.

and use cases	
Research and Technologies	IM2 covers the following research areas: • Unconstrained speech recognition • Language understanding • Computer vision • Machine learning • Multimodal scene analysis • Model of individual and group dynamics • Sociology and social-psychology • Structure, index, summarize communication scenes • User interfaces
Benchmarking of project results	Each of the following technology module is evaluated in international benchmark initiatives (NIST, DARPA, …): • ASR: Automatic speech recognition • KWS: keyword spotting • SEG: speaker segmentation • ID/LOC: identification and • localization/tracking • FOA: focus of attention • GAA: gesture and action recognition IM2 provides a huge corpus with recorded meetings for internal and external evaluation and benchmarking. IDIAP has shown the good performance of their computer vision technology in the ImageCLEF 2007 evaluation for the medical annotation task.

4. SOA OF EXISTING BENCHMARKING INITIATIVES + WHO IS PARTICIPATING IN WHAT (EU&NI)

4.1. Introduction and WG2 objectives:

When addressing audio-visual search engine challenges, we have identified benchmarking and evaluation issues as a critical topic.

Despite the availability of many effective multimedia retrieval methodologies created by the research community, few commercial products currently incorporate such techniques. It is not obvious which technique is the best for a given problem. It is clear, however, that to cope with the rapid growth in the production of and access to digital multimedia content, evaluation campaign will help to facilitate the wider use and the dissemination of multimedia retrieval research results.

Benchmarking efforts are usually intended to be precise and measure carefully how systems or algorithms perform with respect to a dataset, a task and an evaluation metric. Thus, to be scientifically valid, they have to be specific such that results are unambiguous and measurable. This makes benchmarks necessarily very narrow in focus and they often exclude much research. The goal is to find research questions that are of general interest, where a number of researchers are working on pretty much the same goal, and then evaluate this work. In this context, the benchmarking will format all the research work in the community letting people working on the same tasks and necessarily limit the innovation. During ACM Multimedia Information Retrieval, a panel was organized on "Diversity in Multimedia Information Retrieval Research" were the question: "Does benchmarking kills innovation?" was discussed[5]. The panel paper[6] and slides are also available on this web page.

Besides TRECVID, benchmarking initiatives are becoming numerous: ImageCLEF, Pascal, ImagEval, ... This definitely shows that no existing single initiative could be by itself satisfactory by offering the context to test all the tasks addressed by our community. Also, due to the richness of the scientific objectives of multimedia search engines corresponding to growing and evolutionary use-cases and user needs, benchmark initiatives should be able to follow the field dynamic.
Nevertheless, benchmarking remain necessary and valuable for the community as it provides objective reference among the numerous technical academic and industrial solutions. But it should be carefully set with clear and fair rules, and wide consensus of the community regarding definition of tasks, evaluation parameters, performance measures, ground truth setting, conflict of interest avoiding, ... Joining all these conditions remain quite challenging and the bottleneck of some initiatives.

In the ideal conditions, we believe that winning a benchmark is worth thousand publications for academia and thousand press releases for industrials and represent a "moment of truth" among all what technology providers (academia/industrial) can argue on their work and results.

[5] http://riemann.ist.psu.edu/mir2006/index.html

[6] *JAMES Z. WANG, NOZHA BOUJEMAA, ALBERTO DEL BIMBO, DONLAD GEMAN, ALEXANDER G. HAUPTMANN AND JELENA TESIC. (2006). Panel: Diversity in Multimedia Information Retrieval Research. In Proceedings of the 8th ACM international workshop on Multimedia information retrieval. 5-12.*

Chorus activities in this topic are conducted within the WG2. The related web site is: http://www.ist-chorus.org/wg2---evaluation-benchmarking-an.php
Having a common understanding of evaluating multimedia retrieval systems would allow technology users and companies to orient themselves to select the retrieval technique most suitable for their own specific needs. The problem is that current evaluation initiatives are disparate and run independently of each other, and there is a lack of coordination of these initiatives.

Among our objectives within Chorus is to address this topic by putting together these dispersed initiatives on the benchmarking and evaluation of multimedia information retrieval to establish a clear understanding of the current situation and determine how best to move forward in a unified and cooperative way. During our first year effort, we have organized two events allowing experience sharing among benchmarking communities around the existing initiatives during Chorus Rocquencourt workshop[7] and also during CBMI'07[8] panel. The programs of these two events as well as links to the presentations are provided in the annex.

By bringing together organizers of existing multimedia evaluations in the Chorus events, we allow sharing experiences and plan for the next period of Chorus to put forward best practices to improve the existing evaluation initiatives. The following initiatives have participated to Chorus events: TRECVID, INEX, ImageCLEF, CLEF, ImageEval, SHREC, MIREX, ELDA, Robin, and the Pascal Challenges. Most of the existing evaluation campaigns workflow is typically similar (registration of participants, distribution of data, submission of results, creation of ground truth, evaluation, dissemination of results during workshop/conference. The communalities could be analyzed to identify how existing evaluations efforts could be mutualized such as databases collections maintenance and ground truth generation. Also, another benefit would be to avoid that the scientific community is requested several times for participation to different campaigns where some tasks are very close even using different data collections.

The evaluation method developed by the TREC (TExt Retrieval) conference is considered the standard methodology for large scale evaluation of information retrieval systems [VOORHEES,1998]. Most of the benchmark initiatives described in this document are to some extent based on this model. Subsequently, an evaluation typically consists of the following phases:

- **Establishing a common dataset** In order to prevent biases all participants work on exactly the same dataset. In general, a difference is made between the *training set*, which is used by the participants to train their systems and the *test set***,** which is similar to the training set and used for the final evaluation. This difference is necessary for systems to prevent bias towards the training set. Furthermore, a dataset is typically selected for a particular task or track. This may include generating or tailoring the dataset to a specific task, but most of often the dataset is an excerpt from real life data, which is representative for the problem domain.

- **Definition of the task to be performed** All participants perform exactly the same task, of which the results are evaluated and compared. Typically a task reflects a real life need within a particular domain. In general organizers of benchmark initiatives try to find a balance between relatively easy tasks they know is supported by state of the art technology and challenging tasks that are not yet covered.

[7] http://www.ist-chorus.org/chorus-wg2.php

[8] http://www.ist-chorus.org/bordeaux---june-25-07.php

- **Establishing of the ground truth** In order to evaluate results provided by participants for a particular task, the correct response or ground-truth, should be known. Although this may appear trivial, establishing the ground-truth can be a rather complicated task because of the quantity and complexity of the dataset or the ambiguous nature of the response. Often the ground-truth is established manually by domain experts, which is typically a rather labor intensive task.

- **Assessing results relative to the ground truth.** The results submitted by a participant for a particular task are evaluated relative to the ground truth. Typically this is an automated process that produces a metric, which allows comparison with other participants. Sometimes, however, submitted results are judged (and cross-validated) by human experts. Although the objective of benchmarking is to establish a quality metric for technology within a particular domain, most initiatives emphasize the benchmark as a platform of discussion rather then a competition.

The objective of this document is to raise awareness between researchers on the availability of the different benchmarking initiatives and to make available description of their activities and properties.

4.2. Overview of existing benchmark initiatives

In order to map the landscape of currently active benchmark initiatives CHORUS organized a workshop (14-3-2007, INRIA, Rocquencourt) for which it invited representatives of the major multimedia benchmarks, who all gave an presentation about their respective benchmark initiative. This initial meeting was followed up by a panel discussion (26-6-2007, CBMI, Bordeaux) on benchmark initiatives. Based on the initial workshop and panel discussion we established 5 dimensions that we use to compare the benchmarks initiatives:

- **Definition of tracks and tasks** denotes a short description of the tracks and task a participant can compete in. A track refers to the "theme" of comparison, such as copy-detection for video or artist identification for music, whereas a task refers to a particular assignment the participant has to complete. Although most initiatives cover multiple tracks, a participant does not necessarily need to compete in all of them.

- **Evaluation metrics of task** denotes the metric that was used to evaluate a task. The standard metrics used in information retrieval include Mean Average Precision (MAP), Binary Preference (Bpref), Mean Reciprocal Rank (MRR) and Geometric Mean Average Precision (GMAP). However, some tasks are unsuited for evaluation using these measures. In this case, we indicate the evaluation metric for the specific initiative.

- **Type and size of the data used** describes the quantity and quality of the data is used for the benchmark initiative.

- **Method used for generating the ground-truth** describes the method used to obtain the ground-truth, which is used as a measurement to evaluate the submitted results of the participants.

- **Participation statistics** denotes for the last three years (2007, 2006, and 2005) the number of registrations of intended participation and the number of registered participants that submitted results.

In addition we represent in the overview the:

- **URL,** which denotes the address of the initiatives website.
- **Conclusion,** which denotes a partial conclusion from the perspective of the initiative.

Find below the overview of the benchmark initiatives we address:

<table>
<tr><td colspan="3">TrecVid</td></tr>
<tr><td colspan="3">The TREC conference series is sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies. The goal of the conference series is to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. In 2001 and 2002 the TREC series sponsored a video "track" devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. Beginning in 2003, this track became an independent evaluation (TRECVID) with a 2-day workshop taking place just before TREC.
TRECVID is coordinated by Alan Smeaton (Dublin City University) and Wessel Kraaij (TNO Information and Communication Technology). Paul Over and Tzveta Ianeva provide support at NIST.</td></tr>
<tr><td>Definition of tracks and tasks (2007)</td><td colspan="2">• Shot detection
• Semantic Concept Features
• Automatically create MPEG-1 summary
Maximum duration to be determined
Shows the main objects (animate and inanimate) and events from “rushes”
Evaluated using simple play and pause controls
Need not be series of frames directly from the video
Summaries can contain picture-in-picture, split screens
• Search
• Interactive
• Manual
• Automatic</td></tr>
<tr><td>Evaluation metrics of tasks</td><td colspan="2">Mean Average Precision</td></tr>
<tr><td rowspan="2">Type and size of data used</td><td>2005</td><td>Data unavailable</td></tr>
<tr><td>2006</td><td>158 hours Arabic, Chinese, English Broadcast News
Common speech recognition, translation, annotations</td></tr>
</table>

	2007	100 hours Dutch TV shows, Common speech recognition, translation; several groups provided low-level features and (unverified) semantic concept detection results 100 hours BBC "Rushes" - raw stock footage, natural sound, highly repetitive, ong segments, reusable shots of people, objects, events, locations, etc.
Method used for the generation of the ground-truth	• Researchers submit experiment results on test collections • Cut off at certain rank for each system • Top results from systems are pooled (redundancy removed) • Pooled results manually judged for relevance • All systems submission results are scored for all results • With manual truth assumed to be 'complete' • Works well with good variety of system approaches • Cost effective and scalable	
Participation statistics	2005	Registrations: 63 Search submissions: 42 Search runs submitted: 112
	2006	Registrations: 70 Search submissions: 54 Search runs submitted: 123
	2007	Registrations: 71 Search submissions: N.A. Search runs submitted: N.A.
URL	http://www-nlpir.nist.gov/projects/trecvid/	
Conclusion from TrecVid perspective	• Standardized evaluations and comparisons – improve science • Weed out many hypotheses from small, idiosyncratic data • Test on common large collection and some common metadata • Failures are not embarrassing and can be presented at the TRECVID workshops! • Virtually all work is done on one extracted keyframe per shot • Anyone can participate • Sign promise to use the data for research only	

ImageClef

ImageCLEF is the cross-language image retrieval track which is run as part of the Cross Language Evaluation Forum (CLEF) campaign. The ImageCLEF retrieval benchmark was established in 2003 with the aim of evaluating image retrieval from multilingual document collections. Images by their very nature are language independent, but often they are accompanied by texts semantically related to the image (e.g. textual captions or metadata). Images can then be retrieved using primitive features based on pixels with form the contents of an image (e.g. using a visual exemplar), abstracted features expressed through text or a combination of both. The language used to express the associated texts or textual queries should not affect retrieval, i.e. an image with a caption written in English should be searchable in languages other than English.

Besides textual and multimodal tasks, ImageCLEF offers two purely visual tasks for image classification or object detection/retrieval.

Note: A pre-conference workshop[9] was organized together with the MUSCLE network of excellence the day before the workshop for the past three years with high-quality keynote speakers on visual information retrieval evaluation and related topics.

Definition of tracks and tasks (2007)	• Ad-hoc retrieval with query in different language from the annotation or multilingual image annotations (2003-2007) • Object classification/retrieval task; purely visual (2006-2007) • Medical image retrieval task (2004-2007) • Medical image classification task; purely visual (2005-2007) • Interactive image retrieval (2004-2006) • Geographic retrieval from image collections (2006)	
Evaluation metrics of tasks	• Mean Average Precision as a lead measure • BPref, P(10-50) used for comparison • Many ideas on how to find better measures No resources to pursue this	
Type and size of data used	• IRMA collection for medical image classification (11'000 images) • ImageCLEFphoto collection (IAPR TC 12) (20'000 images) • ImageCLEFmed collection (~70'000 images) • Varying degree off annotations and languages • Realistic collections for this specific task (containing image of varying quality, majority of English annotations, domain-specific vocabularies and abbreviations, spelling errors, …)	
Method used for the generation of the ground-truth	• Classification Collections used were classified beforehand • Retrieval Pooling is used with varying number depending on submissions Judgment scheme: relevant – partially – non-relevant Double judgments to analyze ambiguity • Interactive Participants evaluate themselves (time, Nrel)	
Participation statistics	2005	36 registrations 24 submissions 300 runs

[9] http://muscle.prip.tuwien.ac.at/ws_overview_2007.php

	2006	47 registrations 30 submissions 300 runs
	2007	51 registrations 38 submissions >1'000 runs
URL	http://www.imageclef.org/	
Conclusion	• ImageCLEF creates important resources and is acknowledged in the field (50 registrations) • Discussions at workshop are regarded as very stimulating • Lack of participation for interactive retrieval • Lack off funding is a major problem to professionalize it and analyze all data • Resource sharing could really help!	

ImageEval

In 2005, the Steering Committee of ImagEVAL had the opportunity of proposing evaluation campaigns for funding by the French "Techno-Vision" program. The ImagEVAL project relates to the evaluation of technologies of image filtering, content-based image retrieval (CBIR) and automatic description of images in large-scale image databases

The objective of ImagEVAL is double:

- to organize important evaluations starting from concrete needs and using professional data collections
- to evaluate technologies held by national and foreign research laboratories, and software solutions

Definition of tracks and tasks (2007)	• Transformed image recognition • Combined text/image strategies for image retrieval • Text area detection • Object detection (e.g. Car, tree, …) • Extraction of attributes (e.g. indoor/outdoor, day/night, natural/urban, …)
Evaluation metrics of tasks	• MAP : Mean Average Precision (main metric) and complementary Precision/Recall based metrics • Mean Reciprocal Rank (for a sub-task of the transformed image recognition) • Christian Wolf's metric (for the text area detection): this metric (implemented in DetEVAL tools) is mainly based on the metrics used in ICDAR evaluation, nevertheless it enables a clever evaluation of the classical over and low segmentation problem that appear when dealing with bounding boxes for both results and ground truths.

Type and size of data used	• Old postcards (~7600 images) • Black & white, color photographs (~50 000 images) • Transformed image recognition : 42 500 images • Combined text/image strategies for image retrieval : 700 web pages • Text area detection : 500 images • Object detection (e.g. Car, tree, …) : 14 000 images • Extraction of attributes (e.g. indoor/outdoor, day/night, natural/urban, …) : 23 500 images	
Method used for the generation of the ground-truth	Ground truth files build by two professionals that annotated each image.	
Participation statistics	2005	*Data unavailable*
	2006	20 registrations 11 submissions
	2007	*Data unavailable*
URL	http://www.imageval.org/	
Conclusion	• Very interesting and challenging data provided by professionals that actively participated to the creation of the campaign • ImagEVAL is a part of the solution answering the lack of evaluation in the computer vision community • Correct participation level for a first edition but need to attract more international labs and companies • We need to collaborate with other evaluation campaigns, share experiences and elaborate a coherent planning to avoid overlapping • Define more focused evaluation problems according to end users feedbacks and potential overlapping with other evaluation campaigns	

TechnoVision-ROBIN	
Technovision is a recent program of the French Ministry of Research and Technology that will fund evaluation projects in the area of computer vision. Many vision algorithms have been proposed in the past, but comparing their performance has been difficult owing to the lack of common datasets. Technovision aims to correct this by funding the creation of large, representative image datasets. ROBIN is a Technovision proposal covering the evaluation of object retrieval algorithms	
Definition of tracks and tasks (2007)	• multi-class objects detection • generic objects detection • generic objects recognition • image categorization

Evaluation metrics of tasks	• Detection Recall for maximal Precision: R Precision for maximal Recall: P Equal Precision and Recall: EER Area under the curve: AUC • Discrimination Discrimination at minimal uncertainty rate: D Uncertainty at maximal discrimination rate: U Equal discrimination and uncertainty rate: EDU Confusion matrix at maximal uncertainty: (c, c) • Rejection Equal Rejection Rate: ERR
Type and size of data used	• 6000 images from a static camera and images from a moving vehicle. • Satellite images containing 10000 regions of interest (128x128 pixels) • 6400 Aerial images and 1000 short videos containing vehicles and infrastructure elements • 10000 aerial images with computer synthesized objects • 15000 computer generated images • 1500 multi-sensor aerial images
Method used for the generation of the ground-truth	Manual annotation
Participation statistics	*Data unavailable*
URL	http://robin.inrialpes.fr/
Conclusion	First round is still running, no conclusion available yet

IAPR TC-12 Image Benchmark

IAPR TC-12 Benchmark consists of 20,000 images (plus 20,000 corresponding thumbnails) taken from locations around the world and comprising an assorted cross-section of still natural images, providing the resources to carry out evaluation of visual information retrieval from generic photographic collections (i.e. containing everyday real-world photographs akin to those that can frequently be found in private photographic collections as well). Each photograph is thereby associated with a semi-structured text caption in three languages: English, German and Spanish.

Definition of tracks and tasks (2007)	The IAPR TC-12 Image Benchmark has not been used in a standalone evaluation event yet, but provided the resources for the following tasks: • ImageCLEFphoto (2006-2007): ad-hoc retrieval (with the query language either being identical or different from that used to describe the images) • ImageCLEF object classification/retrieval task (2007), purely visual • MUSCLE Live Retrieval Evaluation Event (2007)
Evaluation metrics of tasks	• MAP as a lead performance measure • bpref, GMAP, P(20) as additional performance indicators

Type and size of data used	• 20,000 still natural photographs of generic content (e.g. people, animals, cities, landscapes) • Detailed semi-structured captions in up to three languages (English, German, Spanish) • 60 query topics in TREC format (topic titles, narratives, and sample images)	
Method used for the generation of the ground-truth	• ImageCLEFphoto, Live Event: Pooling is used with varying number depending on submissions Judgment scheme: relevant – partially – non-relevant Double judgments to analyze ambiguity Interactive Search and Judge to complete pools with further relevant images • ImageCLEF object classification, Live Event: Collections used were classified beforehand	
Participation statistics	2005	*Data unavailable*
	2006	Registrations: 36 (ImageCLEFphoto) Submissions: 12 (ImageCLEFphoto) Runs: 157 (ImageCLEFphoto)
	2007	Registrations: 32 (ImageCLEFphoto), 22 (ImageCLEF object retrieval), 3 (Live Event) Submissions: 21 (ImageCLEFphoto), 7 (ImageCLEF object retrieval), 3 (Live Event) Runs: 616 (ImageCLEFphoto), 38 (ImageCLEF object retrieval), 3 (Live Event)
URL	http://eureka.vu.edu.au/~Egrubinger/IAPR/TC12_Benchmark.html	
Conclusion	• New query topics will be created for 2008 • Evaluation events that will use the IAPR TC-12 Benchmark include: • ImageCLEF 2008 (ad-hoc retrieval task and object annotation task) • GeoCLEF 2008 • MUSCLE Live Retrieval Evaluation Event 2008	

CIVR Evaluation Showcase

Image and video storage and retrieval continue to be one of the most exciting and fastest-growing research areas in the field of multimedia technology. However, opportunities for the exchange of ideas between different groups of researchers, and between researchers and potential users of image/video retrieval systems, are still limited. The International Conference on Image and Video Retrieval (CIVR) series of conferences was originally set up to illuminate the state of the art in image and video retrieval between researchers and practitioners throughout the world. This conference aims to provide an international forum for the discussion of challenges in the fields of image and video retrieval.

Video and image retrieval systems find their way to regular conference demo sessions, but they are never exposed and run simultaneously. The CIVR Evaluation Showcase event aims to fill this lacuna. Specifically, we aim for a showcase that goes beyond the regular demo session: it should be fun to do for the participants and fun to watch for the conference audience. To reach this goal, a number of participants simultaneously do an interactive search task during the showcase event. At the CIVR 2007, three live evaluation events were held for the first time.

Definition of tracks and tasks (2007)	• Video Retrieval (VideOlympics) textual search (e.g "Find shots of a meeting with a large table.") • Image Retrieval text queries (e.g "Find images of snowy mountains"). Visual queries (e.g. "Where is the church shown in the example image?") • Copy Detection Find real copies of entire long videos (from 1 minute to 3 hours). Find copies of clips that are transformed (e.g Copies are transformed by cropping; fade cuts; flips; insertion of logos etc.
Evaluation metrics of tasks	• Video Retrieval (VideOlympics) Precision, recall, speed, best system voted by conference attendees, etc. • Image Retrieval For the visual queries, the amount of time taken for the first correct answer to be found was recorded. For the text queries, the ratio of correct to incorrect images within the first *N* images returned was calculated. The value of *N* was based on the number of correct images for each query in the ground truth. • Video Copy Detection Quality metric based on number of correct answers returned. Speed metric. Note that the evaluation results will not be published, the emphasis is on demonstrating the capabilities of the technology for a well-defined task that interests many people.
Type and size of data used	• Video Retrieval (VideOlympics) TRECVid 2006 test data (160 hrs of Arabic, Chinese, and US broadcast news). • Image Retrieval Extended IAPR TC12 dataset (21000 images) • Video Copy Detection Newly created dataset containing web video clips, TV archives and movies(~100 hours of video)

Method used for the generation of the ground-truth	• Video Retrieval (VideOlympics) Manual relevance judgements. • Image Retrieval Manual relevance judgements. • Video Copy Detection Videos from which modified versions are generated are known.	
Participation statistics	2005	*Data unavailable*
	2006	*Data unavailable*
	2007	9 participants (VideOlympics) 3 participants (Image Retrieval) 10 participants (Video Copy Detection)
URL	http://www.civr2007.com/showcase.php	
Conclusion	• Live retrieval evaluation includes • Effect of the user interface. • Speed / efficiency of retrieval of the system. • Skill of the user • Currently no metrics exists to measure this.	

SHREC (3D)

The Network of Excellence AIM@SHAPE is taking the initiative to organize a 3D shape retrieval evaluation event: SHREC - 3D Shape Retrieval Contest. The general objective is to evaluate the effectiveness of 3D-shape retrieval algorithms. The contest is organized in conjunction with the SMI conference (Shape Modeling International) where the evaluation results will be presented.

Definition of tracks and tasks (2007)	• Watertight models (object models represented by seamless surfaces) • Partial matching • protein models • CAD models • Relevance feedback • Similarity measures • 3D faces
Evaluation metrics of tasks	• Relevance measure (highly relevant, marginally relevant) • Precision, Recall • First, Second Tier • (Normalized) (Discounted) Cumulated Gain • Average Dynamic Recall
Type and size of data used	Princeton Shape Benchmark (1814 classified polygonal models)
Method used for the generation of the ground-truth	Manually established
Participation statistics	*Data unavailable*
URL	http://www.aimatshape.net/event/SHREC

Conclusion	3D media have specific properties/requirements, which justifies a 3D benchmarking initiative. However, the conceptual framework is similar to other benchmarks initiatives, suggesting closer cooperation can be beneficial.

MIREX

The Music Information Retrieval Evaluation eXchange (MIREX) is a community-based formal evaluation framework coordinated and managed by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the University of Illinois at Urbana-Champaign (UIUC). IMIRSEL has been funded by both the National Science Foundation and the Andrew W. Mellon Foundation to create the necessary infrastructure for the scientific evaluation of the many different techniques being employed by researchers interested in the domains of Music Information Retrieval (MIR) and Music Digital Libraries (MDL).

For the past two years MIREX participants have met under the auspices of the International Conferences on Music Information Retrieval (ISMIR). The first MIREX plenary convened 14 September 2005 in London, UK, as part of ISMIR 2005. The second plenary of MIREX 2006 was convened in Victoria, BC on 12 October 2006 as part of ISMIR 2006. Some of the tasks, such as "Audio Onset Detection," represent micro level MIR/MDL research (i.e., accurately locating the beginning of music events in audio files, necessary for indexing). Others, such as "Symbolic Melodic Similarity," represent macro level MIR/MDL research (i.e., retrieving music based upon patterns of similarity between queries and pieces within the collections).

Definition of tracks and tasks (2007)	• Audio Artist Identification • Audio Classical Composer Identification • Audio Artist Identification subtask • Audio Genre Classification • Audio Music Mood Classification • Audio Music Similarity and Retrieval • Audio Onset Detection • Audio Cover Song Identification • Real-time Audio to Score Alignment (a.k.a Score Following) • (Postponed to possibly 2008) • Query by Singing/Humming • Multiple Fundamental Frequency Estimation & Tracking • Symbolic Melodic Similarity	
Evaluation metrics of tasks	• Human listening tests on similarity denoted on a broad scale (3 classes) and a fine scale (10 classes). • Objective statistics based on meta-data	
Type and size of data used	5000 music files, 9 genres	
Method used for the generation of the ground-truth	Evaluated by human judgments	
Participation statistics	2005	41 submissions 72 runs
	2006	46 submissions 92 runs
	2007	*Data unavailable*
URL	http://www.music-ir.org/mirex2007	

Conclusion	• Challenges for Music Retrieval Benchmarking Data and access to it sufficient size real-world sufficient quality • Metadata high-quality labels (production-style) ground truth annotation • Evaluation automatic vs. human evaluation

INEX

The aim of the Initiative for the Evaluation of XML Retrieval (INEX), launched in 2002, is establish an infrastructure and provide means, in the form of a large XML test collection and appropriate evaluation metrics, for the evaluation of content-oriented XML retrieval systems. INEX has a strong international character; participants from over 80 organisations, distributed across Europe, America, Australia, Asia, and Middle-East have so far contributed to INEX. The main INEX Ad Hoc task focuses on text-based retrieval of XML fragments. The INEX Multimedia track is concerned with other types of media that can also be found in XML collections. Existing research on multimedia information retrieval has already shown that it is far from trivial to determine the combined relevance of a document that contains several multimedia ob jects. The objective of the INEX MM track is to exploit the XML structure that provides a logical level at which multimedia ob jects are connected, to improve the retrieval performance of an XML-driven multimedia information retrieval system. INEX MM ran a pilot evaluation study in 2005 and has been established as an INEX track in 2006 and 2007.

Definition of tracks and tasks (2007)	**MMfragments task:** The objective of this retrieval task is to find relevant multimedia XML fragments (i.e., XML elements or passages that contain at least one image) given a multimedia information need, which may contain visual or structural hints. Within the MMfragments task, there are three subtasks: • Focused: return a ranked list of elements or passages to the user. • Relevant In Context: return relevant elements or passages clustered per article to the user. • Best In Context:return articles with one best entry point to the user. **MMimages task**: The objective of this retrieval task is to find relevant images given a multimedia information need, that may contain visual hints. The requirement is to to return a ranked list of documents (=image + metadata) from this collection. In this task, the type of the target element is defined, so it is basically closer to an image (or a document) retrieval task, rather than XML element or passage retrieval.
Evaluation metrics of tasks	**MMfragments task**: Since the relevance assessments are performed at the sub-document level, systems are compared using effort-precision/gain-recall graphs, the eXtended Cumulated Gain (XCG) metrics used in many INEX tasks. The summary statistic of these, i.e., mean average effort precision, is also reported. **MMimages task**: mean average precision and recall precision graphs.

<table>
<tr><td>Type and size of data used</td><td colspan="2">The resources used for the multimedia track are based on Wikipedia data:
Wikipedia XML collection: A Wikipedia crawl converted to XML consisting of 659,388 XML documents with image identifiers added to the <image > tags for those images that are part of the Wikipedia image XML collection. This is the target collection for the MMfragments task.
Wikipedia image collection: A subset of 171,900 images referred to in the Wikipedia XML collection is chosen to form the Wikipedia image collection.
Wikipedia image XML collection: This XML collection is specially prepared for the multimedia track. It consists of XML documents containing the images in the Wikipedia image collection and their meta-data. This is the target collection for the MMimages task.
Image classification scores: For each image, the classification scores for the 101 MediaMill concepts are derived by University of Amsterdam.
Image features: For each images, the set of 120D feature vectors that has been used to derive the image classification scores is also available. These feature vectors can be used to build a custom CBIR-system, without having to pre-process/access the image collection.</td></tr>
<tr><td>Method used for the generation of the ground-truth</td><td colspan="2">For both tasks, the topics are generated by the participants in INEX MM track and the relevance assessments are also performed by them.
MMfragments task: It requires assessments at the sub-document level, a simple binary judgement at the document level is not sufficient. Still, for ease of assessment, retrieved fragments are grouped by document. Once all participants have submitted their runs, the top N fragments for each topic are pooled and grouped by document. Assessors look at the documents in the pool and highlight the relevant parts of each document. The assessment system stores the relevance or non-relevance of the underlying XML elements.
MMimages task: TREC style document pooling of the top N documents (= images + metadata) and binary assessments at the document level.</td></tr>
<tr><td rowspan="3">Participation statistics</td><td>2005</td><td>7 registrations
5 submissions
21 runs</td></tr>
<tr><td>2006</td><td>20 registrations
4 submissions
31 runs</td></tr>
<tr><td>2007</td><td>16 registrations
4 submissions
30 runs</td></tr>
<tr><td>URL</td><td colspan="2">http://inex.is.informatik.uni-duisburg.de</td></tr>
</table>

Synthesis and conclusion	• Realistic and sizable document collection Interesting additional resources Easy entry point for IR/DB researchers (no image analysis needed) • Few participants Top performing runs use no visual information Too little data to be conclusive • Re-usable test collection Inter assessor agreement high No submission bias

Cross-Language Speech Retrieval (CL-SR)

The CLEF Cross Language Speech Retrieval (CL-SR) benchmark test evaluates spoken document retrieval systems in a multilingual context. In 2006 the CL-SR track included search collections of conversational English and Czech speech using six languages (Czech, Dutch, English, French, German and Spanish). In CLEF 2007 additonal topics were added for the Czech speech collection, and additonal speech recognition results were available for the English speech collection. Speech content was described by automatic speech transcriptions manually and automatically assigned controlled vocabulary descriptors for concepts, dates and locations, manually assigned person names, and hand-written segment summaries. Additional resources of word lattices and audio files can be made available.The track was coordinated by U. Maryland (US), Dublin City U. (IE) and Charles U. (CZ).

Definition of tracks and tasks (2006)	○ Task 1: retrieve pre-defined topics in ASR decoded speech archive (American English – spontaneous speech) ○ Task2: retrieve pre-defined topics in ASR decoded speech archive (Czech – spontaneous speech) ○	
Evaluation metrics of tasks	○ Mean uninterpolated Average Precision (MAP (using the Trec_val program from NIST: http://trec.nist.gov/trev_val/)	
Type and size of data used	○ English task: ○ The resulting test collection contains 8,104 segments from 272 interviews totaling 589 hours of speech ○ 63 search topics ○ 8.104 cohorent segments (equivalent of "documents" in a classic IR task) ○ 30.497 relevance judgements ○ ASR transcripts were provided by one partner (IBM for English)	
Method used for the generation of the ground-truth	• The collection from the Shoah Visual History Foundation contains a 10,000 hour subset for which manual segmentation into topically coherent segments was carefully performed by subject matter experts.	
Participation statistics	2005	○ 7 participants

	2006	○ English task: 6 participants ○ Czeck task: 3 paricipants
	2007	○ unknown
URL	http://www.clef-campaign.org/2007/2007agenda.html http://clef-clsr.umiacs.umd.edu/	

NIST Spoken Term Detection

The STD task is to find all of the occurrences of a specified "*term*" in a given corpus of speech data. For the STD task, a term is a sequence of one or more words. The evaluation is intended to help develop technology for rapidly searching very large quantities of audio data. Although the evaluation actually uses only modest amounts of data, it is structured to simulate the very large data situation and to make it possible to extrapolate the speed measurements1 to much larger data sets. Therefore, systems must be implemented in two phases: indexing and searching. In the indexing phase, the system must process the speech data without knowledge of the terms. In the searching phase, the system uses the terms, the index, and optionally the audio to detect term occurrences.

Definition of tracks and tasks (2006)	The STD task is to find all of the occurrences of a specified "*term*" in a given corpus of speech data. For the STD task, a **term** is a sequence of one or more words. Terms will be specified only by their orthographic representation. Example terms are "grasshopper", "New York", "in terms of", "overly protective", "Albert Einstein", and "Giacomo Puccini".
Evaluation metrics of tasks	Systems will be evaluated for both speed and detection accuracy. Speed and accuracy will be measured for a variety of conditions, for example as a function of term characteristics (such as frequency of usage and acoustical features) and corpus characteristics (such as source type and signal quality). Basic detection performance will be characterized in the usual way via standard detection error tradeoff (DET) curves of miss probability (PMiss) versus false alarm probability (PFA). Miss and false alarm probabilities are functions of the detection threshold, q, and will be computed separately for each search term.
Type and size of data used	The development and evaluation corpora will include three languages and three source types. - The three languages will be Arabic (Modern Standard and Levantine), Chinese (Mandarin) , and English (American). - The three source types will be Conversational Telephone Speech (CTS), Broadcast News (BNews), and Conference Room (CONFMTG) meetings i.e., goal oriented, small group, roundtable meetings. - 1-3 hours per language and source type
Method used for the generation of the ground-truth	Search queries are labeled manually for the test corpus

Participation statistics	2006	○ 9 submissions
	2008	○ Planned
		○
URL	http://www.nist.gov/speech/tests/std/	

Nist Rich Transcription

The Rich Transcription evaluation series is implemented to promote and gauge advances in the state-of-the-art in several automatic speech recognition technologies. The goal of the evaluation series is to create recognition technologies that will produce transcriptions which are more readable by humans and more useful for machines. As such, a set of research tasks has been defined which are broadly categorized as either Speech-to-Text Transcription (STT) tasks and Metadata Extraction (MDE) tasks.

The evaluation series was started in 2002 and continues to this day. The meeting recognition community is expanding the scope of the RT evaluations to include multimodal research including audio and video.

Definition of tracks and tasks (2007)	○ Evaluation of quality of automatic indexing of meeting recordings (4 measures): ○ Speech-to-text (STT) transcription rate ○ Diarization 1: Who spoke when ○ Diarization 2: Speech Activity Detection ○ Diarization 3: Source Localization
Evaluation metrics of tasks	○ Word error metric für STT task ○ The Diarization Error Rate (DER) metric is used to assess SPKR system performance. DER is the ratio of incorrectly attributed speech time, (either falsely detected speech, missed detections of speech, or incorrectly clustered speech) to the total amount of speech time, expressed as a percentage ○ Diarization "Speech Activity Detection" (SAD) rate ○ Speaker Localization and Tracking Rate
Type and size of data used	○ Speech recordings from lecture rooms ○ Speech recordings from meeting rooms
Method used for the generation of the ground-truth	○ Manual annotation

Participation statistics	2005	○ 9 participants (also partners from European projects: CHIL, AMI, ○ Not all sites participates in all 4 tasks
	2006	○ Unknown
	2007	○ Unknown
URL	http://www.nist.gov/speech/tests/rt/index.htm http://www.nist.gov/speech/publications/papersrc/rt05sresults.pdf	

4.3. Conclusion

A willingness to cooperate has already been demonstrated through several common events (i.e. MUSCLE/ImageCLEF workshops, Chorus evaluation session). By bringing together a number of these initiatives into a single entity, a cross-disciplinary approach to multimedia retrieval benchmarking can be developed. Already, common evaluation tasks have been identified over the different initiatives that will allow joining forces. Still many open issues remain and need much work and discussion within the community.

The outputs from several meetings dig out some hard issues which need deeper investigation and are summarized below.

- Technology assessment vs user satisfaction: Best evaluated system may not be usable. Existing commercial systems often evaluate poorly. On the other hand, users are satisfied with commercial systems. Are we missing?
  Accurate performance measures? (make existing ones better)
  Relevant perf. measures? (find new ones).
  More should be done on including the user perspective in evaluation

- CLEF2005 interrogation: Why as we have good results on cross lingual evaluation, none of the best systems have a commercial success?
  Tentative answer: conditions of test do not reflect the real use of the systems

- Requirements for a user oriented evaluation
  Key issue: non-intrusive approach
  Real "subjects"
  Real applications
  Simulation (Wizard of Oz)

What is needed is:

- Basic Research Evaluation (validate research direction)
- Technology Evaluation (assessment of solution for well defined problem)
- Usage Evaluation (end-users in the field)
- Impact Evaluation (socio-economic consequences)
- Program Evaluation (funding agencies)

Our future plans in the project next stages include: the mapping of the current landscape of existing benchmark initiatives assessing their differences and common properties to put together our efforts to better address the remaining hard problems. We plan to continue our investigations to provide recommendations for the best practices for methods and systems evaluation.

Annex I: Evaluation efforts (and standards) within ongoing EU Projects and National Initiatives

In this section we tried to collect participation of ongoing European projects and national initiatives to evaluation campaign through a questionnaire. We have partial information coming from Sapir, Tripod, Vitalas, Aim@shape, Vidivideo and MultimediaN.

Project names: MultimediaN & VidiVideo
Arnold Smeulders & Marcel Worring

1- Internal technical evaluation within WPs
- Test Corpora Type (Text, audio, video…):
  Video from TRECvid
  Video from surveillance internally
  ALOI static database of objects
  MediaMill challenge

- Test Corpora Size:
  TRECvid: Hundreds of hours partially annotated in TRECvid manner
  ALOI: 100 different recordings of 1000 objects = 100.000 images
  MediaMill challenge: 101 concepts with ground truth and models based on TRECvid data.

- Performance measures (Mean precision,…)
  TRECvid style: Mean Average Precision
  ALOI: recognition rates
  Video Olympics: number of retrieved items in a five minute period, pleasant interface by voting of potential users.

2- Participation in open national/European/international Benchmark initiatives:
- Name and level (european?) of the initiative:
  TRECvid: worldwide
  ALOI: scientific
  VOC: worldwide
  VideoOlympics: worldwide

- Nbr of participants
  TRECvid: 60 participants – all international – and growing
  ALOI: downloads
  Video Olympics: 9 participants

- How is generated the ground truth?
  TRECvid style: basic annotation supplemented by parties
  ALOI: fully documented at scanning
  Video Olympics: fully annotated for target questions

- Are you the organizer?

No, of TRECvid, but the NIST is.
Yes, of ALOI, see Int Journal Comp Vision Geusebroek & Smeulders
Yes, of the MediaMill challenge, see ACM Multimedia 2006
Yes, of the Video Olympics, see www.videolympics.org for information and a video impression of the first edition.

3- User Trials (feedback with real end-users, no relation with the provided technologies)

We think this is not a very useful question at this point. We work closely with the national video archive of the Netherlands in MultimediaN, VidiVideo and other projects. When there is a real need we will engage real end-user at the first instance.

However, we do are busy developing user group question types.

4- Participation in standardization effort:
- Label and name (MPEG7, JPSearch, XMLx…)
- Others: …
- More infos on this standardization context and objective:
- Abstract of your contribution

XML Dublin Core storage format of detected results.

Project name: SAPIR

Yosi Mass

1- Internal technical evaluation within WPs
- Test Corpora Type – we use FlickrXML files extracted from Flickr. Each file contains text metadata as appear in Flickr as well as 5 MPeg-7 Visual Descriptors (Scalable Color, Color Structure, Color Layout, Edge Histogram and Homegenous Texture) extracted from the image.
- Test Corpora Size: - 40M images. We plan to grow to 100M images.
- Performance measures – currently we don't have automatic measures. We use a UI to search for images that are similar to a given image possibly combined with Text.

2- Participation in open national/European/international Benchmark initiatives:
- No

3- User Trials (feedback with real end-users, no relation with the provided technologies)
We defined 5 possible scenarios that can benefit from large scale content based search in audio-visual data. The 5 scenarios are – Tourist, Journalist helper, Music&Text, Advanced home messaging and Hollywood&Home. The scenarios can be found on the project site at http://www.sapir.eu. We then run some focus groups to evaluate the scenarios

- Number of these external users: - 5-7 per scenario
- Do the users belong to different communities? : Yes, some are novice and some are professional. For example for the Journalist scenario we interviewed some journalists.

- Trials protocol: We did some UI Sketches for the scenarios and then interviewed the participants in the focus groups
- User's satisfaction criteria: We measured along 3 dimensions – effectiveness, efficiency and satisfaction. We used the following criterias –

- **Perceived effectiveness**
  Are you able to precisely formulate your request?
  Do you get the requested results?
  Do you get sufficient recall information to judge the value of the result?
  Do you get sufficient precision information to judge the value of the result?

- **Perceived efficiency**
  Do you formulate precise queries with minimal efforts?
  Do you get the results within reasonable time?
  Does the ranking and presentation of the results fit the intention of your quest?

- **Perceived satisfaction**
  Do you find the service easy to use? E.g. no hazzle, no errors, logical structure.
  Do you find the service enjoyable to use (pleasant, comfortable, nice design, etc)
  Do you get sufficient supported? E.g. during the installation phase or when errors or unexpected situations occurs.
  Do you find the cost/benefit ratio reasonable?
  Do you trust the providers of the service?
  Do you find the service accessible? E.g. mobility issues

The results of the findings from the Focus groups are part of a deliverable that will be put towards the YE on the project web site.

4- Participation in standardization effort:
- Label and name (MPEG-7, MPEG-A, MPEG-21, OMA)
- More infos on this standardization context and objective: to be supplied toward the YE
- Abstract of your contribution: to be supplied until the YE

Project name: Tripod
Mark Sanderson

1- Internal technical evaluation with related WPs
- Test Corpora Type (Text, audio, video…): Image collection
- Test Corpora Size: Several thousand
- Performance measures (Mean precision,…): Not entirely determined yet, some classic retrieval effectiveness measures; for caption creation, maybe the bleu or rouge measures.

2- Participation in open national/European/international Benchmark initiatives:
- Name and level (european?) of the initiative: geo-CLEF, possibly in the follow on to MUSCLE
- Web site of the initiative: http://www.clef-campaign.org/; www.**muscle**-noe.org

- Nbr of participants: ~15
- Nbr and Title of tasks: geoimage track
- Performance measures: Standard retrieval measures
- How is generated the ground truth? Relevance assessors
- How are maintained the test data collections? CLEF maintain the data
- Are you the organizer? Co-organiser

3- User Trials (feedback with real end-users, no relation with the provided technologies)
- Number of these external users: Still to be determined
- Do the users belong to different communities? : Large public? Professionals? (Which are…)
- Trials protocol: Still to be determined
- User's satisfaction criteria: Still to be determined

4- Participation in standardization effort:
- Label and name (MPEG7, JPSearch, XMLx…) Tripod will build on the XMP standard
- Others: …
- More infos on this standardization context and objective:
- Abstract of your contribution

Tripod will evaluate two aspects of its outputs. 1) It will evaluate the quality of the image captions that it outputs; 2) it will evaluate the search engine that searches over the enhanced images. Evaluation of summaries will be conducted by creating a range of existing manually captioned images and comparing a different set of automatically captioned images with the manual set. Retrieval evaluation will be conducted in a classic IR test collection approach. We plan to be strongly involved in CLEF and in the follow on from the MUSCLE network of excellence. Our involvement will be in providing data sets to those exercises and in contributing to the experimental design.

Project name: AIM@SHAPE
Michela Spagnuolo

1- Internal technical evaluation with related WPs
- Test Corpora Type (Text, audio, video…): *digital 3D objects*
- Test Corpora Size: *depending on the object represented*
- Performance measures (Mean precision,…):

2- Participation in open national/European/international Benchmark initiatives:
- Name and level (european?) of the initiative: *SHREC: 3D Shape Retrieval Contest, international initiative*

- Web site of the initiative: *http://www.aimatshape.net/event/SHREC/*
- Nbr of participants & Nbr and Title of tasks: *the contest is organized in tracks, each for a specific 3D retrieval task, either in terms of retrieval method (eg, partial/global) or shape type (eg protein/CAD models)*
  1- Watertight models. Eight groups initially registered, five groups actually participated.
  2- CAD models. Nine groups initially registered, four groups actually participated
  3- Partial matching. Five groups initially registered, only two actually participated.
  4- Protein models. Three groups participated.
  5- 3D face models. Seven groups initially registered, three actually participated.
- Performance measures:

For each query there exists a set of highly relevant items and a set of marginally relevant items. Therefore, most of the evaluation measures have been split up as well according to the two sets. Measures used: true and false positives, true and false negatives, first and second tier, precision, recall, average precision, average dynamic recall, cumulated gain vector, discounted cumulated gain vector, normalized cumulated gain vector (see Section 4 of the attached SHERC06.PDF for a complete description of the performance measures)

- How is generated the ground truth? *Manually, by track organizers*
- How are maintained the test data collections? *In the first two SHREC contests, they have been maintained by the organizers; we are considering the possibility to maintain them directly in the ShapeRepository of the AIM@SHAPE project (see shapes.aimatshape.net)*
- Are you the organizer? *AIM@SHAPE is organizing the contest and more precisely, Remco Veltkamp, UU (email: remco.veltkamp@uu.nl*

3- User Trials (feedback with real end-users, no relation with the provided technologies)
- Number of these external users:
- Do the users belong to different communities? : Large public? professionals? (which are…)
- Trials protocol:
- User's satisfaction criteria:

not applicable, the contest is meant for a scientific audience

4- Participation in standardization effort: *none*
- Label and name (MPEG7, JPSearch, XMLx…)
- Others: …
- More infos on this standardization context and objective:
- Abstract of your contribution

Project name: VITALAS

Nozha Boujemaa

1- Internal technical evaluation
- Test Corpora Type (Text, audio, video…): image, audio, video, text
- Test Corpora Size used to develop the VITALAS system:
  - ~ 1000 professional images + textual metadata per image
  - ~ 100 hours of broadcast archive video + metadata per program

- Test Corpora Size used for large scale retrieval using the VITALAS system:
  - ~ 3 million professional images + textual metadata per image
  - ~ 10.000 hours of broadcast archive video + metadata per program
- Performance measures (Mean precision,…): Mean average precision and recall graphs

2- Participation in open national/European/international Benchmark initiatives:
- Name and level (european?) of the initiative:
  - INEX Multimedia – international (Theodora Tsikrika (CWI) (a VITALAS partner) is one of the two organisers of the INEX Multimedia track).
  - TRECVID - international
  - ImageCLEF - international

3- User Trials (feedback with real end-users, no relation with the provided technologies)
- Number of these external users: *not defined yet*
- Do the users belong to different communities? : professionals (Jouranalists, Documentalists)
- Trials protocol: *not defined yet*
- User's satisfaction criteria: *not defined yet*

4- Participation in standardization effort:
- JPSearch, XQuery
- More infos on this standardization context and objective: The VITALAS project aims to offer significant contributions to the development of European and International Standards. The areas in which the project can make a substantial contribution include content representation, query languages for cross-media retrieval, and the evaluation of multimedia / cross-media retrieval systems.

Annex II: Related Chorus Events to Benchmarking and Evaluation

Below is the program of Chorus Roquencourt workshop. Slides of all presentation are available on the web site: http://www.ist-chorus.org/chorus-wg2.php
Short abstract with link to each benchmark initiatives are available on:
http://www.ist-chorus.org/benchmark-initiatives-for-multim.php

NAVS Chorus cluster, March 14th 2007

Agenda Chorus WG2 meeting - 14:30-17:30

Evaluation and Benchmarking of Multimedia Content Search Methods

The objective is to make the point on ongoing evaluation initiatives.

14:30 - 14:45 TrecVid - Alex Hauptmann (CMU - USA)
14:45 - 15:00 ImageClef - Henning Müller (UHG - Switzerland)
15:00 - 15:15 ImagEval - Pierre Alain Moellic (CEA - France)
15:15 - 15:30 Pascal Challenge & Robin - Frédéric Jurie (INRIA Rhône-Alpes)

15:30 - 15-45 Short Statements: IAPR-TC12 - Marcel Worring (UvA - Netherlands);
CIVR Evaluation Showcase - Allan Hanbury (VUT - Austria)

15:45 - 16:00 Coffee break

16:00 - 16:15 SHREC (3D) - Michela Spagnuolo (CNR - Italy)
16:15 - 16:30 MIREX - Andreas Rauber (VUT - Austria)
16:30 - 16:45 INEX - Thijs Westerveld (CWI - Netherlands)

16:45- 17:30 **Panel discussion:**
1- Why so many benchmarch inititives? Is there communalities?
2- How can they work closer together?
3- What are the main difficulties encountered: data collections, data annotation, task definition, task evaluation, participation...?
4- How can we face the identified problems

Meeting Closer: Next steps in Chorus WG2 activities - Nozha Boujemaa (INRIA - France)

Program and slides of CBMI'2007 panel are available on the web site:
http://www.ist-chorus.org/events_0.php

CBMI Chorus Panel: June 25th 2007 Bordeaux

CBMI homepage

Topic: Benchmarking Multimedia Search Engines

Panel Chair: Nozha Boujemaa INRIA -France (slides)
Panelist:

- Stéphane Marchand-Maillet - University of Geneva, Switzerland (slides)
- Christian Fluhr - CEA, France (slides)
- Kahlid Choukri - ELDA, France (slides)

With the contributions from Henning Mueller (SIM - Geneva), Paul Clough (Univ. Sheffield)

The panel will address the following questions:

1. "Role of the user in the evaluation process of multimedia retrieval techniques; How much difficult taking the user in the evaluation process?"
2. "How to measure search engines performance/success: user satisfaction or technology accuracy?"
3. "How to quantify the success in each situation? How much is it dependent from scenarios and context (application)?"
4. "Are the best performance systems the most successful commercially?"
5. With the ending question: "How useful the evaluation is? Pushing a head the knowledge or killing the innovation?"

5. P2P SEARCH, MOBILE SEARCH AND HETEROGENEITY

5.1. Introduction

Multimedia search appears to have at least the following subclasses: P2P search and mobile search. They are both important: P2P search could very well be a model for searching the whole web as audiovisual content is becoming dominant and mobile search will cater to the increasing number portable devices. In this part of the document we will address these two areas showing their importance, their state of the art and the main research players.

5.2. P2P search

5.2.1. Introduction

A P2P application is different from the traditional client/server model because it acts both as a client and a server. That is to say, while they are able to request information from other servers, they also have the ability to respond to requests for information from other clients, at the same time. **Figure 1** shows the architecture of a P2P network, where each node acts as a user interface, service provider, message router, and –possibly partial- resource repository. The links between nodes tend to be dynamic. The advantage of a peer-to-peer architecture compared to traditional client-server architectures is that a machine can assume the role that is most efficient for the performance of the network. This implies the load on the server is reduced/distributed, which allows for more specialized services.

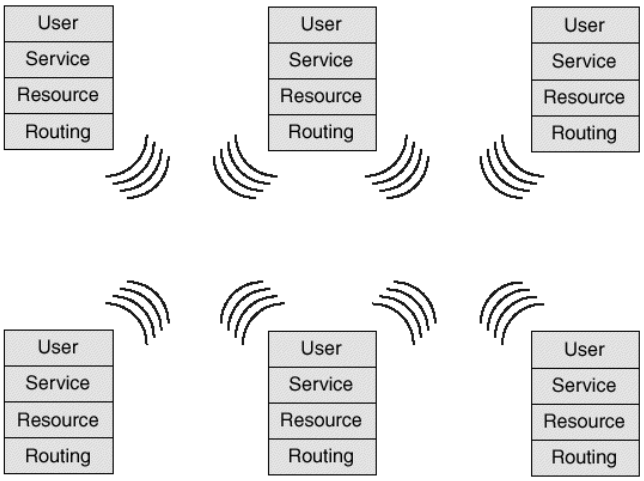


Figure 1: Architecture of a P2P network

A typical peer-to-peer application has the following key features:

- **Peer discovery.** The application must be able to find other applications that are willing to share information. Historically, the application finds these peers by registering to a central server that maintains a list of all applications currently willing to share, and giving that list

to any new applications as they connect to the network. However, there are other means available, such as network broadcasting or discovery algorithms.

- **Querying peers for content**. Once these peers have been discovered, the application can ask them for the content that is desired by the application. Content requests typically come from users, but it is possible that the peer-to-peer application is running on its own and performing its query as a result of some other routed network request.
- **Sharing content with other peers.** In the same way that the peer can ask others for content, it can also share content after it has been discovered.

The social classification or collaborative tagging component of P2P is relatively important: some research work on the social aspects of P2P search related to the different kinds of folksonomies has been carried.
Social search, in general, takes into account all user input to refine the search: social bookmarking and tagging, sharing personal item lists, etc…

Social selection needs the active participation of the user. She shares compiled lists or tagged items, so that the content slowly grows. Specialized search engines only need an initial setup and can be refined after that. Users give useful information to the search engine by writing in search keywords. There is no need for additional input.

5.2.2. Context

Web search is almost exclusively under the control of centralized search engines. Lately, various projects have started building and operating a P2P web search network, but so far these endeavours are fairly small in scale.

Ironically, Web search and Internet scale file content search seem to be perfect candidates for a P2P approach, for several reasons:

1) The data is originally highly distributed, residing on millions of sites (with more and more individuals contributing, e.g., through their blogs)

2) A P2P network could potentially dwarf even the largest server farm in terms of processing power and could thus enable much more advanced methods for computational intensive tasks, such as linguistic data analysis, statistical learning, or ontology based background knowledge and reasoning (all of which are out of the question when you have to serve hundred millions of queries per day on a, however big but centralized, server farm).

3) There is growing concern about the world's dependency on a few quasi monopolistic search engines and their susceptibility to commercial interests, spam or distortion by spam combat, biases in geographic and thematic coverage, or even censorship. These issues have led to postulate that "the Web should be given back to the people".

The peer-to-peer (P2P) approach, which has become popular in the context of file-sharing systems such as Gnutella or KaZaA, allows handling huge amounts of data in a distributed and self-organizing way. In such a system, all peers are equal and all of the functionality is shared among all peers, so that there is no single point of failure and the load is evenly balanced across a large number of peers. These characteristics offer enormous potential benefits for search capabilities powerful in terms of scalability, efficiency, and resilience to failures and dynamics. Additionally, such a search engine can potentially benefit from the intellectual input (e.g., bookmarks, query logs, etc.) of a large user community. One of the key difficulties, however, is to efficiently select promising peers for a particular information need.

Effective discovery methods rely on the information published for a particular resource. Commonly used discovery methods include:

- *Flooding broadcast queries*

When a peer makes a query, the query is then broadcasted to all the neighbour peers. If its neighbour peers could not solve the query, then the query is broadcasted to neighbour's neighbour peers. If a resource is found, that peer will send a message to the original sender of the query, indicating it can solve the query, and then establish a peer-to-peer connection. The original Gnutella implementation is an example of a flooding broadcast discovery mechanism.

Each query has a time-to-live (ttl) counter. Typically, the ttl is set between 5 and 7, and the value is decremented by each node as it relays the message. Another counter tracks the number of hops. Once the ttl counter reaches zero, the query will be discarded.

Due to the broadcast nature of each query, the system does not scale well (o(n2)); the bandwidth network assumption grows exponentially with a linear increase in the number of peers. Raising the number of peers in the system will cause the network to quickly reach bandwidth saturation.

This type of method has the advantage of flexibility in the processing of queries. Each peer can determine locally how it will process the query and respond accordingly. It is simple to design and efficient. Unfortunately, it is suitable only for small networks. As well as that, this type of mechanism is very susceptible to malicious activity, rogue peers can send out large number of queries, which produce a significant load on the network.

- *Selective forwarding systems*

Instead of sending a query to all peers, it is selectively forwarded to specific peers who are considered likely to locate the resource. Peers will become super peer automatically if they have sufficient bandwidth and processing power, i.e. if a peer has broadband connection and higher processing power. Peers with dial-up connection (low bandwidth) will make queries to super peers. This type of systems use flow control algorithm (fca), which tries to apply a form of intelligent flow control in terms of how a peer forwards request and response messages and a sensible priority scheme, as well as how it drops messages that won't fit into the connections. Selective forwarding systems are more scalable than flooding broadcast systems.

This approach greatly reduces bandwidth limitations to scalability. But it is susceptible to malicious activity: a rogue peer can insert itself into the network at the various points and misroute queries, or discard them altogether.

Each peer must also contain some amount of information used to route or direct queries received. The size of this information is negligible in a small network, but in large networks, this overhead may grow to levels that are unacceptable, hence it is not suitable for a large peer network.

- *Decentralized hash table networks*

In decentralized hash table networks, each file stored within the system is given a unique ID, typically a sha-1 hash of its content, which is used to identify and locate a resource. Given this unique ID, a resource can be located quickly despite the size of the network. Since this key identifies each resource, it is impossible to perform a fuzzy or keyword search within the network. If a peer is looking for a file from another peer, it must obtain this key first in order to retrieve the file.

These systems are also susceptible to malicious activity by rogue peers: they may discard a query, insert large amount of frivolous data to clutter the key space, or flood the network with queries to

degrade the performance.

- *Centralized indexes and repositories*

Indexes of all peers and their resources are kept on a main server. A query is sent to a server, then the server will look-up the index, if the query can be solved, then the server will send a message to the original query sender explaining where he can get the file. Napster uses centralized indexed and repositories system.

Centralized indexes have provided the best performance for resource discovery. The server in centralized indexes and repositories system is expensive, the bandwidth and hardware required to support large networks of peers are expensive.

If the server in the system fails to function properly, it brings down the whole network. In the case of Napster, it has a cluster of servers, so that if one server fails, the rest of the servers will continue supporting the network.

- *Distributed indexes and repositories*

The idea of distributed index is that each content broker in the network keeps an index of local files as well as an index of some files stored in some neighbouring content broker. When a content broker receives a query from a peer, it first checks to see if the query can be satisfied locally. If it cannot, it uses the local index to decide which content broker to forward the request to. The index on each server is not static and changes as files move through the system. With this approach, we could eliminate the need for expensive centralized servers.

If well designed, distributed indexes and repositories provide currently one of the best performances and scalability. In addition, it has a high sngle point failure tolerance, because a content broker only contains a relative small number of indexes in comparison to the centralized server, so that if one content broker goes down, the network will still function properly.

The problem with this type of indexing system is that if a file is changed locally by a peer, then the content broker will not be aware of this fact. Subsequently, when another peer requests that particular file, the content broker will return an out-of-date copy of that file. The overhead in keeping everything up-to-date and efficiently distributed is a major detriment to scalability.

Peers joining and leaving the network from time to time,Also when a peer leaves the network, all the resources indexes stored in that peer will become unavailable to other peers.

Distributed indexing systems, as they currently exist, cannot provide robust discovery in large networks.

- *Relevance driven network crawlers*

Relevance driven network crawlers use a database of existing information the peer has accumulated to determine which resources it encounters may or may not be relevant or interesting to the peers.

Over time, a large amount of information is accrued, which is analysed to determine what common elements the peer has found relevant. The crawler then traverses the networks, usually consisting of html documents for new information, which matches the profile distilled from the previous peer information.

The time required for the crawler to traverse a large amount of content is very long; it is not suitable for large networks.

5.2.3. Main players

Except filesharing applications, like bittorennt, Gnutela, Edonkey, etc. there are only a dozen of companies working on P2P search engines, some have already a tool, others are in research or in development. Some like Minerva and Yahoo are developing specific algorithms, other like Faroo, Yacy, Open search are fully distributed P2P search engines. Most companies have distributed crawlers but a central index : Majestic, GPU, Grob and Boitho.

5.2.4. The State of the Art

There are many workshops and papers around the subject of P2P IR (see references). Among principal recent worshop in this area we can name: Distributed IR at SIGIR 2004, P2PIR at SIGIR 2004, Heterogeneous and Distributed IR at SIGIR 2005, P2PIR 2005 and 2006 at CIKM, Large-Scale Distributed Systems for IR at SIGIR 2007, Adversarial IR on the Web, at WWW 2007 and IPTS. Most have good reports on line giving historical context.

There are isolated solutions today for P2P text search and for P2P similarity search but for single AV features (e.g. for color or shape). There are no efficient solutions combining both text and multiple features (e.g. both color and shape). Of course all solutions have to deal with scalability.

5.2.4.1. *The scalability viewpoint*

Comprehensive Web search based on a P2P network has been considered infeasible from a scalability viewpoint. Recent work, however, indicates that the scalability problems could be overcome, either by distributing a conceptually global keyword index across a DHT style network or by having each peer compile its local index at its own discretion (using the peer's own "native" data sources or performing thematically focused Web crawls and other data extraction according to the peer's interest profile). In addition, various acceleration techniques can be employed. For example, one pursues a multilevel partitioning scheme, a hybrid between partitioning by keyword and partitioning by document. Another uses view trees for result caching to improve the P2P search efficiency.

From a query processing and IR viewpoint, one of the key issues is query routing: when a peer poses a query with multiple keywords and expects a high quality top10 or top100 ranked result list, the P2P system needs to make a judicious decision on which other peers the query should be forwarded.

This decision needs statistical information about the data contents in the network. It can be made fairly efficiently in a variety of ways, like utilizing a DHT based distributed directory, building and maintaining a semantic overlay network (SON) with local routing indexes, or using limited forms of epidemic gossiping.

However, efficiency of P2P query routing is only one side of the coin. Of course, we also expect good search result quality, that is, good effectiveness in IR terminology, measured in terms of precision and recall. The goal is to be as good as the best centralized search engines, but the P2P approach faces the challenge that the index lists and statistical information that lead to good search results are scattered across the network.

For example, consider two or three keyword queries such as "Michael Jordan", "native American music", or "PhD admission".

A standard, efficient and scalable, approach would decompose each of these queries into individual terms such as "native" and "American" and "music", identify the best peers for each of the terms separately, and finally combine them, e.g., by intersection or some form of score aggregation in order to derive a candidate list of peers to which the query should be forwarded. The result of this "factorization" would often lead to mediocre results as the best peers (and files located on those

peers) for the entire query may not be among the top candidates for any of the individual keywords. The root cause of the above problem is that the outlined "factorized" method for P2P query routing and processing has no way of taking into account the correlation between the keywords in the query. We miss out on the fact that, for example, "PhD" and "admission" are statistically correlated in the corpus, and, even worse, that the best matches for the entire query should exhibit a higher than average frequency of both terms (ideally within some proximity window). Standard search engines do not necessarily consider these correlations either, but they process index lists on the overall document space directly, whereas the P2P system first needs to identify other peers for query routing in order to access index lists and then sees only partitions of the global index space. Thus, the necessarily coarser aggregation granularity of routing indexes or the distributed directory causes an additional penalty for a P2P approach. On the other hand, directly simulating the centralized algorithms in the P2P network would incur undue communication costs.

One may argue that critical correlations of the above kind typically occur in composite names or phrases, as suggested by our examples. Although this is indeed often the case, the observation alone does not provide a solution. It is virtually impossible to foresee all phrases or names or correlated term pairs that will appear in important user queries, and brute force pre-computation of statistical measures for all possible pairs of terms is not a viable option.

5.2.4.2. File sharing and Text only

- *Searching in unstructured P2Ps*

In an unstructured P2P system, no rule exists that strictly defines where data is stored and which nodes are neighbours of each other. To find a specific data item, early work such as the original Gnutella used flooding, which is the Breadth First Search (BFS) of the overlay network graph with depth limit D. D refers to the system-wide maximum TTL of a message in terms of overlay hops. In this approach, the querying node sends the query request to all its neighbours. Each neighbour processes the query and returns the result if the data is found. This neighbour then forwards the query request further to all its neighbours except the querying node. This procedure continues until the depth limit D is reached. Flooding tries to find the maximum number of results within the ring that is centred at the querying node and has the radius: D-overlay-hops. However, it generates a large number of messages (many of them are duplicate messages) and does not scale well.

Many alternative schemes have been proposed to address the problems of the original flooding.

These works include iterative deepening, k-walker random walk, modified random BFS, two-level k-walker random walk, directed BFS , intelligent search, local indices based search, routing indices based search, attenuated bloom filter based search, adaptive probabilistic search, and dominating set based search . They can be classified as BFS based or Depth First Search (DFS) based. The routing indices based search and the attenuated bloom filter based search are variations of DFS. All the others are variations of BFS.

In the iterative deepening and local indices, a query is forwarded to all neighbours of a forwarding node. In all other schemes, a query is forwarded to a subset of neighbours of a forwarding node. The searching schemes in unstructured P2P systems can also be classified as deterministic or probabilistic. In a deterministic approach, the query forwarding is deterministic. In a probabilistic approach, the query forwarding is probabilistic, random, or is based on ranking. The iterative deepening, local indices based search, and the attenuated bloom filter based search are deterministic. The others are probabilistic.

Another way of categorizing searching schemes in unstructured P2P systems is regular-grained or coarse-grained. In a regular-grained approach, all nodes participate in query forwarding. In a coarse-grained scheme, the query forwarding is performed by only a subset of nodes in the entire

network. Dominating set based search is coarse-grained, because the query forwarding is performed only by the dominating nodes in the CDS (Connected Dominating Set). All the others are regular-grained.

Another taxonomy is blind search or informed search. In a blind search, nodes do not keep information about data location. In an informed search, nodes store some metadata, a process that facilitates the search. Blind searches include iterative deepening, k-walker random walk, modified random BFS, and two-level k-walker random walk. All the others are informed search.

- *Iterative deepening*

Yang and Garcia-Molina borrowed the idea of iterative deepening from artificial intelligence and used it in P2P searching. This method is also called expanding ring. In this technique, the querying node periodically issues a sequence of BFS searches with increasing depth limits. The query is terminated when the query result is satisfied or when the maximum depth limit D has been reached. Iterative deepening is tailored to applications where the initial number of data items returned by a query is important. However, it does not intend to reduce duplicate messages and the query processing is slow.

- *k-walker random walk and related schemes*

In the standard random walk algorithm, the querying node forwards the query message to one randomly selected neighbour. This neighbour randomly chooses one of its neighbours and forwards the query message to that neighbour. This procedure continues until the data is found. Consider the query message as a walker. The query message is forwarded in the network the same way a walker randomly walks on the network of streets. The standard random walk algorithm uses just one walker. This can greatly reduce the message overhead, but causes longer searching delay.

In the k-walker random walk algorithm, k walkers are deployed by the querying node. That is, the querying node forwards k copies of the query message to k randomly selected neighbours. Each query message takes its own random walk. Each walker periodically “talks” with the querying node to decide whether that walker should terminate. Nodes can also use soft states to forward different walkers for the same query to different neighbours. K-walker random walk algorithm attempts to reduce the routing delay. On average, the total number of nodes reached by k random walkers in H hops is the same as the number of nodes reached by one walker in kH hops. Therefore, the routing delay is expected to be k times smaller.

Another similar approach, called the modified random BFS, was proposed. The querying node forwards the query to a randomly selected subset of its neighbours. On receiving a query message, each neighbour forwards the query to a randomly selected subset of its neighbours (excluding the querying node). This procedure continues until the query stop condition is satisfied.

Three approaches are considered and evaluated using k-walker random walk: owner replication, path replication, and random replication. All three schemes replicate the object found, when a query is successful. The owner replication replicates an object only at the requesting node. The path replication creates copies of an object on all nodes on the path, from the providing node to the requesting node. The random replication places copies on the p randomly selected nodes that were visited by the k walkers. The path replication implements the square-root replication. The random replication has slightly less overall search traffic than the path replication, because path replication intends to create object copies on the nodes that are topologically along the same path. Both the path replication and the random replication have less overall search traffic than the owner replication.

- *Directed BFS and intelligent search*

The basic idea of directed BFS approach is that the query node sends the query message to a subset of its neighbours that will quickly return many high-quality results. These neighbours then forward the query message to all their neighbours just as in BFS. To choose "good" neighbours, a node keeps track of simple statistics on its neighbours, for example, the number of query results returned through that neighbour, and the network latency of that neighbour.

Based on these statistics, the best neighbours can be intelligently selected using the following heuristics:

- The highest number of query results returned previously
- The least hop-count in the previously returned messages (i.e. the closest neighbours)
- The highest message count (i.e. the most stable neighbours)
- The shortest message queue (i.e. the least busy neighbours)

By directing the query message to just a subset of neighbours, directed BFS can reduce the routing cost in terms of the number of routing messages. By choosing good neighbours, this technique can maintain the quality of query results and lower the query response time. However, in this scheme only the querying node intelligently selects neighbours to forward a query. All other nodes involved in a query processing still broadcast the query to all their neighbours, as in BFS. Therefore, the message duplication is not greatly reduced.

There is also a similar approach called intelligent search. The query type considered is called the keyword query: a search for documents that contain desired keywords listed in a query. A query is represented using a keyword vector. This technique consists of four components: a search mechanism, a profile mechanism, a peer ranking mechanism, and a query similarity function.

When the querying node initiates a query, it does not broadcast the query to all its neighbours. Instead, it evaluates the past performance of all its neighbours and propagates the query only to a subset of its neighbours that have answered similar queries before and therefore will most likely answer the current query. On receiving a query message, a neighbour looks at its local datastore. If the neighbour has the desired documents, it returns them to the querying node and terminates.

Otherwise, the neighbour forwards the query to a subset of its own neighbours that have answered similar queries before. The query forwarding stops when the maximum TTL is reached.

The cosine similarity model is used to compute the query similarity. Based on this model, the similarity between two queries is the cosine of the angle between their query vectors. To determine whether a neighbour answered similar past queries, each node keeps a profile for each of its neighbours. The profile of a neighbour contains the most recent queries that were answered by that neighbour. The profile is created and updated using two schemes. In one scheme, each peer continuously monitors the query and query response message. Queries answered by a neighbour are stored in the profile for that neighbour. In the second scheme, the peer that replies to a query message broadcasts this information to all its neighbours. Neighbours are ranked to facilitate the selection.

- *Local indices based search*

The local indices intend to get the same number of query results as scoped-flooding with less number of nodes processing a query. In local indices, each node keeps indices of data on all nodes within k-hop distance from it. Therefore, each node can directly answer queries for any data in its local indices without resorting to other nodes. All nodes use the same policy P on the list of depths, at which the query should be processed. The nodes whose depths are listed in P check their local indices for the queried data and return the query result, if the sought data is found. These nodes also

forward the query message to all their neighbours, if their depths are not equal to the maximum depth limit. All other nodes, whose depths are not listed in P, just forward the query message to all their neighbours, once receiving it, and do not check their local indices. When the depth limit is reached, the query is terminated even if the query result is not satisfied. Note that all nodes in a P2P system organized using local indices play equal roles. The local indices are updated when a node joins, leaves, or modifies its data.

The local indices approach is similar to iterative deepening. Both broadcast the query message based on a list of depths; however, in iterative deepening, all nodes within the maximum depth limit process the query. In local indices, only nodes whose depths are listed in the policy P process the query. In addition, the iterative deepening approach spreads the query message iteratively with increasing TTL; the local indices approach spreads the query message once with the maximum TTL.

- *Routing indices based search*

Routing indices is similar to directed BFS and intelligent search in that all of them use the information about neighbours to guide the search. Directed BFS only applies this information to selecting neighbours of the querying source (i.e. the first hop from the querying source.) The rest of the search process is just as that of BFS. Both intelligent search and routing indices guide the entire search process. They differ in the information kept for neighbours. Intelligent search uses information about past queries that have been answered by neighbours. Routing indices stores information about the topics of documents and the number of documents stored in neighbours. Routing indices considers content queries, queries based on the file content instead of file name or file identifier. One example of such a content query is: a request for documents that contain the word “networks”. A query includes a set of subject topics. Documents may belong to more than one topic category. Document topics are independent. Each node maintains a local index of its own document database based on the keywords contained in these documents.

The goal of a Routing Index (RI) is to facilitate a node to select the “best” neighbours to forward queries. A RI is a distributed data structure. Given a content query, the algorithms on this data structure compute the top m best neighbours. The goodness of a neighbour is application dependent. In general, a good neighbour is the one through which many documents can be quickly found.

A routing index is organized based on the single –hop routes and document topics. There is one index entry per route (i.e. per neighbour) per topic. An RI index entry, (networks, B), at node A stores information about documents in the topic: networks that may be found through the route (A-> B). This entry gives hints on the potential query result, if A forwards the query to B (i.e. the route A -> B is chosen); hence, the name Routing Index. A routing index entry is very different from a regular index entry. If (networks, B) were the regular index entry, it would mean that node B stores documents in the topic: networks. By organizing the index based on neighbours (routes) instead of destinations (indexed data locations), the storage space can be reduced.

Three types of RIs, compound RI, hop-count RI, and exponentially aggregated RI, are proposed. They differ in RI index entry structures. A compound RI (CRI) stores information about the number of documents in each interesting topic that might be found, if a query is forwarded to a single-hop neighbour.

The goodness of a neighbour for a query in CRI is the number of desired documents that may be found through that neighbour.

- *Attenuated bloom filter based search*

The attenuated bloom filter based search assumes that each stored document has many replicas

spread over the P2P network; documents are queried by names. It intends to quickly find replicas close to the query source with high probability. This is achieved by approximately summarizing the documents that likely exist in nearby nodes. However, the approach alone fails to find replicas far away from the query source.

Bloom filters are often used to approximately and efficiently summarize elements in a set. A bloom filter is a bit-string of length m that is associated with a family of independent hash functions. Each hash function takes as input any set element and outputs an integer in [0, m). To generate a representation of a set using bloom filters, every set element is hashed using all hash functions. Any bit in the bloom filter, whose position matches a hash function result, is set to 1. To determine whether an element is in the set described by a bloom filter, that element is hashed using the same family of hash functions. If any matching bit is not set to 1, the element is definitely not in the set. If all matching bits in the bloom filter are set to 1, the element is probably in the set. If the element is indeed not in the set, this is called a false positive.

Attenuated Bloom Filters are extensions to bloom filters. An attenuated bloom filter of depth d is an array of d regular bloom filters of the same length w. A level is assigned to each regular bloom filter in the array. Level 1 is assigned to the first bloom filter. Level 2 is assigned to the second bloom filter. The higher levels are considered to be attenuated with respect to the lower levels.

To route a query for a file, the querying node hashes the file name using the family of hash functions. Then the querying node checks level-1 of its attenuated bloom filters. If level-1 of an attenuated bloom filter for a neighbour has 1s at all matching positions, the file will probably be found on that neighbour (1-hop distance from the query source). We call such a neighbour a candidate. The querying node then forwards the query to the closest one among all candidates. If no such candidate can be found, the querying node will check the next higher level (level-2) of all its attenuated bloom filters similarly to checking level-1. If no candidate can be found after all levels have been checked at the query source, this indicates that definitely no nearby replica exists. On receiving the query, a neighbour of the querying node looks up its local data store. If the data is found, it will be returned to the query source. If not, this neighbour will check its attenuated bloom filters similarly. During the query processing, if a false positive is found after d (the depth of the attenuated bloom filter) unsuccessful hops, the attenuated bloom filter based search terminates with a failure. No back tracking is allowed.

The attenuated bloom filter approach can be combined with any structured approach to optimize the searching performance. We can use the attenuated bloom filters to try locating nearby replicas. If no nearby replica exists, we switch to the structured approach to continue the lookup.

The hop-count RI is similar to the attenuated bloom filter approach. Both summarize the documents at some distance from the querying source. There are two differences between them. One is that the attenuated bloom filter is a probabilistic approach while the hop-count RI is a deterministic approach if omitting the document change. The other is that the attenuated bloom filter provides information about a specific file while the hop-count RI provides the number of documents on each document category but not a specific file.

- *Adaptive probabilistic search*

In the Adaptive Probabilistic Search (APS), it is assumed that the storage of objects and their copies in the network follows a replication distribution. The number of query requests for each object follows a query distribution. The search process does not affect object placement and the P2P overlay topology.

The APS is based on k-walker random walk and probabilistic (not random) forwarding. The querying node simultaneously deploys k walkers. On receiving the query, each node looks up its local repository for the desired object. If the object is found, the walker stops successfully.

Otherwise, the walker continues. The node forwards the query to the best neighbour that has the highest probability value. The probability values are computed based on the results of the past queries and are updated based on the result of the current query. The query processing continues, until all k walkers terminate either successfully or fail (in which case the TTL limit is reached).

To select neighbours probabilistically, each node keeps a local index about its neighbours. There is one index entry for each object, which the node has requested or forwarded requests for through each neighbour. The value of an index entry for an object and a neighbour represents the relative probability of that neighbour being selected for forwarding a query for that object. The higher the index entries value the higher the probability. Initially, all index values are assigned the same value. Then, the index values are updated as follows. When the querying node forwards a query, it makes some guess about the success of all the walkers. The guess is made based on the ratio of the successful walkers in the past. If it assumes that all walkers will succeed (optimistic approach), the querying node pro-actively increases the index values associated with the chosen neighbours and the queried object. Otherwise (pessimistic approach), the querying node proactively decreases the index values. Using the guess determined by the querying node, every node on the query path updates the index values similarly when forwarding the query.

The index values are also updated when the guess for a walker is wrong. Specifically, if an optimistic guess is made and a walker terminates with a failure, then the index values for the requested object along that walker's path are decreased. The last node on the path sends an update message to the preceding node. On receiving the message, the preceding node decreases the index value for that walker and forwards the update message to the next node on the reverse path. This update procedure continues on the reverse path until the querying node receives an update message and decreases the index value for that walker. If the pessimistic approach is employed and a walker terminates successfully, the index values for the requested object on the walker's path are increased. The update procedure is similar. To remember a walker's path, each node appends its ID in the query message during query forwarding and maintains a soft state for the forwarded query. If a walker A passes by a node, which another walker B stopped by before, the walker A terminates unsuccessfully. The duplicate message was discarded.

Compared to the k-walker random walk, the APS approach has the same asymptotic performance in terms of the message overhead. However, by forwarding queries probabilistically to most promising neighbour(s) based on the learned knowledge, the APS approach surpasses the k-walker random walk in the query success rate and the number of discovered objects.

The APS uses the same guess for all objects. This imprecision causes more messages. Therefore, the swapping-APS (s-APS) constantly observes the ratio of successful walkers for each object and swaps to a better update policy accordingly. The weighted-APS (w-APS) includes the location of objects in the probabilistic selection of neighbours. A distance function is embedded in the stored path of the query and is used in the index update. When the pessimistic guess is made for a walker and the walker succeeds, the index values for neighbours closer to the discovered object are increased more than those for distant neighbours.

- *Dominating set based search*

In this approach, routing indices are stored in a selected set of nodes that form a connected dominating set (CDS). A CDS in a P2P network is a subset of nodes which are connected through direct overlay links. All other nodes that are not in the CDS can be reached from some node in the CDS in one-hop. Searching is performed through a random walk on the dominating nodes in the CDS.

The construction of the CDS uses solely the local information: a node's 1-hop and 2-hop neighbours. The construction consists of two processes: marking followed by reduction. The marking process marks each node in the P2P system as either a dominating node or a non

dominating node. The marker T represents a dominating node, while the marker F represents a non-dominating node. A node is marked using T, if two of its neighbours are not directly connected (i.e. these two neighbours are not neighbours of each other). At the end of the marking process, all nodes with marker T form the CDS. To reduce the size of the CDS, two reduction rules are applied during the reduction process. Each node in the CDS is assigned a 1-hop ranking value.

This ranking value is the sum of the number of documents on a node and the number of documents of the node's neighbour that has the most documents. The first reduction rule specifies that if the neighbours of a node A in the CDS are a proper subset of neighbours of another node B in the CDS and the node A has a smaller 1-hop ranking value than node B, then remove node A from the CDS.

The second reduction rule states that a node C is removed from the CDS, if the following three conditions are satisfied:

1) Two neighbours A and B of the node C are also dominating nodes.

2) The neighbour set of C is a proper subset of the union of the neighbour sets of A and B.

3) The node C has a 1-hop ranking value that is smaller than the values of both A and B.

Searching is conducted on the CDS as follows: if the querying source is not a dominating node, the source forwards the query to its dominating neighbour with the highest 1-hop ranking value. If the querying source is a dominating node, it forwards the query to its dominating neighbour with the highest 1-hop ranking value. This querying source also forwards the query to a non dominating neighbour, if that neighbour has the most documents among all neighbours of the querying source. On receiving a query request, a dominating node looks up its local database for the searched document and performs the query forwarding similarly to a querying source that is a dominating node. On receiving a query request, a non-dominating node only looks up the local database and does not forward the query any further. All found documents are returned from the hosting nodes to the querying source along the reverse query paths. The query stops when the TTL limit is reached or a node is visited the second time.

The dominating set based approach intends to get the greatest number of documents by forwarding queries primarily on dominating nodes, which are well-connected and have many documents themselves or whose neighbours have many documents. The construction of the CDS does not incur more overlay links, as often occurs in super peers. The cost of creating and maintaining the CDS is lower than that of routing indices.

- *Searching in strictly structured P2Ps*

In a strictly structured system, the neighbour relationship between peers and data locations is strictly defined. Searching in such systems is therefore determined by the particular network architecture. Among the strictly structured systems, some implement a distributed hash table (DHT) using different data structures. Others do not provide a DHT interface. Some DHT P2P systems have flat overlay structures; others have hierarchical overlay structures.

A DHT is a hash table whose table entries are distributed among different peers located in arbitrary locations. Each data item is hashed to a unique numeric key. Each node is also hashed to a unique ID in the same key space. Each node is responsible for a certain number of keys. This means that the responsible node stores the key and the data item with that key or a pointer to the data item with that key. Keys are mapped to their responsible nodes. The searching algorithms support two basic operations: lookup(key) and put(key). Lookup(k) is used to find the location of the node that is responsible for the key k. put(k) is used to store a data item (or a pointer to the data item) with the key k in the node responsible for k. In a distributed storage application using a DHT, a node must publish the files that are originally stored on it, before these files can be retrieved by other nodes. A file is published using put(k).

Different non-hierarchical DHT P2Ps use different flat data structures to implement the DHT. These

flat data structures include ring, mesh, hypercube, and other special graphs such as de Bruijn graph. Chord uses a ring data structure. Pastry [uses a tree-based data structure that can be considered as a generalization of a hypercube. A d-dimensional toroidal space is used to implement the DHT in CAN. The space is divided into a number of zones. Each zone is a hyper-rectangle and is taken care of by a node. The zone boundaries identify the node responsible for that zone.

The systems Koorde, Viceroy, and Cycloid have overlays with constant degrees. Koorde embeds a de Bruijn graph on the Chord ring for forwarding lookup requests. The overlay of Viceroy is an approximate butterfly network. The butterfly level parameter of a node is selected according to the estimated network size. Cycloid integrates Chord and Pastry and imitates the cube-connected-cycles (CCC) graph routing. Cycloid performs better than Koorde and Viceroy in large-scale and dynamic P2P systems.

- *Searching in hierarchical DHT P2Ps*

All hierarchical DHT P2Ps organize peers into different groups or clusters. Each group forms its own overlay. All groups together form the entire hierarchical overlay. Typically, the overlay hierarchies are two-tier or three-tier. They differ mainly in the number of groups in each tier, the overlay structure formed by each group, and whether or not peers are distinguished as regular peers and super peers/dominating nodes. Super peers/dominating nodes generally contribute more computing resources, are more stable, and take more responsibility in routing than regular peers. We will focus on Kelips and Coral.

- *Kelips*

Kelips is composed of k virtual affinity groups with group IDs. Inside a group, a file is stored in a randomly chosen group member, called the file's home node. Thus Kelips offers load balance in the same group and among different groups.

- *Coral and related schemes*

Coral is an indexing scheme. It does not dictate how to store or replicate data items. The objectives of Coral are to avoid hot spots and to find nearby data without querying distant nodes. A distributed sloppy hash table was proposed to eliminate hot spots. In DHT, a key is associated with a single value that is a data item or a pointer to a data item. In a DSHT, a key is associated with a number of values which are pointers to replicas of data items.

- *Other hierarchical DHT P2Ps*

In Kelips and Coral, all peers play equal roles in routing. The differences among peers, such as processing power and storage capacity, are not considered. The nodes with more contributed resources are called *super peers*. Otherwise, they are called *peers*. A super peer may be demoted to a peer. A peer may also become a super peer. The system architecture consists of two rings: an outer ring and an inner ring. The outer ring is a Chord ring and consists of all peers and all super peers. The inner ring consists of only super peers.

- *Searching in non-DHT P2Ps*

The non-DHT P2Ps try to solve the problems of DHT P2Ps by avoiding hashing. Hashing does not keep data locality and is not amenable to range queries. There are three big kinds of non-DHT P2Ps: SkipNet, SkipGraph, and TerraDir. SkipNet is designed for storing data close to users. SkipGraph is intended for supporting range queries. TerraDir is targeted for hierarchical name searches. Searching in such systems follows the specified neighbouring relationships between nodes.

- *Searching in loosely structured P2Ps*

In loosely structured P2Ps, the overlay structure is not strictly specified. It is either formed based on hints or formed probabilistically. In Freenet and Phenix, the overlay evolves into the intended structure based on hints or preferences. In Symphony the overlay is constructed probabilistically. Searching in loosely structured P2P systems depends on the overlay structure and how the data is stored. In Freenet, data is stored based on the hints used for the overlay construction. Therefore, searching in Freenet is also based on hints. In Phenix, the overlay is constructed independent of the application. The data location is determined by applications using the Phenix. Therefore, searching in Phenix is application dependent. In Symphony, the data location is clearly specified but the neighbouring relationship is probabilistically defined. Searching in Symphony is guided by reducing the numerical distance from the querying source to the node that stores the desired data.

5.2.4.3. Audiovisual search

Indexing is essential for achieving efficiency in the management and querying of multimedia data. Moreover, index sharing is an essential aspect of the scalability objective, by ensuring a reasonable scaling of network resource consumption by distributed queries. In order to cope with the exponential growth of digital data, scalable and distributed storage structures need to be developed. By dynamically adding new computational and storage resources, such structures would distribute the data so that no centralized nodes are used for both search and maintenance transactions.

Provided enough reliable computational power is available, this approach is able to solve the scalability problem through parallel execution of queries. The performance can even be tuned to the needs of specific applications by load balancing and properly adjusting the capacity of computational resources.

Multimedia features can be indexed by assuming the metric space model of similarity. In this respect, SAPIR proposed four methods for similarity searching based on the P2P communication paradigm, often referred to in the literature as Scalable and Distributed Data Structures (SDDS). Specifically, the first two, designated GHT* and VPT* structures follow the basic generalized hyperplane and ball partitioning principles. The other two apply transformation strategies, where the metric similarity search problem is transformed into a series of range queries executed on existing distributed hash tables (DHT), for exact (range) matching over traditional attribute-like data. Following the well known designations of the underlying structures, they are called the MCAN and the M-Chord.

Each of the four structures is able to execute similarity queries for any metric and they all exploit parallelism during query processing. All of them have experimentally been implemented over the same computer network and tested on several synthetic and real-life datasets. Preliminary results are very encouraging and basically confirm the hypothesis of constant scalability of such implementations. SAPIR aims at defining standard APIs for connecting and querying the distributed indices.

The main objective is to achieve multi-feature similarity ranking based on P2P similarity indices developed for single features. The basic lesson learned is that the similarity score (or grade) a retrieved object receives as a whole depends not only on the scores it gets for individual predicates, but also on how such scores are combined. All these aspects influence the query execution costs. In order to understand the problem, consider a query for objects with circular shapes and red colour. In order to find the best match, it is not enough to retrieve the best matches for the colour features and the shapes. Naturally, the best match for the whole query need not be the best match for a single (colour or shape) predicate. To this aim, Fagin has proposed the so-called A0 algorithm that solves the problem. There have been several extensions of this work, but they don’t deal with similarities over different medias. They do not consider distributed environments as well.

Complex similarity query execution over multiple distributed single-feature overlays represents an important challenge of SAPIR, because a naïve solution might result in overwhelming increase of communication costs in the underlying computer network. In principle, our approach will be based on the incremental nearest-neighbour algorithm executed on individual peers, coordinated by a modified Threshold Algorithm (TA) to efficiently obtain the global result.

In particular, SAPIR will exploit properties of our P2P overlay networks, which pose some difficulties for a distributed execution of complex similarity queries, but at the same time they offer new structural properties that can be for such query execution exploited. Supposing multiple single feature overlays over the same physical P2P network, the routing processes of individual overlays can take advantage of sharing paths or at least some parts of them. At the same time, once a peer with potentially qualified items of feature one is reached and the items tested, the peer can also test the relevance of items belonging to feature two, provided they are derived from the same object. Naturally, this can be generalized to an arbitrary number of features. Such architecture can capitalize on independence of peers resulting in parallel query execution.

5.3. Mobile search

5.3.1. Introduction

Mobile search is the means people use on the ir portable devices to find content on or off portal directly by browsing or by entering a search query via the mobile version of an online Internet search engine, or by using a specialized mobile search function provided by an operator or other service provider and usually based on a white-label solution. This section is primarily focused on search functionality rather than browsing.

Some people argue there is no difference between mobile search and traditional search. Others think there are substantial differences in the way results are presented to the user, essentially because of constraints on the size of the screen. Personalisation and localisation of mobile devices are also other important points to address. In this section we define some areas where these differences appear significantly in search. We will also assume the searched content isn't specifically mobile, i.e. the search concerns the regular web. As P2P mobile isn't a lot addressed in the research, P2P mobile search will be defined in a further deliverable and not in this section.

- *Personalized search and Context information*

A mobile device is indicative of personalized services offered to each user. Mobile search can be personalized taking into account both the device characteristics (screen analysis, memory capabilities, applications installed), as well as the user's history and the user's contextual information. Personalized search in mobile environments has the advantage of focused results to match the user's interests, as well as limiting the amount of results to cope with, considering the limited mobile capabilities concerning memory, bandwidth and processing power.

As far as context is concerned, factors such as time and location can be employed to assist the mobile search. Better results can be yielded that are more relevant to the user's general interests, but also to their short term interests. For example, for a user searching for musical concert tickets, the system should take into account the user's location (country, city) and either fetch results with concert tickets in that location or present the "local" results on top on the list.

- *Results page layout*

Mobile search results typically render the results in one long column as opposed to the multiple column layout that is often used to present traditional search results on PC browsers. Consequently, this makes different types of results, such as sponsored links, harder to spot, even when they are labelled, because they appear inline with the ordinary results. In an attempt to improve the usability

and appeal of their product, many mobile search engines design their search engine like a portal, with links directly to specific information. This reduces the amount of typing necessary for the user to find what he or she is looking for.

- *Local & vertical results*

The major mobile search engines are competing to create the best user experience possible. In many instances, doing so involves the search engines surmising the user's search goal and presenting the user with those specific search results first. For that reason, mobile search engines put a higher focus on local and vertical (classical) results, frequently featuring them much more prominently than traditional web results. These can include: maps, local results, links to official sites, images, weather and even sports scores. These results are even more important to consider in the mobile web, because of their premium placement on limited mobile results pages.

- *Character limits*

As you might expect, mobile search results are frequently truncated versions of what would normally appear in the traditional results page. If you are optimizing a mobile-specific site, there is a whole new set of character limits to work with when optimizing metadata. If you are optimizing an existing site to be found in both mobile and traditional search, you should abide by the character limits in traditional search, while at the same time remaining conscious of what will be omitted in the mobile search results.

- *URL display*

In traditional search results, complete URLs are always provided for each search result, but this is not always the case in mobile search engines. Some mobile search engines will eliminate the 'http://' from the URL, or display only the domain in the search results, even though the result links to a deeper page on the site. Optimized sub-domains can be very useful in traditional SEO, but might be even more useful in mobile search engines, when everything after the domain extension (.com/.net/.co.uk etc.) is eliminated. Since savvy users sometimes evaluate display URLs to determine which result they will click on, the architecture of the URL can be used to influence that decision. To make this more concrete, consider a person looking for the results of a football game on a mobile phone. Which URL seems like it is the most likely to get you the information in the fewest number of clicks:

A ESPN.com

B NFL.ESPN.com

C Football-Scores.ESPN.com

D FootballScores.com

The correct answer is likely a tie between options 'C' and 'D.' While ESPN is clearly an authority site, FootballScores.com and ESPN.com may lure some viewers away because of their simplicity. Optimized sub-domains are a good idea in some cases, but even in mobile SEO they are not always the best option. In some instances, users are more likely to click on simpler URLs, and other times they are not.

- *Recommendation*

Terms recommendation or results recommendation is employed in many known search engines. In mobile environments, this could be useful in order to save users time from typing. Recommendation

can either be used in terms of collaborative filtering, where recommended results are produced based on what other users have searched for, when searching for a specific concept, as well as from the user's past behaviour in terms of a history log.

5.3.2. Context

There is a great diversity in mobile search engines. While the goal of all the mobile engines is the same, their approaches vary considerably. In this section we will present their main differences and the impact of these differences on Search engines optimisation.

- *Presentation of results*

One of the more frustrating differences between the mobile search engines is the number of results they present on the main results page, and the number of results that they will present on the secondary 'web results' page. Since mobile search engines are designed more like portals rather than traditional search engines, they have all come up with a variety of ways of presenting the information that is yielded from a search result. This can be handy for users, but makes tracking and comparison a bit trickier.

In general, mobile search engines provide vertical results, ordered by relevance. Windows Live provides two mobile web results on the main results landing page, Google Mobile and AOL Mobile provide six, and Yahoo provides ten. An exception is Google iPhone, which presents eight web results but providing tabs along the top if the user needs to access local or vertical results.

- *Search box location*

The AOL mobile landing page provides a search box at the top and bottom of the page, but only on the bottom of the results page. Conversely, Yahoo OneSearch provides a search box at the top of the landing page, and a search box at the bottom and the top of the results page. Windows Live provides one search box at the top of the search landing page, and one at the bottom of the results page. Google iPhone provides only one search box at the top of the landing page and the top of the results page.

- *Local & vertical results*

Some mobile search engines, like AOL and Google iPhone will break local and vertical results into different tabs along the top of the page. Others present a mixed landing page with vertical results such as maps, weather forecasts, images and sports scores provided inline with web results. Google Mobile and Yahoo OneSearch both maintain results pages where the main focus is web results, but they do integrate some vertical results inline with web results. Conversely, AOL Mobile and Windows Live both provide mixed results that do not focus on any particular type of result.

- *Location setting*

It won't be long before GPS enabled mobile devices set and update a user's location automatically, but for now setting your location is still a manual process. While Google Mobile, AOL Mobile and Windows Live all allow you to set your location, Google iPhone and Yahoo OneSearch do not. Google and AOL Mobile both have options on the main search page to change your location. Google Mobile will allow you to set your location by city or zip code, but AOL takes it a step further and lets you specify your location down to the street address.

Windows Live does not have links on the main search page to change your default location; instead, they update the user's default location whenever the user searches for a specific geographic location, so if your default location is set to Denver, but you want information about a restaurant in

Houston you can search for 'PapaMia Houston' and your default location will be updated to Houston for subsequent searches. Unfortunately, there are no options or instructions for changing the default location on the main search page, so users are left to figure this out on their own.

Location settings can impact the local and vertical results that you are presented, and in the future may also affect the mobile web rankings as well. Currently, Google, AOL Mobile and Windows Live are tailoring the local and vertical results by the user's default location, but are not tailoring web results by location.

- *Keyword bolding*

Traditional search engines will sometimes put the keyword(s) that you have searched for in bold to help your eye key into the most relevant results. Most of the mobile search engines, (all but Windows Live) have adopted this practice to varying degrees as well. Yahoo OneSearch will bold keywords in the title line, description and URL, while all of the Google driven engines, including Google Mobile, Google iPhone and AOL Mobile will only bold terms when they are located in the description part of the results. Windows live is the only engine evaluated that is not bolding any keywords in search results pages.

- *User agent detection*

Currently, Google Mobile, AOL Mobile and Microsoft OneSearch incorporate user agent detection to determine exactly what type of mobile device you are using to access their search engine. They will then use that information to optimize the results pages for viewing on your specific mobile device. This is done primarily to ensure images, maps and other graphics to are sized to fit the screen without right-to-left scrolling. In the future, this information could be integrated into the search algorithm to improve the ranking for pages that display well on your specific mobile device.

- *Transcoding*

Google Mobile, AOL Mobile and Windows Live all integrate transcoding software to re-arrange web pages that are designed for the traditional web and to make them viewable on a smaller screen. This is good news for sites that have yet to begin optimizing the user experience for the mobile web, but can also cause problems. Forms or JavaScript may be rendered un-usable on the transcoded version of the site, and the transcoded page may not provide adequate idea arrangement of the elements on the page.

While transcoding improves the usability of the site in the short term, it may hinder SEO and can make interacting with the site more difficult. The transcoded page is hosted temporarily on the search engine server and domain, rather than on the original website. It is unclear weather transcoding impacts Google's evaluation of the activity on your site, but it definitely makes it harder to get accurate links to the site because the URLs are re-formulated in the transcoding. Many of the mobile search engines have indicated that they recognize the 'handheld' style sheet, and will use it to render the site when it is available, but it is not always the case. In all cases, you can choose to view the html version of the site by clicking on a link at the bottom of the page, or simply performing your search in the traditional version of the search engine, rather than the mobile version.

- *the Impact on mobile Search Engine Optimizer?*

All of the differences that we can see amongst the mobile search engine players are simply an indication that the industry is still in its infancy, and has yet to develop standards. Mobile search engines are still determining how they can provide users with the best experience, and SEOs are still figuring out how to compare such variable results. The main conclusions that can be drawn is

that mobile SEO is different from traditional SEO, but not so different that everything must be re-learned. Mobile SEOs must be patient for the mobile web and the mobile search experience to catch up with the traditional web that we have become so used to. It is an exciting time in mobile search, when things are constantly changing, standards are slowly being formed and nothing is taken for granted.

5.3.3. Main players

For the reasons explained in the last chapter the mobile search market is very fragmented.
Expectations for mobile search and local mobile search in particular are rising. As mobile ad networks form, mobile M&A activity heats up and the search engines pour greater attention and resources into their mobile offerings. One could say we are on the cusp of a new mobile era. Indeed, as much as we can be reluctant to use the term, one could dub the forthcoming mobile Internet "Web 3.0."

Of course people have been saying and predicting the emergence of the mobile Internet for almost 10 years. Forecasts and predictions rarely come true in their original time frames, but they typically do come true eventually. And today, the resources, infrastructure and consumer demand make a mobile Internet more tangible and much closer to reality.

What took the desktop Internet roughly a decade to develop is happening in a much more condensed period of time in mobile. And for all its complexity and fragmentation, there are numerous companies working on making content access and delivery on mobile devices a much more intuitive and user-friendly experience. User experience is the key to mobile services, because once users adopt the mobile Internet (or variations thereof) in meaningful numbers, which is starting to happen, the ad dollars will flow and real money will be made.

Right now the "mobile Internet" is really four separate areas that will eventually blend to varying degrees. Each of the four areas has big players we will try to classify in.

- *Nouveau Directory Assistance & Voice Search*

This category grows out of tried and true "directory assistance," the original form of local mobile search. In 2006 there were roughly 6.5 billion calls to 411 in the United States and many more billions around the world. Because of the Internet and other factors (e.g., corporations blocking 411), directory assistance continues to shift to mobile phones.

So-called "operator assisted yellow pages" (live agents helping users finding listings and other information) were repeatedly tried and failed. However, today, ad-supported directory assistance appears here to stay. Eg mobile.Yell.com, V-enable.com,180srch.com.

- *Text-Based Local Search*

After directory assistance and its more sophisticated cousin voice search, the volume of usage in text messaging. Depending on whose numbers you believe, anywhere from 35 percent to 70 percent of U.S. mobile consumers send and receive text messages (with varying degrees of frequency). This is clearly where the volume of mobile data usage is today, as opposed to WAP browsing. However, text is arguably the least sexy mode of accessing information on a mobile device (if the most practical). One of the leaders in this category is 4Info, which is doing some impressive things and getting some very impressive CPM rates. The company, partly owned by newspaper publisher Gannett, is not exclusively about local but local is an important piece of what it's doing. And many of the voice search options in the first segment allow content and contact details to be received via text message in addition to audio. Eg 4info.com, ask.com, nownow.com.

- *WAP Local Search*

WAP usage, while numbering in the millions is still in an early stage of development and has much less adoption for many reasons, including hardware limitations, separation of text and mobile Internet pricing plans and so on. All the major search engines and portals, yellow pages sites and local search pure plays now have WAP sites.

Yahoo's oneSearch is something of a standout in this category. Among others are wapreview.com and wapmcnearky.com.

- *Local Mobile Applications*

All major search providers also have downloadable applications, many of which are being pre-loaded on phones.

Then there are interesting alternative content and search applications, represented by the Where and ZenZui "platforms."

Applications offer by far the best and richest user experience. The problem for search engines (and users) is that they must be downloaded and so represent the smallest segment of the market with intrinsic barriers to adoption. Thus the challenge is to get applications preloaded on the next phone the user buys and/or to bring the application experience into a WAP environment.Eg maporama, local.com, mojopages.com.

- *Bringing It All Together*

Google's "diversified" approach is a metaphor for the challenges and fragmentation of the mobile market right now: the company has an offering in each of the above segments. There are numerous other companies, including Microsoft and Yahoo that have comparable offerings in most or all of the segments.

The mobile market, just because of the proliferation of different handsets, will always be fragmented to some degree. But we can expect to see increasing integration of the types of functionality that are currently largely separated -- the blending of voice interfaces, text and WAP and, potentially, applications that come preloaded on phones (e.g., Google Maps on the iPhone).

Speaking of the iPhone, it has done a great service to the market, refocusing the industry on the user experience and general usability in mobile. Consumers fundamentally want local information on the go and thus consumer demand is "pent up." Mobile usability and the "mobile Internet" now just have to catch up to the consumer.

5.3.4. The State of the Art

As the mobile user base expands, so do device storage capacities and wireless services. Not only are these phones accumulating more device-resident data such as email, appointments and photos, but they are also increasingly used as front-end interfaces to ever-larger external data sets, including web sites, traffic information, and Yellow Pages data. Many query-answer systems and web browser interfaces that target mobile platforms have debuted within the last year, including offerings from every major search engine.

In many systems in the literature, emphasis is shown on interfaces that serve search engine services in mobile devices. While existing solutions do cater to small screens and low bandwidth, they are modelled after desktop web search, posing three primary usability issues for the mobile setting. First, they rely on text entry as the method of input, even though the persistent trend toward smaller phones is directly at odds with the goal of achieving the efficiency of full-size keyboard text entry.

Second, the focus has been on search off the device, under-utilizing the device's expanding processing power and storage capabilities and thus unnecessarily impoverishing the search UI. Finally, both the SMS and web search models support directed search tasks, but are less appropriate for browsing and exploratory search scenarios ("sense-making") that are quite complementary to the mobile setting.

- *Interfaces*

Many information access interfaces present data attributes (metadata) that users can include in queries to large data sets, rather than expecting users to remember them. Dynamic query interfaces (Shneiderman) encourage iterative composition and refinement of complex queries by providing continuous visual update of the query results as users restrict data attributes included in the query.

Standard bookmarks and saved queries (De Luca et al.) help speed page revisitation, but most systems rely on device-specific text entry methods for ad hoc keyword search. Word prediction and completion algorithms have the potential to reduce the number of entered characters, but also have the drawback that most fail for non-dictionary words, and may still require users to select from several alternatives.

Karlson et al. present a novel approach for searching large data sets from a mobile phone. Existing interfaces for mobile search require keyword text entry and are not suited for browsing. They propose an alternative approach which uses a hybrid model that is based on iterative data filtering rather than on the tedious keyword entry. More specifically their approach involves navigation and selection of hierarchical metadata (facet navigation) with incremental text entry to further narrow the results. Information seeking strategies take many forms depending on the task at hand, user knowledge, and target data set. According to the task of search, they provide two definitions: a directed search is one in which the user knows the precise target in advance, while a browsing task is one characterized by specifying criteria that describe a data need, and which may evolve during search. They mention that such a task can assist in selecting between the traditional keyword search and their own facets-based approach. As far as facets are concerned, the use of data attributes (metadata) can be organized into orthogonal dimensions (facets) as a means not only to structure search results, but as a tool to guide users in formulating powerful Boolean queries. This approach not only reduces cognitive load through recognition, but allows users to reliably restrict results by attribute values rather than by keyword alone.

Their structural philosophy is counteracting the limitation of most mobile phones concerning the lack of touch screens. Their system, FaThumb, is optimized for keypad interaction. The Facet Navigation region is intentionally designed to map spatially to numbers 1 though 9 on the numeric keypad. While this design restricts the branching factor of the hierarchy (with a maximum 8 at each level), its depth and balance are dictated only by the target data set. For any domain, we believe consistency of facet location is crucial to promoting user mastery of the hierarchy. Thus they opted for a terminating tree, meaning users must return to the top of the tree to explore paths that lead from other top-level facets. On the other hand, as data sets grow, it may be appropriate to dynamically generate nodes within the facet tree to provide more efficient distribution of data among the available zones (e.g., consider date facets labelled by day for high frequency data, but by year for low frequency data). Dynamic strategies may be most effective at lower levels of the tree, since users may be more willing to browse a less familiar but more useful set of choices once they have already narrowed the data set by a few familiar initial selections.

An interesting work is done by Google. Kamvar and Baluja present a study of search patterns on Google's mobile search interface. They examine search queries and the general categories under which they fall. Useful conclusions for mobile search interface can be drawn by observing users interaction, referring to the time users spend inputting a query, viewing the search results and how often they click on a result. They provide insight through large scale log analysis. A comparison between Google XHTML and PDA interfaces takes places, where it was found that the number of

keywords forming a query is quite similar between desktop, PDAs and XHTML, while mobile users are a bit briefer. Concerning the categories of interest these seem to be more in the entertainment category (ringtones, adult content, celebrity searches) in the case of XHTML, since users consider their cell phone as a more personal and private device, whereas PDAs topics of search tend to be more business-oriented. The click-through rate across all categories was consistently low which suggests users are relying heavily on snippets in wireless search for their information. They believe users requesting the search results from the same query may be confusing the "Search" button for the "Next" link. The next link on the wireless page is much smaller and shown with much less context than its desktop equivalent. This in-depth examination of wireless search patterns seems suggests that the search interface should be alter ed concerning the design of mobile search engines interfaces.

- *Community-based search*

Church et al. point out that limited screen real-estate and restricted text input capabilities, from which mobile devices suffer, affect the usability of many mobile Internet applications. Most attempts to provide mobile search engines have involved making only simplistic adaptations to standard search interfaces. For example, fewer results per page are returned and the 'snippet' text associated with each result may be truncated. They attempt to deal with the snippet text issue proposing the I-SPY system. The I-SPY system can track and record past queries that have resulted in the selection of a given result page and we argue that these related queries can be used to help users understand the context of a search result in place of more verbose snippet text.

More specifically, the I-SPY search engine focuses on community-based search by recording the search histories (queries and result selections) of communities of like-minded individuals. Their concept can actually be considered to belong to the broader machine learning technique of collaborative filtering, as was mentioned in the introductory section concerning mobile search. This information is stored in a query-result hit-matrix that records the number of user selections that a result p_j has received in response to a query q_i and the information is used to adapt future result-lists for similar queries by promoting results that have been selected in the past. Thus, I-SPY gradually adapts to the learned needs of communities of individuals and this has been previously shown to significantly improve overall search performance.

As a conclusion, mobile internet search engines need an economic way to summarise the contents of their search results. Traditional snippet text is simply too verbose. In the I-SPY system the suggestions made include using previously successful queries as an alternative and provision of some preliminary empirical evidence that implies that these queries may be as informative as snippet text. These resultant queries take up less than half the space of snippet text and can also be used as a simple way for users to launch further more elaborated searches. All of these benefits suggest that related queries could be quite valuable in the mobile search domain.

- *Results Classification*

Hierarchical classifications have been used previously in search interfaces. Search results that include hierarchical labels can help users identify relevant items or further refine (or broaden) searches. Search engines such as Yahoo and OpenDirectory order results by relevance, but display each with a human-assigned hierarchical category; following the category hyperlink allows users to browse via the hierarchical category directory. Other systems help users quickly identify appropriate subsets from voluminous results by organizing results into hierarchical categories.

Nadamoto et al. deal with the problem of web results organization introducing a way of restructuring web search results for passive viewing. Their restructuring method is to classify search results dynamically into several groups, which they call carousels. By analyzing Web pages, their method classifies web pages of search results into four groups, based on similarity, difference, detail, summary relationships. The user can select a carousel and transit among carousels by a few

interactions. The contents of each carousel are presented automatically and repeatedly. The user watches and listens to a carousel and does the simple interaction for a carousel transition.

- *Contextual search*

All of the above research work seems to focus on the technical limitations of mobile devices and their viewing capabilities. Most approaches develop methods of smart interfaces with navigation schemes or different type of information used for viewing that supplements the viewing of traditional search engines. Another issue is raised by Flanagan et al., where the exploitation of the user's context seems to be an additional step so that mobile search engines make the difference.

The user's context refers to information related to the situation of an entity (person, object, place) relevant to the interaction between a user and an application:

- The user's profile (explicit/implicit preferences, current activity in the device, history).
- The user's general activity (location, date-time, orientation, acceleration).
- The user's environment (temperature, humidity, sound, light).
- The user's social environment (nearby people, current social situation).

Flanagan et al. propose an approach for using user's context for adapting the mobile device profile. In this approach, the low-level context is captured through on-board sensors in the mobile device (user's physical environment is directly monitored). More specifically, they monitor the user's activity, i.e. orientation, stability, acceleration, in hand and environmental conditions, such as ambient or artificial illumination, noise, air humidity, temperature. The user does not provide any explicit feedback. The signals obtained from the sensors are recorded and various feature extraction algorithms are applied to generate low-level context. The low-level information is used to determine a higher-level context in a context hierarchy. The representation of low-level context is made using symbols. The generation of higher-level context is base on clustering the symbols (fusion from various sources) with a Symbol String Clustering algorithm. A profile of the device is constructed and is adapted automatically as the user passes through different contexts. Thus, the mobile device responds to the user's context by changing its profile according to it.

A context-aware mobile application on mobile devices for mobile users is implemented by Coppola et al. Their system is constructed based on a distributed architecture for sending mobile application to the mobile device using context. Their system is constituted by two main modules: the MoBeSoul module which captures and handles user context, and a MoBeLet application, which is downloaded and executed on the mobile device. The concrete context is captured through physical sensors (noise, light level, temperature), "virtual" sensors (date, time, alarm time), explicit user actions (communication, profile selections) and context History (previous user's actions). An inferential mechanism is implemented to derive abstract context (higher-level) using the concrete context (low-level). Both contexts have a probability measure representing how likely they are for the user. Contexts are divided to public (user's approximate location) and private (user's exact position, or other personal data). The automatically collected context, along with user's demographic information and explicitly denoted user preferences are stored into databases (User Profile and Usage & Download Statistics). Then, the public context descriptors are transmitted to the MoBe Descriptors Server. The MoBe Descriptors Server selects the MoBeLet applications that are more relevant to the user's context. The descriptors of the selected MoBeLet applications are transmitted to the mobile device and are filtered using the private context descriptors. The finally selected MoBeLet applications are downloaded to the mobile device in order to be executed.

A client-server software architecture, implemented in the mobile terminal, for adapting the mobile device profile, enhancing mobile applications and sending appropriate information to the mobile device by exploiting the user's context is presented by Korpipää et al. The context information they take into account is provided by: sensors for capturing sound, light, acceleration, temperature,

touch, etc, applications currently running, time information, or explicit user actions such as scheduled tasks, preferences, and network resources for communication activities. The captured context is processed for the extraction of the useful features (Resource-server). The extracted information is represented as concepts in a contextual ontology consisting of a Schema representing the structure and properties and a client usable, extendable vocabulary for describing context information (Context manager). Each context expression contain a type and value features. The low-level contexts participate in a reasoning process (using naïve Bayes classifier) for generating higher-level contexts. During application of this system, the device profile is changed, or the currently running application is enhanced according to the generated high-level context.

Apart from the results representation, context search offers improvements on the basic functionality of mobile search engines. The searching process is also affected in the work of Su and Lee. They refer to an approach which adapts the searching process according to the user's active context. In order for the retrieval results to be more related to the user's active task, the query is expanded by terms from the active document. The text processing method selects from the document the top ranked N words to expand the query. The search engine retrieves results with high similarity to the active documents besides of the query itself. They have experimented for investigating if similar (in terms of words) documents to the active user context are equal to useful documents. The evaluation methodology included the ranking of similar documents in comparison to a user provided document (pseudo context). They included different modes of query formulation: 5 keywords of current research area and topic, 5 keywords of the methods-algorithms in the topic and random combination of the two sets. The search results were the top ten ranked documents. The users were asked to manually rate the documents in terms of similarity, relevancy and usefulness. In the evaluation results it was shown that users judged similarities very differently to the program scores. In most cases users judged the similarity, relevance and usefulness of a document to a similar level. Documents with low similarity and high usefulness could provide users with additional knowledge. Documents with low usefulness but high similarity of words did not carry the key information (in the words) that the users were looking for.

5.4. References

A. Crespo, and H. Garcia-Molina, "Routing indices for peer-to-peer systems", Proc. of the 22nd International Conference on Distributed Computing (IEEE ICDCS'02), 2002.

Baeza-Yates, R. Castillo, C. Junqueira, F. Plachouras, V. Silvestri, F. Challenges on Distributed Web Retrieval, ICDE, Istambul, 2007

Barroso, Dean, and Hoelzle, Web search for a planet: The Google cluster architecture, Micro IEEE, 2003.

Bender, Michel, Triantafillou, Weikum, *Design Alternatives for Large-Scale Web Search: Alexander was Great, Aeneas a Pioneer, and Anakin has the Force*, SIGIR Workshop on Large Scale Distributed Systems for IR, 2007.

Bhattacharjee, Chawathe, Gopalakrishnan, Keleher and Silaghi*, Efficient Peer-To-Peer Searches Using Result-Caching*, IPTPS, 2003.

Bragante and Melucci*, Homepage Finding in Hybrid Peer-to-Peer Networks, 8th RIAO Conference on Large-Scale Semantic Access to Content*, 2007.

Brin and Page, The anatomy of a large-scale hypertextual Web search engine, 1998.

Brown Parallel and Distributed IR in Modern information retrieval Baeza-Yates and Ribeiro-Neto, 1999.

B. Yang, and H. Garcia-Molina, Improving search in peer-to-peer networks, Proc. of the 22nd IEEE

International Conference on Distributed Computing (IEEE ICDCS'02), 2002.

Cacheda, Carneiro, Plachouras, Ounis, *Performance analysis of distributed information retrieval architectures using an improved network simulation model*, Information Processing Management, 2007.

Callan, Distributed IR, 2000.

Church, K., Keane, M. T., and Smyth, B. An Evaluation of Gisting in Mobile Search. In Proceedings of the 27th European Conference on Information Retrieval, Santiago de Compostela, Spain, 2005.

Coppola, P., Della Mea, V., Di Gaspero, L., Mizzaro, S., Scagnetto, I., Selva, A., Vassena, L., Riziò, P.Z. *MoBe: Context-Aware Mobile Applications on Mobile Devices for Mobile Users*, Proceedings of the 1st International Conference on Exploiting Context Histories in Smart Environments (ECHISE2005), Munich, 2005.

D. Tsoumakos, and N. Roussopoulos, "Adaptive probabilistic search in peer-to-peer networks", *Proc. of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS'03)*, 2003.

D. Tsoumakos, and N. Roussopoulos, "*Adaptive probabilistic search in peer-to-peer networks*", technical report, CS-TR-4451, 2003.

D. Tsoumakos, and N. Roussopoulos, "*A comparison of peer-to-peer search methods*", Proc. of 2003 International Workshop on the Web and Databases, 2003.

De Luca, E.W. and Nurnberger, A. *Supporting information retrieval on mobile devices*. Proc. Mobile HCI, 347-348, 2005.

Flanagan, J. A., Himberg, J. and Mäntyjärvi, J. A *Hierarchical Approach to Learning Context and Facilitating User Interaction in Mobile Devices*, Artificial Intelligence in Mobile System 2003 (AIMS 2003) in conjunction with Ubicomp 2003, Seattle, USA, 2003.

Kamvar, M., and Baluja, S. *A large scale study of wireless search behavior: Google mobile search*, Proceedings of the SIGCHI conference on Human Factors in computing systems, , Montréal, Québec, Canada, 2006.

Karlson, A., Robertson, G., Robbins, D., Czerwinski, M. & Smith, G. FaThumb*, A facet-based interface for mobile search*. Proc. CHI 2006, ACM Press, 711—720, 2006.

Korpipää, P., Mäntyjärvi, J., Kela, J., Keränen, H., Malm, E-J. *Managing Context Information in Mobile Devices*, IEEE Pervasive Computing Vol. 2, Is. 3, pp. 42–51, July-Sept, 2003.

Li, Loo, Hellerstein, Kaashoek, Karger and Morris, *On the Feasibility of Peer-to-peer Web Indexing and Search*, 2003.

Long and Suel, *Three-Level Caching for Efficient Query Processing in Large Web Search Engines*, WWW Conference, 2005.

Lu and Callan, *Content-based peer-to-peer network overlay for full-text federated search*, 8th RIAO Conference on Large-Scale Semantic Access to Content, 2007.

Luu, Klemm, Podnar, Rajman, Aberer*, ALVIS Peers: A Scalable Full-text Peer-to-Peer Retrieval Engine*, P2PIR, 2006.

M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer, *"P2p content search: Give the Web back to the people,*" February 2006, international Workshop on Peer-to-Peer Systems (IPTPS).

Meng Yu, and Liu, *Building efficient and effective metasearch engines*, ACM Computing Surveys, 2002.

Moffat, Webber, Zobel and Baeza-Yates, *A pipelined architecture for distributed text query evaluation,* 2007.

Nadamoto, A., Kondo, H., and Tanaka, K. *WebCarousel: Restructuring Web search results for passive viewing inmobile environments*, Proceedings of the Seventh International Conference on Database Systems for Advanced Applications (DASFAA 2001), Hong Kong, China, 2001.

Parreira, Michel, Bender*, Size Doesn't Always Matter: Exploiting PageRank for Query Routing in Distributed IR*, P2PIR, 2006.

Puppin, Silvestri and Laforenza, Query-Driven Document Partitioning and Collection Selection, Infoscale 2006.

Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker, "Search and replication in unstructured peer to-peer networks", Proc. of the 16th ACM International Conference on Supercomputing

(ACM ICS'02), 2002.

S. C. Rhea, and J. Kubiatowicz, "Probabilistic location and routing", *Proc. of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'02)*, 2002.

Shneiderman, B. *Dynamic queries for visual information seeking.* IEEE Software, 11, 6 (1994), 70-77.

Skobeltsyn, Aberer, *Distributed Cache Table: Efficient Query-Driven Processing of Multi-Term Queries in P2P Networks*, P2PIR, 2006

Su, J. and Lee, M. *An Exploration in Personalized and Context-Sensitive Search*, Proceedings of the 7th Annual UK Special Interest Group for Computational Linguists Research Colloquium, 2003.

Tang and Dwarkadas*, Hybrid Global-Local Indexing for Efficient Peer-to-Peer Information Retrieval*, NSDI'04, 2004.

V. Kalogeraki, D. Gunopulos, and D. Zeinalipour-yazti, *A local search mecha nism for peerto-peer networks*, *Proc. of the 11th ACM Conference on Information and Knowledge Management (ACM CIKM'02)*, 2002.

Wang, Reinders, Lagendijk, Pouwelse, *Self-organizing distributed collaborative filtering*, SIGIR, 2005.

Wray Buntine, *Open source, distributed and Peer to Peer IR,* ESSIR 2007

Xiuqi Li and Jie Wu, "*Searching Techniques in Peer-to-Peer Networks*", CRC press, 2005.

Zhang and Suel, *Optimized Inverted List Assignment in Distributed Search Engine Architectures*, IPDPS, 2007.

Zhang and Suel, *Efficient Query Evaluation on Large Textual Collections in a Peer-to-Peer Environment*, IEEE International Conference on Peer-to-Peer Computing, 2005.

6. ECONOMIC AND SOCIAL ASPECTS OF SEARCH ENGINES

The purpose of this chapter is to present a number of socio-economic aspects and pinpoint issues of interest for further investigation by IPTS in the second year. It is not intended to give a full overview of all socio-economic aspects. Next period's work will be devoted to understanding the details of the business models from which we intend to derive some pathways for the future and understand their policy implications.

6.1. Introduction

Major search engine providers are large multinationals, offering far more than just a search tool for internet surfers. Google, Yahoo! and Microsoft's MSN Live search have introduced and continue to propose a series of services. Key elements of their ´core business´ and major adjacent services are sketched in Figure 2. The free email accounts, instant messages, and voice over IP services of these multinationals are communication that both complement and compete with traditional ways of communication. Search engine providers are also owners of popular social network sites, like YouTube or Flickr, whose members not only upload large amounts of audio-visual content, but they do also classify (tag) and filter information (e.g. ranking by voting). Making use of social networks, search engine providers get control over huge amounts of structured and unstructured audio-visual content. Although not all this content is valuable as resource for (semi-)automatic tagging and mark-up and further processing, such content together with proprietary content could be packaged and specifically delivered to users. An example how search engines can act as information providers is news syndications. News syndication can be generated automatically by search engines, like Google News or Yahoo News, or in combination with human expertise. Companies, like the Finish M-Brain,[10] use a search engine to pick articles from the internet. Media analysts and experts then select the relevant information, summarize it, and provide it to the clients.

Summarizing, search engine providers have a pivotal role for the digital economy and knowledge society not only because of their famous search tools and because of running huge online advertising business, but also because of their role as enablers of content creation, as information providers and as communication facilitators. These roles are intertwined. The purpose of this paper is not to discuss this complex structure in detail, but rather to discuss some particular socio-economic issues.

[10] http://www.m-brain.fi/english/

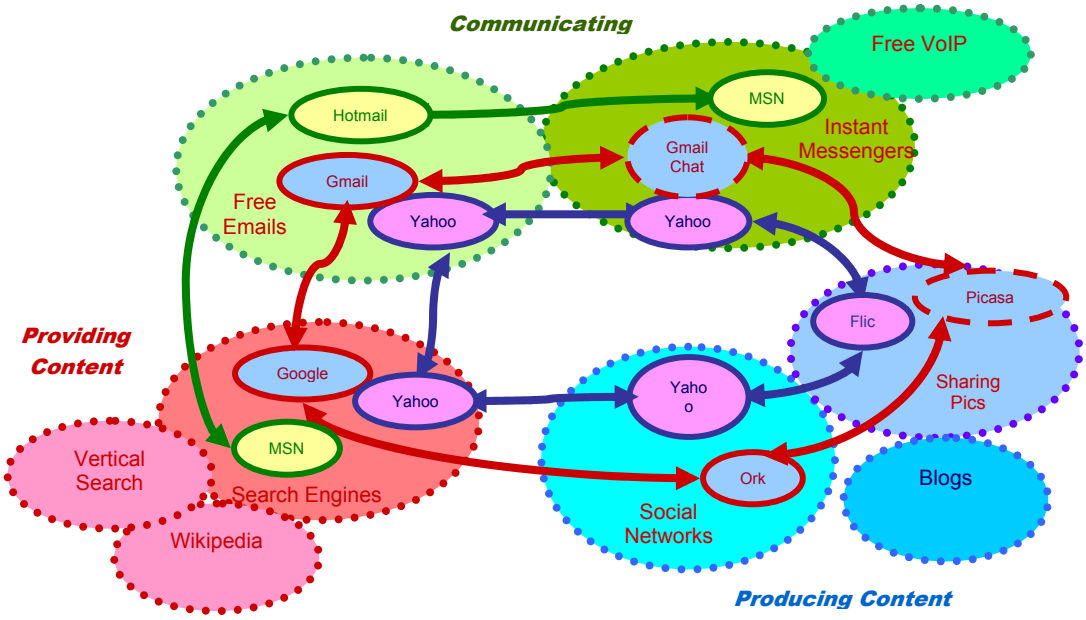


Figure 2: Google, Yahoo and Microsoft operate a number of services that render them key players as enables of content producers, as information providers and as communication facilitators.

6.2. Economic Aspects

In 1973, Daniel Bell predicted that the economic of goods would be replaced by the economics of information.[11] The amount of information created, stored and replicated in 2006 was estimated to be about 161 billion gigabytes – equivalent to three million times the information in all books ever written. That figure is expected to reach 988 billion gigabytes by 2010.[12] This data comes in a variety of formats, and content has evolved far beyond pure text description.Following Bell's prediction, does more information also mean more value? Not necessarily, as information needs also to be 'useful'. From an economical point of view, information becomes valuable only if it is both relevant and new to the user and here is where search engines come into play. As there is an abundance of digital information, search engine add value by filtering relevant and new content for the user. As the degree of relevance and novelty of the information is a critical issue, a main objective of search engine providers is to gather the freshest contents, and to prioritize information following the priority criteria perceived by the user. To this aim, search engines have a set of innovations both from the technological as from the business point of view.

6.2.1. An Innovation-based Business

In an econometric study, Prusa and Schmitz examined empirically whether 'first-movers' become market leaders.[13] They conclude that new firms in the PC software industry have an advantage over

[11] *The Coming of Post-Industrial Society"* Daniel Bell, Harper Colophon Books, New York 1974.

[12] *See* Andy McCue, Businesses face data 'explosion', *ZDNet*, 23rd May 2007, *at* http://news.zdnet.co.uk/itmanagement/0,1000000308,39287196,00.htm (last visited: 18th December, 2007), referring to IDC/EMC Study *The expanding Digital Universe*. The data explosion has been estimated also by other studies, such as the previous *"How Much Information"*, by Peter Lyman and Hal R. Varian (http://www.sims.berkeley.edu/how-much-info-2003)

[13] *"Are new firms an important source of innovation?"* Prusa, Thomas J. and James A. Schmitz, Jr., *Economic Letters*, 35, 1991 339-342.

incumbents in development new software, while incumbents can have a comparative advantage in product improvement of existing categories.

The search engine market evolution is not a story of a 'first-mover' advantage. Yahoo!, Altavista, Inktomi, or Lycos started early but they were unsuccessful to maintain the initial advantage. Google entered the market relative late but employed a far better technology for ranking relevant results. In addition, users appreciated also Google´s less intrusive advertising strategy and other features like their solution to spamming. In a way it is the story of a 'second-mover improvement'.[14] Early players had an advantage, but their technology was not good enough to compete and the 'brand name' advantage declined over time. Of the first wave of search engines, Yahoo! is the only one still maintaining a prominent role, possibly because it provided continuously a good service and technology. While in early times the quality of the technology alone determined the survival of a search engine, this is no longer the sole factor. In fact, today, many users can hardly perceive any notable quality differences amongst the major engines, while brands (to the point that "googling" has become a sinomym of web search) and adjacent services do play a more important role.

Though, that a 'latecomer' would be able to overthrow former market leaders was not foreseeable. This market dynamism makes believe that the search engine market is not a 'winner-takes-it-all' situation, unlike PC operating systems, desktop applications (like Office), or Internet browsers (Netscape first, and Explore later). Although, there is a concentration to few major players, the search engine market is not comparable to the dominance of Amazon for book sales or eBay for auctions. It is a business requiring a steady flow of technological and business innovation. Google has become market leader because it offers an excellent search tool and runs an extraordinary efficient advertising business model. In addition, they have introduced numerous innovative products and attractive services which have been well perceived by the public. In fact, over the past years the sources of revenue are roughly equally divided between advertising on the search portal itself (i.e. Adwords) and the affiliated sites (i.e. AdSense), see the Google Web sites and Google network sites in Table 1. Google's revenues other than advertising, such as licensing (i.e. business search solutions), contribute only minor to the overall result.

In Thousands US$	2003	2004	2005	2006
Advertising in Google web sites	792,063	1,589,032	3,377,060	6,332,797
Advertising in Google Network web site	628,600	1,554,256	2,687,942	4,159,831
Total advertising revenues	1,420,663	3,143,288	6,065,002	10,492,628
Licensing and other revenues	45,271	45,935	73,558	112,289
Total Revenue	1,465,934	3,189,223	6,138,560	10,604,917

Table 1: Revenue for Google in the period 2003 in thousand US$. Source: Google Annual Report 2005 and 2006. Information facilitated to the US securities and Exchange Commission.

Recently, the share of advertising revenues from the Google web sites (60% in 2006) seems to raise with respect to the advertising revenues from the Google Network web sites (39% in 2006) as can be seen from Table 1. In the future, the ratio between the two revenues sources may shift in view that Google's acquired the online advertisement company Doubleclick (the acquisition still needs approval from the competition authorities). Anyhow, Google's web site roughly contributes to approximately half of Google's searches and revenues. The other half derives from subscribed affiliated sites, embedding the Google search technology (advertising platform or pay-per-click business model) in their sites. In principle, these sites might relatively easy shift to a competitor, if

[14] *"Google: What it is and what it is not"*, Michael A Cusumano, Communications of the ACM Vol 48, p15 ff. 2005

they consider another search engine being more convenient for them.[15] In practice, there are few real alternatives in Europe.

Revenues	2003	2004	2005	2006
Advertising: Google web sites	54	50 %	55 %	60 %
Advertising: Google Network web sites a	43	49 %	44 %	39 %
Licensing and other revenues	3%	1 %	1 %	1 %

Table 2: Google's advertising and licensing revenues by share. Source: Google Annual Report 2005 and 2006. Information facilitated to the US securities and Exchange Commission.

For new entrants the entry barrier to become a fully-fledged player (offering the whole value chain) is currently huge. A new search engine provider would need considerable investments to set-up a state-of-the are the infrastructure, including server farms, and cover operational costs, before they can get into the advertising business. And such state-of-the-art infrastructure is necessary to offer a good search experience returning relevant results to a very large audience. For sake of illustration let us assume that in online advertising the average click-through rate might is 2% and the average purchase rate is also 2%.[16] This means that in the best case only four out of thousand people who see an advertisement will buy the product. As the purchase rate is low, advertisers need to reach large audiences to sell their products. For this purpose they establish alliances, buy social network sites, etc. In addition, they try to increase the click-through rate by tailoring ads to target users. For this they analyse user search patterns trying to gather the highest degree of user or group profiles.

The search engine business is highly competitive and resource intensive. On one side, operative costs to maintain a good service is very high. On the other side, the costs for a user to switch from one search engine to a competing one is very low; just one mouse click away. In fact, Fallows[17] points that 56% of users employ more than one search engine and it is likely to assume that users would change if they are not satisfied with the quality of the search. Similarly, advertisers too are loosely tight to a single search engine and would switch to the one providing them with the largest possible audience and the best offer to place their ads.

Low switching costs for users and advertisers provide the basis for a sane competition amongst search engines. At the same time, the huge investments (infrastructure and operational costs) the requirement of a mass market, and an advertiser supported business model suggests that the equilibrium market for general purpose search engines is one with few large competitors.[18] This is similar market structure to national newspapers, where few large companies, compete for readers supported by advertising. A major difference between newspapers and web search engines is, of course, that the newspaper market is less language or country specific. Also, it is more straightforward to adapt experiences in search engine applications learnt in the Anglo-Saxon environment to other languages and countries.

---

[15] *"Google: What it is and what it is not"*, Michael A Cusumano, Communications of the ACM Vol 48, p15 ff. 2005

[16] These are only average figures. In the praxis the click though rate and purchase rate is context dependent. Following AGOF, the conversion rate –i.e. the multiplication of both looking for a product and buying it over the internet– depends highly on the products or service. For instance the highest conversion rates are for books (36.1%), followed by theatre and cinema tickets (31.4%) and flight and train tickets (29.9%), while food (2.4%) and beverages (2.7%) are on the other extreme [AGOF 2007]. See *Berichtsband – zur internet facts 2007*, Arbeitsgemeinschaft Online-Forschung e.V., August 2007 available at www.agof.de/if-2007-i-teil-1-online.download.6033aa53fd516aa8e75adb6e40408d3e.pdf

[17] *'Search Engine Users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve.'*, D. Fallows 2005, PEW Internet & American Life Project

[18] It has been argued that in online businesses, the market structure and the entry barriers may lead to a situation were only few actors can survive. The main argument is that there are inherent limitations of human attention and for some internet-based services, amongst those possibly also the search engine, network effects lead to winner-take-all situation. In other words: in the long-run there is room only for limited number of Googles, eBays, Explorers or Wikipedias to survive for each of their respective sectors, i.e. search engines, online auctions, browsers or encyplopaedias.

Search engine providers have been successful in attracting larger circles of audience by diversifying beyond their core business and offering attractive services. The range and the rate of innovative services have been impressive. Major players, are offering search options for emails, search for mobile phones, or short messaging service of mobile wireless devices. In addition, they integrate novel services to their offers. Google offers print services to search online books, images from satellites, chart groups, news syndication, a tool to perform prices comparison on the web (Froogle). The Google Video store has already 3000 music videos and 300 television programs for sale. According to projections of the research firm IDC, by 2009, more than 30 million wireless subscribers will be watching commercial TV and video on a handheld device.[19] The ensemble of these services and innovations render where users flow to the portal site because of habit, market power and indirect externality.

Summarising, as switching costs for users and advertisers are low, text search engines are forced to innovate continuously on different 'fronts'. First they have been improving their technology. Second, they need to adapt their revenue model. Third, they need to take a series of measures to attract more users. All three factors are important, but not necessarily equally. One search questions is to determine the relative weighting for each three factors and whether there is a change expected in the future. Things may look different in the future. For instance, a major barrier to entry are the expensive server farms needed to support today's main technology approach. Alternative less expensive technical infrastructures for search engines, like P2P, are currently under exploration. If successful, this may decrease the investment costs and give more room for competition. Further, the AV search market does not need to be as monolithic as the current one for text search. Many players may offer complementary and competitive services, where searchers will be chosing different AV search engine providers because of their particular strengths, e.g. for image or audio search, or services, like e.g. better personalization of the interface. Given the user habits in current web search, it seems likely to believe that also the AV market revenue would be based on advertising, although the pricing may differ.

6.2.1.1. Online Advertising

Search engines offer both traditional advertising services and innovative internet-based services. Traditional services include display advertising, like banners or buttons appearing on the search engine's page, or classified advertisements, like ads listings in a directory. Today, search-specific advertising is dominant.. When a query is introduced into a search engine, the user receives two results lists delivered. The first list, is a web search provided for free in a pull mode, whose ranking is by relevancy. It is usually called organic result. The second is an advertising list whose ranking is auctioned. Search-specific advertisement is highly efficient, as the user informs the engine what he/she is looking for, unlike traditional advertisement, e.g. newspaper or TV. Merchants would spend less for marketing and be able to offer cheaper services or products to end-user.[20]

Possible pricing models include display advertising, paying for the delivery of a targeted visitor to the advertiser's website and Pay-per-click (PPC). In PPC, the advertiser pays upon the number of clicks on the hyperlink. Today's most diffused pricing model is given by the 'click-through rate'. In contrast to the 'price-per-click', where the number of user click on a specific ad are counted, the 'click-trough rate' measures how often ads prompts a response from users. An advertiser would be prepared to pay more for a click if the click-through rate is high. Part of the success search-specific advertising is that it allows even small businesses to advertise in the global market, as costs can be as less than $5 to open an account. Similar to the eBay business model, the aim is to capture also parts of the long-tail.[21] An interesting issue is that following the opinion of some observers,[22] the

[19] *"Google becomes an entertainment company"*, Michael Macedonia, January 2006 Computer.org

[20] *'The good, the Bad and the Ugly of the search business'* Kamal Jain, Microsoft Research

[21] All major search engines offer similar advertising models to Google, including Microsoft Ad Center and Yahoo Search Marketing. Google's ad programs are called AdSense and AdWord. Website operators enrol in the AdSense program to enable text, image and video advertisements on their sites. Google administers these ads and generate

technology gap between the leader Google and competitors has significantly narrowed to the point that no significant difference in search quality can be observed amongst the major players, while, over the past periods, the leader's market share continues to increase particularly in Europe. This may indicate that Google is getting into an attractor position in which the search engine's exposure to large audiences attracts more advertisers, who generate more money to provide more services to enlarge the audience.

Following Interactive Advertising Bureau (IAB) and PricewaterhouseCoopers, internet marketing spending in the US totalled $16.9 billion in 2006 [23], and advertising revenues were nearly $10 billion for the first six months of 2007, (nearly 27% increase over same period of 2006). De facto, internet advertising revenues have always grown in two digit rates over the past last years. More importantly the biggest share of the online advertising business is in the hands of search engine providers. Search advertising formats has 41% share, followed by display (rich media, banners, display ads, sponsorships and slotting fees) with 31% and classified ads with 17%.[24] The world market for search-related advertising is estimated to rise over $8 billion for 2007, up from $7 billion in 2006.[25] These estimations may even be higher in view of the recent acquisitions of online advertisement firms by search engine providers. In particular, in spring 2007 Google bought DoubleClick for $3.2 billion, Yahoo! RightMedia for $680 million and Microsoft aQuantive for $6 billion. The huge sums spent for these acquisitions seem to reflect the search engines provider´s optimism regarding online advertising as an expanding market. This optimism seems to be shared by Nielsen/NetRatings reporting that the number of online ad campaigns have increased by 35% in the period April 2006 to April 2007. In addition, combining intelligently search ads and display ads may enhance each other. The role of brand awareness in how users respond to search ads is also gaining attention. Yahoo claims that consumers are more likely to click on a search ad if they had already been exposed to some brand building banner advertising from the same company.[26]

User-generated, user-complemented and user-volunteered content are taken by the search engines at no direct cost from the IPR owner. This content includes also collective property generated by social networks, like metadata generation through file tagging or data sorting. In exchange these companies provide servers, software and a set of rules for enabling users to share content with providers have generated through advertisement. Value, therefore derives both from search engines providers and the users. In literature there is a discussion if this is equally fair for both parts and if it is sustainable business model also in the long-term. Given that owners of high-value content are reluctant to place their content on the web, alternative business models may appear in the future, that better suits the interest of content owners.

6.2.1.2. The Web Search Engines Landscape

Web search is –after sending emails– the second most favourite activity on the internet. For example, 85.9% of German internet users use search engines slightly less than sending emails

---

revenue on either a PPC basis. AdSense has become popular because these ads are less intrusive than most banners and the keyword-based concept makes the ad content of the relevant to the website. The auction-based advertising programme AdWord, specific keywords can be auctioned for a specific time period. Whenever a user types this keyword into the search, the ad will be displayed in the results list as a sponsored link.

[22] 'The Good, the Bad and the Ugly of the Search Business' Kamal Jain, Microsoft Research

[23] http://www.directtraffic.org/OnlineNews/Internet_marketing_20075115473.html

[24] Interactive Advertising Bureau (IAB) and PricewaterhouseCoopers (PwC) www.iab.net/news/pwc2007.asp

[25] *"Wikipedians Promise New Search Engine"* 16 March 2007, http://www.technologyreview.com/Biztech/18394/page1/?a=f

[26] *"Search Advertising"* Financial Times, 11th July 2007

86.1% and far more often than any other activity, like reading newspapers online, chatting or participating in social networks.[27]

Currently close to hundred search engines are operational,[28] but the bulk of the searches are performed by few service providers only. Following the consultancy firm Nielsen/Netratings, the first three operators control more than eighty percent of the market. In particular, online searches by engine performed in the US in August 2007 were executed by Google 53.6%, Yahoo! 19.9%, MSN 12.9%, AOL 5.6%, Ask 1.7% and the rest 6.3%. These searches include local searches, image searches, news searches, shopping searches and other type of vertical search activity. More than 5.6 billion searches were carried out only in that month (August 2007).[29] The ranking of the top three players is undisputed. According to comScore Networks in December 2006, Google sites captured 47.3% of the U.S. search market, Yahoo! 28.5% and Microsoft (10.5 percent). Americans conducted 6.7 billion searches in December 2006. With respect of the same month a year ago, this represents an annual growth rate in search query volume of a 30% increase. This growth rate is considerable and explains the high expectations of online advertisement of search engines as a promising growth market.

Google is the uncontested leader in web search and advertising revenue. Yahoo!, which faced a notable decline time ago, appears slowly recuperating some popularity. Some experts believe that this popularity is due to the new advertising strategy and the success of some recently launched services, such as Yahoo! Answers. MSN appears to move in a slow but constant decline. Any other search engine are far from the top three. A comparison amongst the three companies is not easy, as some interesting data, such as margins, is not publicly available. Moreover, financial data about MSN Live, is embedded in the overall Microsoft account. For sake of simplicity, let us compare Google and Yahoo! as of autumn 2007. Google had a market capitalization of 152.79b$, 5.680 employees, generating a revenue of 9.32b$ and a net income of 2.42 b$. Yahoo! for its part, a market capitalization of 39.36 b$, 9.800 employees, 6.22b$ revenues and 1.17b$ net income (for an overview see Chapter 6.4.1.) Google's revenues and earning have been sky-rocketing over the past three years, and also Yahoo!'s earnings have been increasing, but to a lesser extent.

European internet users make also massive use of search engines as their counterparts on the other side of the Atlantic. The intensive use of search engines explains with they are amongst he most visited pages on the internet and attract a lot of traffic. Google is the most visited search engine in practically all countries of the European Union. For instance in June 2007, Google reached 88.8% of the UK, 69.5% of the French and 69% of the German online population. The internet audience is notably higher than for the MicroSoft sites (83.3 UK, 62.3 France, 54% Germany) and Yahoo! (65.9% UK, 39.6% France and 36% Germany) according to the internet audience measuring company comScore.[30]

The search engine market consolidation becomes evident when observing the evolution of hits over a longer time periods. Figure 3 and Figure shows the evolution of the share for Germany and France, respectively. The evolution of Germany and France is similar to other European Member States. In particular, less than a handful search engine providers have a market share of over ninety percent and Google being much bigger than its followers.

[27] *Berichtsband – zur internet facts 2007*, Arbeitsgemeinschaft Online-Forschung e.V., August 2007 available at www.agof.de/if-2007-i-teil-1-online.download.6033aa53fd516aa8e75adb6e40408d3e.pdf

[28] For an updated list as of 18th October 2007 see Chapter **Erreur ! Source du renvoi introuvable.** on page 3

[29] see www.nielsen-netratings.com

[30] comScore Press releases, available at www.comscore.com

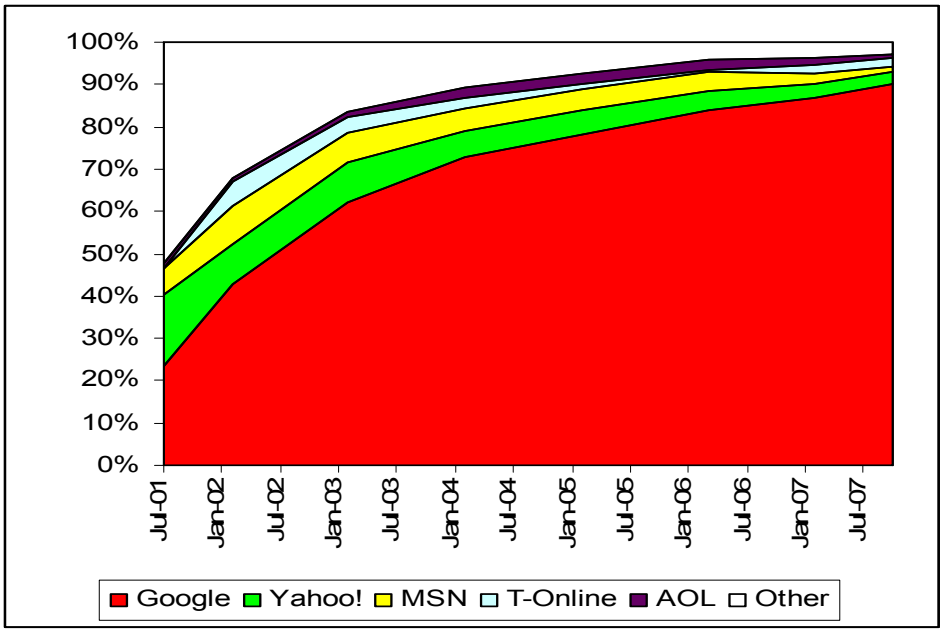


Figure 3 Evolution of WebHits for search engines in Germany in the period 2001 to 2007. Source: WebBarometer,[31] [Speck 2007] and own calculations

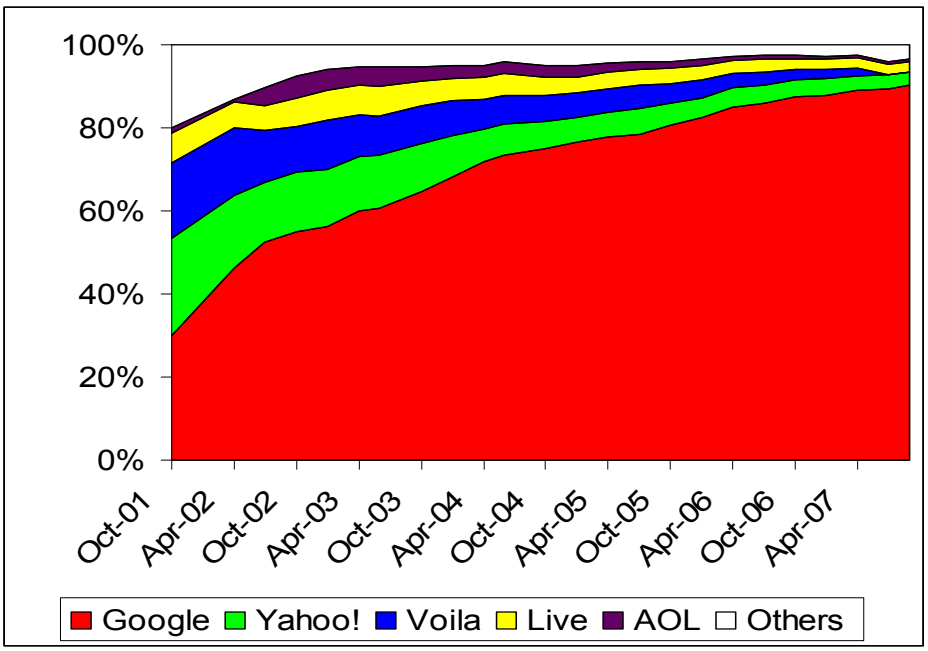


Figure 4: Evolution of WebHits for search engines in France in the period October 2001 to September 2007. Source: Baromètre Secrets2Moteurs[32] and own calculations

These data highlight that the market of the search engine providers is highly concentrated and the way of using them has also penetrated our lives. The average German –for instance– uses Google more than forty times a month[33] and three quarters of the internet users get to internet offers through

[31] http://webhits.de

[32] www.secrets2moteurs.fr

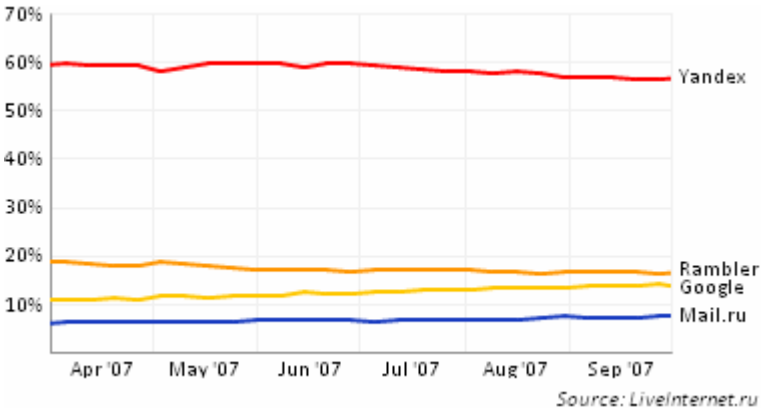
[33] comScore German data June 2007

search engines.[34] Although the traffic amongst the search engine providers may vary from one country to another, the user experience is similar for most western countries (see Chapter 6.2.1.2).

Consultancy firms metering the market share, such as Nielson/NetRatings, Compete, Hitwise or comScore, retrieve data for measuring the search behaviour by installing real-time meters the computes to web surfers (Nielsen states 500,000 people worldwide). The market share retrieved by these consultancy firms may differ to a certain extend for each of the search engines, due to the fact that they employ different metrics for measurement and the accuracy of the data is not sufficiently clear. This may partially explain why comScore's traffic data for Germany and France differs from the hits counts by WebHits.de (Germany) and Secrets2Moteurs.fr (France).

Although the measurement method is not standardized and values may vary amongst consultancy firms, there is consistency with regard to the search engines' top rankings and long-term trends. Though, how the internet audience is measured is not an academic curiosity. Small differences in market shares make a difference and have implications for business decisions. Page views a widely used audience measure used to advertisers to decide where to spend their money are becoming less significant amid the growing use of audio and video on the internet and website ability to automatically update content. Nielsen's methodology is to add 'total minutes' and 'total sessions' information to better measure the degree to which websites engage their users. This way, Nielsen thinks to measure the use of website in a more adequate way. The 'Interactive Advertising Bureau' that represents many of the biggest online publishers in putting together guidelines with the definition of unique users, time spend and other online measures.[35]

The concentration of the web search engine market appears to be a general trend in the USA and the most EU Member States. Why Google is far more dominant in Europe than in the USA is not may have multiple reasons, including, national marketing strategies, better adaptation to market size, better technological adaptation to language, lack of powerful national search engines, etc. An interesting case –although not being of the European Union– is Russia, where Google is only third by market share after Yandex and Rambler (see Figure below). Yandex claims to have a superior technology as it masters better the declinations and conjugations of the Russian language that other search engine. Other Slavic search engines, like the Czech Morfeo or the Polish NetSprint, also claim in their corporate web sites to have an advantageous technology.



How much the Yandex high market share of over 55% in Russia can be explained by better linguistic performance is, however, not obvious as the same search engine provider achieves only 16% in the Ukraine, although the Russian and Ukraine are linguistically speaking very close.

One factor that has certainly favoured Google's dominant position is the rate under which innovative services have been introduced. Many of these have been proposed to the audience at development phase (beta versions), rather than offering finished services to the users. This user

[34] *Internetverbreitung in Deutschland: Potenzial vorerst ausgeschöpft?* Birgit van Eimeren, Heinz Gerhards and Beaste Frees, Media Perspektiven, Vol 8, page 350 - 370

[35] *"Search Advertising"* Financial Times, 11th July 2007

involvement in the development stage is part is the company's culture of learning-by-doing. The company be have benefit from using the internet dominating language English when testing services and applications in the huge Anglo-Saxon environment, before introducing and adapting these into other cultures. The question arises how European cultural diversity may be turned into an advantage.

6.2.2. Issues with the Advertising Model

As advertising is the business model of all major search engines, some of threats and challenges, like conflicting interests between actors in the field, have some commonalities with its traditional pendant. Others issues, however, are arise from the auctioning model, which gains, dominance in the internet worlds.

6.2.2.1. Conflicting Interests

When a merchant subscribes to a ad programme for given key word, it is recommendable that the sponsored list does not conflict with the organic result of the query. For instance, if merchant auctions the term 'cell phone', it is not in its interest that the sponsored link appears in a response of a the ´adverse´ query like 'cell phone radiation danger'. A search engine may choose not to show links conflicting with the advertiser. The potential conflict is between the user and advertiser and it does in practice cause little problems because there is no financial conflict between the two.

The nature of the issue changes when the conflict of interest has financial implication, as in the following case. Every search engine provider is aware that if a merchant does not appear on the search engine result list, then it does -de facto- not exist on the web at all. The search engine may be motivated to decrease intentionally the quality of the search engine for the commercial category to force merchants to buy advertisements.36 This would cause considerable negative consequences for advertisers and users. The bid prices would keep increasing to the point only those merchants with large marketing budgets would appear while more less powerful merchants no matter how good they would not be presented to the audience. At the extreme, powerful merchants who sell at inflated prices could afford large marketing budget. The injured parties of such a scenario are not only the companies who would pay excessive prices for advertising, but also the users, who would have at the end to carry the costs of excessive advertising through the price of the acquired products. A problem is that there is way to identify if search engines do intentionally decrease the quality of for 'commercial' category. Such an abuse of the search mechanism would even be more extreme in case of a monopolistic position of a search engine in which users hardly would hardly have a possibility to change provider. Unfortunately –in view that the search algorithm is not public– there is no easy way to check if the quality of search engines for the 'commercial' category has intentionally decreased to force merchants to buy advertisements.

6.2.2.2. The content quality problem

The biggest asset of a conventional library is not its index (although it is very important), but most notably the books available in the library. With regard to a library, the commercial value between index and content seems to be somewhat inversed in the internet environment. While search engines are highly profitable, many content owners make little or no money. Search engine companies do neither share the revenue from the ads on the index directly with the content owner. User-generated, user-complemented and user-volunteered content are taken by the search engine providers at no direct cost. Search engines take also for free other valuable goods, including personal information or file meta-data generated by community file tagging and data sorting. But it is not only a taking, search engine also a giving. They provide at no direct cost for the user servers capacity (storage, processing power, etc.), software and a set of rules for enabling users to share content. Value,

[36] *The good, the Bad and the Ugly of the search business'* Kamal Jain, Microsoft Research

therefore derives both from search engines providers and the users. This interplay has facilitated certainly the amount of content available stored on the internet, but how much it has contributed to high-quality of the content is less clear. As content owners are often not direct beneficiaries their intellectual property, many IPR holders do chose not uploading quality content on the internet.

If Europe wants to shift quicker towards a knowledge-based economy it would be advisable to improve not only the quantity but also the quality of content on the web. The actual model has been successful, and it may become even more so if the potential quality problem becomes a limitation. In a more general way, it would be worth reflecting how the internet economy could share best benefits amongst their stakeholders. Although this may be a too ambitious undertaking, the search engine market may be an important case to study possible model. Some former concepts, which were proposed in the past and could not be implemented at the time, could be reassessed under the current market environments and technological possibilities. As a matter of illustration we may cite Laudon's proposal to establish a (national) 'information market for property rights of individuals'.[37] Laudon explored the idea that individuals may sell their own property rights in personal information on markets. As Laudon emphasized already in 1996 there is already a large market in personal information, but the property rights are held by those who collect and compile information about individuals and not by the individuals themselves. These third parties buy and sell information that can impose cost on those individuals, without the individuals being directly involved in the transactions. Laudon proposed that pieces of individual information could be aggregated into bundles that would be leased on a public market, which he refers as National Information Market. For instance, a person might offer information about himself to a company that aggregated it with other persons with similar demographic and marketing characteristics. Groups of this kind could be targeted as "youngster, male, interested in online computer games" or "30-40 year old males looking for family cars in Andalusia".[38]

Search engines and other companies who wanted to make use of such group information could purchase rights to use these mailing lists for limited periods of time. The payments they made would flow back to the individual as "dividends". Individuals who found the annoyance cost of being on such lists greater than the financial compensation could remove their names. Individuals who felt appropriate compensated would remain on the list. Although many practical details would need to be solved to implement Laudon´s market, it is important to recognize that information about individuals is commonly brought and sold already by third parties in market like environment.[39] Such a national or EU-wide information market might contribute individuals to gain an economic stake in those transactions in which they are concerned but they currently do not have.

In addition it would be worthwhile investigating other policy options to support the generation of content. For instance, a kind of web yellow pages could be encouraged that provide a catalogue of companies with website directions and topic hierarchy; ideally the list would comprise services within a proper ontology. This list might be contributed by companies during registration or feeded by the databases of governmental bodies. Another policy measure could be to push for standards for web services for local transport and mapping services so that citizens make take advantage of it on future mobile applications.

---

[37] *"Markets and Privacy"* Kenneth C. Laudon, 1996 Communications to the ACM 39(9), 92-104

[38] It is worth observing that the Fair Information Practices Principles would automatically be implemented if the property rights in individual information resided solely with individuals: secret information archives would be illegal, individual could demand the right of review before allowing information about themselves to be used and those who wanted to utilise individuals information would have to explicitly request that right form the individual in question or and agent acting on this behalf.

[39] *'Economics and Search"* Hal R Varian, SIGIR August 1999 and *"Economic aspects of personal privacy"*, Hal R Varian, December 1996 both available at http://people.ischool.berkeley.edu/~hal/Papers/privacy/

6.2.2.3. Self-bidding and Click-Fraud

Search algorithms are well-kept secrets and will remain so, because this assures companies a competitive advantage. It prevents also parties with vested interest to manipulate the advertising search engine results in their interest if they would know the details of the algorithms used to rank the results. At the same time, non-transparent auction systems have an inherent risk of fraud through self-bidding. If an eBay seller bids on its own listings through a proxy account, eBay considers this a fraud. Similarly, self-bidding in the search engines domain would also be possible, but difficult to prove because of the complex auction system, which some observers consider to be opaque. The opacity results from non-revealing exact terms under which the auction bid is awarded. When bidding for a keyword the price is an important criterion, but not the only one. Moreover, Google Checkout customers get about 20% discount on Google adwords. This inflates the bids of discount getting bidders. In the case the discount getting bidder does not win the top slot, then other advertisers end up paying the Google checkout subsidy, instead of Google itself, who becomes the beneficiary in two ways.[40]

Another important issue is 'click-fraud'. Search engine companies sell specific keywords to advertisers. When a user searches enters this specific term, a link to the advertiser is displayed in the results page. The advertiser then pays the search engine company a fixed amount for each user that clicks on the advertiser's link. This have given rise to the so-called 'click-fraud' phenomenon, whereby a person, automated script, or computer program repeatedly clicks on the competitor's advertisements in order to drive up the advertising costs paid by their competitors.[41] The average price-per-click for popular keywords is in the order of $1.70 and in some rare cases it can raise as high as $50. It is estimated that click fraud has generated the losses as high as $3.8 billion annually.[42] With regard to click-fraud, search engines have a dual role as advertising networks and publishers on their own search engines. A search engine loses money to undetected click fraud when it pays out to the publisher. In turn it generates revenue when it collects it from the advertiser. It is believed, but not proven, that as a search engine more collects than what it pays out, thus click fraud indirectly benefits search engines.

6.2.3. Adjacent Markets

Web search engines are economic drivers, whose technology and business have given raise to other adjacent markets. The dynamic sector of search engine optimisation is direct spill-over from the web search sector and the technology attractive also for enterprise search solutions and future mobile search.

6.2.3.1. Search Engine Optimization

Search Engine Optimisation (SEO) is a trend that has raised considerable dynamism and is possibly the biggest side-markets around the main search engine landscape. SEO aims at improving both the volume and quality of traffic to a web site from search engines via search results in order to get a better chance for sites to appearing highly ranked. SEO can target image search, local search, and industry-specific vertical search engines. Common for all is that for increasing a site's relevance, SEO needs to consider how search algorithms work and what people search for. Search engine providers have guidelines on how to take care site's coding and structure in order to facilitate search engine indexing crawlers to spider efficiently the site. Apart from these 'legal' ways to optimize websites to be ranked, some SEO use also spamdexing techniques. Spamdexing or so-called black

[40] *'The good, the Bad and the Ugly of the search business'* Kamal Jain, Microsoft Research, available at www.idei.fr/doc/conf/sic/papers_2007/jain.pdf

[41] *"Click Fraud – An overview"*, Jessie C Stricchiola Alchemist Media. www.alchemistmedia.com/CPC_Click_Fraud.htm

[42] *"Click Fraud looms as Search Engine Threat"*, Michael Lidtke Associated press, 11 feb 2005

hat methods (examples include link farms and keyword stuffing) aim at increasing the sites ranking at the expense of search engine user experience, as they may be directed to less relevant sites. Therefore sites employing these techniques may remove from the search engine listings.

Some marketing experts report that people are increasingly ignoring conventional online advertising.[43] Therefore, considerable effort is spent to make advertising more effective in terms of manpower and investment. This has boosted the SEO area. Being rated high in the organic results list and to pay to appear in the sponsored list are two distinct ways to gain visibility for merchants. The fact that many merchants spend considerable amounts for SEO, rather than spending directly on advertising, may indicate means that they consider it as necessary (and possibly the better) option. One reason may be that the organic results list may be perceived by users as 'neutral' and more prone to their interests. This may give rise to a kind of economic discrimination, since the richest providers would be in the position to put more money into SEO techniques than financially weaker ones, and consequently they will be more likely to get return on investment.

The increased level of sophistication in search marketing has pushed also the barrier of entry for new entrants. These entry costs include high costs for the technology and (outside) professional support needed do manage online campaigns.[44] This explains why search engine optimization is an expanding market, worth $1.5 billion worldwide in 2005, according to Forrester Research. By 2010, European marketers will spend almost €3bn, up from €856m in 2004, on search marketing.[45] The SEO market is very fragmented and the profile of the companies being active in this sector is generally, small but specialised enterprises.

Search algorithms are well-kept secrets also with the aim to prevent potential spammers to manipulate the search engine results ranking of the query results. Also undisclosed is the way the auctioning systems. For the auction system, search engines use –apart of the price– a number of other actors before deciding awarding to be ranked in the sponsored list. As the parameters of auctions are undisclosed, -if my ad loses- I do not know the reason and do not learn how to optimize better. The advertiser can hardly determine the way search engines decide how to rate adverts in their systems. The search engine's undisclosed qualitative assessments are basically the root of the 'opaque' search engine optimisation business.

6.2.3.2. Business Search Solutions

In the past, companies have invested largely in the IT infrastructure and in particular the hardware for information storage and handling. They have gathered the necessary resources and technologies to capture, store and transfer the information the enterprise needs for its operation. A remaining bottle neck is to provide a consolidated user-centred view for employees to ease their jobs and render them more efficient. This shift from a basically storage oriented infrastructure to an information consumption, goes along with a user-centric model rather than a technology based one. Providing an efficient, interactive and secure way to present user-specific content is complex, because it has to take into account different operational systems, file formats, schemas, etc.

Therefore, tailored search solutions for business and enterprises are becoming an emerging field. The aim is to identify and enable specific content across the enterprise to be indexed, searched, and displayed to authorized users. Following a study by the consulting firm IDC, the worldwide market for enterprise search and retrieval software in 2005 was $976m. This is a growth of 32% with respect to the previous year. The size of this sector is notably smaller then the aforementioned web search advertisement market. The three big players, Google, Yahoo! and Microsoft, have some activity in the field, but their revenues from licensing technology are minor. Though, business

[43] *"Internet advertising: Is anybody watching?"*, Xavier Drèze François-Xavier Hussherr, Journal of Interactive Marketing, 17 Vol 4, p8

[44] *"Search Advertising"* Financial Times, 11th July 2007

[45] http://www.searchmarketeers.com/Default.aspx?tabid=927

search solutions may be an interesting case study for Europe, as many of the key players are European, including FAST (Norway), Autonomy (United Kingdom) and Expert System SpA (Italy). For more company info see Chapter 6.4.4. Some of their products comprise knowledge management modules on top of the search function. This way, it is intended to uncover meaning arising from any enterprise information including documents, emails, entries in relational databases, etc.

Today, the market for 'knowledge management' tools is very distinct from web search engine market. The more the search engines move from text-based search to audio-visual search, the more the technological interest will overlaps, as need developing solutions for conceptual search, document classification, text mining and information analysis and correlation. This may drive current web search engines to penetrate more the 'knowledge management' market. The fact that Microsoft has made an offer to buy FAST may be an indicator of this trend.[46]

6.2.3.3. Mobile Search

Mobile Search refers to information retrieval services accessible through mobile devices like phones or PDA.[47] European telecom operators do provide some search options for their 2G, 2.5G and 3G services. For this, telecom operators rely on technology provided by companies like Google or FAST[48], alternatively users can access the URL of search engines offering a dedicated interface for handheld services, like MetaGer.[49]

Although still in its creation, the mobile search market is likely to differ significantly from web search engine market. The technological context (e.g. small screens, limited bandwidth), the reduced amount of suitable content for mobile devices, the role of the market players (e.g. as telecom operators as a provider to the internet by mobiles do have a more powerful role, than internet service providers have for accessing the internet via a computer), the user behaviour (e.g. type of search requested on the move), might beg for a different search engines business model. Walled-garden markets seem to be the currently prevailing model, but it may become more open in the future. There are discussions if a flat-rate pricing is possible of if bandwidth restrictions will force payment by bit download. This make would make a difference not only for bandwidth intensive downloading such as video (e.g. There may pricing by video per resolution), but would have also implications on location-based services which are regarded to be very promising and would allow to find the nearest restaurant typing the question to or simply speaking into our mobile telephone. .

[46] http://www.01net.com/editorial/368946/microsoft-s-achete-la-place-de-numero-un-de-la-recherche-en-entreprise/

[47] Although being a mobile device, laptops are not considered within this category as their technical characteristics are more similar to PC than mobile telephones or PDA, in terms of accessing and displaying audio-visual content.

[48] www.fast.no

[49] www.metager.de is a Metasearch Engine with a specific Palm browser option

6.3. Social Aspects

Using search engines is the second most common activity on the internet, only preceded by sending emails. 85.9% of all German internet users make queries with search engines slightly less than sending emails 86.1% [50] and more often than any other activity, like reading newspapers, chatting or participating in social networks. Citizens in other European countries are similarly often search engines. The intensive use of search engines explains with they are amongst he most visited pages on the internet and attract a lot of traffic. Google is the most visited property in most countries of the European Union. For instance in June 2007, Google reaches 88.8% of the UK, 69.5% of the French and 69% of the German online population. The internet audience is notably higher than for the Microsoft sites (83.3 UK, 62.3 France, 54% Germany) and Yahoo! (65.9% UK, 39.6% France and 36% Germany) following the internet audience measuring company comScore.[51]

These figures highlight that the market of the search engine providers is highly concentrated and the way of using them has also penetrated our lives. The average German –for instance– uses Google more than forty times a month [52] and three quarters of the internet users get to internet offers through search engines.[53] Although the traffic amongst the search engine providers may vary from one country to another, the user experience is similar for most western countries. The user experience and behaviour has been analysed in recent study whose main messages will be presented in the following chapter.

6.3.1. Patterns

6.3.1.1. User behaviour patterns

In a recent telephone interviews with about 2200 adults, Pew Internet & American Life project investigated the internet user behaviour with regard to the use of search engines. They conclude that the average user in the USA is content, dependent and naïve.[54]

Their survey found that 84% of internet users have used search engines and 56% of them use search engines on any given day. This data is in line with the analysis of major consulting firms measuring internet data traffic (see previous chapter). Also interesting is the high level of dependency on search engines as perceived by the users. 35% of the searchers use a search engine daily and 47% of searchers will use a once a week. Interestingly, 32% consider themselves "addicts" and say they cannot live without search engines. The dependency is focalized with respect to providers. 44% of searchers say they regularly use one single search engine, 48% will use just two of three search engines and only 7% will use more than three. This explains partially the market concentration around Google, Yahoo! and MS live Search. One explanation why users are loyal to few search engines is that internet users are generally very positive about their online search experiences. In particular, 87% of the internet users say they have successful search experiences most of the time.

More worrying, however, is that fact that 68% of users say that search engines are a fair and unbiased source of information (while only 19% say they do not place that trust in search engines) [55]

---

[50] *Berichtsband – zur internet facts 2007*, Arbeitsgemeinschaft Online-Forschung e.V., August 2007 available at www.agof.de/if-2007-i-teil-1-online.download.6033aa53fd516aa8e75adb6e40408d3e.pdf

[51] comScore Press releases, available at www.comscore.com

[52] comScore German data June 2007 available at www.comscore.com

[53] *Internetverbreitung in Deutschland: Potenzial vorerst ausgeschöpft?* Birgit van Eimeren, Heinz Gerhards and Beate Frees, Media Perspektiven, Vol 8, page 350 - 370

[54] *'Search Engine Users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve.'*, D. Fallows 2005, PEW Internet & American Life Project

[55] *'Search Engine Users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve.'*, D. Fallows 2005, PEW Internet & American Life Project

[56]. Most users may be naïve about search engines or simply do not fully realize that and how search engines make money. An explanation in this regard is that many users interviewed did not realize that search engines make money through advertising. While practically all interviewees can discriminate between regular programming and its infomercials in TV, only a slightly more than third of search engine users are aware of the difference between the paid or sponsored results, on one side, and the unpaid or 'organic' results, on the other, presented by search engines. Overall, only about one in six searchers say they can consistently distinguish between paid and unpaid results.[57]

With regard to the distribution of searchers by gender and age, this follows largely the pattern of internet users. Generally, speaking men and younger users are more plugged into the world of search than women and older users. In earlier times, when internet was dominated by young men, two of the most popular search topics were sex[58] and technology. Nowadays, search landscape has changed because of the demographic enlargement of the internet user population, their more diverse interest and the huge growth of online content. A recent study examining search trends finds the proportion of searches for especially sex and pornography has declined since 1997 while searches of tamer topics of commerce and information have grown.[59]

6.3.1.2. Product Search

Practically all internet users perform online search of products. In Germany alone, 37.5 million users have informed this medium to get information about products; this is 97.3% of the online population.[60] The motivation is to prepare the acquisition of products, may it be on the traditional way or over the internet. More than half of the internet users search information about flight and train tickets (58,9%), holiday planning and last-minute offers (57,8%), books (56,6%), hotels (54.1%), tickets for cinema, theatre or other (52,9%), cars (52,3%), music CD (49,0%), telecommunication products (48,9%), DVD and video (39,9%). How many searches finally materialize into acquisitions depends of the sectors and the specificity of the products. For instance, books have a conversion rate of 70%, while cars achieve hardly 16%. In most of the cases the initial search to buy any product starts at the level of a search engine provider, which point to the service provider that will offer the product we search.

Internet users have traditionally performed product comparison on specialized sites like like Billiger[61], mySimon[62], Bonprix[63], Pricegrabber[64] which used their software agents to gather product price information and compare to compare them. Search Engine providers are also entering also this domain, like Yahoo! Shopping or more recently Google Product Search. Given their huge indexes and their expertise in search technology it is a natural market for them.

---

[56] It seems that there is a growing lack of trust in news media and Americans believe that news organisatons a biased. See*´ Voters Believe Media Bias is Very Real'* Zogby Pool, 14th March 14, www.zogby.com

[57] *'Search Engine Users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve.'*, D. Fallows 2005, PEW Internet & American Life Project

[58] All time hits are searches include attractive celebrities. Britney Spears and Pamela Anderson have been on the Lycos top 50 for 277 weeks in a row.

[59] *'Web Search: Public Searching of the Web'* Amanda Spink and Bernard J Jansen, Springer Publishers, 2004

[60] *Berichtsband – zur internet facts 2007*, Arbeitsgemeinschaft Online-Forschung e.V., August 2007 available at www.agof.de/if-2007-i-teil-1-online.download.6033aa53fd516aa8e75adb6e40408d3e.pdf

[61] www.billiger.de

[62] www.mysimon.com

[63] www.bonprix.com

[64] www.pricegrabber.com

6.3.1.3. Vertical Search Engines

General purpose search engines, such as Google or Yahoo!, are very effective when users search for web sites, web pages, or general information. For search within a specific medium or in specific content categories, specialized search engines are better performing. Users are increasingly using these so-called vertical search engines for search in specific categories or media. Examples of category-focused vertical search engines include search engines for shopping (e.g. Froogle or NexTag), for government (e.g searchgov.com), for legal (e.g. law.com and lawcrawler), for traveling (e.g. travelocity and Expedia), financial (e.g. business.com and Hoovers), or business (e.g. knuru) Media-focused search engines -on the other hand- focuses on within specific online media. These search engines are used for discussion boards, forums, groups, or answer pages (e.g. Omgili and board-tracker), for scanning news worldwide. (e.g. bincrawler, Google groups, knuru), for searching the blogosphere (e.g. Technorati, knuru, and Blog-search-engine) for search in mailing lists (e.g. E-Zine List), or for search on chat rooms (e.g. e.g. Chatsearch, Search IRC). A more detailed compendium of search tools is given in the annex.

Specialization goes also along with a personalized search. Continuously personalized experience for each user is a core driver for search engines. This is applies for any type of search engines, but may become the key differentiation factor for vertical search engines. The user experiences is key, irrespective is a job seeker is looking for a new employment, if a client is looking for a integrative travel package, a television viewers selecting the right news segments of a shop keeper to advice on the best accessory. One asset is interactivity with the search medium to increase the search experience. This may change the way we search. For instance the large video proliferation may raise the possibility to video syndications (similar to netvibes), where new pieces of work may result from picking video fragments and recompiling them in a creative way.

Two phenomena seem to occur at the same time. One is the emergence of specialised search engines in different domains. The other one a consolidation of general purpose search engines, triggered by the fact that few search engines that can effectively compete in the tough advertising market. These phenomena are not necessarily excluding. General purpose engines could introduce features (e.g. Directories or separate tools) that cover also specialized areas.

6.3.2. The Web 2.0 Context

6.3.2.1. Communities developing Search Engines

The term web 2.0 refers to a second generation of web-based communities and hosted services which aim to facilitate collaboration and sharing between users. Examples of such collaborative services are social-networking sites, wikis and folksonomies. Basically, there are two facets of search engines within the Web 2.0 context: the first one is what the web-based community can do for search engines and the second what search engines can offer for (future) web 2.0 applications. Chris Sherman clusters these applications in different categories, namely shared bookmarks and web pages;[65] tag engines, tagging and searching blogs and RSS feeds;[66] collaborative directories;[67] personalized verticals or collaborative search engines;[68] collaborative harvesters; [69] Social Q&A sites.[70]

[65] Such as Del.icio.us, Shadows, or Furl

[66] Such as Technorati, or Bloglines

[67] Such as Open Directory Project, Prefound, Zimbio and Wikipedia

[68] Such as Google Custom Search, Eurekster, Rollyo

[69] Such as Digg, Netscape, Reddit and PopUrl

[70] Such as Yahoo Answers, and Answerbag

Of particular interest are those projects and services constructed and maintained in a sustainable manner by a community of volunteers. One example is the Open Directory Project (ODP), also known as dmoz, a multilingual open content directory.[71] In this collaborative directory, the web is catalogued by user community, which has established a system on how to handle, organize and prioritize millions of inputs. In a way, these web communities have established an operational ´authority model´ for their domains, similar to other traditional communities the ´impact factor´ of academic journals.

Some web communities are already getting together to provide personalized search engine by offering results from a user selected collection of trusted sites on any given topic. Rollyo,[72] for example, does this by searching those sites that have been chosen by an inscribed used after carrying our search query. Eurekster's Swicki[73] is another collaborative search results aggregator, whose concept is to adapt a search engine to your own needs. For this, a swiki user has to provide information about the topic of interest by selecting relevant keywords, websites, site search, etc. Based on click patterns the information is used to learn which results users like the most and move them to the top. Over time user feedback will modify the search queries. Learning from the behaviour of your swicki's users which search results are relevant and which filtering techniques work the best for your topic. For their operation, both Rollyo and Eurester are using Yahoo! index. Recently, Google offers also the possibility to tailor the search engine specifically to user's needs, like non-profit, government, or educational organisations.

In the above examples, search engines get personalized through the adaptation of the query algorithm, but they still operate a server-based network principle. Many bottom-up approaches developed by the web community, however, operated on principles of decentralised technological resources.[74] Making use of is discussed in literature and some beta-version are being tested already. Examples include OpenSearch,[75] YaCy[76] and Faroo[77] are examples search engines currently being tested that operate under peer-to-peer principles.

One of the major motivations of web communities to develop a search engine is their fear to be manipulated or suffer censorship by dominant search engine providers. Therefore, their technology offers more transparency about the search process and complies with high privacy standard. Most of these collaborative projects follow wiki-principles and use open source software or reveal their code.[78] They intend also to use the user's search patterns behaviour for the user's own benefit (rather than for adapting advertising strategies of the search engine providers). Wikipedia is a successful example how web communities can effectively collaborate together, Wikia Search to create an open global search engine is another.

6.3.2.2. Communities tagging and filtering audiovisual content

Search engines are at the heart of popular multimedia sites like as wikipedia, Flickr, or YouTube. The steady increase of creation, storage and interchange of audio-visual material renders search engines even more interesting. Making use of user generated preferences, like Chacha or

[71] See Wade Roush, New Search Tool Uses Human Guides, Technology Review, February 2, 2007, at http://www.techreview.com/Infotech/18132.

[72] www.rollyo.com

[73] www.eurekster.com/swickibuilder/dir.aspx

[74] See for example www.golem.de/0411/34880.html

[75] www.open-search.net

[76] www.yacy.de

[77] www.faroo.com

[78] For example Wikia Search http://search.wikia.com

WikiaSearch[79] (the project announced in 2006 by Wikipedia founder with a investment backing of over $4 million capital), are just emerging.

Basically there are two major ways to carry out AV search, through content-based search and metadata search. Content-based search is a considerable technological challenge. The EU funded projects gathered under the umbrella of the CHORUS coordination action offer a nice view of the spectrum of scientific challenges. If successful, speech and pattern technologies would be able to automatise many search processes. The creation of meta-data can be automatized to a certain extent only. Researchers are pursuing the development of software that automatically tags audio-visual content. In spite of the efforts, it seems unlikely that that getting rid complete of any human input will be possible. The cognitive abilities of humans and semantic understanding make people hardly replaceable by machines.

In early times, search engines operated with human edited directories, e.g. Yahoo! or Lycos. Today, practically all leading search engine providers perform search by an automated process –including user behaviour by clicks, popular URLs, and link structure)– and manual input is limited. Having people paid to introduce meta-data on audio-visual content is financially unviable option at large scale. However, there will always need a certain level of human input, particularly audio-visual search is likely to dependent on humans as long as tagging will be necessary. Here, social networks and web communities emerge as an unexpected ally. In Web 2.0 environments shared bookmarks and web pages,[80] tag engines, tagging and searching blogs and RSS feeds[81] are common. Web communities members are very active members do provide meta-data for free and this information is largely available for search engine providers. The exploitation of these freely available metadata will a focus of future search engine providers in order to offer a better search experience that prioritizes by reflecting the user's relevance. In addition, audio-visual content on social networks – like in Flickr- could be used as data to train high-level automatic object recognisers in image search.

6.3.3. Privacy, Security and Personal Liberty

6.3.3.1. Profiling of Individuals

Whenever a query is introduced, the search engine stores the query and associates it to an IP address and a cookie, from which the user's computer might be identified. The more additional (non-search) services a search engine offers, the more personal information they can gather and combine. The threat is that users can be identified, and their habit, hobbies, believes and political views could be monitored. The problem is that many users are too naïve or not aware of the data stored about them. How much better information campaigns may contribute to raise awareness is unclear.

The popular assumption seems to be that privacy has already been irrevocably eroded,[82] because of some prominent negative experiences. Recording the search queries of users can easily be used to the identification of the searcher, as a prominent American Online (AOL) case shows. On 4th august 2006, AOL released a data file on search queries. It contained 20 million search keywords introduced by some 650,000 users over a 3-month period. Each user on this list was numbered by a unique sequential key, and the user's search history was compiled. The file did not include any personal information per se, but certain keywords could contain personally identifiable information, like user typing in their own name, their address, social security number or by other data.

[79] http://search.wikia.com/wiki/Main_Page

[80] Such as Del.icio.us, Shadows, or Furl

[81] Such as Technorati, or Bloglines

[82] *"The future of the internet is not the internet: open communications policy and the future wireless grid(s)"* Lee W McKnight, NSF/OECD Workshop, Washington 31st January 2007

Although intended for research purposes only, this data file was widely diffused into the blogosphere and on popular sites. The list got into the hands of some New York Times journalist, who tested whether it was possible to identify and locate individuals from the 'anonymous' search records. Shortly after, the New York Times discovered the identity of several searchers by simply cross referencing the data with phonebooks or other public records. AOL took consequences of this privacy breach by firing some responsible. More importantly, the AOL case demonstrates that data collected by search engines can lead to the identification of the user and can be misused to infringing the private sphere. In fact, to target ads better search engine providers keep the user query data indefinitely without giving any control to users.[83] Even worse, the user's information stored is not limited to the search query only. The more additional (non-search) services a search engine offers, the more personal information they can gather and combine. Some examples:

In October 2004, Google introduced Desktop Search, which indexes the content on personal computers including files, emails or web search tracking (optional). The potential –but also the threat- of such a programme is that permits personalised search. This may tie users to the software provider's solutions. A battle is starting around search behaviour and its technology.[84]

Google offers a service called "My Search History" which allows users to retrieve and store former searches. Recoding over long time periods search histories may provide insights on what someone is doing, his interests and thinking. The search engine provider would be able to provide advertisers with far more sophisticated consumer profiles if it maintains a comprehensive database of search histories that can be sorted by individual user.

A danger is that such a monitoring may be employed to monitor and eventually suppress political opponents. Such an erosion of the personal liberty is not implausible scenario, given that major search engine companies have already given in political pressures in the past, like the filtering of internet content in China.[85]

6.3.3.2. Censorship

The common perception has been that the internet is an unstoppable force for democratization, a force for liberation that cannot be tamed by local governments. While the internet has undeniably contributed to making citizens getting access to information in many parts of the world, this cannot be generalized everywhere. Some search engines have been accused of censorship. The accusers assume political and economic motivation, as the following examples show.

Yahoo! Google China, Microsoft, AOL, Baidu and others, are accused to have cooperated with the Chinese government in order to implementing a system of Internet censorship in mainland China. In fact, Google's Chinese search engine (www.google.cn) filters information perceived to be harmful by the government of the People's Republic of China, including content relating to the Tiananmen Square protests of 1989, sites supporting the independence movements of Tibet and Taiwan, the Falun Gong movement, or more recently the Chinese demonstrations against Japans more recent attempts at revisionists history.[86] J. Zittrain and B. Edelman from Harvard Law School, who studying exclusions from search engine search results all over the world, report that China is not the

[83] *'The good, the Bad and the Ugly of the search business'* Kamal Jain, Microsoft Research

[84] *"Google: What it is and what it is not"*, Michael A Cusumano, Communications of the ACM Vol 48, p15 ff. 2005

[85] *'Esse est indicato in Google: Ethical and Political Issues in Search Engines'*, Lawrence M Hinman, International Review of Information Ethics, Vol 3 p 19, June 2005

[86] The list of words censored by search engines in the People's Republic of China are regularly updated. They are available –for instance– in Wikipedia
http://en.wikipedia.org/wiki/List_of_words_censored_by_search_engines_in_the_People%27s_Republic_of_China

only country performing censorship, similar filtering of internet documentation is practiced also in Saudi Arabia.[87]

Although China has little economic influence on and no political power over Google, it seems that the US search engine provider has accommodated to the wishes of the Chinese government. Some US observers, amongst those Prof. L. Hinman at the University of San Diego, are worried that Google could eventually be much more strongly influenced by the United States governments which has far greater economic and political impact on Google than does the government of China.[88] In fact, the power of search engines lies that –due to its key role for the internet– it may contributes preventing citizens for accessing certain sites on the internet. Such a scenario is a potentially frightening aspect for Europeans, whose values include a maximum of personal liberty, and access to uncensored information.

Freedom of speech on the web has meaning only if the speech could be communicated to the interested audience. In view that every search engine may have a both an intentional and a unintentional bias, it would be suitable to be able to discriminate between both. Possibly, software algorithms to detecting unintentional bias could be help in for such a purpose. A piece of work in towards this aim is CenSEARCHip[89]. This tool explore the differences in the results returned by different countries' versions of the major search engines. Web search and image search functions are available for the four national sites (United States, China, France, and Germany) of Google and Yahoo! When clicking the "Image Search" button, each side of the display shows images returned in the first page of search results only by that country's search engine.

Through the agreements with the search engine providers, the Chinese government has successfully restricted their citizen's access to non-desired politic sites. In addition, the government is very strict with citizens, trying to circumvent their internet policies. Following information by Open Search[90] "in April 2005, Shi Tao, a journalist working for a Chinese newspaper, was sentenced to 10 years in prison by the Changsha Intermediate People's Court of Hunan Province, China (First trial case no 29), for "providing state secrets to foreign entities". The "secret", as Shi Tao's family claimed, refers to a brief list of censorship orders he sent from a Yahoo! Mail account to the Asia Democracy Forum before the anniversary of Tiananmen Square Incident".

6.3.3.3. Racism and the Protection of Youth

Major search engines apply internal rules of conduct to protect against forbidden information or youth endangering content. Apart of this industry self-regulation there is at least one case of industry-government co-regulation in the EU. In Germany, all major search engine providers have subscribed to a code of conduct that obliges them not to display those URL that have been marked as endangering by the German Authority for Youth Protection (BPjM - Federal Department for Media Harmful to Young Persons).[91] The working principle is the following: search engines providers become members of FSM ('Freiwillige Selbstkontrolle Multimedia-Diensteanbieter (FSM)' an a registered association founded in 1997 by e-commerce and web-operating companies dedicated to the protection of the youth and minors. The FSM operates a hotline where any person or organisation may report on illegal or harmful web content. The governmental BjM and the FSM are in close contact and members about harmful content whose sites are then taken blanked by the members. Content subject to restricted distribution under German law on harming young people

---

87 *"Empirical Analysis of Internet Filtering in China*" Jonathan Zittrain and Benjamin Edelman, Harvard Law School, http://cyber.law.harvard.edu/filtering/china/

88 *'Esse est indicato in Google: Ethical and Political Issues in Search Engines'*, Lawrence M Hinman, International Review of Information Ethics, Vol 3 p 19, June 2005

89 http://homer.informatics.indiana.edu/censearchip/

90 www.open-search.net/Opensearch/WhyOpenSearch

91 www.bundespruefstelle.de/

include sites with explicit incitement to hate or violence against a group of people (proscribed by the criminal law such as Volksverhetzung), instructions on how to commit a crime, glorification or trivialization of violence, incitement to racial hatred, content glorifying war or showeing minors in an unnatural/harmful situation.

Although all EU countries have regulations and law protecting the minor, Germany seems to be the only Member State within the EU where co-regulation for search engine providers is currently in place. This does not mean, however, that these countries do not pay attention, on preventing minors to have access to harmful content.

The laws on youth protection and against racisms, as well as the freedom of expression may vary form country to country, explaining differences in search results. For instance, in many EU Member States, anti-Semitic websites are illegal. Therefore, Google.de and google.fr do not list these anti-Semitic sites [92], while this is not the case in the US. In a fact, when querying the term 'jew' several of the top ranked sites in Google.com are anti-semitic. The Google management is aware of the issue and relased a note explaing the companies policy in repect and noting that anti-semitic sites do not typically appear in a search for 'jewish people', 'jews' or 'judaism', but only in the search of the singular word 'jew'. [93] This points also to another more general problem, namely that harmful or illegal content may be hidden / appear after querying on unrelated or naïve terms.

6.3.4. Search Engine Result Manipulation

Search engines tend to penalize sites when they detect that methods are used that not conform to their guidelines. Search engine providers can reduce their rankings or eliminating completely their listings from the research results. The prominent disputes of the past to are a result of opposing interests between search engine providers and search engine optimizers. Search engine providers have argued that by penalizing black cheeps they are defending user's interest not to get a distorted ranking. One potential threat of the practice to down rank sites is that it may be on an arbitrary way or misused by search engines providers in order to force commercial sites to subscribe to the search engines advertising programs.

In February 2006, Google found that BMW's German website influenced search results to ensure top ranking when users searched for "used car." BMW's German website, which is reliant on javascript code unsearchable by Google, used text-heavy pages liberally sprinkled with key words to attract the attention of Google's indexing system. Google considered that spiking doorway pages with keywords, was not complying with Google's guideline not to present different content to search engines than displaying to users. Therefore Google reducing BMW's page rank to zero, ensuring the car manufacturer's site no longer appeared at the top.[94] Similar, accusations of manipulating page ranks have been reported in the past, including SearchKing,[95] Ricoh Germany, or 'September 11th Truth'. [96] Page rank manipulations are not stricted to Google, Baidu has also been told to have decreased the rank of the blogging service Sina, since Sina published several negative reports on Baidu.[97]

The aforementioned BMW case could be considered as a consequence of the fierce battle for marketing of site by the search engine optimization (SEO) industry (see also chapter 6.2.3.1). SEO

[92] A search for the German expression 'Jude' or 'Juden' or the French 'Juif' delivers millions of entries, but the first pages of top ranked sites are not anti-Semitic.

[93] www.google.com/explanation.html

[94] See for example: http://news.bbc.co.uk/2/hi/technology/4685750.stm

[95] www.pandia.com/sw-2002/40-google.html

[96] *"Google Doesn't Like 911truth.org",* 25th September 2007, available at http://www.911truth.org/article.php?story=2007092200814732

[97] www.cwrblog.net/413/baidu-manipulates-search-rank.html

aims at improving site architecture in such a way search engines can index it well in and by optimizing keyword phrases in the site content. The objective is to get high rankings in search engines organic results. SEO make use of techniques that search engines recommend as part of good design (so-called white hat), but may use spamdexing techniques that search engines do not approve (so-called black hat). At the first glance this distinction appears to clear, but at second sight this may neither easy to implement nor always to be objective. Therefore search engine providers use automatic but also manual procedures to counter effect misdoings. At the same time this leaves room for search engine providers to commit injustice.

Basically we have to understand that if a merchant does not appear on the search engine ranking, then I does no exist on the web. Search engine providers are aware of this and may be tempted use it to their advantage. A frightening scenario is that market leaders do intentionally decrease the quality of search for the "commercial" category in other to force merchants to subscribe to their advertisement programmes.[98] Such an abuse of the dominating role would most likely distort the market with serious consequences. Merchants would be obliged to increase the bids for advertising. In the long-term, only merchants that can afford a large marketing budget might survive.

6.3.5. The public responsibility of search engines

6.3.5.1. Education and Learning

On one hand, search engines have become crucial to society because they are used by many millions of people. On the other hand, search engines are owned by private companies, whose objective is to make profit. This creates a tension between the corporate mission of the shareholder's interest and the public role of search engines.

Computers have not only entered our homes but also our schools from which the internet can be accessed. Many EU Member States have programmes aiming at connecting schools to the internet, to increase the student's IT literacy, introducing e-learning programmes, internet based life long learning programmes, etc.[99] In a nutshell, education and learning patterns have drastically changed of the past decade. The services provided by search engines have become central to education and an indispensable tool for pupils and students. Geographic search on online maps have displaced traditional search in paper atlas. Online reference database, like Wikipedia have displaced traditional encyclopaedia. Bibliographic search in libraries have been replaced by online search like Google Scholar. In addition, Google's project to scan books and making them publicly available has been an additional asset for accessing information.

Undoubtedly, search engines have greatly contributed to make information available for pupils and students. Today, probably many students search Google far more often than consulting in books for information or going to the library. While offering free access to information is positive, the concentration of information in few locations controlled by very few companies bears some risks. One potential risk is manipulation; another bias. The latter can be voluntary (e.g. by systematic by omission) or involuntary.

6.3.5.2. Are Search Engines a public good?.

The web has become the principal source for research information and news for many people in the developed world. It is a predominantly increasing way to get informed about the news of the world. The vast amount of information available on the web would be practically useless without search

[98] *'The good, the Bad and the Ugly of the search business'* Kamal Jain, Microsoft Research

[99] *"The Future of ICT and Learning in the Knowledge Society",* Yves Punie, Marcelino Cabrera, Marc Bogdanowicz, Dieter Zinnbauer, Elena Navajas, April 2006 IPTS publication EUR Number: 22218 EN
www.jrc.es/publications/pub.cfm?id=1407

engines. As search engines are gatekeepers of the web, guiding people to reach their desired destinations, the question arises how much search engines fulfil a public responsibility and if a universal service must be assured. Similar to telecommunication providers having to offer a minimal universal service to any citizen requiring it, there is an ongoing discussion if access to the internet would also need to be included in a future universal service. In such a future scenario it is not unlikely to believe that search engine providers would have to take their stake to offer such a universal service.

Defenders of a public good view, like Lucas D Introna and Helen Nissenbaum, see the web as a conveyor of information is getting the elements of a public good. And the way search engines perform the news syndication influences the view of the news.[100] For them the ideal web would facilitated associations and communications that could empower and give voice to those who traditionally have been weaker and ignored. They consider that society would need to protect public interest against encroaching commercial interests. As a consequence they consider public support for developing more egalitarian and inclusive search mechanisms and fore reach into search and meta-search technologies that would increased the transparency and access.[101]

[100] *'Esse est indicato in Google: Ethical and Political Issues in Search Engines'*, Lawrence M Hinman, International Review of Information Ethics, Vol 3 p 19, June 2005

[101] *'Shaping the Web: Why the Politics of Search Engines Matters'*, Lucas D Introna, Helen Nissenbaum, 2000, The Information Society 16:3, p169 ff

6.4. Annex: Profiles of Selected Search Engine Providers

6.4.1. Overview

The Pandia website provides extensive list of tools to search the internet, which on the 18th October comprised over 100 engines.[102] The list comprise tools for web search, directory search, custom search, local search, search in databases, social search and search in reference material and dictionaries and is presented in Table 3.

Search Engines	Multisearch services	Metasearch
Google Yahoo! Yahoo!/Alta Vista Yahoo!/All The Web Ask Windows Live Search (MSN) Exalead GigaBlast WiseNut Snap iWon Seekport UK SearchUK Aesop.com FyberSearch factbites MoJeek Accoona Yoono more...	MsFreckles Trexy A9 **Local search** Goolge Local Yahoo! Local MSN City Guides Ask Local AOL local Windows Live Local Infospace Local Local Search Guide Pandia Plus Regional Search Engines **Maps** Google Maps Local Windows Live Local Yahoo! Maps	Pandia Metasearch Search.com Metacrawler Mamma Dogpile RedeSearch.com Kartoo Ixquick Vivisimo HotBot Comet way Pagebull visual search SearchTheWeb2 (the long tail) more... **Custom Search** Directories over vertical and custom search engines: Custom Search Engines CSE Links CustomSearchEngine.com vErtical sEarch
Directories	**Directories**	**Databases**
About the best... Pandia Plus [Q] Yahoo! [Q] Best of the Web Mahalo UKPlus ChaCha Femina Cybergrrl Backwash Joe Ant Goguides.org Tygo.com Small Business Directory v7n Seven Seek	About.com Argus Librarians' index BUBL UK Infomine Open Directory [Q] Academic Info Gimpsy IllumiRate Skaffe Alive Undum Rubber Stamped Aviva Directory of Directories	Beaucoup! Search Engine Links Congoo **Social search** Wink del.icio.us tailrank fanpop Shadows Yahoo! MyWeb PreFound zimbio Rollyo digg LookSmart's Furl

102 www.pandia.com/powersearch/index.html

WOW	more...	Shoutwire Netscape reddit Metafilter Pageflakes diigo Sproose collarity similicio.us

Table 3: List of relevant search engines by area of operation. Source Pandia (www.pandia.com). Accessed 18/10/2007.

With regard to search engines several engines for finding audio-visual material, there are several tools for this purpose, see Table 4. The list does no distinguish between content-based search and meta-data search technology. Some engines have been discussed in the body of this document.

Images and media	TV and video	Music & MP3
Google Images A.Vista Images Pagebull AlltheWeb Pictures AlltheWeb Video Yahoo gallery Yahoo Image Search FindSounds.com Ditto images Webseek Footage.net picsearch Pixsy more...	Blinkx TV Yahoo! Video Google Video Singing Fish A.Vista Video Windows Live Video YouTube Search for Video MySpace Video AOL Video Guba Veoh Metacafe ClipBlast Blip TV ClipShack Juicecaster.com Stickam ZippyVideos Vidiac Putpile Live Video more...	P = fee based music-map (for related artists) A.Vista Audio/MP3 Audiogalaxy P Napster P Magnatune eMusic P AllofMP3 P Mp3.com Y! Music Apple iTunes (requires download) P more mp3... Yahoo! Audio Search more...
Radio		**Podcasting**
Pandia Radio Search Radio Locator (MIT) BRS Radio Directory vTuner Live Radio All Radio Radio Spy Virtual Tuner Real Guide		Podscope Podcast.net Podcast directory the podcast network Yahoo! Podcasts feedster.com PodZinger podanza

Table 4: Table audio-visual search engines. Source Pandia (www.pandia.com) accessed 18/10/2007.

In spite of the large number of search engines, only few of them are larger companies. The concentration effect has been discussed in the economic chapter. As a matter of illustration,

Search Engine Provider	**Google**	**Yahoo!**	**MSN LiveSearch**	**Ask Excite CitySearch**	**Baidu**
Parent Organization			Microsoft	IAC Search & Media	
Headquarters	USA	USA	USA	USA	China
Market Cap:	$160.79b	$31.74b	$275.77b	$8.73b	$7.10b
Employees:	10.674	11.400		16.000	3,113
Revenue (ttm):	$13.43b	$6.65b	$51.12b	$6.42b	$157.05m
Gross Margin (ttm):	60.26%	60.20%	79.08%	48.43%	66.87%
EBITDA (ttm):	5.78B	2.07B	20.48B	946.57M	68.13M
Oper Margins (ttm):	32.45%	12.99%	37.23%	7.23%	32.06%
Net Income (ttm):	3.69B	730.19M	14.07B	194.23M	57.59M

Table 5: Comparison of the major search engine providers in terms of financial data. Note that data are for the parent organization or the search engine provider, which in the case of Microsoft and IAC have also other important business operations. Source: Yahoo! Finance, and Company Reports

In the following, the most business summary of some selected companies will be presented. The information is taken form their own sites or are the information provided to the financial portal of Yahoo!

The word-wide players mentioned underneath have considerable business role most Member States of the European Union. Google is the market leader in all countries we have investigated so far, these include the United Kingdom, France, Germany, The Netherlands, Italy and Spain. Due to the supremacy of US players globally, we have include Baidu (China), Yanex (Russia) and Rambler (Russia) as examples of champions in their respective markets. Finally, we have included a section with a selection of European companies with interesting technology.

6.4.2. World-wide Players

6.4.2.1. Google (USA)

Business Summary

Google, Inc. provides targeted advertising and Internet search solutions worldwide. It offers intranet solutions via an enterprise search appliance. The company's products and services include Google.com that offers Google Base, which lets content owners submit content that they want to share on Google Web sites; personalized homepage and search; and Google Video and YouTube that lets users find, upload, view, and share video content, as well as Web, image, book, and literature search. It offers communication, collaboration, and communities, such as Gmail that is Google's Web mail service that comes with built-in Google search technology for searching emails; orkut that enables users to search and connect to other users through networks of trusted friends; Blogger, a Web-based publishing tool that lets people publish to the Web using Weblogs; and Google Docs & Spreadsheets, which allow users to create, view, and edit documents and spreadsheets using a browser. The company also offers Google GEO that offers earth and local maps; Google Labs that tests product prototypes and solicits feedback on how the technology could be used or improved; and Google Mobile that lets people search and view both the mobile Web,

consisting of pages created specifically for wireless devices, and the entire Google index, including products like Image Search. In addition, it offers AdWords, an online self-service program that enables advertisers to place text-based ads on Google Web sites; AdSense, a program through which Google distributes its advertisers' ads for display on the Web sites of its Google Network members; and Google Checkout, an online shopping payment processing system for consumers and merchants. Further, the company licenses its Web search technology along with Google AdSense service for search to companies.

Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043, USA
Phone: 650-253-0000
Fax: 650-253-0001
Web: www.google.com

6.4.2.2. Yahoo! (USA)

Business Summary

Yahoo! Inc. provides Internet services to users and businesses worldwide. It offers online properties and services to users; and various tools and marketing solutions to businesses. The company's search products include Yahoo! Search, Yahoo! Toolbar, and Yahoo! Search on Mobile, Yahoo! Local, Yahoo! Yellow Pages, and Yahoo! Maps that allow user to navigate the Internet and search for information from their computer or mobile device. It also offers marketplace products that comprise Yahoo! Shopping, Kelkoo, and Yahoo! Auctions for shopping; Yahoo! Real Estate for real estate information; Yahoo! Travel, an online travel research and booking site and Yahoo! FareChase, a travel search engine; Yahoo! Autos to price and compare cars online; and Yahoo! Personals and Yahoo! Personals Premier for online dating. Yahoo! provides information products, such as Yahoo! News that aggregates news stories; Yahoo! Finance that offers financial resources; Yahoo! Food, an online food destination; Yahoo! Tech that offers information on consumer electronics; and Yahoo! Health, a healthcare destination. Its entertainment offerings comprise Yahoo! Sports, Yahoo! Music, Yahoo! Movies and Yahoo! TV, Yahoo! Games, and Yahoo! Kids; communications products include Yahoo! Mail and Yahoo! Messenger with Voice; communities offerings include Yahoo! Communities and Yahoo! Photos; and front door products comprise Yahoo! Front Page and My Yahoo!. In addition, it offers Yahoo! Broadband, Yahoo! Digital Home, Yahoo! Mobile, and Yahoo! PC Desktop to access its content and communities across Internet-enabled devices. Further, it provides Yahoo! HotJobs, an online recruitment solution; Yahoo! Small Business to purchase products on the Internet; and Yahoo! Local that offer businesses a service to post company information. It has strategic partnerships with Seven Network Limited; eBay; AT&T, Inc.; and Verizon Communications, Inc. The company was founded in 1994 and is headquartered in Sunnyvale, California.

Yahoo! Inc.
701 First Avenue
Sunnyvale, CA 94089, USA
Phone: 408-349-3300
Fax: 408-349-3301
Web Site: www.yahoo.com
Employees: 11.400

6.4.2.3. MSN Live Search, Microsoft (USA)

Business Summary

Microsoft Corporation engages in the development, manufacture, licensing, and support of software products for various computing devices worldwide. It operates in three divisions: Platforms and Services, Microsoft Business, and Entertainment and Devices. The Platforms and Services division comprises Client, Server and Tools, and Online Services Business segments. Client segment offers operating systems for servers, personal computers (PCs), and intelligent devices. Server and Tools segment offers Windows Server operating systems. Its Windows Server products include the server platform, operations, security, applications, and collaboration software. It also builds software development lifecycle tools for software architects, developers, testers, and project managers; and provides consulting, and training and certification services. Online Services Business segment provides personal communications services, such as email and instant messaging; and online information offerings, such as MSN Search, MapPoint, and the MSN portals and channels. The Microsoft Business division includes Microsoft Office system of programs, services, and software solutions. It also provides financial management, customer relationship management, supply chain management, and analytics applications. The Entertainment and Devices division offers the Xbox video game system, such as consoles and accessories, third-party games, and games published under the Microsoft brand, as well as Xbox Live operations, research, and sales and support. It provides PC software games, online games, and other devices; and consumer software and hardware products, such as learning products and services, application software for Macintosh computers, and PC peripherals. The division also develops and markets products that extend the Windows platform to mobile devices and embedded devices. Microsoft was founded in 1975 by William H. Gates III and is headquartered in Redmond, Washington.

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399, USA
Tel +1 425-882-8080
Fax: +1 425-936-7329
Web www.microsoft.com

6.4.2.4. Ask.com, Excite, CitySearch, (USA)

Business Summary

IAC is a conglomerate operating more than 60 diversified brands in sectors being transformed by the internet, online and offline. Within the internet media and advertising IAC operates the brands Ask.com; CitySearch; Excite, Evite. Employees: Approximately 20,000 full-time employees as of December 2006.

IAC/InterActiveCorp
555 West 18th Street
8th Floor
New York, NY 10011
Tel +1 212-314-7390
Fax: +1 212-632-9621
Web www.iac.com

6.4.3. Regional Champions

6.4.3.1. Baidu (China)

Business Summary

Baidu.com, Inc. provides Chinese language Internet search services. Its services enable users to find relevant information online, including Web pages, news, images, and multimedia files through its Web site links. The company offers a Chinese language search platform, which consists of Web

sites and certain online application software, as well as Baidu Union, which is a network of third-party Web sites and software applications. Its products include Baidu Web Search that allows users to locate information, products, and services using Chinese language search terms; Baidu Post Bar and Baidu Knows, which provide users with a query-based searchable community; and Baidu News that provides links to an extensive selection of local, national, and international news. The company also offers Baidu MP3 Search that provides algorithm-generated links to songs and other multimedia files provided by Internet content providers; Baidu Image Search, which enables users to search millions of images on the Internet; Baidu Space to create personalized homepages in a query-based searchable community; Baidu Encyclopedia; and other online search products and software tools. Baidu.com designs and delivers its online marketing services to its P4P and tailored solutions customers based on their requirements. The company's auction-based P4P services enable its customers to bid for priority placement of their links in keyword search results. Baidu.com primarily serves small and medium enterprises, large domestic corporations, and Chinese divisions or subsidiaries of large multinational corporations in the e-commerce, information technology services, consumer products, manufacturing, health care, entertainment, education, financial services, and real estate and other industries. The company was founded in 2000 and is headquartered in Beijing, China.

Baidu.com, Inc.
12th Floor Ideal International Plaza
No 58 West-North 4th Ring
Beijing, 100080
Tel: +86 10 8262 1188
Fax: +86 10 8260 7007
Web www.baidu.com

6.4.3.2. Yandex (Russia)

Yandex (Russian: Я́ндекс)[103] is a Russian search engine and one of the biggest Russian Web portals. It has been online since 1997. Its name can be explained as "Yet Another iNDEXer" (yandex) or "Языково́й (language) Index". Besides the Russian word "Я" corresponds to the English pronoun "I", "Яndex" looks a little bit like translation.

According to research studies conducted by Gallup Media, FOM and Comcon, Yandex is the largest resource and largest search engine in Russian Internet, based on the audience size and internet penetration.

Yandex LLC became profitable in November of 2002. In 2004 Yandex sales increased to $17M, which was 10 times greater than the company revenues just 2 years earlier, in 2002. The net income of the company in 2004 constituted $7M. In June of 2006 the weekly revenue of Yandex.Direct context ads system exceeded $1M. All of Yandex accounting measures have been audited by Deloitte & Touche since 1999.

The closest competitors of Yandex in the Russian market are Rambler and Mail.ru. Although services like Google and Yahoo! are also used by Russian users and have Russian interfaces, Google has about 21-27% of search engines generated traffic to Russian sites and Yandex has around 42-49% (Mar 2007). In Ukraine Yandex enjoys 16 percent share of the search traffic while Google has 40 percent share.

One of the biggest Yandex advantages for Russian-language users is understanding Russian inflection in search queries.

[103] From Wikipedia

In March 2007 Yandex acquired social networking site Moikrug.ru - - a Russian social network to search and support professional and personal contacts

Yandex Яндекс
Address: 1, building 21, Samokatnaya St.,
Moscow 111033
tel. +7 495 739-70-00,
fax +7 495 739-70-70

6.4.3.3. Rambler (Russia)

Rambler Media's main website is Rambler.ru, a leading and the oldest Russian language internet portal, which combines search with email/communication and community activities and media and entertainment services. Rambler.ru aggregates the best of class internet media and services in Russia and enables mass audiences to navigate to specific pages according to their interests.

Rambler Media generates revenues primarily from advertising, which includes banner or display advertising, context display advertising and sponsored key word searches, e-commerce referral and product placement. Rambler Media incorporates a full-service wholly-owned advertising agency called Index 20 in charge of generating sales from display advertising. In 2005, Rambler Media introduced “sponsored links search” and &lqduo;context advertising” through Begun (meaning “Runner” in Russian). Begun is one of Russia's leading search and contextual text based advertising platforms with a network of over 35,000 individual advertisers and over 50,000 partner distribution sites. Rambler Media has a 25.1% interest in Begun.

Rambler Media has been publicly traded on the AIM market of the London Stock Exchange (LSE: RMG) since June 2005.

In December 2006, Prof-Media, one of Russia's largest media holding groups and a major private investor in most sectors in the Russian media market, became Rambler Media's majority shareholder by acquiring approximately 55% of Rambler Media.

Rambler
Leninskaya sloboda, 2
115280, Moscow, Russia
Phone/Fax: +7 (495) 745-3619
E-mail: info@ramblermedia.co

6.4.4. European Actors

6.4.4.1. Fast (Norway)

Business Summary

FAST's Business is Enterprise Search. Since we set up our company in Norway back in 1997. We are the market leader in Enterprise Search and number one in revenue growth. We have no debt. We have been profitable, exceeding our projections, for every quarter during the last 4 years. And we have made these profits while investing a quarter of our income back into R&D. Performance like this gives us the freedom to invest in innovation and win on value and financial return.

Headquarters Oslo Offices: Helsinki, Paris, Frankfurt, München, Milano, Rome, Tromsø, Madrid, Zürich, Amsterdam, London.

6.4.4.2. Exalead (France)

Founded in 2000 by search-engine pioneers, Exalead is a global provider of software that is designed to simplify all aspects of information search and retrieval for organizations of all sizes.

Based on the first and only unified technology platform for desktop, intranet or Web search, Exalead offers easier deployment, administration and use than any other enterprise-type search software. This is true whether for one or thousands of desktops, a small business or global enterprise, and conforms to any technology environment. It also adapts to user habits for a uniquely satisfying search experience.

Exalead software is used by leading banking and financial services, media, consumer packaged goods, research, retailing sports entertainment and telecommunications companies. Exalead is an operating unit of Qualis, an international holding company.

6.4.4.3. NetSprint (Poland)

NetSprint.pl, formerly XOR Internet, was established in Warsaw in 2000. From the very launch of operation, NetSprint has been focused on creating precise and efficient search engines. The goal and ambition of NetSprint is to provide users with a quick and intuitive tool for retrieval of any type of information, both on the Internet and in closed archives.

Thanks to the experience of our IT team and our focus the needs of local users and customers, the solutions offered by NetSprint are more effective and more attractive in terms of price than the corresponding products of our global competitors.

In November 2004 Netsprint.pl ranked seventh in the Rising Stars category of the prestigious Fast 50 Deloitte ranking of the fastest-growing technology companies in Central Europe.

The Netsprint search engine is available on the Polish and Lithuanian markets. In Poland, it is used on the NetSpint.pl site, the Wirtualna Polska portal and on over 140 other big Internet sites. NetSprint is also accessible on several thousand amateur pages in its amateur version (the so called "skin"). In April 2004 the NetSprint search engine won the 4-th edition of the "Internet Now" competition in the category of "Data-base, catalogues, search engines". In June 2005 NetSprint.pl won again in the 5-th edition of "Internet Now" competition.

Market: Poland and Lithuania (Netsprint.lt)
Websites which use the engine: NetSprint.pl, wp.pl, and 140 others.
NetSprint.pl Sp. z o.o.
ul. Bieżanowska 7
02-655 Warszawa, Poland
tel. (022) 844 49 90, fax (022) 852 20 60
http://firma.netsprint.pl/

6.4.4.4. Morfeo (Czech Republic)

The search engine Morpheo (www.morfeo.cz) was developed by scientists related to Charles University in Prague, mostly: Martin Mares (http://mj.ucw.cz/) and Robert Špalek (http://www.ucw.cz/~robert/index-en.html). The development has been sponsored by the advertising company Netcentrum s.r.o. (http://www.netcentrum.cz) which is also one of the most important users and works as an exclusive distributor of the commercial version.

Back in 1997, Martin Mareš wrote the first version called Sherlock 1.0 as his term project at MFF UK but it somehow escaped from his control soon – in October 1997 it was indexing the whole .cz domain in cooperation with the Bajt company. The time slowly passed by, the author was busy working on other stuff, Bajt had its own problems and the whole project would have been almost forgotten weren't it for people from Netcentrum who were building a new Czech portal, wanted to use Sherlock for searching and were willing to sponsor its further development. After several years of successfully running Sherlock 1.2 on a couple of servers, Robert Spalek joined the "team" and together we decided to rewrite the whole project from scratch and change the whole architecture (confirming the ancient wisdom that every good program including TeX has to be

rewritten at least once in its lifetime :)). Unfortunately, we have been forced to delay the public release of this version for some time. So was it back in 2001. In September 2002, we have resurrected the freely distributable version of Sherlock, but in the meantime Apple started distributing another program of the same name as part of their OS X, so we decided to rename the whole package to Sherlock Holmes (or Holmes) to avoid both confusion and trademark problems.

Market: Czech Republic, Slovakia, Poland
Websites which use the engine: onet.pl, morfeo.cz, morfeo.sk
Netcentrum S.R.O
Drtinova 557/10
15000 Praha 5, Czech Republic
Phone : +420 227 018 100
Fax : +420 227 018 104
Web site : o.centrum.cz

6.4.4.5. Autonomy (United Kingdom)

Autonomy is the acknowledged leader in the rapidly growing area of Meaning-Based Computing (MBC). Founded in 1996 and utilizing a unique combination of technologies borne out of research at Cambridge University, the company has experienced a meteoric rise and currently has a market cap of $4 billion and offices worldwide. Autonomy's position as industry leader is widely recognized by analysts including Gartner Group, Forrester Research and Delphi, which calls Autonomy the fastest growing public company in the space. Autonomy's revenues are twice that of its nearest rival.

Meaning-Based Computing extends far beyond traditional methods such as keyword search which simply allow users to find and retrieve data. Keyword search engines for example cannot comprehend the meaning of information; these products were developed simply to find documents in which a word occurs. Unfortunately, this inability to understand information means that other documents that discuss the same idea (i.e. are relevant) but use different words are overlooked. Equally, documents with a meaning entirely different to that which the user searches for are frequently returned, forcing the user to alter their query to accommodate the search engine.

In addition, some of the key functionality of Meaning-Based Computing such as automatic hyperlinking and clustering are simply not available in keyword search engines. For example, automatic hyperlinking which connects users to a range of pertinent documents, services or products that are contextually linked to the original text requires that the meaning of the original document is fully understood. Similarly for computers to automatically collect, analyse and organize information computers have to be able to extract meaning. Only Meaning-Based Computing Systems can do this.

Revenue USD 250.1 million (2006), 116% higher compared to 2005
Employees 1,300
Autonomy Corporation plc
Cambridge Business Park
Cowley Rd
Cambridge CB4 0WZ, United Kingdom
Tel: +44 (0) 1223 448000
Fax: +44 (0) 1223 448001
www.autonomy.com

6.4.4.6. *Expert System (Italy)*

Expert System S.p.A
Founded Modena, Italy (1989) Products Cogito
Employees 140 (2007)
Expert System S.p.A
Via Virgilio, 56/Q - Staircase 5
41100 Modena – Italy
Tel: +39 059 894011
Fax: +39 059 894099
info@expertsystem.net
www.expertsystem.net

7. SEARCH ENGINES FOR AUDIO-VISUAL CONTENT: LEGAL ASPECTS, POLICY IMPLICATIONS & DIRECTIONS FOR FUTURE RESEARCH

7.1. Introduction

We are currently witnessing a trend of data explosion**.** In June 2005, the total number of Internet sites was believed to be in the order of 64 million, with two digit annual growth rates. This data comes in a variety of formats, and content has evolved far beyond pure text description. It can be assumed that search engines, in order to cope with this increased creation of audiovisual (or multimedia) content, will increasingly become audio-visual (AV) search engines.

By their nature, audio-visual search engines promise to become a key tool in the audio-visual world, as did text search in the current text-based digital environment. Clearly, AV search applications would be necessary in order to reliably index, sift through, and 'accredit' (or give relevance to) any form of audiovisual (individual or collaborative) creations. AV search moreover becomes central to predominantly audiovisual file-sharing applications. AV search also leads to innovative ways of handling digital information. For instance, pattern recognition technology will enable us to search for categories of images or film excerpts. Likewise, AV search could be used for gathering all the past voice-over-IP conversations in which a certain keyword was used. However, if these key applications are to emerge, search technology must transform rapidly in scale and type. There will be a growing need to investigate novel audio-visual search techniques built, for instance, around user behaviour. Therefore, AV search is listed as one of the top priorities of the three major US-based search engine operators - Google, Yahoo! and Microsoft. The French Quaero initiative, for the development of a top-notch AV search portal, or the German Theseus research programme on AV search, provide further evidence of the important policy dimension.

This paper focuses on some legal and policy challenges for European content industries emanating from the development, marketing and use of AV search applications. As AV search engines are still in their technological infancy, drawing attention to likely future prospects and legal concerns at an early stage may contribute to improving their development. The paper will thus start with a brief overview of trends in AV search technology and market structure.

The central part of this paper emphasises the legal, regulatory and policy dimension of AV search. The possibility exists that existing regulation is lagging behind technological, market and social developments: search engines may either fall between the mazes of existing legal regulation, or the application of existing law to search engines may be sub-optimal from the viewpoint of policy-makers. In order to assess the situation, a variety of EU directives and selected national laws have been screened, including:

- *intellectual property rights* (trademarks, copyright, patents)
- *competition law* (horizontal & vertical integration, joint dominance)
- *media law* (transparency, oversight, media pluralism, content regulation)
- *e-commerce law* (liability, self- and co-regulation, codes of conduct)
- *communications law* (EU electronic communications package)
- *law of obligations* (consumer protection, anti-spyware/spam, security defects)
- *criminal law* (e.g. anti-terrorism)
- *constitutional law & fundamental rights* (freedom of expression, property, privacy)

A fully-fledged analysis of all those legal obligations is beyond the scope of this paper. However, bearing in mind the complete set of obligations, the paper considers a select number of laws in more detail. The search engine landscape consists of three main parts. First, there is a large number of content providers that make their content available for indexing by the search engine's crawlers. Second, there are the advertisers that provide most of the income for the search engine activity. Finally, search engines interact with users, and the relevance of their search results depends to a large extent on the user data they gather.

The relation between search engines and content providers is regulated by means of copyright law. Copyright law, with its dual economic and cultural objectives is a critical policy tool in the information society because it takes into account the complex nature of information goods. It seeks to strike a delicate balance at the stage of information creation. Copyright law affects search engines in a number of different ways, and determines the ability of search engine portals to return relevant organic results.[104] Courts across the globe are increasingly called on to consider copyright issues in relation to search engines. This paper analyses some recent case law relating to copyright litigation over deep linking, provision of snippets, cache copy, thumbnail images, news gathering and other aggregation services (e.g. Google Print).

The relation between search engines and advertisers is regulated by means of trademarks law. Trademarks are important for search engines. If they cannot sell keywords freely, they are not worth their market valuation. If competitors are allowed to buy ad keywords that contain registered trademarked names, then the search engine may be diverting some of the income streams away from the owners of the trademarked words toward their competitors. There has been intense litigation on this issue on both sides of the Atlantic. US courts are currently undecided but leaning towards giving leeway to search engines; EU courts, on the other hand, seem to be in favour of giving TM holders broad rights in relation to the use of their registered TM by search engines. This paper considers issues involving the use of TM terms in meta-tag for search engine optimisation, in search engine advertising auctions, and in organic results.

The relation between search engines and their users depends to a large extent on data protection law. Recently, search engine providers have been confronted with a series of significant complaints regarding the logging of user data The question arose whether these practices are in compliance with existing EU data protection and data retention obligations, and more generally, whether search engine regulation is in line with the fundamental right to protection of private life. This paper considers the potential impact of data protection and privacy laws on the development of a thriving European AV search engine market. The paper includes a brief overview of the manner in which search engines profile as well as the commercial and other reasons behind these profiling activities. The paper reviews recent high profile cases in the US (COPA, AOL) and EU (WP 29 debate). It discusses the likely application of current legal regulatory obligations to search engines, and considers the response of search engines both in terms of technological change as well as proposals to amend existing legal regulation.

The laws are not the same for the whole of Europe. Though they are harmonized to a certain extent, there are differences in each EU Member State. It is not the intention of this paper to address particular legal questions from the perspective of a particular jurisdiction or legal order. Instead, the analysis tackles the various questions from the higher perspective of European policy. The aim is to inform European policy in regard to AV search through legal analysis, and to investigate how specific laws could be viable tools in achieving EU policy goals.

Finding the proper regulatory balance in each of these areas of regulation will play a pivotal role in fostering the creation, marketing and use of AV search engines. For instance, too strong copyright,

[104] Organic (or natural) results are not paid for by third parties, and must be distinguished from sponsored results or advertising displayed on the search engine portal. The main legal problem regarding sponsored results concern trademark law, not copyright law.

trademark or data protection laws may hamper the development of the AV search market; it may affect the creation and availability of content, the source of income of AV search engine operators, as well as their capacity to improve and personalise search engine resuls. Conversely, laws which are unduly lenient for AV search engine operators may inhibit the creation of sufficient content, put their advertising income at risk, or instill fear of pervasive user profiling and surveillance. The paper refers each time to relevant developments in the text search engine sector, and considers to what extent the specificities of AV search warrant a different approach.

Section 2 briefly describes the functioning of web search engines and highlights some of the key steps in the information retrieval process that raise copyright issues. Section 3 reviews the market context, and business rationales. Section 4 offers the main legal questions and arguments relating to copyright (relation with content providers), trademarks (relation with advertisers), and data protection (relation with users). Section 5 places these debates in the wider policy context and infers three key messages. Section 5 offers some tentative conclusions.

7.2. Search Engine Technology

For the purposes of this paper, the term 'web search engine' refers to a service available on the Internet that helps users find and retrieve content or information from the publicly accessible Internet.[105] The best known examples of web search engines are Google, Yahoo!, Microsoft and AOL's search engine services. Web search engines may be distinguished from search engines that retrieve information from non-publicly accessible sources. Examples of the latter include those that only retrieve information from companies' large internal proprietary databases (e.g. those that look for products in eBay or Amazon, or search for information inside Wikipedia), or search engines that retrieve information which, for some reason, cannot be accessed by web search engines.[106] Similarly, we also exclude from the definition those search engines that retrieve data from closed peer-to-peer networks or applications which are not publicly accessible and do not retrieve information from the publicly accessible Internet. Though many of the findings of this paper may be applicable to many kinds of search engines, this paper focuses exclusively on publicly accessible search engines that retrieve content from the publicly accessible web.

Likewise, it is better to refer to search results as "content" or "information", rather than web pages, because a number of search engines retrieve other information than web pages. Examples include search engines for music files, digital books, software code, and other information goods.[107]

In essence, a search engine is made up of three essential technical components: the crawlers or spiders, the (frequently updated) index or database of information gathered by the spiders, and the query algorithm that is the 'soul' of the search engine. This algorithm has two parts: the first part defines the matching process between the user's query and the content of the index; the second (related) part of this algorithm sorts and ranks the various hits. The process of searching can roughly be broken down into four basic information processes, or exchanges of information: a) information gathering, b) user querying, c) information provision, and d) user information access.

---

[105] *See* for a similar definition, James Grimmelmann, *The Structure of Search Engine Law* (draft), October 13, 2006, p.3, *at* http://works.bepress.com/james_grimmelmann/13/.

[106] Part of the publicly accessible web cannot be detected by web search engines, because the search engines' automated programmes that index the web, crawlers or spiders, cannot access them due to the dynamic nature of the link, or because the information is protected by security measures. Although search engine technology is improving with time, the number of web pages increases drastically too, rendering it unlikely that the 'invisible' or 'deep' web will disappear in the near future. As of March 2007, the web is believed to contain 15 to 30 billion pages (not sites), of which one fourth to one fifth is estimated to accessible by search engines. *See* and compare www.pandia.com/sew/383-web-size.html and http://technology.guardian.co.uk/online/story/0,,547140,00.html.

[107] Search engines might soon be available for locating objects in the real world. *See* John Battelle, *The Search: How Google and its rivals rewrote the rules of business and transformed our culture (2005)*, p 176. See James Grimmelmann, *supra*.

7.2.1. Four Basic Information Flows

7.2.1.1. Search Engines Gather and Organise Content

In the beginning of the search engines' life cycle, web masters were encouraged to submit information directly to the search engines operators.[108] Though this is still one possible method, today's major search engines do not require any extra effort to submit information, as they are capable of finding pages via links on other sites. The web search process of gathering information is driven primarily by automated software agents called robots, spiders, or crawlers that have become central to successful search engines.[109] The agents do not actually visit the pages or content repositories. The process is not so different from what a browser does: the software agent exchanges information with the content provider.

7.2.1.2. Users Query the Search Engine: From 'Pull' to 'Push'

The second major information flow that determines search results is the series of queries the user inputs in the search box. User queries may be divided in three categories: navigational (the user wants to find specific information), informational (the user is looking for new data or facts), and transactional (the user is seeking to purchase something).[110] The query is usually made of a couple of keywords. A number of new search engines are being developed at the moment that propose query formulation in full sentences,[111] or in audio, video, picture format.

Most search engines start recording (or logging) the user information in order to offer better search results. One trend is, for instance, the provision of increasingly personalized search results, tailored to the particular profile and search history of each individual user.[112] Another major trend is the development by search engines of information gathering services regarding news, and other types of information. At the intersection of these trends lies the development of proactive search engines that crawl the web and 'pushes' information towards the user according to this user's search history and profile.

---

[108] It is acknowledged that Google and Yahoo still offer submission programs, while some search engines, including Yahoo!, even operate paid submission services that assure the inclusion into the database, but do not secure any specific ranking within the search results. But these practices are now no longer mainstream.

[109] There are of course alternatives on the market, such as the open directory project whereby the web is catalogued by humans, or search engines that tap into the wisdom of crowds to deliver relevant information to their users, such as Wiki Search, the wikipedia search engine initiative (http://search.wikia.com/wiki/Search_Wikia), or ChaCha (http://www.chacha.com/). *See* Wade Roush, New Search Tool Uses Human Guides, *Technology Review*, February 2, 2007, *at* http://www.techreview.com/Infotech/18132.

[110] *See* Andrei Broder, *A Taxonomy of Web Search*, 36 ACM SIGIR Forum, no.2 (2002), *at* http://www.acm.org/sigs/sigir/forum/F2002/broder.pdf.

[111] *See* Stefanie Olson, Spying an Intelligent Search Engine, ZDNet, August 21, 2006, *at* http://www.zdnet.com.au/news/communications/soa/Spying_an_intelligent_search_engine/0,130061791,139267128,00.htm

[112] *See* Your Google Search Results Are Personalised, http://www.seroundtable.com/archives/007384.html. *See* also Kate Greene, A More Personalized Internet?, *Technology Review*, February 14, 2007, *at* http://www.technologyreview.com/Infotech/18185/. This raises intricate data protection issues. *See* Boris Rotenberg, Towards Personalised Search: EU Data Protection Law and its Implications for Media Pluralism. In Machill, M.; M. Beiler (eds.): *Die Macht der Suchmaschinen / The Power of Search Engines*. Cologne [Herbert von Halem] 2007, forthcoming. Profiling will become an increasingly important way for identification of individuals. It will raise concerns in terms of privacy and data protection. This interesting topic is however outside the scope of this paper (information can be found elsewhere. *See* Clements, B, Maghiros I, Beslay L, Centeno C, Punie Y, Rodriguez C, Masera M, "Security and privacy for the citizen in the Post-September 11 digital age: A prospective overview" 2003, EUR 20823 available at www.jrc

7.2.1.3. Search Engines Return Results

The third information flow is the provision of relevant search engine results by the search engine to its user. This is often an iterative process in the sense that the user may want to refine his or her query according to the results that are returned by the search engine. Better search engines provide more relevant results without the user's need to insert too many queries.

The key here for search engines is to determine relevance of specific content for a given query. In the past, search engines relied uniquely on the text of web sites. Over time, however, search engines have become more sophisticated, integrating metadata (data about the pages or content), tags, user click stream data, as well as the link structure. The latter involves information about which pages link in and out of which pages. Link structure analysis is helpful, for instance, in determining the popularity of content. Search engines thus make use of complex ranking algorithms with more than 100 factors for ranking content. Every search engine has its own recipe on the factors to evaluate the ranking of web pages. For instance, Google makes use of the well-known PageRank concept.[113] Given that every search engine uses hundreds of factors for the ranking, whose composition and weight can change continually, and because their respective algorithms are also different, results are likely to be quite distinct between competing search engines. A web page that ranks high in a particular search engine can rank lower in another search engine or even on the same search engine some days later.

Because of its importance in returning relevant results and giving search engines a competitive edge, the ranking algorithms are widely considered the soul of the search engine. Generally speaking, details on their algorithms and architecture – particularly for the crawlers, indexers, and ranking – are kept behind vaulted doors as business secrets.[114]

One important point that needs to be stressed here is the fact that the process is increasingly automated, with as little human intervention as possible. The process of mechanically making sense of the masses of information that is available on the Internet is now reaching a high level of sophistication. This can be seen at the stage of gathering information, user querying, and returning of relevant results.

7.2.1.4. Users obtain the Content

The line between search engines and content providers is increasingly blurred. Many providers of online services provide search engines for their own services. The same holds for sites that are aggregations of user produced content. Likewise, decentralized peer-to-peer networks use the same resources provided by users (computing power, bandwidth, storage) to retrieve and provide content to its community of users.

In addition, a number of search engines provide content directly to their users. They store content on their cache, in order to make it easier for the user to retrieve the information. They archive content, enabling users to receive the information, even when the original content is no longer available. For visual information, it is now common practice for many search engines to provide thumbnails (or smaller versions) of pictures.

Simply put, search engines are powerful intermediaries that determine or facilitate the connection or information exchange between content or information providers, and users. Each such connection

[113] PageRank is an algorithm that weighs a page's importance based upon the incoming links. PageRank interprets a link from Page A to Page B as a vote for Page B by Page A. It then assesses a page's importance by the number of votes it receives. PageRank also considers the importance of each page that casts a vote, as votes from some pages are considered to have greater value, thus giving the linked page greater value. In other words, the PageRank concept values those links higher which are more likely to be reached by the random surfer.

[114] Search engines also increasingly learn from the large volumes of user data. Query histories provide valuable information by which search engines can improve the relevance of their results.

may be detrimental to the users, content providers, or third users (be they competing content providers, regulators, or advertisers) who would rather not have such connection occur.[115]

7.2.2. Search Engine Operations and Trends

7.2.2.1. Indexing

Once the crawler has downloaded a page and stored it on the search engine's own server, a second programme, known as the indexer, extracts various bits of information regarding the page. Important factors include the words the web page or content contains, where these key words are located and the weight that may be accorded to specific words and any or all links the page contains. The index is further analysed and cross-referenced to form the runtime index that is used in the interaction with the user.

A search engine index is like a big spreadsheet of the web. The index breaks the various web pages and content into segments. It stores where the words were located, what other words were near them, and analyses the use of words and their logical structure. By clicking on the links provided in the engine's search results, the user may retrieve from the server the actual version of the page. Importantly, the index is not an actual reproduction of the page or something a user would want to read.

7.2.2.2. Caching

Most of the major search engines now provide "cache" versions of the web pages that are indexed. The search engine's cache is, in fact, more like a temporary archive. Search engines routinely store for a long period of time, a copy of the content on their server. When clicking on the "cache version", the user retrieves the page as it looked the last time the search engine's crawler visited the page in question. This may be useful for the user if the server is down and the page is temporarily unavailable, or if the user intends to find out what were the latest amendments to the web page.

7.2.2.3. Robot Exclusion Protocols

Before embarking on legal considerations, it is worth recalling the regulatory effects of technology or code. Technology or 'code' plays a key role in creating contract-like agreements between content providers and search engines. For instance, since 1994 the robot exclusion standard has allowed newspapers to prevent search engine crawlers from indexing or caching certain content. Web site operators can do the same by simply making use of standardised html code. Add '/robots.txt' to the end of any site's web address and it will indicate the site's instructions for search engine crawlers. Similarly, by inserting NOARCHIVE in the code of a given page, web site operators can prevent caching. Each new search engine provides additional, more detailed ways of excluding content from its index and/or cache. These methods are now increasingly fine-grained, allowing particular pages, directories, entire sites, or cached copies to be removed.[116]

Standardising bodies are currently working on implementing standardised ways to go beyond the current binary options (e.g. to index or not to index). Right now content providers may opt-in or opt-out, and robot exclusion protocols also work for keeping out images, specific pages (as opposed to entire web sites), but many of the intermediate solutions are technologically harder to achieve. Automated Content Access Protocol (ACAP) is a standardized way of describing some of the more fine-grained intermediate permissions, which can be applied to web sites so that they can be decoded by the crawler. ACAP might – for instance – indicate that text can be copied, but not the

[115] *See* about an attempt to offer more transparency: Google Webmaster Central Adds Link Analysis Tool, February 6, 2007, *at* http://www.seroundtable.com/archives/007401.html.

[116] *See* for a detailed overview Danny Sullivan, Google releases improved Content Removal Tools, *at* http://searchengineland.com/070417-213813.php.

pictures. Or it could say that pictures can be taken on condition that photographer's name also appears. Demanding payment for indexing might also be part of the protocol.[117] This way, technology could enable copyright holders to determine the conditions in which their content can be indexed, cached, or even presented to the user.

7.2.2.4. From Text Snippets & Image Thumbnails to News Portals

Common user queries follow a 'pull'-type scheme. The search engines react to keywords introduced by the user and then submit potentially relevant content.[118] Current search engines return a series of text snippets of the source pages enabling the user to select among the proposed list of hits. For visual information, it is equally common practice to provide thumbnails (or smaller versions) of pictures.

However, search engines are changing from a reactive to a more proactive mode. One trend is to provide more personalized search results, tailored to the particular profile and search history of each individual user.[119] To offer more specialized results, search engines need to record (or log) the user's information. Another major trend is news syndication, whereby search engines collect, filter and package news, and other types of information. At the intersection of these trends lies the development of proactive search engines that crawl the web and 'push' information towards the user, according to this user's search history and profile.

7.2.2.5. Audio-visual search

Current search engines are predominantly text-based. They gather, index, match and rank content by means of text and textual tags. Non-textual content like image, audio, and video files are ranked according to text tags that are associated with them. While text-based search is efficient for text-only files, this technology and methodology for retrieving digital information has important disadvantages when it is faced with other formats than text. For instance, images that are very relevant for the subject of enquiry will not be listed by the search engine if the file is not accompanied with the relevant tags or textual clues. Although a video may contain a red mountain, the search engine will not retrieve this video when a user inserts the words "red mountain" in his search box. The same is true for any other information that is produced in formats other than text. In other words, a lot of relevant information is systematically left out of the search engine rankings, and is inaccessible to the user. This in turn affects the production of all sorts of new information.[120]

There is thus a huge gap in our information retrieval process. This gap is growing with the amount of non-textual information that is being produced at the moment. Researchers across the globe are currently seeking to bridge the gap. One strand of technological developments could provide a solution on the basis of text formats by, for instance, developing intelligent software that

---

[117] *See* Struan Robertson, Is Google Legal?, *OUT-LAW News*, October 27, 2006, *at* http://www.out-law.com/page-7427

[118] A number of new search engines are being developed at the moment that propose query formulation in full sentences, or in audio, video, picture format.

[119] *See* Your Google Search Results Are Personalised, http://www.seroundtable.com/archives/007384.html. *See* also Kate Greene, A More Personalized Internet? *Technology Review*, February 14, 2007, *at* www.technologyreview.com/Infotech/18185/. This raises intricate data protection issues. *See* Boris Rotenberg, Towards Personalised Search: EU Data Protection Law and its Implications for Media Pluralism. In Machill, M.; M. Beiler (eds.): *Die Macht der Suchmaschinen / The Power of Search Engines*. Cologne [Herbert von Halem] 2007, pp.87-104. Profiling will become an increasingly important way for identification of individuals, raising concerns in terms of privacy and data protection. This interesting topic is however beyond of the scope of this paper (information can be found elsewhere. *See* Clements, B, et al., "Security and privacy for the citizen in the Post-September 11 digital age: A prospective overview" 2003, EUR 20823 available at www.jrc.es

[120] *See* Matt Rand, Google Video's Achilles' Heel, *Forbes.com*, March 10, 2006, *at* http://www.forbes.com/2006/03/10/google-video-search-tveyes-in_mr_bow0313_inl.html.

automatically tags audio-visual content.[121] Truveo is an example of this for video,[122] and SingingFish for audio content.[123] Another possibility is to create a system that tags pictures using a combination of computer vision and user-inputs.[124]

AV search often refers specifically to new techniques better known as content-based retrieval. These search engines retrieve audio-visual content relying mainly on pattern or speech recognition technology to find similar patterns across different pictures or audio files.[125] These pattern or speech recognition techniques make it possible to consider the characteristics of the image itself (for example, its shape and colour), or of the audio content. In the future, such search engines would be able to retrieve and recognise the words "red mountain" in a song, or determine whether a picture or video file contains a "red mountain," despite the fact that no textual tag attached to the files indicate this.

This sector is currently thriving. Examples of such beta versions are starting to reach the headlines, both for visual and audio information. Tiltomo[126] and Riya[127] provide state-of-the-art content-based image retrieval tools that retrieve matches from their indexes based on the colours and shapes of the query picture. Pixsy[128] collects visual content from thousands of providers across the web and makes these pictures and videos searchable on the basis of their visual characteristics. Using sophisticated speech recognition technology to create a spoken word index, TVEyes[129] and Audioclipping[130] allow users to search radio, podcasts, and TV programmes by keyword.[131] Blinkx[132] and Podzinger[133] use visual analysis and speech recognition to better index rich media content in audio as well as video format. The most likely scenario, however, is a convergence and combination of text-based search and search technology that also indexes audio and visual information.[134] For instance, Pixlogic[135] offers the ability to search not only metadata of a given image but also portions of an image that may be used as a search query.

Two preliminary conclusions may be drawn with respect to AV search. First, the deployment of AV search technology is likely to reinforce the trends discussed above. Given that the provision of relevant results in AV search is more complex than in text-based search, it is self-evident that these will need to rely even more on user information to retrieve pertinent results. As a consequence, it

---

[121] *See* about this James Lee, Software Learns to Tag Photos, *Technology Review*, November 9, 2006, *at* http://www.technologyreview.com/Infotech/17772/.

[122] http://www.truveo.com

[123] SingingFish was acquired by AOL in 2003, and has ceased to exist as a separate service as of 2007. See http://en.wikipedia.org/wiki/Singingfish

[124] *See* Michael Arrington, Polar Rose: Europe's Entrant Into Facial Recognition, *Techcrunch*, December 19, 2006, *at* http://www.techcrunch.com/2006/12/19/polar-rose-europes-entrant-into-facial-recognition.

[125] Pattern or speech recognition technology may also provide for a cogent way to identify content, and prevent the posting of copyrighted content. See, Associated Press, MySpace launches pilot to filter copyright video clips, using system from Audible Magic, *Technology Review*, February 12, 2007 *at* http://www.technologyreview.com/read_article.aspx?id=18178&ch=infotech.

[126] http://www.tiltomo.com.

[127] http://www.riya.com.

[128] http://www.pixsy.com.

[129] http://www.tveyes.com; TVEyes powers a service called Podscope (http:// www.podscope.com) that allows users to search the content of podcasts posted on the Web.

[130] http://www.audioclipping.de.

[131] *See* Gary Price, Searching Television News, *SearchEngineWatch*, February 6, 2006, *at* http://searchenginewatch.com/showPage.html?page=3582981. See

[132] http://www.blinkx.com.

[133] http://www.podzinger.com.

[134] *See* Brendan Borrell, Video Searching by Sight and Script, *Technology Review*, October 11, 2006, *at* http://www.technologyreview.com/read_article.aspx?ch=specialsections&sc=personal&id=17604.

[135] http://www.pixlogic.com.

seems likely that we will witness an increasing trend towards AV content 'push', rather than merely content 'pull'. Second, the key to efficient AV search is the development of better methods for producing accurate meta-data that describe the AV content. This makes it possible for search engines to organise the AV content optimally (e.g. in the run-time index) for efficient retrieval. One important factor in this regard is the ability of search engines to have access to a wide number of AV content sources on which to test their methods. Another major factor is the degree of competition in the market for the production of better meta-data for AV content. Both these factors (access to content, market entry) are intimately connected with copyright law.

7.3. Market Developments

The technology does not operate in a vacuum. By virtue of the Internet's development, search engines have become vital players. But they can only carry out their mission through their interaction with content or information providers, advertisers, and users. This section will first consider the pivotal role of search engines in the information society. Second, it will provide a brief description of the search engine landscape – that is, the various players involved and their relation with search engines. Finally, it will concisely show how the centrality of search has led a number of players in the digital economy to adapt their business models to this new reality.

7.3.1. The Centrality of Search

Although dominated by three US-based giants (i.e. Google, Yahoo! and Microsoft), the search engine market is currently extremely active. The search engine space spans across all sorts of information. We currently witness the deployment of search engines for health, property, news, job, person, code or patent information. They will increasingly be able to sift through information coming from a wide range of information sources (including emails, blogs, chat boxes, etc.) and devices (desktop, mobile). Search engines are able to return relevant search results according to the user's geographic location or search history. Virtually any type or sort of information, any type of digital device or platform, may be relevant for search engines.

Search is also increasingly a central activity that has become the default manner to interact with the vast amounts of information that are available on the Web. For most users the search box is the entry door into the digital environment. Many queries or intentions in the user's mind, whether navigational, transactional, or informational, take the shape of a few words in the search box. Some commentators therefore consider search functionality the core to the development of the emerging application platform. That emerging platform supports server side, AJAX-based online applications that can run smoothly within a web browser.[136]

This centrality is evident from the vast amounts of traffic that flows through the major search engines. Search engines are heavily used intermediaries. The search volume for January 2007 is more than 7.19 billion searches in the USA alone.[137] The volume and the market shares may vary slightly by the method the investigation has been carried out, but the ranking is clear: Google comes on top, followed by Yahoo!, MSN, AOL and Ask.[138] Web search is thus responsible for most web

[136] *See* Stephen E. Arnold, THE GOOGLE LEGACY. HOW GOOGLE'S INTERNET SEARCH IS TRANSFORMING APPLICATION SOFTWARE, (2005); John Battelle, THE SEARCH: HOW GOOGLE AND ITS RIVALS REVROTE THE RULES OF BUSINESS AND TRANSFORMED OUR CULTURE (2005).

[137] Top Search Providers for January 2007, Nielsen/Netratings 28/02/2007, *at* http://ww.netratings.com/pr/pr_070228.pdf.

[138] At present, more than 60 search engines are operational, but the bulk of the searches are performed by few service providers only. Following the consultancy firm Nielsen/Netratings, the first three operators control more than eighty percent of the market. In particular, in January 2007 online searches in the US were executed by Google 49.2%, Yahoo! 23.8%, MSN 9.6%, AOL 6.3%, Ask 2.6 and all others together 8.5%. For the same month, comScore Networks sees Google sites capturing 47.5% of the U.S. search market, Yahoo! 28.1% and Microsoft 10,6%, Ask 5.4% and AOL 4.9% (see http://www.comscore.com/press/release.asp?press=1219). Variations between Nielsen/Netratings, comScore and other rating / traffic measuring service providers are a consequence of the measurement methods. However, the

traffic and both Google as Yahoo! offer two digits growth rates. This growth rate is considerable and explains the high expectations of online advertisement of search engines as a promising growth market.

7.3.2. The Adapting Search Engine Landscape

The search engine landscape consists of three main parts. First, there is a large number of content providers that make their content available for indexing by the search engine's crawlers. Second, there are the advertisers that provide most of the income for the search engine activity. Finally, new players have arisen whose livelihood depends on the business model of search engines.[139]

The content providers' market is in a very dynamic condition at the moment, with a number of business models competing with one another. While technology gives content providers a number of technological tools for controlling the accessing, using and sharing of content created or owned by them, the need to use of so-called Digital Rights Management (DRM)[140] tools is increasingly questioned, and currently highly contentious.[141] For instance, by January 2007 the last publisher to use DRM for audio CDs stopped doing so because the cost of implementing DRM did not measure up to the results. In the Internet music industry, an increasing amount of music is sold without DRM protection, and major players have called upon the industry to remove DRM protection.[142] Major players are also gradually discovering that giving away content "for free," may spur another type of business models that may turn out to be more profitable on the World Wide Web.[143] A number of major players are arising, for instance, in regard video sharing, such as YouTube, MySpace, or Joost and NetFlix.[144] In other words, content may well be moving from closed environment to an open environment in which being available, reachable, is of paramount importance: survival in this brave new world depends on being found by (prominent) search engines.

---

ranking amongst the search engines is stable. Equally interesting is that fact that the number of search queries increases annually by 30%. The major beneficiary is Google, which also increased its market share and saw its profits rocketing in 2006 by 110% to $3.07bn. *See* http://technology.guardian.co.uk/news/story/0,,2003373,00.html; http://business.timesonline.co.uk/article/0,,9075-2578425,00.html.

[139] Namely, these are, on the one hand, the players that offer a set of services and techniques for content providers to be ranked high in the organic results (search engine optimization), and, on the other hand, the players that fraudulently take advantage of the pay-per-click advertising model to make money.

[140] Digital rights management (DRM) tools is an umbrella term that refers to the collection of technologies used by copyright owners for protecting digital content against unwanted copying. With DRM, clients need to be authenticated to access contents. The authentication process controls the access rights clients have paid for and assures that it is delivered. Through DRM technology it is also possible to choose the level of access to the selected song, i.e. listening to the song only once, permission to save, permission to copy, to use in another media, etc. See http://en.wikipedia.org/wiki/DRM.

[141] Recently Sony Uk and Sony France have a lost a case against a consumer rights organisation because they did not inform consumers about the lack of interoperability of their products and services to other devices. See http://www.edri.org/edrigram/number5.1/drm_sonyfr (January 17, 2007). See for the judgment of December 15, 2006: http://www.tntlex.com/public/jugement_ufc_sony.pdf. A similar case is on-going against Apple's iPod in France, Germany and Norway; see Associated Press, German, French Consumer Groups Join Nordic-Led Drive Against Apple's iTunes Rules, *Technology Review*, January 22, 2007, *at* http://www.technologyreview.com/read_article.aspx?id=18098&ch=biztech, Apple DRM Illegal in Norway: Ombudsman, *The Register*, January 24, 2007; *at* http://www.theregister.co.uk/2007/01/24/apple_drm_illegal_in_norway.

[142] *See* Chris Nuttall, Apple Urges End to Online Copy Protection, *Financial Times*, February 6, 2007, *at* http://www.ft.com/cms/s/5469e6ea-b632-11db-9eea-0000779e2340.html.

[143] *See* Eric Pfanner, Internet Pushes Concept of Free Content, *Herald tribune*, January 17, 2007 *at* http://www.iht.com/articles/2007/01/17/yourmoney/media.php. See also Cory Doctorow, EMI abandons CD DRM, January 8, 2007; *at* http://www.boingboing.net/2007/01/08/emi_abandons_cd_drm.html

[144] *See* Brendan Borrell, Joost Another YouTube?, *Technology Review*, January 29, 2007; *at* http://www.techreview.com/Biztech/18111/; See The Economist Editorial, The Future of Television – What's On Next, *The Economist*, February 8, 2007; *at* http://economist.com/business/displaystory.cfm?story_id=8670279.

An important normative question regarding the relation between search engines and content providers is in how far and when content providers may have control over the search engines' basic functions. The trend, however, seems to be that a content provider may prevent a search engine from indexing or caching some of the content it provides through the use of standardised automatic robots (robot exclusion protocols). Most major search engines routinely agree to respect such exclusions.[145] Some search engines have decided to go even further. Google recently introduced Sitemaps, a new tool for content providers, which aims to give websites more control over what content they do or don't want included in Google News.[146]

The second type of players with which search engines interact on a daily basis are the advertisers. The predominant business model for search is advertising.[147] The leading search engines generate revenue primarily by delivering online advertisement. The importance of advertising for search engines is evident, also, from their spending. In 2006, Google was planning to spend 70% of its resources on search and advertising related topics.[148] A few years ago, advertising on search engine sites was very much like in analogue media. This included mainly banner advertising,[149] and sometimes paid placement, whereby ads were mixed with organic results.[150] But many users considered these too intrusive and not sufficiently targeted or relevant to the search or web site topic, and not taking advantage of the interactive nature of the Web. By contrast, online advertising differs from traditional advertising that traceability of results is easier. Mainstream search engines now mainly rely on two techniques. These are advertising business models that rely on actual user behaviour: pay-per-click (advertiser pays each time the user clicks on the ad) and pay-per-performance (advertiser pays each time the user purchases or prints or takes any action that shows similar interest).[151]

Finally, players that depend on search engines (of which there are many) have been adapting their activities in order to take advantage of the centrality of search. Both the content and advertising markets have thus been adapting rapidly to the prominence of search engines. As regards the ranking of content or information by relevance in the organic results, a range of strategies and techniques are being employed to get links from other sites. These are called search engine optimisation (SEO), and aim at raising the relevance of certain content or web site for a given query. They include two broad categories. First there are techniques that search engines recommend as part of good design and that are considered desirable because they increase the efficiency of information retrieval and lower transaction costs. But there are also those techniques that search engines do not approve of and attempt to minimize the effect of, referred to as spam-dexing.[152] Of

---

[145] *See* above, in the technology section.

[146] *See* Aoife White, Court to Hear Google-Newspaper Fight, *CBS News*, November 23, 2006, *at* http://www.cbsnews.com/stories/2006/11/23/ap/business/mainD8LITLI00.shtml.

[147] Another source of revenue is selling search functionality for business. The revenues from licensing are however modest relative to their income from advertising *see* http://investor.google.com/releases/2006Q4.html

[148] More specifically, 20% is spent on local search, Google Earth, Gmail, Google Talk, Google Video, Enterprise solutions, Book Search, Adsense, Desktop search and mobile search and the remaining 10% for Orkut, Google Suggest, Google Code, Adsense Offline, Google Movies, Google Readers, Google Pack and Wifi. *See* Jonathan Rosenberg, Google Analyst Meeting 2006, *at* http://investor.google.com/pdf/20060302_analyst_day.pdf.

[149] In this technique, the advertiser pays the search engine or platform provider each time the user sees the ad.

[150] The idea is that the bidder who values the high ranking most will pay the price for it, and as a result users will encounter information in an efficient manner.

[151] At present, pay-per-click seems to strike the best balance. On the one hand, the system provides an incentive for search engines or affiliate sites to target ads correctly; on the other hand, those advertisers who value their ranking most will be prepared to pay the most for a given set of keywords, during the online auctions. Most leading search engines provide the ads in a separate column. These ads are generated using similar algorithms as for organic search results. That is, the sponsored ad depends on the user's key words, advertiser's willingness to pay, and the popularity of this ad with other users. This selection process continuously adapts itself according to the circumstances and developments.

[152] Some industry commentators classify these methods, and the practitioners who utilize them, as either "white hat SEO", or "black hat SEO". Black hat SEO includes hiding popular keywords invisibly all over the page, or showing the

course, it is not always easy to draw a line between accepted and non-accepted optimisation techniques, and it is contentious to what extent search engines should be allowed or expected to fiddle with the results brought up by the sole functioning of the algorithm.[153] The above-depicted advertising techniques have also generated their own type of fraud, referred to as click fraud. Click-fraud refers to the situation in which a competitor to a given advertiser creates a program whereby the ads of the advertiser are clicked repeatedly, thereby artificially inflating the figures and the bill for the advertiser. Another type of click-fraud arises when a player registers as an affiliate and then repeatedly clicks on the ads he himself has served, thereby making profit.

Since they are dependent on good content and advertising income that relies on accurate measurement of user behaviour, search engines have an interest in fighting these types of malpractices. Search engines have been engaging in a technological arms race with both content and advertising fraudsters. The paradox is thus that, while search engines have an interest in keeping the image of being transparent and objective, the algorithms that determine the ranking of both the organic results and the ads remain kept behind sealed doors. This is one of the recurrent tensions underpinning search engine policy.

7.3.3. Extending Beyond Search

Search engines affect the business model of content owners by placing targeted advertising on affiliate sites.[154] Search engines use their unique capacity to link relevant ads with relevant key-words and content. Through affiliate networks, they are seeking to reach out, extending their "tentacles" deep into the fabric of the Internet.[155] In doing so, search engines may affect media players filtering and accreditation power by taking over some of their editorial functions in terms of relevance, ranking, etc. They may affect newspapers' business model by caching content and thus diminishing their sales of archived content, or by directing traffic round their front page and thus potentially curbing their advertising income. They may affect trademark owners by directing traffic to competitors, depending on their trademark policy. This highlights the power of search engines to determine or affect the business model of the various players with which it is interaction.

Some of the players have been competing head-on with the search engine to keep their share of a given market. There is such a clash between search engines and application providers, as well as between search engines and content providers. At the level of the application layer, we are currently witnessing a high degree of technological convergence: more and more creators of technology integrate search engines in their applications. Apple OSX treats search as a basic functionality of the operating system. Almost every application now on the Internet includes some sort of search functionality. But search engines also integrate new types of applications in their functionality. For instance, a number of search engines are providing open APIs with the aim of providing the next generation OS or platform with search functionality at its core.

At the level of the content layer too, search engines are increasingly starting to compete with classic media services. Search engines are now populating the many applications that they themselves provide with appropriate content; good examples are Google News, Google Print, Google Earth.

---

search engine another page than the one shown to users. Given the importance of link structure, prominent black hat SEO now includes the creation of so-called "linkfarms" with thousands of sites and pages that point to each other, giving the sense of a community of users.

[153] *See* Rachel Williams, Search engine takes on web bombers, *Sydney Morning Herald*, January 31, 2007, *at* http://www.smh.com.au/articles/2007/01/30/1169919369737.html.

[154] Search engines were instrumental in the development of banner and pay-per-click advertising. See John Battelle, THE SEARCH: HOW GOOGLE AND ITS RIVALS REVROTE THE RULES OF BUSINESS AND TRANSFORMED OUR CULTURE (2005).

[155] These "tentacles" taken together create affiliate networks, whereby the advertiser pays for each event, while the middlemen (search engines) and the site on which the ad appears share the revenue. For instance, the largest of the advertising networks are Google's AdWords/AdSense and Yahoo! Search Marketing. Other important ad networks include media companies and technology vendors.

We confuse some of the video sharing sites with the search engines, because the major search engines are transforming rapidly into full-scale platforms that also provide content. Moreover, the distinction between classic search engines that respond to a particular user query from its cache or index, and an aggregation service that provides a collection of information online is bound to become smaller in the future with the move toward ever more personalized services. The Yahoo Pipes service is one more piece of evidence of this growing trend toward proactive search services that are tailored to the user profile.[156]

At the initial of the search engines' development, we denoted a marked difference in approach between portals and search engines. Google portrayed itself as a search engine, while Yahoo! with its directory of information was considered to be more like a portal. Gradually, those two approaches have been converging with Yahoo! integrating a powerful search engine at its core, and Google providing a flurry of applications around its main search functionality, and entering the content provision market. But one of the possible consequences of this initial divide may be the fact that Yahoo! does not object so much to being considered a media player. Google, on the other hand, stresses the fact that it is merely providing a tool that facilitates access to information, all kinds of digital information.

The same tensions are defining the environment within which AV search engines unfold. The same players are competing for a share of this important market. As noted previously, we denote a rising importance of AV content online. Evidence can be garnered from 2006 figures concerning the use of YouTube and MySpace for sharing and downloading videos online. 2006 was a banner year for YouTube. The video sharing site launched in February 2005 and had claimed over 40 percent of the online video market share by May 2005. By October 2005, YouTube was logging more than 100 million video downloads per day and by the end of the year had become the sixth most popular site on the Internet.[157] We see traditional content providers such as broadcasters making deals with the online video sharing sites for the provision of their content.[158]

There is thus nothing more logical than to expect AV search to rise in importance with the explosion of AV content online. According to some analysts, image search, for instance, is the fastest growing search category on the Internet today.[159] This paper argues that legal regulation will help determine the extent to which AV search technology is able to fulfil its promise.

The next section will briefly consider some high profile copyright cases that have arisen. It will discuss the positions of content owners and search engines on copyright issues, and provide an initial assessment of the strengths of the arguments on either side.

7.4. Legal aspects

7.4.1. Copyright in the Search Engine Context

Traditional copyright law strikes a delicate balance between an author's control of original material and society's interest in the free flow of ideas, information, and commerce. Such a balance is

[156] *See* http://pipes.yahoo.com/pipes/

[157] *See* Clement James, BBC and YouTube discuss content deal, IT News.com.au, January 25, 2007, *at* http://www.itnews.com.au/newsstory.aspx?Cianid=44892; See also Gates: Internet to revolutionize TV in 5 years, *C|NET News*, January 27, 2007, http://news.com.com/2100-1041_3-6154009.html.

[158] *See* Clement James, BBC and YouTube discuss content deal, *IT News.com.au*, January 25, 2007, *at* http://www.itnews.com.au/newsstory.aspx?CIaNID=44892&r=hstory. Jane Wardell, BBC signs program deal with YouTube, *Associated Press*, March 2, 2007, *at* http://news.findlaw.com/ap/f/66/03-02-2007/a0e30018d027da47.html.

[159] Among search verticals, image search enjoyed the strongest year over year growth in February 2006, increasing 91 percent. See Nielsen/Netratings, March 30, 2006, *at* http://www.nielsen-netratings.com/pr/pr_060330.pdf.

enshrined in the idea/expression dichotomy which states that only particular expressions may be covered by copyright, and not the underlying idea.[160]

In US law, the balance is struck through the application of the "fair use" doctrine. This doctrine allows use of copyrighted material without prior permission from the rights holders, under a balancing test.[161] Key criteria determining whether the use is "fair" include questions as to whether it is transformative (i.e. used for a work that does not compete with the work that is copied), whether it is for commercial purposes (i.e. for profit), whether the amount copied is substantial, and whether the specific use of the work has significantly harmed the copyright owner's market or might harm the potential market of the original. This balancing exercise may be applied to any use of a work, including the use by search engines.

By contrast, there is no such broad catch-all provision in the EU. The exceptions and limitations are specifically listed in the various implementing EU legislations. They only apply provided that they do not conflict with the normal exploitation of the work, and do not unreasonably prejudice the legitimate interests of the right-holder.[162] Specific exemptions may be in place for libraries, news reporting, quotation, or educational purposes, depending on the EU Member State. At the moment, there are no specific provisions for search engines, and there is some debate as to whether the list provided in the EU copyright directive is exhaustive or open-ended.[163] In view of this uncertainty, it is worth analysing specific copyright issues at each stage of the search engines' working.

The last few years have seen a rising number of copyright cases, where leading search engines have been in dispute with major content providers. Google was sued by the US Authors' Guild for copyright infringement in relation to its book scanning project. Agence France Presse filed a suit against Google's News service in March 2005. In February 2006, the Copiepresse association (representing French and German-language newspapers in Belgium) filed a similar law suit against Google News Belgium.

As search engines' interests conflict with those of copyright holders, copyright law potentially constrains search engines in two respects. First, at the information gathering stage, the act of indexing or caching may, in itself, be considered to infringe the *right of reproduction*, i.e. the content owners' exclusive right "to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part" of their works.[164] Second, at the information provision stage, some search engine practices may be considered to be in breach of the *right of communication to the public*, that is, the content owners' exclusive right to authorise or prohibit any communication to the public of the originals and copies of their works. This includes making their works available to the public in such a way that members of the public may access them from a place and at a time individually chosen by them.[165]

---

[160] For a more exhaustive analysis of copyright issues, see Boris Rotenberg & Ramón Compañó, Search Engines for Audio-visual Content: Copyright Law & Its Policy Relevance, in Justus Haucap, Peter Curwen & Brigitte Preissl, *forthcoming* (2008).

[161] A balancing test is any judicial test in which the importance of multiple factors are weighed against one another. Such test allows a deeper consideration of complex issues.

[162] *See* Art.5.5, Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, *OJ L 167, 22.6.2001*.

[163] *See* IVIR, The Recasting of Copyright & Related Rights for the Knowledge Economy, November 2006, pp.64-65, *at* www.ivir.nl/publications/other/IViR_Recast_Final_Report_2006.pdf. Note, however, that Recital 32 of the EUCD provides that this list is exhaustive.

[164] *See* Art.2, Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, *OJ L 167,* 22.6.2001.

[165] *Ibid.*, Art.3.

7.4.1.1. Right of reproduction

7.4.1.1.1 Indexing

Indexing renders a page or content searchable, but the index itself is not a reproduction in the strict sense of the word. However, the search engine's spidering process requires at least one initial reproduction of the content in order to be able to index the information. The question therefore arises whether the act of making that initial copy constitutes, in itself, a copyright infringement.

Copyright holders may argue that this initial copy infringes the law if it is not authorized. However, the initial copy is necessary in order to index the content. Without indexing the content, no search results can be returned to the user. Hence it appears search engine operators have a strong legal argument in their favour. The initial copy made by the indexer presents some similarities with the reproduction made in the act of browsing, in the sense that it forms an integral part of the technological process of providing a certain result.

In this respect, the EU Copyright Directive states in its preamble that browsing and caching ought to be considered legal exceptions to the reproduction right. The conditions for this provision to apply are, among others, that the provider does not modify the information and that the provider complies with the access conditions.[166]

The next section considers these arguments with respect to the search engine's cache copy of content.

7.4.1.1.2 Caching

The legal issues relating to the inclusion of content in search engine caches are amongst the most contentious. Caching is different from indexing, as it allows the users to retrieve the actual content directly from the search engines' servers. The first issues in regard to caching relate to the reproduction right.

The question arises as to whether the legal provision in the EU Copyright Directive's preamble would really apply to search engines. One problem relates to the ambiguity of the term ‘cache’. The provision was originally foreseen for Internet Service Providers (ISPs) to speed up the process. It may give the impression that content is only temporarily stored on an engine's servers for more efficient information transmission. Search engines may argue that the copyright law exception for cache copies also applies also to search engines. Their cache copy makes information accessible even if the original site is down, and it allows users to compare between live and cached pages. However, cache copies used by search engines fulfill a slightly different function. They are more permanent than the ones used by ISPs and can, in fact, resemble an archive. Moreover, the cache copy stored by a search engine may not be the latest version of the content in question.

In US law, the legal status under copyright law of this initial or intermediate copy is the subject of fierce debate at the moment.[167] For instance, in the on-going litigation against Google Print, publishers are arguing that the actual scanning of copyrighted books without prior permission constitutes a clear copyright infringement.[168]

---

[166] *See* EUCD, *supra*, Recital 33.

[167] *See*, for instance, Frank Pasquale, Copyright in an Era of Information Overload: Toward the Privileging of Categorizers, *Vanderbilt Law Review*, 2007, p.151., *at* http://ssrn.com/abstract=888410; Emily Anne Proskine, Google Technicolor Dreamcoat: A Copyright Analysis of the Google Book Search Library Project, 21 *Berkeley Technology Law Journal* (2006), p.213.

[168] Note that this is essentially an information security argument. One of the concerns of the publishers is that, once the entire copy is available on the search engines’ servers, the risk exists that the book become widely available in digital format if the security measures are insufficient.

In the EU, however, the most important issue appears to relate to the use of particular content, or whether and how it is communicated to the public. In the *Copiepresse* case, the Court made clear that it is not the initial copy made for the mere purpose of temporarily storing content that is under discussion, but rather the rendering accessible of this cached content to the public at large.[169]

7.4.1.2. Right of communication to the public

7.4.1.2.1 Indexed Information

(i) Text Snippets

It is common practice for search engines to provide short snippets of text from a web page, when returning relevant results. The recent Belgian *Copiepresse* case focused on Google's news aggregation service, which automatically scans online versions of newspapers and extracts snippets of text from each story.[170] Google News then displays these snippets along with links to the full stories on the source site. Copiepresse, an association that represents the leading Belgian newspapers in French and German, considered that this aggregation infringed their copyright. The argument is that their members - the newspapers - have not been asked whether they consent to the inclusion of their materials in the aggregation service offered by the Google News site.[171]

Though it is common practice for search engines to provide short snippets of text, this issue had not raised copyright issues before. However, this may be a matter of degree and the provision of such snippets may become problematic, from a copyright point of view, when they are pro-actively and systematically provided by the search engines. One could argue either way. Search engines may argue that thousands of snippets from thousands of different works should not be considered copyright infringement, because they do not amount to one work. On the other hand, one may argue that, rather than the amount or quantity of information disclosed, it is the quality of the information that matters. Publishers have argued that a snippet can be substantial in nature – especially so if it is the title and the first paragraph – and therefore communicating this snippet to the public may constitute copyright infringement. One might also argue that thousands of snippets amount to substantial copying in the qualitative sense.

The legality of this practice has not yet been fully resolved. On 28th June 2006, a German publisher dropped its petition for a preliminary injunction against the Google Books Library Project after a regional Hamburg Court had opined that the practice of providing snippets did not infringe German copyright because the snippets were not substantial and original enough to meet the copyright threshold.[172]

---

[169] *See* Google v. Copiepresse, Brussels Court of First Instance, February 13, 2007, at p.38.

[170] *See* Google v. Copiepresse, Brussels Court of First Instance, February 13, 2007, at p.36. The *Copiepresse* Judgment is available at http://www.copiepresse.be/copiepresse_google.pdf. *See* Thomas Crampton, Google Said to Violate Copyright Laws, *The New York Times*, February 14, 2007, *at* http://www.nytimes.com/2007/02/14/business/14google.html?ex=1329109200&en=7c4fe210cddd59dd&ei=5088&partner=rssnyt&emc=rss.

[171] *See* Latest Developments: Belgian Copyright Group Warns Yahoo, *ZDNet News*, January 19, 2007, at http://news.zdnet.com/2100-9595_22-6151609.html; Belgian Newspapers To Challenge Yahoo Over Copyright Issues, *at* http://ecommercetimes.com/story/55249.html. A group representing french- and german-language belgian newspaper publishers has sent legal warnings to yahoo about its display of archived news articles, the search company has confirmed. (They complain that the search engine's "cached" links offered free access to archived articles that the papers usually sell on a subscription basis.) See also Yahoo Denies Violating Belgian Copyright Law, *Wall Street Journal*, January 19, 2007, *at* http://online.wsj.com/.

[172] *See* Germany and the Google Books Library Project, Google Blog, June 2006, *at* http://googleblog.blogspot.com/2006/06/germany-and-google-books-library.html.

By contrast, in the above mentioned *Copiepresse* case, the Belgian court ruled that providing the titles and the first few lines of news articles constituted a breach of the right of communication to the public. In the court's view, some titles of newspaper articles could be sufficiently original to be covered by copyright. Similarly, short snippets of text could be sufficiently original and substantial to meet the 'copyrightability' threshold. The length of the snippets or titles was considered irrelevant in this respect, especially if the first few lines of the article were meant to be sufficiently original to catch the reader's attention. The Belgian court was moreover of the opinion that Google's syndication service did not fall within the scope of exceptions to copyright, since these exceptions have to be narrowly construed. In view of the lack of human intervention and fully automated nature of the news gathering, and the lack of criticism or opinion, this could not be considered news reporting or quotation. Google News' failure to mention the writers' name was also considered in breach of the moral rights of authors. If upheld on appeal, the repercussions of that decision across Europe may be significant.

(ii) Image Thumbnails

A related issue is whether the provision by search engines of copyrighted pictures in thumbnail format or with lower resolution breaches copyright law. In *Arriba Soft v. Kelly*,[173] a US court ruled that the use of images as thumbnails constitutes 'fair use' and was consequently not in breach of copyright law. Although the thumbnails were used for commercial purposes, this did not amount to copyright infringement because the use of the pictures was considered transformative. This is because Arriba's use of Kelly's images in the form of thumbnails did not harm their market or their value. On the contrary, the thumbnails were considered ideal for guiding people to Kelly's work rather than away from it, while the size of the thumbnails makes using them, instead of the original, unattractive. In the *Perfect 10* case, the US court first considered that the provision of thumbnails of images was likely to constitute direct copyright infringement. This view was partly based on the fact that the applicant was selling reduced-size images like the thumbnails for use on cell phones.[174] However, in 2007 this ruling was reversed by the Appeals Court, in line with the ruling on the previous *Arriba Soft* case. The appeals court judges ruled that "Perfect 10 is unlikely to be able to overcome Google's fair use defense."[175] The reason for this ruling is the highly transformative nature of the search engine's use of the works, which outweighed the other factors. There was no evidence of downloading of thumbnail pictures to cell phones, nor of substantial direct commercial advantage gained by search engines from the thumbnails.[176]

By contrast, a German Court reached the opposite conclusion on this very issue in 2003. It ruled that the provision of thumbnail pictures to illustrate some short news stories on the Google News Germany site did breach German copyright law.[177] The fact that the thumbnail pictures were much smaller than the originals, and had much lower resolution in terms of pixels, which ensured that enlarging the pictures would not give users pictures of similar quality, did not alter these findings.[178] The court was also of the view that the content could have been made accessible to users without showing thumbnails – for instance, indicating in words that a picture was available. Finally, the retrieving of pictures occured in a fully automated manner and search engines did not

[173] *See* Kelly v. Arriba Soft, 77 F.Supp.2d 1116 (C.D. Call 1999). *See* Gasser, Urs, Regulating Search Engines: Taking Stock and Looking Ahead, 9 *Yale Journal of Law & Technology* (2006) 124, p.210; *at* http://ssrn.com/abstract=908996.

[174] The court was of the view that the claim was unlikely to succeed as regards vicarious and contributory copyright infringement. *See* Perfect 10 v. Google, 78 U.S.P.Q.2d 1072 (C.D. Cal. 2006).

[175] *See* Perfect 10, Inc. v. Amazon.com, Inc., (9th Cir. May 16, 2007), judgment available *at* http://lawgeek.typepad.com/LegalDocs/p10vgoogle.pdf.

[176] *See* p. 5782 of the judgment.

[177] *See* the judgment of the Hamburg regional court, available at http://www.jurpc.de/rechtspr/20040146.htm, in particular on pp.15-16. See on this issue: http://www.linksandlaw.com/news-update16.htm

[178] *Ibid.*, p.14.

create new original works on the basis of the original picture through some form of human intervention.[179]

The German Court stated that it could not translate flexible US fair doctrine principles and balancing into German law. As German law does not have a fair use-type balancing test, the Court concentrated mainly on whether the works in question were covered or not by copyright.[180] Contrary to text, images are shown in their entirety, and consequently copying images is more likely to reach the substantiality threshold.[181] It may therefore be foreseen that AV search engines are more likely to be in breach of German copyright law than mere text search engines.

A related argument focuses on robot exclusion protocols. The question arises as to whether not using them can be considered by search engines as a tacit consent to their indexing the content. The court's reaction to these arguments in relation to caching is significant here. These issues are thus considered below.

7.4.1.2.2 Cached Information

The second set of issues related to the caching of content revolves around the right of communication to the public. When displaying the cache copy, the search engine returns the full page and consequently users may no longer visit the actual web site. This may affect the advertising income of the content provider if, for instance, the advertising is not reproduced on the cache copy. Furthermore, Copiepresse publishers argue that the search engine's cache copy undermines their sales of archived news, which is an important part of their business model. The communication to the public of their content by search engines may thus constitute a breach of copyright law.

The arguments have gone either way. Search engines consider, that information on technical standards (e.g. robot exclusion protocols), as with indexing, is publicly available and well known and that this enables content providers to prevent search engines from caching their content. But one may equally argue the reverse. If search engines are really beneficial for content owners because of the traffic they bring them, then an opt-in approach might also be a workable solution since content owners, who depend on traffic, would quickly opt-in.

Courts on either side of the Atlantic have reached diametrically opposed conclusions. In the US, courts have decided on an opt-out approach whereby content owners need to tell search engines not to index or cache their content. Failure to do so by a site operator, who knows about these protocols and chooses to ignore them, amounts to granting a license for indexing and caching to the search engines. In *Field v Google*,[182] a US court held that the user was the infringer, since the search engine remained passive and mainly responded to the user's requests for material. The cache copy itself was not considered to directly infringe the copyright, since the plaintiff knew and wanted his content in the search engine's cache in order to be visible. Otherwise, the plaintiff should have taken the necessary steps to remove it from cache. Thus the use of copyrighted materials in this case was permissible under the fair use exception to copyright. In *Parker v Google*,[183] a US court came to the same conclusion. It found that no direct copyright infringement could be imputed to the search engine, given that the archiving was automated. There was, in other words, no direct intention to infringe. The result has been that, according to US case law, search engines are allowed to cache

[179] *Ibid.*, p.15.

[180] *Ibid.*, p.19

[181] *Ibid.*, p.16.

[182] *See* Field v. Google, F.Supp.2d, 77 U.S.P.Q.2d 1738 (D.Nev. 2006); judgment available at http://www.eff.org/IP/blake_v_google/google_nevada_order.pdf

[183] *See* Parker v. Google, Inc., No. 04 CV 3918 (E.D. Pa. 2006); judgment available at http://www.paed.uscourts.gov/documents/opinions/06D0306P.pdf.

freely accessible material on the Internet unless the content owners specifically forbid, by code and/or by means of a clear notice on their site, the copying and archiving of their online content.[184]

In the EU, by contrast, the trend seems to be towards an opt-in approach whereby content owners are expected to specifically permit the caching or indexing of content over which they hold the copyright. In the *Copiepresse* case, for instance, the Belgian Court opined that one could not deduce from the absence of robot exclusion files on their sites that content owners agreed to the indexing of their material or to its caching.[185] Search engines should ask permission first. As a result, the provision without prior permission of news articles from the cache constituted copyright infringement.[186]

7.4.2. Trademark Law

7.4.2.1. Early Litigation and Importance of Trademark Law

The issue of search on trademarked terms is one of the most litigated issues in the search engine context.[187] Trademarks are important for search engines. If search engines cannot sell keywords freely, they are not worth their market valuation. If competitors are allowed to buy ad keywords that contain registered trademarked names, then the search engine may be diverting some of the income streams away from the owners of the trademarked words toward their competitors. Trademark law has a lot to say about the actual practices of search engines in regard advertising.

Google decided in 2004 to reverse its policy on trademarks. In the US and Canada it permit advertising bids on trademarked items, but forbids the use of the TM in the text of the advertising. Outside the US and Canada, it does not permit the use of trademarked items neither in the ads nor for triggering the ads. Yahoo! on the other hand, explicitly forbids this in its keyword auctions.[188]

7.4.2.2. Scenarios and Legal Questions

To be sure, we need to distinguish between three situations. There is the obvious case in which trademarked terms are being used by a competitor in the text or content of advertising on a search engine portal. However, the really contentious issues relate to situations in which the trademark remains invisible to Internet users. The trademarked item is part of the algorithm in two distinct situations. First, advertisers may use a registered trademark in the metatags of their web site ("meta-tagging scenario"). Search engines rely on keyword and description meta-tags for the selection of relevant results, and the risk exists that they would return a competitor's page among the main results for a user query on that specific trademarked item. The second situation concerns the case in which advertisers bid for a competitors' trademark in advertising auctions of search engines ("search engine auction scenario"). When users type the well-known trademark the risk then exists that the competitors' advertising message will rank higher than the one of the trademark owner. Only the two last situations are considered below. The main focus, however, is on the last issue since this is

[184] *See* David Miller, Cache as Cache Can for Google, March 17, 2006, *at* http://www.internetnews.com/bus-news/article.php/3592251.

[185] *See* Google v. Copiepresse, Brussels Court of First Instance, February 13, 2007, at p.35; s*ee* also the judgment of the Hamburg regional court, *at* http://www.jurpc.de/rechtspr/20040146.htm, p.20.

[186] *See* Struan Robertson, Why the Belgian Court Ruled Against Google, *OUT-LAW News*, February 13, 2007, *at* http://out-law.com/page-7759.

[187] *See* Judge sides with Google in dispute over keywords, CNET News.com, September 29, 2006; Google loses French Trademark Lawsuit, CNET News.com, June 28, 2006; Google loses trademark dispute in France, CNET News.com, January 20, 2005; Google's ad sales tested in court, CNET News.com, February 13, 2006; Google may be liable for trademark infringement, CNET News.com, August 16, 2005.

[188] Danny Sullivan, Paid Search Ads & Trademarks: A Review of Court Cases, Legal Disputes, & Policies, Search Engine Land, September 3, 2007, *at* http://searchengineland.com/070903-150021.php

the only scenario in which search engines may be held liable for (enabling) trademark infringement.[189] This gives rise to three distinct legal questions.

(i) Meta-tags Scenario v Auctioning Scenario

The first question is thus whether the search engines' trademark practices in relation to advertising can be analogised to the meta-tagging scenario from a legal point of view. The trend in mata-tagging cases is in favour of liability of the web site provider who inserted the trademarked items in the meta-tags. Some courts have found that both these conducts should be considered analogous for the purposes of trademark law, while other courts consider that the metat-tagging scenario gives rise to liability but not the keyword auctioning scenario.[190] It appears that the two situations are not totally analogous for two reasons. First, it is always possible to see the trademarked items in metatags, either because they are in the text on the web site, or because they appear in the source code of the web site. By contrast, users cannot see the trademarked items in search engine auctions. Second, in the meta-tagging scenario the link comes up in the organic results, while in the auctioning scenrio the results come up in the advertising results. Given that consumers are more likely to expect some connection between the trademarked terms and the organic content or source, than between the trademarked term and the advertising message, one may argue that more caution and consequently stronger trademark protection is warranted in the meta-tags scenario.

(ii) Trademark Infringing Use

The second question is whether the search engines' keywording practice constitutes "infringing use" in the meaning of trademark law. The trademark use criterion is very complex, since there is more than one way in which one may consider that a trademark has been "used". The European Court of Justice considered that infringing use is a use by a third party that "is liable to affect the functions of the trademark, in particular its essential function of guaranteeing to consumers the origin of the goods."[191] In other words, infringing use refers to the use of a trademark in a way which would take away (some of) the goodwill created by the trademark owner.

But of course it is possible to argue either way. One may claim that the concept should be broadly interpreted. Given that trademark owners invest huge sums of money in creating goodwill for the brand, and making it unique in the eyes of the consumer, they should also be the one reaping the benefits thereof. Conversely it is obvious that the connection between the trademark holder and the consumer is triggered or created by means of visible information. Therefore, one may equally hold that if advertisers do not display the trademark or information to consumers in any form, they cannot be said to be using the mark. In sum, the understanding of the terms "infringing use" is subject to diverging interpretations. Depending on one's view the scope of the trademark owner's rights may either be broad and relate to a number of uses of the trademark, or may be restricted to control over the purely visible or "informational" use of the mark towards the consumer.

(iii) Likelihood of Consumer Confusion

The third question is whether the search engines' trademark practices bring with them "likelihood of consumer confusion."[192] For a start, this criterion is not universal. While it is a necessary criterion in US law under section 32(1) of the Lanham Act, there is no such statutory requirement in many EU member States. For instance, in German law the finding of "likelihood of confusion" is presumed in certain cases, such as when an identical mark is used for goods or services that are in the same class as that for which the trademark is registered.

---

[189] For a good overview, see Eric Goldman, Deregulating relevancy in internet trademark law, 54 *Emory Law Journal*, 507 (2005); Zohar Efroni, Keywording in Search Engines as Trademark Infringement: Issues Arising fro *Matim Li v. Crazy Line*, Max Planck Working Paper, November 2006.

[190] *See* for specific case law on US and Germany, Efroni, p.9.

[191] *See Arsenal v Matthew Reed*, ECJ, C-206/01 (12.11.2002), para.51.

[192] This question is of course irrelevant if the previous question relating to infringing use is answered negatively.

However, due to the conceptual difficulty in determining the exact meaning of the above criterion relating to "trademark use", most jurisdictions appear to take the likelihood of consumer confusion into account, either implicitly or explicitly. In order to bring the necessary balancing elements, Efroni advocates greater reliance on the likelihood of confusion test as a presumption indicating trademark use. It would then be up to the advertiser to rebut that the use of the trademark is infringing trademark law.[193] This would mean that not every likelihood of confusion is actionable. This approach would also lead to a much more flexible test in wich a number of elements can be balanced against one another. Important elements are the interest of having free competition between advertisers, innovation in search engine advertisin, the right and benefits of comparative advertising, or the right to freedom of expression in the form of advertising.

(iv) Wrongful Advantage

As regards the jurisdictions that rely to a large extent on the finding of infringing use, some balancing might be introduced by having regard more closely to the issue of whether search engines gain wrongful advantage from the keyword auctioning business. Obviously search engines can be said to benefit somehow from the goodwill created by the brandowners. However, evidence is needed in each specific case as to whether this advantage may be considered wrongful, so as to avoid ending up with a limitless right for trademark owners. At the same time, this wrongful advantage test may bring the necessary flexibility in the application of trademark law in the search engines context. It is important to bear in mind the fact that search engines bring great benefit to society, and that they rely tro a large extent on the advertising business to offer their services from which many parties benefit (users, advertisers, and content providers).

7.4.3. Data Protection Law

7.4.3.1. Increasing Data Protection Concerns

On 17th March 2006, Google, the major web search engine, won a partial victory in its legal battle against government. In an attempt to enforce the 1998 Child Online Protection Act, government had asked it to provide one million web addresses or URLs that are accessible through Google, as well as 5,000 users' search queries. In *Gonzales v. Google*, a California District Court ruled that Google did not have to comply fully with the US government's request. Google need not disclose a single search queries, and shall provide no more than 50,000 web addresses[194]. However, it soon appeared that Microsoft, AOL and Yahoo! had handed over such information requested by government in that specific case,[195] and in the course of this case all search engines publicly admitted massive user data collection. It turns out that all major search engines are able to provide a list of IP addresses with the actual search queries made, and vice versa.[196]

Not even 5 months later, AOL's search engine logs were responsible for yet another round of data protection concerns. There was public outcry when it became known that it had published 21 million search queries, that is, the search histories of more than 650,000 of its users. While AOL's intentions were laudable (namely supporting research in user behaviour), it appeared that making the link between the unique ID supplied for a given user and the real world identity, was not all that difficult.[197]

---

[193] Zohar Efroni, *supra*, p.17.

[194] Broache, A.: Google Wins Porn Probe Fight. In: *CNET News*, March 20, 2006; available at http://news.zdnet.co.uk/internet/0,1000000097,39258371,00.htm

[195] Hampton, M.: Google in bed with US intelligence, February 22, 2006, available at http://www.homelandstupidity.us

[196] Sullivan, D., Which Search Engines Log IP Addresses & Cookies – And Why Care?, 2006; available at http://blog.searchenginewatch.com/blog/060206-150030

[197] McCullagh, D.: AOL's disturbing glimpse into users' lives. In: *CNET News*, August 7, 2006, available at http://news.com.com/2100-1030_3-6103098.html; Barbaro, M., T. Zeller: A Face is Exposed. In: *New York Times*,

Even more recently, the Article 29 Working Party had a public exchange of views with Google about its data retention policies, i.e. the logging of user data for indefinite periods of time. The Working Party questioned the legality of this practice in light of the data protection laws.[198]In July, Google said that it would start deleting identifying information after 18 months. Other operators such as Yahoo! and Ask followed suit, the latter even giving its user the option to prevent their data from being stored in the first place.[199] The last news came from Google's side, when it advocated the introduction of global privacy standards based on the APEC privacy framework.[200]

These cases and public debates are milestones in raising awareness of the importance of data protection as regards web search. Importantly, these cases highlight a genuine need to better understand and analyse data protection issues. This issue is especially critical in a context of increased personalisation of search engines. Personalisation for the purposes of the present paper is the ability to proactively tailor offer to the tastes of individual users, based upon their personal and preference information. Personalisation is critically dependent on two factors: the search engines' ability to acquire and process user information, and the users' willingness to share information and use personalisation services[201].

7.4.3.2. Trends Towards Greater Personalisation

At present, search engines differentiate themselves from their competitors mainly thanks to the quality of their crawlers that gather digital information, and the volume and quality of their index, as well as by means of their algorithm which determines the relevance of search hits. One main consequence of search engine personalisation, however, is the enrichment of the latter process of defining relevance by means of a fourth component: a database containing the user profiles. Such a database is necessary for the search engine to effectively personalise the search results, or in order to rank the hits by "personalised relevance". Generally, search engines upload a cookie program in the computer of the user, during this user's first visit on the search engine site. That cookie bears a unique identifier or serial number, and is linked to the use of that browser on that particular computer. From that moment, every query made on the search engine using that particular browser software will be recorded, together with the Internet address, the browser language, the time and date of the query.

To be sure, personalisation makes sense both from a technological and economic viewpoint. There is a genuine need for user-side information. User information may be used for internal tracking, for improving search engine's response to user queries, and for preventing click-fraud. Likewise, the emerging audio-visual or multimedia search applications hinge very much on user information, given the difficulties encountered in accurately carrying out pattern recognition.

More personalised search also benefits the end-user. It helps the user remember search queries that have been viewed in the past. It may moreover be necessary in a context of proliferation of data. Search engines seek to cope with the explosion of data, formats and content diversity. Many searches are actually undertaken with some kind of answer, and there is currently an imbalance

---

August 9, 2006; available at http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000&en=f6f61949c6da4d38&ei=5090

[198] See for an overview of this saga Google calls for international privacy laws and policies, *OUT-LAW News*, 14/09/2007, at http://out-law.com/page-8470; Our data retention is not data protection watchdogs' business, says Google privacy boss, OUT-LAW News, 06/07/2007, *at* http://out-law.com/page-8233; Data protection watchdogs' letter to Google goes public, OUT-LAW News, 30/05/2007, http://out-law.com/page-8099; Google will delete search identifiers after two years, OUT-LAW News, 20/03/2007, http://out-law.com/page-7888.

[199] See Kevin Allison, Seeking the Key to Web Privacy, *Finanacial Times*, September 23, 2007.

[200] See for the APEC Privacy Framework: http://www.ag.gov.au/www/agd/rwpattach.nsf/VAP/(03995EABC73F94816C2AF4AA2645824B)~APEC+Privacy+Framework.pdf/$file/APEC+Privacy+Framework.pdf

[201] Chelappa, R.K.; R.S. Sin: Personalization versus Privacy: An Empirical Examination of the Online Consumer Dilemma. In *Information Technology and Management*, Vol. 6, 2005, pp.181-202.

between the answer we search for, and getting a list of thousands of documents. As it is unlikely that a two or three word query can unambiguously describe a user's informational goal, and as users tend to view only the first page of results,[202] personalising may be one way to provide the end-user with more relevant hits.[203]

The commercial interest in having more personalised search is equally beyond doubt. Better profiling would bring the search engine operators greater advertising revenue, as it would enable the latter to better price-discriminate. Search is a critical commercially relevant behaviour that indicates near-future user action. Rather than buying bluntly against words and context, personalisation would enable advertisers to buy against people and their likely habits. Thus, more and more personal information is gradually being drawn into the search domain. The harvesting of profiles and user information may rely increasingly on client-side applications. Search functionality now extends to desktop and email, files, notes, journals, blogs, music, photographs, etc. Toolbars, for instance, essentially grant the search engines access to users' hard drives every time they launch a search, which is many times a day. In the future, search may then even become "prospective." Search engines would match a user's record against new information passing through their matching engine. In sum, personalisation of search appears to result in huge benefits for both the commercial players and for the end-user.

Though the idea to personalise search has been around for some time already (e.g. with Hotbot thinking about it as far back as 1996),[204] technological advances as regards storage, processing power, and artificial intelligence,[205] have meant that the drive toward increased personalisation have increased in recent years. There are basically two approaches, which are often combined. The first approach is to let the user define more narrowly the settings of her search engine. This amounts to personalisation of the index or sources from which the search engine will draw results. Examples of this are Rollyo, PSS!, Yahoo! Search Builder, and Google's recently launched Custom Search Engine. In short, this approach allows you to name an engine, include search terms, and web sites you want it to search. It can be very narrow or broad. This can be shared or strictly private. You can invite others to help or just accept volunteers who learn about the search engine.[206] Personalisation is not restricted to the individual users. Thus, Eurekster's social search engine is an example of personalisation of the results ranking according to both the interests and behaviour record of a community of users. The idea is that, in line with the logic of web 2.0 users would tag their search results, make notes, and share these with other users, thus mapping the Web. The second approach, which appears to be more fruitful given that most users will not take the time to set up their customised search engines, is to *automatically* re-rank results provided by search engines, or to show different users different results based on their past behaviour. The most prominent example of this approach is currently A9, an Amazon service which uses the Google index. Other examples are Google's Personalized Search, or Findory, which uses fine-grained information about individual pages the user viewed.

---

[202] Machill, M.; C. Neuberger, W. Schweiner, W, Wirth: Navigating the Internet: A Study of German-Language Search Engines. In: *European Journal of Communication*, Vol 19, Nr. 3, 2004, p.325.

[203] Teevan, J.: S.T. Dumais, E. Horvitz: Beyond the Commons: Investigating the Value of Personalizing Web Search. In: *Workshop on New Technologies for Personalized Information Access*, 2005; available at http://haystack.lcs.mit.edu/papers/teevan.pia2005.pdf

[204] Gasser, U.: Regulating Search Engines: Taking Stock and Looking Ahead. In: *Yale Journal of Law and Technology*, Vol.9, 2006, pp.204; available at http://ssrn.com/abstract=908996

[205] Olsen, S.: Spying an Intelligent Search Engine. In: *CNET News*, August 18, 2006; available at http://news.com.com/Spying+an+intelligent+search+engine/2100-1032_3-6107048.html

[206] Sherman, C.: Google Launches Custom Search Engine Service, October 24, 2006; available at http://searchenginewatch.com/showPage.html?page=3623765; Hafner, K.: Google Customizes Search Tool to Cut through Web Noise. In: *The New York Times*, October 24, 2006; available at http://www.iht.com/articles/2006/10/24/business/google.php; Bradley, P.: Your Search, Your Way, September 19, 2006, available at http://searchenginewatch.com/showPage.html?page=3623434

Some have argued that current efforts toward personalisation of search is the wrong way to go, on the ground that people have changing interests, and because you cannot read the mind of a user by means of a few keywords entered in a search box.[207] However, the fact is that efforts toward personalisation are currently being undertaken. There is moreover little doubt that the current trend of gathering a maximum amount of user data shall continue. Given dramatic increases in processing power and storage capacity, there is no reason to believe that major players in the search engine market will not log all the personal information. Given that advertising is the biggest income source for many if not most search engines, and given that advertisers seem to appreciate the trend toward personalisation,[208] it makes sense to forecast an increasing reliance on personalised search. Information will thus be logged by search engines unless society makes a deliberate, concerted effort preventing this. It is consequently necessary to understand why and how this logging activity may need to be halted.

7.4.3.3. Data Protection Implications

Search engines conjure up the image of people being able to gain knowledge about other people's private lives using search engines.[209] This paper considers an arguably more important privacy debate. Namely, it questions whether the various search engines' logging activities are in line with EU data protection laws, and highlights the importance of this debate for media pluralism.

As a starting point, it is important to bear in mind that responses to data profiling by search engines may take many forms: law, technology, social norms and market. One example of a technological response to surveillance by search engines is TrackMeNot, a tool which produces a lot of 'noise' and obfuscates the actual web searches in a cloud of false leads.[210] Another example is Tor, a technology that allows users to mask their IP address by means of a proxy server.[211]

Search engine logging raises two related types of legal regulatory issues. The first type of privacy is privacy of communications, which covers the security and privacy of emails, and other forms of digital communication. Directive 2002/58/EC provides certain privacy protections for data gathered in the course of communications using publicly available electronic communications networks and services. In particular, recital 25 of the preamble states that cookies are legitimate provided that the users are given adequate information, and have the ability to refuse the cookie. This Directive is not particularly compelling for search engines, and is not dealt with any further here.

The second type relates to information privacy, or the actual collection and handling of personal data. In this respect, the EU Data Protection Directive (Directive 95/46/EC) defines private data as any information relating to an identified or identifiable natural person. An identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number (Art.2(a)) .

The first question that arises is thus whether the data that are being recorded by search engines constitute personal data in the meaning of EU data protection legislation. Some of the queries made by the user may contain the name, telephone number of address of a given person. For instance, an increasing number of user tries to see what information is available about himself, by typing his name into the search engine box (vanity searches). Though in all of the policies regarding users'

---

[207] Valdes-Perez, R.: Why Search Personalisation is a Dead End, 2006; available at http://vivisimo.com/docs/personalization.pdf

[208] Odlyzko, A.: Privacy, Economics, and Price Discrimination on the Internet. In: *ACM International Conference Proceeding Series*, Vol. 50, 2003; available at http://www.dtc.umn.edu/~odlyzko/doc/privacy.economics.pdf

[209] Tavani, H.T.: Search Engines, Personal Information and the Problem of Privacy in Public. In: *International Review of Information Ethics*, Vol.3, 2005, pp-39-45; available at http://www.i-r-i-e.net/inhalt/003/003_tavani.pdf

[210] See Howe, D.C.; H. Nissenbaum: TrackMeNot, 2006, available at http://mrl.nyu.edu/~dhowe/trackmenot

[211] See Christopher Soghoian, The Problem of Anonymous Vanity Searches, p.5; available at http:// ssrn.com/abstract=953673

search histories there are clear indications as to how one may get rid of one's search history, it is not clear at all whether the information is wiped out completely, also at the end of the search engine.[212]

Some have argued that none of the information thus recorded by search engines appears to constitute, in itself, personally identifiable information. This is because it is not actually possible to assert with a high degree of certainty who actually made the searches. Indeed, someone else might have typed your personal information in the search box, or two people might use the same browser engine to search using the same computer. Likewise, the actual information that is recorded in the digital dossier or profile will be (at best) a patchy overview of someone's life, given that the person may be using different browser software and/or search engines. In two recent court cases in France, user's IP addresses were considered not to be personally identifyable information in the sense of existing data protection legislation.[213] At the same time, the leading view across Europe

In addition, it is important to note that there may sometimes be ways to link search query information to a particular person's computer by comparing the records of the search engine company with the logs of the Internet Service Provider (ISP). All major search engines are currently encouraging users to proactively help them with the building of the database, and they are providing other online applications and services. There is little doubt, for instance, that Google may have a reasonably good sense of a user's real world identity if that person is logged in to one of the Google applications – say, Gmail – and is simultaneously conducting search queries on the Google search engine.[214] Furthermore, the AOL case gives us a good idea of the actual ease with which it is possible to assert the real identity behind a list of search queries tied to unique ID numbers. In these circumstances, all of the above-mentioned data protection obligations would fall on the search engine operators.

Finally, it is increasingly recognized that, contrary to popular belief, it is not the principle of secrecy which lies at the centre of data protection but the principle of autonomy. Data protection includes not only the right to keep personal matters out of the public eye, but also and foremost the right to be left alone, to be free from intrusion – to have some degree of autonomy over one's acts. Data and information regarding one's past activities are an important element in this debate. Data protection refers to the fact that I need to have some degree of control, autonomy, over the way my personal data are being processed. In this view, it is not so important whether you know the real world identity of the user who entered the search terms, or whether the information can be linked to a particular real world identity.[215] Surveillance by market players is intended to induce (as opposed to suppress) users into buying behaviour, but it is no less invasive of our autonomy than government control that may want to prevent users from certain behaviour. The fact that we are often watched by machines which seem less invasive from a secrecy point of view does not make it less problematic from a data protection point of view. While secrecy and autonomy were in many ways one and the same concept in physical space, this is not true in the digital environment where my personal data may well be secret to the search engines, but these may nonetheless severely affect my autonomy.

In other words, it appears increasingly clear that search engines ought to comply with various provisions enshringed in the national laws implementing the data protection directive. Specficially, personal data should be processed fairly and lawfully (Art.6(1)(a)), they are to be collected for

---

[212] Sullivan, D., (b) Private Searches Versus Personally Identifiable Searches, 2006; available at http://blog.searchenginewatch.com/blog/060123-074811

[213] See Paris Appeals Court Decision – Anthony v. SCPP (27.04.2007), *at* http://www.legalis.net/jurisprudence-decision.php3?id_article=1954, and Paris Appeals Court Decision – Henri v. SCPP (15.05.2007) http://www.legalis.net/jurisprudence-decision.php3?id_article=1955, discussed in http://www.edri.org/edrigram/number5.17/ip-personal-data-fr.

[214] Goldberg, M.A.: The Googling of Online Privacy: Gmail, Search-Engine Histories and the New Frontier of Protecting Private Information on the Web. In: *Lewis & Clark Law Review*, Vol.9, 2005, pp.253

[215] Dan Solove, *The Digital Person. Technology and Privacy in the Information Age*. New York, NY [NYU Press] 2004

specified and legitimate purposes (Art.6(1)(b)). In addition, the data processing in question needs to be relevant (Art.6(1)(d)), and not excessive in relation to the purpose for which they have been collected (Art.6(1)(c), Artt.7-8). Finally, the data need to be kept accurate and up-to-date, when necessary with the help of data subjects (Art.7(a)), ought to be stored no longer than necessary for attainment of the objective; and may be disclosed only with the consent of the data subject (Art.7(a)).[216]

7.5. Policy Issues: Three Key Messages

7.5.1. Increasing Litigation in AV Search Era: Law as a Key Policy Lever

7.5.1.1. Sharp Tensions Surrounding Search Engine Operations

In each of the debates discussed above, it is possible to spot similar trends. The view of content providers, advertisers, and consumer and civil organisations is straightforward. They argue that search engines are free riding on their creations, their goodwill, or the user's data without appropriate remuneration, or without taking care of data protection obligations.

The content generated by the providers is used by search engines in two distinct ways. First, search engines can become fully-fledged information portals, directly competing with the content providers that provide their very content.[217] Second, search engines use the content providers' creations as the source upon which they base their (sometimes future) advertisement income. Therefore, content providers are increasingly unwilling to allow search engines to derive benefits from listing or showing their content without remuneration. Brand owners are of the view that the goodwill created by them may be used by search engines to derive income. Users are increasingly concerned that the information that is held about them may be used.

Search engines have a diametrically opposed view. They emphasise their complementary role as mere conduits in directing web-traffic to content providers, money to advertisers, and relevant content to their users. A recent report by the consulting company Hitwise shows that US newspapers' web sites receive 25% of their traffic from search engines.[218] Consequently, the search engines' view is that the relationship is mutually beneficial, in that search engines indirectly pay content providers through the traffic they channel to them, provide advertisers with a unique platform for increasing their brand name and commercial sales, and bring the most relevant to the users for free.

7.5.1.2. Unclear Legal Status

Search engines are gradually emerging as key intermediaries in the digital world, but it is no easy task to determine whether their operations, which are to a large extent automated, constitute copyright, trademark or data protection infringements. Due to their inherent modus operandi, search engines are pushing the boundaries of existing law. Issues are arising which demand a reassessment of some of the fundamentals of law.

With regard to copyright law search engines raise a flurry of novel questions: does scanning books constitute an infringement of copyright, if those materials were scanned with the sole aim of making them searchable? When do text snippets become substantial enough to break copyright law if they are reproduced without the content owners' prior permission? With regard to trademark law, it is

---

[216] On the other hand, Google's argument to the effect that the two years period followed from the data retention Directive was quickly rebutted by the Art. 29 Working Party, on the ground that the obligation to keep the data for two years applies to providers of public electronic communications networks and services, which search engines are not.

[217] *See* Google v. Copiepresse, Brussels Court of First Instance, February 13, 2007, at p.22.

[218] *See* Tameka Kee, Nearly 25% of Newspaper Visits Driven by Search, *Online Media Daily*, Thursday, May 3, 2007, *at* http://publications.mediapost.com/index.cfm?fuseaction=Articles.showArticleHomePage&art_aid=59741.

unclear whether the use of trademarked items to trigger ads constitutes "use of a trademark" in the sense of the law, or whether consumers are likely to be confused. With respect to data protection law, it is clear that user data are of fundamental importance in the development of improved search engines, but the question arises to what extent the data gathered by search engines constitute personal information in the sense of data protection law, and what may be the most appropriate means for balancing the various interests involved.

7.5.1.3. The Role of Technology & Market Transactions

Automation is inherent to the Internet's functioning: the question thus arises whether permission and agreement should equally be automated, or governed by technological standards. A good example comes from the copyright debate. In that context, search engines argue that if content providers prefer not to be included in the index or cache, they simply have to include the robot exclusion protocols in their website. Asking each content providers for prior permission would be unfeasible in practice. Content providers, on the other hand, argue that not including robot exclusion protocols in their websites cannot be considered as an implicit permission to use their content, since robot exclusion protocols cannot be regarded as law. There is currently no law in force stating that the non-use of robot exclusion protocols is equal to implicitly accepting indexing and caching.

On the one hand, developments which aim to increase flexibility are welcome, because there is probably no one-size-fits-all solution to the copyright problem. Technology may fill a legal vacuum, by allowing parties at distinct levels of the value chain to reach agreement on the use of particular content. This approach has the advantage of being flexible.

On the other hand, the question arises as to whether society wants content providers to exert, through technological standards, total control over the use of their content by players such as search engines. Such total control over information could indeed run counter to the aims of copyright law, as it could impede many new forms of creation or use of information. This is a recurrent debate. For example in the DRM debate, many commentators are skeptical about technology alone being capable of providing the solution.

Another regulatory modality is the market, or contractual deals amongst market players. For instance, there have been a number of market deals between major content providers and major search engines. In August 2006, Google signed a licensing agreement with Associated Press. Google also signed agreements with SOFAM, which represents 4,000 photographers in Belgium, and SCAM, an audio-visual content association. Initially, both SOFAM and SCAM were also involved in the Copiepresse litigation. On 3 May 2007, the Belgian newspapers represented by Copiepresse were put back on Google news. Google agreed to use the no-archive tag so that the newspapers' material was not cached On 6 April 2007, Google and Agence France Presse reached an agreement concerning licensing.

Consequently, as regards policy, the question arises as to whether there ought to be any legal intervention at all, since the market may already be sorting out its own problems. A German Court supported this view in its decision on thumbnails.[219] As it is a non-consolidated business and information is scarce, it is currently difficult to judge whether there is a market dysfunction or not. One of the salient facts here is that the exact terms of the deals were not rendered public, but in each one Google was careful to ensure that the deal was not regarded as a licence for the indexing of content. Google emphasised the fact that each deal will allow new use of the provider's content for a future product.[220] Some commentators see the risk that, while larger corporations may have plenty

[219] *See* the judgment of the Hamburg regional court, *at* http://www.jurpc.de/rechtspr/20040146.htm, p.20.

[220] Distinction between AFP/AP and copiepresse case. More difficult to remove AFP/AP content from Google news since hundreds of members are posting these stories on their site; comparatvely there are far fewer sources of Copiepresse content. In addition, AFP and AP are also different from classic news site because they get the bulk of their revenue from service fees from their subscribers, and derive little direct benefit from traffic from Google

of bargaining power to make deals with content owners for the organisation of their content, the legal vacuum in copyright law may well erect substantial barriers to entry for smaller players who might want to engage in the organisation and categorisation of content. "In a world in which categorizers need licenses for all the content they sample, only the wealthiest and most established entities will be able to get the permissions necessary to run a categorizing site." [221] The same is true to some extent as regards branding. Brand owners may reach exclusivity agreements with the biggest and wealthiest search engines, thereby excluding upcoming players in the sector.

This may become particularly worrying for emerging players. Concrete examples are emerging methods for categorizing and giving relevance to certain content, like the decentralised categorisation by user-participation. Although automatised, search engines are also dependent on (direct or indirect) user input. The leading search engines observe and rely heavily on user behaviour and categorisation. A famous example is Google's PageRank algorithm for sorting entries by relevance which considers the number clicks, and ranks the most popular URLs according to the link structure. There is a multitude of other sites and services emerging, whose main added value is not the creation of content but categorising it. This categorisation may involve communicating to the public content produced by other market players. Examples include shared bookmarks and web pages,[222] tag engines, tagging and searching blogs and RSS feeds,[223] collaborative directories,[224] personalized verticals or collaborative search engines,[225] collaborative harvesters,[226] and social Q&A sites.[227] This emerging market for the user-driven creation of meta-data may be highly creative, but may nonetheless be hampered by an increasing reliance on licensing contracts for the categorisation of content.

In other words, law is not the only policy lever. There are other regulatory, technical and economic means of advancing the interests of the European AV content and AV search industry. However, it is clear from the above discussion that these regulatory means are influenced by copyright, trademark and data protection law which determine the permissible uses of certain content, brand names, or user data by search engines. Specifically, the law may have an impact on the use of certain technologies and technological standards; and the law may influence the conclusion of agreements between search engines and content providers, advertisers and users.

7.5.1.4. A Matter of Degree

As a result, we denote one common pattern across the various bodies of law analysed. Issues relating to trademark law will become more acute in the audiovisual search context, given that the ads that can be served using AV search technology are likely to have a more powerful influence on consumer habits than the presently predominant text-based ads. The more audio-visual – rather than solely text-based – content is put on the Internet, the more we may expect copyright litigation problems to arise with respect to AV search engines. The reason is that premium AV content is generally more costly to produce and commercially more valuable than text-based content. Finally, given that it is already difficult to return pertinent results for text-based content, AV search engines will have to rely even more on user profiling; those user profiles will by the same token enable search engines to target users directly and thereby compete with traditional media and content owners. In sum, in comparison with pure text-based search, trademark, copyright and data protection litigation in the AV search environment may be expected to increase.

---

[221] Frank Pasquale, *supra*, pp. 180-181.

[222] For instance, Del.icio.us, Shadows, Furl.

[223] For instance, Technorati, Bloglines.

[224] For instance, ODP, Prefound, Zimbio and Wikipedia.

[225] For instance, Google Custom Search, Eurekster, Rollyo.

[226] For instance, Digg, Netscape, Reddit and Popurl.

[227] For instance, Yahoo Answers, Answerbag.

In sum, the analysis highlights two aspects. First, no radically new legal problems are to be expected in the AV search context, as compared to the existing text-based environment. Second, law is a key policy lever in the search engine context, whose importance may moreover be expected to increase as we move on to an AV search environment.

7.5.2. Combined Effect of Laws: Need to Determine Default Liability Regime

7.5.2.1. Search Engines as Key Intermediaries

Search engines have become indispensable organisers and categorizers of data. They enable users to filter huge amounts of data and thus play an increasingly pivotal role in the information society. Search engines' main contribution is producing meta-data, for instance when indexing material. The above discussion indicates a number of unresolved issues in applying various laws to search engines. One important issue with respect to AV search engines relates to the copyright status of producers of meta-data, i.e. information (data) about particular information (data).[228]

7.5.2.2. Focusing on Individual Law is Insufficient

This section develops the following two points. First, each of the individual laws affects search engines and other emerging intermediaries in the digital environment. Second, focusing on each law individually may not yield the best result – there is a need to consider the laws together, and their combined effect on the market for those new intermediaries.

Let us consider copyright law to make this point. Copyright law originates from the 'analogue era' with rather limited amounts of data. In those times, obtaining prior permission to reproduce materials or to communicate them to the public was still a viable option. Nowadays with huge amounts of data, automation is the only efficient way of enabling creation in the digital era. Automation raises intricate and unforeseen problems for copyright law. In addition, the automatic collection and categorisation of information by search engines and other meta-data producers is all-encompassing. Search engine crawlers collect any information they can find, irrespective of its creative value. They do this in a fully automated manner. The result may eventually be that search engines are forced to comply with the strictest copyright standard, even for less creative content. There are various policy dimensions here: (i) amending the law, and (ii) relying on the market.

7.5.2.2.1 Legal Regulation

Changing (slightly) the focus of EU copyright law could have positive *economic* effects. Today's main exceptions to copyright law are the right to quotation, review, or the special status granted to libraries. Automatic organization and filtering of data are not the focus of current copyright law. The above view suggests, however, that there is value in an efficient and competitive market for the production of meta-data, where the organisation of information is becoming increasingly critical in environments characterised by data proliferation. Some commentators consider that it would be beneficial to give incentives not only for the creation of end-user information, but also for the creation of meta-data. This could be achieved by including a legal provision in the copyright laws that take into account new methods for categorising content (e.g. the use of snippets of text, thumbnail images, and samples of audiovisual and musical works), some of which even as additional exceptions or limitations of copyright.[229] Increasing clarity on these practices might ease the entry of smaller players into the emerging market for meta-data.

---

[228] Metadata vary with the type of data and context of use. In a film, -for instance- the metadata might include the date and the place the video was taken, the details of the camera setting, the digital rights of songs, the name of the owner, etc. The metadata may both be automatically generated or manually introduced, like tagging of pictures in online social networks (e.g. Flickr).

[229] *See* Frank Pasquale, *supra*, p.179 (referring to Amazon's "look inside the book" application).

Similar arguments also apply to the *cultural or social* dimension, where copyright can be regarded as a driver of freedom of expression through its incentives to people to express their intellectual work. Again, given today's information overload, categorizers of information are also important from a social point of view. First, the right to freedom of expression includes the right to receive information or ideas.[230] One may argue that, in the presence of vast amounts of data, the right to receive information can only be achieved through the organization of information. Second, categorisations – such as the ones provided by search engines – are also expressions of information or ideas. Indeed, the act of giving relevance or accrediting certain content over other content through, for instance, ranking, is also an expression of opinion. Third, the creation or expression of new information or ideas is itself dependent on both the finding of available information and the efficient categorisation of existing information or ideas. EU Copyright Law and the Creation of Meta-Data for AV Search

7.5.2.2.2 Commercial Deals

Content providers and search engines need each other far too much. Search is big business and brings traffic. Content providers have some interest in keeping the search engines working and directing traffic towards their own sites. But search engines are equally useless without available content.[231]

The hope of the news providers in the *Copiepresse* was that, if enough content and copyright owners object to being indexed without compensation then search engines will have substantially less content to index, and will be forced to come to the negotiation table. The case has potentially international ramifications. Google faces parallel case in France and in the US, where Agence France Presse has sued it for copyright infringement in the DC District Court in Washington. The Danish association of newspapers (Danske Dagblades Forening) has delayed the launch of Google News Denmark, arguing that Google will have to make separate agreements with each one of the publishers. The same legal and other negotiation techniques are being employed in relation to the Google Library project regarding the scanning of copyrighted books. Author and publisher organisations in many different countries are suing the search engine. These law suits are thus like strong positioning moves, or business negotiations that are going on in court.[232] Newspapers and other content providers want search engines to continue directing traffic, but they also want search engines to pay for the fact that they receive revenues in part thanks to their content.

Besides answering in court, search engines have had two types of responses in relation to audio-visual content. The first move is one of increased (vertical) integration with online platforms for sharing and viewing audio-visual content, such as YouTube or Google Video.[233] It appears that here

---

[230] *See* Art. 10 European Convention on Human Rights.

[231] A related point is of course that powerful search technology also makes it easier for right-holders to identify content and determine whether illegal copies of copyrighted content have been posted online. *See* Myspace Launches Pilot To Filter Copyright Video Clips, Using System From Audible Magic, *Technology Review*, February 12, 2007, *at* http://www.technologyreview.com/read_article.aspx?id=18178&ch=infotech. *See* Eric Auchard, Google Sees Video Anti-Piracy Tools as Priority, *Reuters*, February 22, 2007, *at* http://today.reuters.com/news/articlenews.aspx?type=technologyNews&storyid=2007-02-23T030558Z_01_N21366907_RTRUKOC_0_US-GOOGLE-YOUTUBE.xml. The technology solution is as follows: all major content providers send their content to Audible Magic to be logged into the database. Audible Magic uses "fingerprinting technology" that can recognise content no matter how this content is tampered with. Acoustic fingerprinting technology, for instance, is about creating a unique code from an audio-wave. This is different from other content identification technologies such as hash codes because the fingerprint is not generated from the binary data in the file. As a result, the acoustic fingerprint will be the same, irrespective of whether the file has been compressed, ripped into a different lower quality format, or amended. *See* http://en.wikipedia.org/wiki/Acoustic_fingerprint.

[232] *See* for this point Jeffrey Toobin, Google's Moon Shot. The Quest for the Universal Library, *The New Yorker*, January 29, 2007, *at* http://www.newyorker.com/fact/content/articles/070205fa_fact_toobin

[233] *See* Michael Liedtke, Google Video suit could signal YouTube trouble ahead, *The Associated Press*, November 8, 2006, at http://www.usatoday.com/tech/news/2006-11-08-google-sued_x.htm; Google Faces Legal Challenges Over

too platform operators will continue to be at odds with the right-holders until they licence the clips.[234]

The second avenue is to conclude contractual agreements with content providers. For instance, it appears that Google sometimes agrees to pay for content. Google agreed to pay The Associated Press for stories and photographs, and settled copyright disputes with 2 groups in Belgium.[235] This strategy appears to bring with it a greater risk of (horizontal) concentration in the search engine sector. At present, it is still easy to switch between providers. Search personalisation has been one strategy of some search engines for tying users to their services. The risk is real that the contractual negotiations on the indexing and caching of copyrighted content may lead to increased barriers to entry.[236] Contractual negotiations are bilateral, and it is not unlikely that an agreement on the part of the search engine to pay for the indexing and caching of valuable content may come together with exclusivity clauses as is customary in other media segments. A contractual settlement between search engines and content providers may well result in distinctions between the types of content that may be retrieved by the various search engines. In some sense, this may signal a departure from the classic horizontal and open market structure that characterises the Internet, as opposed to the broadcast model. If such were the case, this would add another significant barrier to entry, and new start-ups would be less likely to threaten the incumbents in the search engine sphere.

In sum, the copyright regime has a hard task taking into account the search engines' unique role in making information accessible. This might be detrimental not only for a flourishing content sector, but also for development of new search engine technology (intermediaries). The more search engines move toward content aggregation and personalisation, the more likely it is that they will be affected by copyright law. At the same time, the sole application of copyright law in this sphere, and the barriers to entry that may result from contractual negotiations between search engines and content providers, may well require us to consider more closely whether there is a need to introduce some form of media law obligations. This is a debate that may have widespread ramifications and affect the basic fundaments of the Internet as a whole. A differentiation among search engines, which are widely believed to be among the key players of today's Internet, would put into question the basic nature of the Internet as an open, horizontal communications platform.

7.5.2.3. In Search of the Default Liability Regime

Information products and services (i.e. culture) are intrinsically different in nature from—say—beans. A non-functioning media market may have catastrophic effects not only for the media players themselves but for society at large. In Europe and elsewhere, the media and their artefacts

---

Video Copyright, *Reuters*, November 11, 2006, at http://news.com.com/Google+faces+legal+challenges+over+video+copyright/2100-1030_3-6134679.html.

[234] *See* Jefferson Graham, Google Takes Hits From Youtube's Use Of Video Clips, *USA Today*, February 13, 2007, *at* http://www.usatoday.com/tech/news/2007-02-12-google-youtube_x.htm. A French Film producer sued Google for copyright infringement. It asked the court to sentence Google to provide compensation for loss of income. It alleged that Google had not acted as a simple host but as a fully responsible publisher when it made available its film on Googe Video. The film was downloaded 43,000 times in a very short time lapse. Astrid Wendlandt & William Emmanuel, French Film Producer Sues Google France, *Reuters*, November 23, 2006, *at* http://today.reuters.com.

[235] *See* for this Aoife White, Court to Hear Google-Newspaper Fight, *CBS News*, November 23, 2006, *at* http://www.cbsnews.com/stories/2006/11/23/ap/business/mainD8LITLI00.shtml; and Google Settles Copyright Dispute with 2 Groups in Belgium, *International Herald Tribune*, November 24, 2006, *at* http://www.iht.com/articles/2006/11/24/business/google.php.

[236] Some of the other significant barriers to entry in the search engine market are hardware related. Google and its competitors are currently engaged in an arms race toward ever more powerful server capacity. Each of them is believed to have many hundreds of thousands of servers in their server farms or datacentres. This server capacity provides search engines with the capability to speedily answer user queries. Speed is considered a major competitive element for attracting users. The server base may thus be considered a major barrier to entry, as it is unlikely that new entrants could quickly deploy a similar infrastructure. *See* Elinor Mills, Google Says Speed Is King, *C|NET News*, November 9, 2006, *at* http://news.com.com/Google+says+speed+is+king/2100-1032_3-6134247.html.

are thus recognised as deserving special regulatory attention in the interest of freedom of expression and freedom of information. This regulatory intervention takes the form of media law (or public interest regulation). The broadcasting sector, for instance, is one of the most heavily regulated sectors. Broadcasters are granted revocable conditional licences that are then tied to a set of stringent ownership requirements, media concentration rules, and content regulations. It should be stressed that by and large the great majority of media laws originate in the member States.[237] However, due to the lack of clear metrics for assessing, for instance, media pluralism or impact of certain players on audiences, correcting for perceived market failures is a highly complex exercise. Intervention needs to be carried out with caution. This is especially so in fast-paced technology markets.

The important question thus arises to what extent and how traditional media laws are applicable in search engine-related questions. Generally speaking, media law does not talk about search engines; search engines are not in the media law dictionary. Despite their importance in the information society, search engines are systematically left out of sector-specific regulations.[238] This is no different at the European level. For instance, the main media regulatory instrument at the European level is currently the TV Without Frontiers Directive (TVWF).[239] The TVWF Directive explicitly excludes "communication services providing items of information or other messages on individual demand." While the TVWF Directive is currently in the process of being amended, the basic scope of the TVWF Directive does not change in relation to search engines. The on-going discussions seem to make clear that search engines that provide links to audiovisual content shall not be considered audiovisual media services in the sense of the Directive on AVMS.[240] Likewise, though search engines are closely related to EPGs for DTV, the Framework Directive only covers "associated facilities" that relate to the provision of DTV or digital radio as narrowly defined in the specific Directives. Search engines are not regulated under communications law either. The EU communications framework provides that it does not regulate services which provide or exercise editorial control over content transmitted over electronic communications networks. In sum, search engines seem to be beyond the scope of European laws relating to media and communications services.[241]

This regulatory gap is perhaps the result of the particularly complex nature of search engines, and the question arises to what extent they can be compared to current media players. To most people, search engines appear objective because they are fully automated, give content providers the choice whether to be indexed or not, and merely respond to user queries. Search engines also like to portray themselves as such. For instance, Google stresses its objectivity and lack of bias on its very site when declaring that "our search results are generated completely objectively and are independent of the beliefs and preferences of those who work at Google."[242] This desire of complete

---

[237] See, for instance, the debate on the proposed EU Media Pluralism Directive which was due to remove barriers to cross-border activities of media players, by harmonising the media concentration rules across Europe. In the end, however, the Directive was never proposed for mainly political reasons: MS did not want to give up control over their media ownership laws. *See* G. Doyle, From 'Pluralism' to 'Ownership': Europe's Emergent Policy on Media Concentrations Navigates the Doldrums, *Journal of Information, Law and Technology (JILT)* (1997), http://elj.warwick.ac.uk/jilt/commsreg/97_3doyl/

[238] *See* Nico van Eijk, Search engines: Seek and Ye Shall Find? The Position of Search Engines in Law, *IRIS plus* (Supplement to *IRIS - Legal observations of the European Audiovisual Observatory*), 2006-2, *at* www.obs.coe.int/oea_publ/iris/iris_plus/iplus2_2006.pdf.en.

[239] *See* Directive 89/552/EEC of 3 October 1989 on the Coordination of Certain Provisions laid down by Law, Regulation or Administrative Action in Member States Concerning the Pursuit of Television Broadcasting Activities, O.J. L.298/23 of 17 October 1989.

[240] See for this http://www.hieronymi.de/PDF%20Dokumente/376676XM.pdf

[241] *See* Nico van Eijck, *supra*, p.5.

[242] *See* http://www.google.com/explanation.html

impartiality is one of the reasons why search engines are careful when it comes to hand manipulation or intervention in the results.[243]

At the same time, search engines have stressed their subjectivity in their relation with web masters in regard to search engine optimisation,[244] or in disputes with content providers. For instance, search engines have argued in recent law suits over ranking that ranking is a subjective statement of opinion about page quality, which falls under the right to freedom of expression.[245] Also, for reasons of search fraud, Google and other search engines cannot be totally passive conduits. They have an interest in preventing fraud, otherwise they may risk that users turn to other search engines that provide better, more relevant, search results. This subjectivity is logical: very much like media players, search engines are trying to maximize user satisfaction, and thus they must include some sort of subjectivity. In other words, the different with classic media players may be a mere matter of degree.[246] Due to the vast amounts of information, automated processes have become common place, and direct editorial intervention by humans as regards the results of the algorithmic selection is the exception. In this view, search engines have some degree of subjectivity like other media players, but their editorial choices are enshrined in the actual algorithm. This consideration is especially important at a time when search engines are moving toward content aggregation, proactively pushing content to the end user, and are thus acting in many ways like personalised broadcasters.

A number of commentators have been debating whether some form of tailored media regulations ought to be enacted that take into account the specificities of search engines. Search engines are very similar to other media players. In fact, they are taking away large amounts of advertising income from classic media players. In recent years search engines have acquired a prominent role in granting users widespread access to information, and in giving the various advertisers even more "tailored eyeballs" than any broadcaster could offer them. Likewise, technologists and ethics scholars have convincingly stressed the fact that technology is not neutral, but that it has values and bias embedded in it.[247] Examples of proposed media law-type measures are increased transparency and various labelling and signalling measures by trusted third parties,[248] or public investment in alternative search engines.[249]

At least one other commentator argues, on the one hand, that it is unavoidable that search engines make editorial judgments. But those editorial judgments are both desirable and necessary. This is so

---

[243] *See* Rachel Williams, Search engine takes on web bombers, *Sydney Morning Herald*, January 31, 2007, *at* http://www.smh.com.au/articles/2007/01/30/1169919369737.html.

[244] *See* James Grimmelmann, *supra*, p.27

[245] *See* KinderStart v. Google, Case 5:06-cv-02057-JF (N.D. Cal. motion to dismiss granted July 13, 2006. This is reminiscent of the case law that pitted broadcasters against cable operators. Broadcasters argued that they should be granted access to the cable network on grounds of freedom of expression, while cable operators argued that they too enjoyed the right to freedom of expression which included the right not to broadcast certain views. *See* Turner Broadcasting System, Inc.. v. F.C.C. (93-44), 512 U.S. 622 (1994).

[246] The recent *Copiepresse* case also evidenced this interesting tension. In the beginning of the judgment, Google argued that Google News was a specialised search engine and not an information portal. As such it did not compete with the newspapers' sites. But when it came to the exceptions, Google argued that its service fell under the fair use exception of news reporting. This tension reflects the problems people have in classifying search engines in the media world.

[247] *See* Lucas Introna & Helen Nissenbaum: Shaping the Web: Why the Politics of Search Engines Matters, *The Information Society*, 16(3), 2000, pp. 169-186. *See* Frank Pasquale, Rankings, Reductionism, and Responsibility**,** *Seton Hall Public Law Research Paper No. 888327,* February 25, 2006, *at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=888327.

[248] Transparency of ownership and sources is a central value of media regulation, and could become central in relation to search engines too. But this transparency should be adapted to the specificities of search engines. Some have argued that one should open the search algorithms to public scrutiny. But this stands in tension with the idea that algorithmic innovation is ensured through secrecy and trade secret protection.

[249] Media pluralism can be sub-divided in two types. *Internal* pluralism rules are measures that seek to ensure that each media outlet gives a fair and complete overview of the range of views on a give topic. *External* pluralism measures, on the other hand, seek to remedy the risk that the media sector be overly concentrated.

because search engines continually fight against spammers and fraudsters. In this view, government regulation will not be any more compelling at deciding which bias, which subjective view, should prevail in the ranking. However, this view rests on two assumptions or dynamics that would curb the bias of search engines. First, the move toward personalisation of search results moots the search engine bias since it breaks the above described snowball effect, and it caters for minority interest. Second, market forces and low switching costs between search engines mean that if a search engine's bias degrades the relevance of search results, users will use alternative search engines.[250]

The question arises whether the move towards AV search engines offers compelling reasons for re-thinking the current situation. One might need to distinguish between audio-visual and other media. Media law history has shown that the degree of media obligations increases as we move from text, to audio, to audio-visual. This may be inferred, first, from the distinct regulatory regimes that apply to radio and television broadcasters. Broadcasting regulation, for instance, is mainly a result of the cogent effects of AV programming on audiences.[251] Second, this has also transpired in some of the case law on media. In *Jersild*, for instance, the European Court of Human Rights accepted that restrictions on the right to freedom of expression may be more stringent in the case of audio-visual (as opposed to print) media when it stated that the latter often have 'a much more immediate and powerful effect.'[252]

In sum, search engines draw so much traffic that search engine web sites have become ideal candidates for advertising. In fact, search engines are key to the pay-per-click business model that is currently dominant. Second, on the basis of their indexing and recording of user queries and profiles, search engines are able to match user interests with the related content that is available on the Internet, and are increasingly converting themselves from mere conduits of information to active information gatherers pushing content to the user. It thus appears that search engines start competing with traditional media players in a number of respects. However, at present few media law obligations are directly applicable to search engines. This is paradoxical, since search engines are central to the new information economy, and as a result the position of search engines in media law is a topic for intense debates. It remains to be seen whether the switch to AV search engines, and the fact that the impact of audio-visual content is considered more cogent than text-based information products, will alter the existing equilibrium.

7.5.3. EU v. US: Law Impacts Innovation In AV Search

Analysts tend to agree that the search engine market is thriving at present. There are a number of innovation trends that can be spotted.[253] The question thus arises to what extent EU law allows for innovation, or may be seen as hampering it. The paper finds markedly different approaches to

---

[250] *See* Eric Goldman, Search Engine Bias and the Demise of Search Engine Utopianism, 9 *Yale Journal of Law and Technology* (2006), pp. 188-200; *at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=893892.

[251] Of course, one may argue that the main reason for this stringent regime was spectrum scarcity. But with many voices out there, it is submitted that with the resulting scarcity of attention, audio-visual media may still necessitate more careful consideration as regards regulation. Note that we will soon be witnessing a related move to audio-visual advertising. Google is expected to develop an audio version of AdSense which would allow any podcast producer to include ads in their shows. *See* Frank Barnako, Google Seen Powering Podcast Ad Growth, February 12, 2007, *at* http://internet.seekingalpha.com/article/26787.

[252] *See Jersild v. Denmark*, Judgment of 23 September 1994, A.298, p.23.

[253] *See* Nitin Karandikar, Top 17 Search Innovations Outside Of Google, May 7, 2007, http://www.readwriteweb.com/archives/top_17_search_innovations.php; *See also* Giorgio Soffiato, Le 17 innovazioni che cambieranno i motori di ricerca http://sitiwebmarketing.boraso.com/motori-di-ricerca-search-marketing/le-17-innovazioni-che-cambieranno-i-motori-di-ricerca.html; See also Emre Sokullu and Richard MacManus, Search 2.0 - what's next?, December 13, 2006, http://www.readwriteweb.com/archives/search_20_what_is_next.php; *See also* Charles Knight, The top 100 alternative search engines, January 29, 2007, http://www.readwriteweb.com/archives/top_100_alternative_search_engines.php (updated May 1, 2007 http://www.readwriteweb.com/archives/top_100_alt_search_engines_april07.php), and The future of search, technology review, July 16, 2007, at http://www.technologyreview.com/Biztech/19050/?a=f.

search engine regulation across the Atlantic. This is evident in copyright, trademarks and data protection law.

7.5.3.1. Copyright Law

Copyright infringement ultimately depends on the facts. Search engines may retrieve and display picture thumbnails as a result of image search, or they may do so proactively on portal-type sites such as Google news to illustrate the news stories. The copyright analysis might differ depending on particular circumstances. The analysis shows how US courts have tended to be more favourable towards search engine activities in copyright litigation. This can be seen, for instance, in the litigation on caching, the displaying of thumbnails, and the use of standardised robot exclusion protocols. The open-ended 'fair use' provision has enabled US courts to balance the pros and cons of search engine activities case by case. However, the balancing test does not confer much legal certainty.

European case law shows that European courts have been rather reluctant to modify their approaches in the wake of fast-paced technological changes in the search engine sector. For instance, they have stuck more to the letter of the law, requiring express prior permission from right-holders for the caching and displaying of text and visual content. This is partly because European copyright laws do not include catch-all fair use provisions. The result is, however, that while US courts have some leeway to adapt copyright to the changing circumstances, the application of copyright law by European Courts is more predictable and confers greater legal certainty.

The paper finds, first, that different courts have reached diametrically opposed conclusions on a number of issues. Second, case law appears to indicate that the closer search engines come to behaving like classic media players, the more likely it is that copyright laws will hamper their activities. Likewise, it appears that the current EU copyright laws make it hard for EU courts to account for the specificities and importance of search engines in the information economy (for instance, increased automatisation and data proliferation).

Comparing EU and US copyright laws in general terms, we can say that EU laws tend to provide a higher degree of legal certainty but its application to search engines may be considered more rigid. US law, on the other hand, is more flexible but may not confer as much legal certainty. Both approaches are not mutually exclusive and a key question for policy makers is how to find a balance between conferring rather rigid legal certainty and a forward-looking more flexible approach in such a fast-paced digital environment.

7.5.3.2. Trademark Law

There has been intense litigation on this issue on both sides of the Atlantic. In the beginning it appeared, relying on the holding in *Brookfield Communications*,[254] that US courts would find that search engines infringed trademark rules when auctioning trademarked words. However, this ruling was about search optimisation, and the use of trademarked items in web site meta-tags. The Court had ruled that it was possible to infringe trademark law by capturing *initial consumer attention*, even though no action is completed as a result of the confusion, may still be an infringement.[255] In *Playboy*,[256] the court applied the *Brookfield* holding to rule that a clear indication on the banner ad of the actual source and sponsor name eliminated likelihood of initial interest confusion. In

[254] Brookfield Communications, Inc v. West Coast Entertainment (DC California 1998 See Internet Business law services (September 29, 2006) The Initial Interest Confusion – Beginning of Liability for Search Engine Companies – relying on Brookfield Communications, Inc v. West Coast Entertainment (DC California 1998)

[255] The court stated that "*To capture initial consumer attention, even though no action is completed as a result of the confusion, may still be an infringement.*"

[256] *Playboy* v Netscape (2004), see on this Gasser, *supra,* p.211.

Geico,[257] a US court ruled that Geico had not presented sufficient factual evidence corroborating the finding that sales of TM to third parties constituted infringement since the ads themselves did not include the TM word and and there was no evidence that this activity alone caused confusion. US courts came to similar conclusions in a flurry of other recent cases.[258]

In the EU, on the other hand, TM litigation seems to follow a different course. In France, the number of trademark lawsuits against Google now number more than 40, and most have gone against the U.S. Internet company. A smaller number of cases have been brought in Belgium and Germany.[259] The position of European courts seems to be that the use of trademarked terms in auctions amounts to a trademark infringement.[260] Google France said that since the case began in 2003, it has implemented a policy barring Internet advertisers from buying search listings under trademarks held by others, as well as a ban on advertising for counterfeit products. There is thus a noticeable difference between EU cases, where TM infringement has been found, and US cases where search engines seem to be more immune.

In sum, it appears that jurisdictions that rely on the likelihood of confusion test, of which the US is the most well-known example, have inherently more flexibility built-in, and consequently more leeway for courts to conduct a balancing test. In this balancing test, Courts will be able to introduce important elements into the calculus such as the interests of competition, advertising innovation, comparative advertising or freedom of expression. In doing so, Courts are also able to bear in mind the importance of search engines in the information society (for all stakeholders), and the role of keyword advertising for funding them.

[257] Geico v Google, 2004, *at* http://blog.ericgoldman.org/archives/geicogoogleaug2005.pdf
[258] Check n go v Google, American Blind v Google (2005); Novak v Overture (2004) ; 800-JR-Cigar v Overture (2000); Newborn v Yahoo Inc. (2005); Rescuecom v. Google (trademark infringement dismissed) (September 28, 2006)
[259] *See* http://www.iht.com/articles/2006/06/28/business/lvmh.php
[260] *See* TGI Paris, 12 juillet 2006, GIFAM et autres v. Google France http://www.juriscom.net/jpt/visu.php?ID=848; CA Paris, 28 juin 2006, SARL Google, Sté Google Inc v. SA Louis Vuitton Malletier http://www.juriscom.net/jpt/visu.php?ID=837; Le Meridien Hotels v Google

7.6. Conclusions

1. Search is an advertising-based industry, relying heavily on well-known brands for its income. It should thus come as no surprise that the first series of cases involving search engines related to trademarks, and concerned the relation between advertisers and search engines. By contrast, the first generation of search engines caused relatively few problems in terms of copyright litigation. Search engines merely retrieved text data from the web, and displayed short snippets of text in reply to a specific user query. Over time, however, search engines started organising and giving users access to more economically valuable content, and copyright infringement claims have come to the fore. Data protection concerns have arisen only in recent times, in relation to the recording and processing of user search queries and user profiling activities.

   Search engines are essential tools in our current information ecosystem. Each of these three debates (copyright, trademarks, data protection) is ultimately about striking the right balance for society in relation to search engines. There is a need, on the one hand, to foster the efficient categorisation and organisation of content by a wide range of players such as search engines, relying on accurate user profiles and funded by advertising. On the other hand, there is equally an interest in incentivising the creation of digital content (copyright), fostering investments in creating goodwill for certain brands (trademarks), and supporting the widespread use of search engine technology (data protection).

   To be sure, law is only one of several possible regulatory modalities determining whether the most appropriate balance is struck. Other essential elements in this debate are technological standardisation (e.g. robot exclusion protocols, privacy enhancing technologies), and commercial deals between market players. Far from being independent from one another, these regulatory modalities impact each other. For instance, copyright law determines the use of robot exclusion protocols. Similarly, the way copyright law is applied may increase or decrease the pressure on search engines to conclude licencing agreements with content owners. However, this paper claims that law is a key policy lever with regard to search engines. The wording of the law, and its application by courts, has a major influence on whether a thriving market will emerge for search engines, including the future AV search engines. Instead of focusing on increased difficulties in applying the law, the shift towards more audio-visual search offers a unique opportunity to rethink trademark law, copyright law and data protection law for the digital environment.

   This paper argues that the legal problems encountered so far in relation to search engines may be expected to increase as we move into the AV search era. Issues relating to trademark law will become more acute in the audiovisual search context. This is because the ads that can be served using AV search technology are likely to have a more powerful influence on consumer habits than the presently predominant text-based ads. Likewise, the more audio-visual content is put on the Internet, the more we may expect copyright litigation with respect to AV search engines. The reason is that premium AV content is generally more costly to produce, and commercially more valuable than text-based content. Finally, it is already difficult for text-based content to return return pertinent results, but AV search engines will have to rely even more on user profiling; those user profiles will by the same token enable search engines to target users directly and thereby compete with traditional media and content owners.

   In sum, the analysis highlights that no radically new legal problems are to be expected in the AV search context, as compared to the existing text-based environment. However, the degree and amount of litigation may be expected to increase as we move on to an AV search environment.

2. Consequently, the switch to AV search appears to require policy makers to bring all of those legal questions in perspective. Trademark law is struggling to come to terms with the use of trademarked terms in the automated ad-triggering mechanisms of the search engine, because the

use of the trademarked term takes place in the background away from the consumer's eyes. With regard to copyright a set of completely new legal issues arises, including those surrounding the caching of content, or the scanning of books with a view to making them searchable. Data protection law has to be re-considered in view of the importance of search engine personalisation in helping users make sense of the vast amounts of information that is available on the Web. Automation and the search engine's unique functionality forces us to reconsider the fundaments of our current legal regime. Legal issues that could still be left aside in the text search era will now need to be addressed.

Over time, we have witnessed a steady transformation of search engines. Storage, bandwidth and processing power have increased dramatically, and automation has become more efficient. Search engines have gradually shifted from a reactive response to the user ('pull') to pro-actively proposing options to the user ('push'). Future search will require increasing organisation and categorisation of all sorts of information, particularly in audio-visual (AV) format. Due to this shift from pure retrievers to categorisers, search engines are in the process of becoming fully-fledged information portals, rivalling traditional media players.

As a result, the position of search engines in law goes beyond the individual laws. There is an increasing need to determine exactly which type of intermediaries search engines are considered to be, and as a result which is the default liability regime search engines should conform to. The least intrusive regulation for search engines is the liability regime laid down in the e-commerce directive. More developed liability and obligations for intermediaries exist in varying degrees in communications and media laws. This default regime is not only important as such, but it also influences the position of courts in relation to legal claims regarding, for instance, copyright, trademark, and data protection. If search engines are analogous to media enterprises, then it follows that they may more easily be held liable for copyright, trademarks, and data protection infringements. Determining the specific nature of search engines, and the default liability regime that applies to them, is a prerequisite for a concerted approach across the various other laws that apply to them. Leaving search engines in a legal vacuum may end up hampering the development of a thriving European search engines sector.

3. Implicitly, the above legal analysis forces us to re-think innovation policy in relation to the search engine context. For instance, the paper claims that copyright's main policy relevance lies in its possible effects on the emerging market for meta-data production. A basic goal of copyright law is to incentivise the creation of content. Given the proliferation of digital content, it becomes more difficult to locate specific content. It becomes comparatively more important to promote the development of methods for accurate organising of AV content than to incentivise creation. This is particularly true in the AV search context, where organising AV content for efficient retrieval is a major challenge, and where many players currently compete to provide the leading technology or method for producing accurate meta-data.

Strong copyright law will force AV search engines to conclude licensing agreements over the organising of content. It supports technology's role in creating an environment of total control whereby content owners are able to enforce licences over snippets of text, images and the way they are used and categorised. By contrast, a more relaxed application of copyright law might take into account the growing importance of creating a market for AV meta-data production and meta-data technologies in an environment characterised by data proliferation. This approach would give incentives for the creation of content, while allowing the development of technologies for producing meta-data.

The analysis suggests that EU and US courts appear to have drawn markedly different conclusions on the same issues as a result of the differences of the respective legal orders. Comparing EU and US copyright law in general terms, we can say that EU copyright law tends to provide a higher degree of legal certainty but its application to search engines may be considered more rigid. US law, on the other hand, is more flexible but may not confer as much

legal certainty. Similarly, US and EU trademark law may well have yielded somewhat different results so far. This could well be a direct result of the fact that US trademark law, relying to a large extent on the "likelihood of consumer confusion|" test, includes more balancing possibilities for Courts than trademark law in many EU Member States with its focus on "trademark use". Finally, it appears equally important to consider the possible effect of data protection laws on innovation. The EU has a much more developed data protection regime than the US, which relies mainly on regulation by technology and regulation by contract (privacy terms and conditions). With a number of high profile debates regarding the logging of user data, and ensuing public concern, there can be little doubt about the importance of addressing this issue. However, it is important to beat in mind the need to address these issues with as minimal impact as possible on critical innovation (such as for instance search engine personalisation).

7.7. Future Research

7.7.1. Social Trends

This section will consider the social aspects of AV search, by placing search in context. The backdrop against which AV search engines will need to be developed is one of increasing user participation (coined web 2.0). The section will thus show how social aspects have always been, and are increasingly revolving around user participation. On the one hand, search engines are at the heart of all of the upcoming web 2.0 applications such as wikipedia, Flickr, or YouTube. On the other hand, search engines are fundamentally dependent on humans, from the early stages onward (e.g. Yahoo was initially a human edited directory). The leading search engines currently observe and rely on user behaviour (clicks, popular URLs, and link structure). There is a multitude of sites and services out there that can be said to offer social search. Chris Sherman sorts them into a number of categories: Shared bookmarks and web pages (Del.icio.us, Shadows, Furl); Tag engines, tagging and searching blogs and RSS feeds (Technorati, Bloglines); Collaborative directories (ODP, Prefound, Zimbio and Wikipedia); Personalized verticals or collaborative search engines (Google Custom Search, Eurekster, Rollyo); Collaborative harvesters (Digg, Netscape, Reddit and Popurl); Social Q&A sites (Yahoo Answers, Answerbag). The section will conclude by asking how and whether the current trends may be expected to increase in the AV search era.

This section will place the search engines within the wider context of access to information and knowledge. Search engines are key tools that help determine to what extent information is accessible at large. It will consider the recognition of search engines' special status in current regulatory initiatives seeking to foster widespread access to knowledge, and will ponder whether this role may be expected to increase in the switch to AV search.

At the same time, it should be remembered that AV search may exacerbate the current trend of increasing centralisation of search engines. This poses deep questions of media pluralism, as a few players seem to become the main entry doors, or access points to the digital world. The section will briefly refer to a number of recent examples, ranging from manipulation of search engines by third parties, to the deliberate intervention of search engines themselves, to cases of censorship. This section concludes by considering whether these issues warrant a more careful approach in the AV era. It revisits the history of media regulation and looks at the distinction between text, audio and video.

7.7.2. Economic trends

The search engine landscape consists of three main parts. First, there is a large number of content providers that make their content available for indexing by the search engine's crawlers. Second, there are the advertisers that provide most of the income for the search engine activity. Finally, new players have arisen whose livelihood depends on the business model of search engines. This section will provide information on the most important player and will consider their respective interests, seeking foremost to give an idea about the various players involved and their respective interests.

The predominant business model for search is currently advertising. The leading search engines generate revenue primarily by delivering online advertisement. The importance of advertising for search engines is self-evident, also, from their spending. In 2006, Google was planning to spend 70% of its resources on search and advertising related topics. A few years ago, advertising on search engine sites was very much like in analogue media. This included mainly banner advertising, and sometimes paid placement, whereby ads were mixed with organic results. But many users considered these too intrusive and not sufficiently targeted or relevant to the search or web site topic, and not taking advantage of the interactive nature of the Web. By contrast, online advertising differs from traditional advertising that traceability of results is easier. Mainstream search engines now mainly rely on two techniques. These are advertising business models that rely on actual user

behaviour: pay-per-click (advertiser pays each time the user clicks on the ad) and increasingly pay-per-performance (advertiser pays each time the user purchases or prints or takes any action that shows similar interest). This section will place the current leading business model based on text advertising in context, and will ponder to what extent the switch to audio-visual search applications warrants/demands a different approach.

Although dominated by three US-based giants (i.e. Google, Yahoo! and Microsoft), the search engine market is currently extremely active. The search engine space spans across all sorts of information. We currently witness the deployment of search engines for health, property, news, job, person, code or patent information. They will increasingly be able to sift through information coming from a wide range of information sources (including emails, blogs, chat boxes, etc.) and devices (desktop, mobile). Search engines are able to return relevant search results according to the user's geographic location or search history. Virtually any type or sort of information, any type of digital device or platform, may be relevant for search engines. Search is thus an increasingly central activity that has become the default manner for many users to interact with the vast amounts of information that are available on the Web. This section will consider the importance of search, highlight current market trends, and ponder to what extent this is likely to change in an AV search context. The developments will be assessed against the possibility of increased market concentration (including initial analysis of barriers to entry, switching costs & network effects)

7.7.3. Further Legal Aspects

Depending on the findings of the research on economic and social trends, a number of additional questions comes up. These include in order of priority:

7.7.3.1. Constitutional law

A. Freedom of expression [Art.10 European Convention on Human Rights]

- What is the role of search engines in fostering the right to freedom of expression (right to be included in index) and access to information (right to have access to a diverse set of information)?
- Do search engines equally have a right to freedom of expression [cf. argument of cable network operators against TV operators]? Does this right to freedom of expression clash with other players' own right to freedom of expression [e.g. to be listed in the organic results, in the advertising results, etc.]
- What might be considered appropriate restrictions to freedom of expression in the context of search engines? [youth protection, blasphemy, national security and terrorism, racism, violent content, etc.]
- Are restrictions more appropriate in the case of AV search, given that audio-visual content is regarded as more powerful, immediate, than text [cf. *Jersild* case]?

B. Right to respect of Private Life [Art.8 European Convention on Human Rights]

- Search engines enable easy access to details about many persons' private lives, and at the same time search engines record a lot of personal information about their users in the act of searching
- Is current regulation in line with the constitutional right to privacy (proportionality of means to ends)?
- What are appropriate restrictions to the right to privacy? [cybercrime, etc.]
- Are those restrictions less appropriate in the case of AV search? Should privacy be more protected in the case of AV search given the nature of AV content?

C. Right to Property [Art.1, 1st Protocol, European Convention on Human Rights]

- Is there a constitutional right to intangible property? [see EU Charter of Fundamental Rights]
- Search engines may give access to all types of information that is protected by some form of intellectual property right [EU database directive, copyright], or the way in which search engines work may enable certain players to make profits at the expense of the owner of a certain IPR [trademark]
- Consequently, what are appropriate restrictions to the right to property – fair use, etc. Are those restrictions applicable in the case of AV search?

7.7.3.2. Intellectual Property Rights

There appears to be more need to research the potential implications of the *sui generis* Database Directive (i.e. EU Directive 96/9/EC on the Legal Protection of Databases) on search engines.

- Indexes are in effect huge databases. Can it be argued that indexes fall within the sui generis database directive?
- What are the practical consequences of the application of this new form of intellectual property to search engines and AV search engines in particular?
- What does this mean in terms of competition?

7.7.3.3. EU Competition Law

The classic competition law analysis is to be carried out for the existing market of text-based search. The key question relating to AV search is then whether possible dominance in the existing market risks to be leveraged into the newly arising AV search market.

A. Single Dominance [Art.82]

- What is the search engines market structure, and what is the relevant market?
- Determine whether likely dominance [study switching costs, barriers to entry]
- will main players in text search leverage market power into AV search? [leveraging]
- focus on abuse of dominance: essential facilities, bundling, tying, leveraging, etc.

B. Anti-competitive agreements and joint (or Collective) Dominance [Art.81]

- is there an oligopoly in the search engine sector? Is there evidence of anti-competitive agreements between market players?
- is "joint dominance" likely in fast-paced market characterized by technological innovation?
- What kind of abuse may be existing? Essential facilities?

7.7.3.4. Media & Communications Law

A. E-Commerce Directive

- Does the e-commerce directive apply and what does this imply in terms of liability of search engine providers?
- Can we identify self-regulation or co-regulation initiatives and existing codes of conduct.
- What is the relation beween e-commerce and TVWF directive?

B. Media Law

- Can search engines be defined as media in the sense of media law?
- Are Search engines biased? If so, is it perceived by users? Are there examples of intentional bias [e.g. BMW, China]?
- Depending on the above, which principles of media law are relevant in relation to text search? [transparency and independence requirements, ownership limits, language and other quotas, etc.]
- Analyse in detail potential application of TV without frontiers Directive to search engines
- Is there any marked difference for AV search?
- What type of EU intervention is warranted/possible, if any?

C. EU Communications Law

- What is the place of search engines in the regulatory package for electronic communications?
- Can we analogise with existing regulations on APIs and EPGs? Why?
- Should we foresee some form of regulation analogous to existing universal service obligations?
- Is intervention warranted on the basis of Significant Market Power [SMP]?
- What about network neutrality? Should we impose different regulatory conditions on players who are responsible for a lot of network traffic?
- What type of standardization, if any, is legally warranted/undertaken?
- Any difference for AV search?

7.7.3.5. Law of Obligations / Liability Law

A. EU Product Liability

- What kind of tort obligations may be imposed on the search engine operator?
- Can these be held liable for not filtering out harmful content? For not giving accurate results? Is a notice at the top of the page sufficient?
- Does the EU product liability Directive apply to search engines?

B. Consumer Protection

- Which types of EU consumer protection regulations apply to search engines?
- May search engines be held liable for spyware, malware, or other software that may be damaging or present on the user's computer as a result of search engine use?
- Is there a filtering obligation on the search engine operator?

C. Anti-Spam Laws

- Web sites and content providers could use images to attract traffic using AV content (falsely claiming to be about that content but in fact being about something totally different (cf, discussion with the use of (invisible) and incorrect metatags on web pages to attract traffic)
- Do anti-spam laws foresee spamdexing and techniques used by content operators to divert traffic this way? If not, should they?

ANNEX to chapter 3: Summary and goals of use cases

Research Effort	Name	UC	Action	Corpora	Method (Requirement)	Product	System Env.	User	User Class	Summary/Notes	Industry/Community
Project	DIVAS	1	Retrieval (General)	Text (Annotations)	Relevance Distance Metric	Unspecified	Unspecified	Unspecified	End User (Vague)	Similarity to user needs	
Project	DIVAS	1	Personalization	Annotations (Profiles)	Unspecified	Profile Manager/Editor	Unspecified	Unspecified	End User (Vague)	Profile creation (by specifying keywords and grouping them by importance); profile editing, sharing, re-use.	
Project	DIVAS	1	Content Delivery	Audiovisual	Unspecified	Targeted Delivery System	Unspecified	Unspecified	Content Provider	Enhanced syndication using profiles: format specification, content protection with DRM	
Project	DIVAS	1	Extraction/Indexing	Annotations (Profiles)	Unspecified	Indexed Content (Profiles)	Unspecified	Unspecified	End User (Vague)	Dynamic building of communities sharing similar interests (classified sets of peer communities)	
Project	DIVAS	1	Extraction/Indexing	Text (Annotations)	Audio-To-Concept	Index Generator (Audio)	Unspecified	Unspecified	Content Provider	Enable the speech-to-text transcription of audio content and subsequent text-to-text matching of lists of keywords.	
Project	DIVAS	2	Content Delivery	Audiovisual	Unspecified	Targeted Delivery System	Unspecified	Unspecified	Content Provider	Enable the use of an information provision service that informs regularly a user of new information.	
Project	DIVAS	3	Extraction/Indexing	Text (Annotations)	Audio-To-Concept	Index Generator (Audio)	Unspecified	Unspecified	Content Provider	Enable the speech-to-text transcription of audio content and subsequent text-to-text matching of lists of keywords.	
Project	DIVAS	3	Analytics (Multimedia)	Audiovisual	Segmentation	Index Generator (General)	Unspecified	Unspecified	Content Provider	Enable alignment-correct synchronization between the audio of video and transcriptions	
Project	DIVAS	3	Analytics (Text)	Text (Annotations)	Statistical Classification (General)	Concordance Generator	Unspecified	Unspecified	Content Provider	Enable statistical analysis of the most frequent spoken words in textual transcriptions of a spoken video.	
Project	DIVAS	3	Classification Systems (General)	Text (Concordances)	Unspecified	Taxonomy Generator (Topical)	Unspecified	Unspecified	Content Provider	Enable creation of structured, hierarchical lists.	
Project	DIVAS	3	Classification Systems (General)	Unspecified	Unspecified	Taxonomy Generator (Interlingual)	Unspecified	Unspecified	Content Provider	Automated word-to-word translation between languages.	
Project	DIVAS	4	Retrieval (Search)	Annotations (Video)	Unspecified	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Locate video from a textual transcription	
Project	DIVAS	4	Retrieval (Search)	Audiovisual	Query by Example	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Locate video from a picture	
Project	DIVAS	4	Retrieval (Search)	Audiovisual	Unspecified	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Locate the original source of a video	
Project	DIVAS	5	Content Delivery	Audiovisual	Query by Controlled Metadata	Targeted Delivery System	Unspecified	Unspecified	Content Provider	Comparison of reference models to incoming feeds to deliver content for 3rd parties (“Picture Matching”); Audience monitoring; Broadcast regulation; Compliance notification	
Project	DIVAS	5	Extraction/Indexing	Audiovisual	Vague	Index Generator (General)	Unspecified	Unspecified	Content Provider	Encoding/analysis of multimedia (reference and incoming streams)	
Project	DIVAS	6	Retrieval (Search)	Audiovisual	Query by Example	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Identify and locate an originating document using an extract of it.	
Project	DIVAS	7	Extraction/Indexing	Unspecified	Unspecified	Indexed Content (Vague)	Unspecified	Unspecified	Content Provider		
Project	DIVAS	7	Retrieval (Search)	Controlled Metadata	Query by Controlled Metadata	Retrieval System (General)	Unspecified	Unspecified	Content Provider	Search for duplicates by comparing index values; output file will be created with found duplicates.	

Project	DIVAS	7	Retrieval (Search)	Controlled Metadata	Query by Controlled Metadata	Retrieval System (General)	Unspecified	Unspecified	Content Provider	Search for duplicates by comparing index values; output file will be created with found duplicates.	
Project	DIVAS	8	Extraction/Indexing	Audiovisual	Vague	Index Generator (General)	Unspecified	Unspecified	End User (Vague)	Online search to automatically create fingerprints of songs and movies	
Project	DIVAS	8	Retrieval (Search)	Audiovisual	Query by Example (Fingerprint)	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Use available fingerprint databases for audio identification	
Project	RUSHES	1	Retrieval (General)	Text (Annotations)	Query by Keyword	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Search by using keywords or semantic concepts.	
Project	RUSHES	1	Retrieval (General)	Annotations (Video)	Query by Controlled Metadata	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Search by using keywords or semantic concepts.	
Project	RUSHES	2	Retrieval (General)	Audiovisual	Query by Example	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Search by visual similarity.	
Project	RUSHES	3	Retrieval (General)	Annotations (Video)	Query by Controlled Metadata	Retrieval System (General)	Unspecified	Journalist	End User (Professional)	Search by location, date and time.	Journalism, Broadcasting
Project	SAPIR	1	Retrieval (General)	Image	Query by Example	Retrieval System (General)	Open	Photography	End User (Simple)	Photo search in social network collections.	Photography
Project	SAPIR	2	Retrieval (General)	Vague	Query by Example	Retrieval System (General)	Unspecified	Consumer	End User (Simple)	Receipts search in social network collections by multimedia queries.	Consumer
Project	SAPIR	3	Retrieval (Browse)	Audio	Query by Example	Retrieval System (General)	Unspecified	Consumer, Music	End User (Simple)	Discover new music using audio excerpts (mobile).	Consumer, Music
Project	SAPIR	4	Extraction/Indexing	Image	Feature Detection	Indexed Content (Images)	Open	Tourism/Heritage	End User (Simple)	Retrieve information on buildings/objects depicted in tourist photos (mobile).	Tourism/Heritage
Project	SAPIR	5	Extraction/Indexing	Audiovisual	Unspecified	Index Generator (General)	Open	Unspecified	End User (Simple, Professional)	Extract information related to films that occur on TV.	
Project	SAPIR	5	Retrieval (General)	Audiovisual	Query by Example	Retrieval System (General)	Open	Unspecified	End User (Simple, Professional)	Search for films/video based on extracted features.	
Project	SAPIR	6	Retrieval (General)	Image	Query by Controlled Metadata	Retrieval System (General)	Open	Unspecified	End User (Simple, Professional)	Utilizes other people's closeness to particular events to retrieve photos.	
Project	TRIPOD	1	Extraction/Indexing	Annotations (Image)	Query by Keyword	Index Generator (General)	Open	Unspecified	Content Provider	Augment photo metadata using captions (i.e., toponyms) to query the web and discover extra information.	
Project	TRIPOD	2	Extraction/Indexing	Controlled Metadata	Query by Controlled Metadata	Index Generator (General)	Open	Unspecified	Content Provider	Create rich photo captions by identifying features using full metadata (Location and Direction) to query geodatabases.	
Project	TRIPOD	3	Retrieval (Browse)	Image (Vague)	Query by Controlled Metadata (Profile)	Recommender System	Closed	Tourism/Heritage	Content Provider	Postcard recommender by profiles (taste, style)	Tourism/Heritage
Project	TRIPOD	3	Personalization	Image (Vague)	Query by Controlled Metadata (Profile)	Indexed Content (Profiles)	Closed	Tourism/Heritage	Content Provider	Postcard recommender by profiles (taste, style)	Tourism/Heritage
Project	VICTORY	1	Retrieval (General)	Unspecified	Vague	Retrieval System (General)	Closed	Automotive, Designers	End User (Professional)	Search for the parts designs using a similarity search.	Automotive, Product Designers

Project	VICTORY	2	Retrieval (General)	Text (Annotations) (Vague)	Vague	Retrieval System (General)	Unspecified	Designers, Decision Makers	End User (Professional)	Collaborative design through query extension and relevance feedback	Designers, Decision Makers
Project	VICTORY	3	Vague	Unspecified	Unspecified	Unspecified	Unspecified	Consumer, Designers	End User (Simple)	Information sharing/exchange.	Consumers, Design Community
Project	VICTORY	4	Unspecified	Vague	Unspecified	Social Sharing System	Open	Open Gaming Communities	End User (Simple)	Customized figure creation sharing.	Open Gaming Communities
Project	VICTORY	5	Retrieval (General)	Vague	Unspecified	Retrieval System (General)	Unspecified	Maintenance/Installation/Support Personnel	End User (Professional)	Access to guideline/catalog repositories.	Maintenance/Installation/Support Personnel
Project	VICTORY	6	Unspecified	Unspecified	Unspecified	Vague	Unspecified	Tourism/Heritage, Travel	End User (Simple)	Information capture and similarity search on objects of interest (mobile).	Tourism/Heritage, Travel
Project	VITALAS	1,1	Extraction/Indexing	Unspecified	Unspecified	Indexed Content (Profiles)	Unspecified	Archivist, Journalist	Content Provider	Build user profiles and personalize user access	Archivist, Journalist
Project	VITALAS	1,2	Retrieval (General)	Image	Query by Controlled Metadata (Profile)	Retrieval System (General)	Unspecified	Archivist, Journalist	End User (Professional)	Find pictures using a search profile.	Archivist, Journalist
Project	VITALAS	1,2	Personalization	Image	Query by Controlled Metadata (Profile)	Retrieval System (General)		Archivist, Journalist	End User (Professional)	Find pictures relevant to a journalist's interests.	Archivist, Journalist
Project	VITALAS	2,1	Extraction/Indexing	Image	Unspecified	Index Generator (General)	Unspecified	Archivist, Journalist	Content Provider	Automated indexing	Archivist, Journalist
Project	VITALAS	2,2	Extraction/Indexing	Image	Visual-to-Concept	Index Generator (General)	Unspecified	Unspecified	Unspecified	Proximity measure: Fusion of visual and textual descriptors.	
Project	VITALAS	2,2	Extraction/Indexing	Image	Statistical Classification (General)	Clustering Algorithm	Unspecified	Unspecified	Content Provider	Classification or non-supervised clustering	
Project	VITALAS	2,2	Retrieval (General)	Image	GUI Development	GUI	N/A	Unspecified	End User (Simple, Professional)	Interactive visualization map	

Project	VITALAS	2,3	Retrieval (General)	Image	Relevance Feedback	Retrieval System (General)	Unspecified	Archivist, Journalist, Art Director	End User (Simple, Professional)	Interactive browsing based on cross modal proximity and interactive relevance feedback	Archivist, Journalist, Art Director
Project	VITALAS	2,3	Retrieval (General)	Image	Cross Modal Proximity	Retrieval System (General)	Unspecified	Archivist, Journalist, Art Director	End User (Simple, Professional)	Interactive browsing based on cross modal proximity and interactive relevance feedback	Archivist, Journalist, Art Director
Project	VITALAS	2,4	Retrieval (General)	Image	Query by Controlled Metadata	Retrieval System (General)	Unspecified	Archivist, Journalist	End User (Simple, Professional)	Retrieve pictures by matching text queries to conceptual links.	Archivist, Journalist
Project	VITALAS	2,4	Retrieval (General)	Image	Feature Detection	Retrieval System (General)		Unspecified		Retrieve pictures of a well known person.	
Project	VITALAS	3,1	Retrieval (General)	Audiovisual	Query by Example	Retrieval System (General)	Unspecified	Archivist, Researcher	End User (Simple, Professional)	Cross modal similarity search to find video extracts.	Archivist, Researcher
Project	VITALAS	3,2	Extraction/Indexing	Audiovisual	Unspecified	Indexed Content (Audiovisual)	Unspecified	Archivist, Researcher	Content Provider	Establish classification, index and retrieve sequences.	Archivist, Researcher
Project	VITALAS	3,3	Retrieval (General)	Audiovisual	Query by Controlled Metadata	Retrieval System (General)	Unspecified	Archivist, Researcher	End User (Simple, Professional)	Retrieval by semantic concept. (Concept Learning?)	Archivist, Researcher
Project	VITALAS	3,3	Extraction/Indexing	Audiovisual	Audio-To-Concept	Index Generator (Audio)		Archivist, Researcher	Content Provider	Retrieval by semantic concept. (Concept Learning?)	Archivist, Researcher
Project	VITALAS	3,3	Extraction/Indexing	Audiovisual	Feature Detection (Face Recognition)	Indexed Content (Images)	Unspecified	Archivist, Researcher	Content Provider	Retrieval by semantic concept. (Concept Learning?)	Archivist, Researcher
Project	VITALAS	4,1	Retrieval (Browse)	Audiovisual	GUI Development	GUI	Unspecified	Archivist, Journalist, Researcher	End User (Professional)	Interactive cartographic visualization for content overview of video with spatial layout with constraints.	Archivist, Journalist, Researcher

Project	VITALAS	4,1	Retrieval (Search)	Audiovisual		Indexing Manager/Editor (General)	Unspecified	Archivist, Journalist, Researcher	Content Provider	Tool development for building homogeneous and normalized textual metadata.	Archivist, Journalist, Researcher
Project	VITALAS	4,1	Extraction/Indexing	Audiovisual	Unspecified	Indexed Content (Audiovisual)	Unspecified	Archivist, Journalist, Researcher	Content Provider	Automatic indexing	Archivist, Journalist, Researcher
Project	VITALAS	4,1	Retrieval (Search)	Audiovisual	Unspecified	Retrieval System (General)	Unspecified	Archivist, Journalist, Researcher	Content Provider	Automatic or assisted identification of the main line of the program (backbone)	Archivist, Journalist, Researcher
Project	VITALAS	4,2	Retrieval (General)	Audiovisual	Query by Example	Retrieval System (General)	Unspecified	Archivist	End User (Professional)	Retrieve news programs with similar subjects (examples can be visual, audio or textual).	Archivist
Project	VITALAS	4,3	Extraction/Indexing	Audiovisual	Speech Recognition	Indexed Content (Audio)	Unspecified	Archivist	Content Provider	Audio detection of logos; Audio recognition of spoken words.	Archivist
Project	VITALAS	4,3	Extraction/Indexing	Audiovisual	Feature Detection	Indexed Content (Audiovisual)	Unspecified	Archivist	Content Provider	Visual object detection (logos, map, text incrustation).	Archivist
Project	VITALAS	4,3	Analytics (Multimedia)	Audiovisual	Segmentation	Multimedia Segments	Unspecified	Archivist	Content Provider	Localization of temporal "markers".	Archivist
National Initiative	iAD	N/A	Retrieval (General)	Vague	Unspecified	Retrieval System (General)	Unspecified	Unspecified	Unspecified		Enterprise Search
National Initiative	iAD	N/A	Analytics (General)	Vague	Unspecified	Retrieval System (General)	Unspecified	Unspecified	Unspecified		Enterprise Search
National Initiative	iAD	N/A	Extraction/Indexing	Vague	Semantic Classification (General)	Unspecified	Unspecified	Unspecified	Unspecified		Enterprise Search
National Initiative	IM2	N/A	Retrieval (General)	Audiovisual	GUI Development	Social Sharing System	Unspecified	Unspecified	End User (Vague)	Development of new meeting browsers.	
National Initiative	IM2	N/A	Analytics (Multimedia)	Audiovisual	Speech Recognition	Multimedia Segments	Unspecified	Unspecified	End User (Vague)	Automatic discourse processing	

National Initiative	IM2	N/A	Analytics (Multimedia)	Audiovisual	Feature Detection	Multimedia Segments	Unspecified	Unspecified	End User (Vague)	Scene analysis, speaker segmentation and tracking, vocabulary speech recognition.
National Initiative	IM2	N/A	Extraction/Indexing	Audiovisual	Audio-to-Concept	Indexed Content (Audiovisual)	Unspecified	Unspecified	End User (Vague)	Collection and full annotation of large amounts (100 hours) of multimodal meeting recordings
National Initiative	IM2	N/A	Extraction/Indexing	Audiovisual	Visual-to-Concept	Indexed Content (Audiovisual)	Unspecified	Unspecified	End User (Vague)	Collection and full annotation of large amounts (100 hours) of multimodal meeting recordings
National Initiative	IM2	N/A	Analytics (Multimedia)	Audiovisual	Segmentation	Multimedia Segments	Unspecified	Unspecified	End User (Vague)	
National Initiative	Multimedia N	N/A	Retrieval (General)	Audiovisual	Vague	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Semantic web search
National Initiative	Multimedia N	N/A	Retrieval (General)	Audiovisual	GUI Development	GUI	Unspecified	Unspecified	End User (Vague)	Develop a multimedia browser (MediaMill).
National Initiative	Multimedia N	N/A	Extraction/Indexing	Audiovisual	Semantic Classification (General)	Metadata Manager/Editor (General)	Unspecified	Unspecified	End User (Vague)	Develop a P2P system for video metadata exchange and indexing (StreetTivo).
National Initiative	Multimedia N	N/A	Analytics (General)	Audiovisual	Unspecified	Metadata Manager/Editor (General)	Unspecified	Unspecified	End User (Vague)	Develop a P2P system for video metadata exchange and indexing (StreetTivo).
National Initiative	MundoAV	N/A	Extraction/Indexing	Audiovisual	Vague	Indexed Content (Audiovisual)	Unspecified	Companies, Audiovisual Professionals	End User (Professional)	Selectively index content by development/deployment of a spider that retrieves indexed data from website by looking in <XML> Index Files, and stores it in a local database.
National Initiative	MundoAV	N/A	Retrieval (General)	Audiovisual	Unspecified	Unspecified	Unspecified	Unspecified	Unspecified	
National Initiative	QUAERO	N/A	Retrieval (General)	Audiovisual	Unspecified	Unspecified	Unspecified	Consumers	End User (Vague)	Search multimedia, including broadcast media, on name, context and metadata annotations.
National Initiative	QUAERO	N/A	Retrieval (General)	Audiovisual	GUI Development	Unspecified	Unspecified	Consumers	End User (Vague)	Enhance user experience of multimedia search services by developing more convenient interfaces.
National Initiative	QUAERO	N/A	Analytics (Multimedia)	Audiovisual	Segmentation	Multimedia Segments	Unspecified	Tourism/Heritage	Content Provider	Digital Heritage: Improve annotation and encoding by a combination of automatic and manual means.
National Initiative	QUAERO	N/A	Analytics (Multimedia)	Audiovisual	Feature Detection	Indexed Content (Audiovisual)	Unspecified	Tourism/Heritage	Content Provider	Digital Heritage: Improve annotation and encoding by a combination of automatic and manual means.

National Initiative	QUAERO	N/A	Analytics (Multimedia)	Audiovisual	Speech Recognition	Indexed Content (Audio)	Unspecified	Tourism/Heritage	Content Provider	Digital Heritage: Improve annotation and encoding by a combination of automatic and manual means.
National Initiative	QUAERO	N/A	Retrieval (General)	Audiovisual	Vague	Digital Asset Manager	Unspecified	Broadcasting	Content Provider	Digital asset management: Multimedia Ingestion, (post) production, aggregation, storage, search and re-purposing.
National Initiative	QUAERO	N/A	Classification Systems (General)	Audiovisual	Semantic Classification (General)	Taxonomy Generator	Unspecified	Businesses	Content Provider	Platform for text/image annotation: Translation, classification and structuring of information into knowledge (including thesaurus construction and named entity recognition).
National Initiative	QUAERO	N/A	Retrieval (Search)	Audiovisual	Query by Example (Fingerprint)	Retrieval System (General)	Unspecified	Broadcasting	Content Provider	Video and audio fingerprinting.
National Initiative	THESEUS	N/A	Extraction/Indexing	Audiovisual	Video segmentation	Multimedia Segments	Unspecified	Unspecified	Content Provider	Video and image recognition.
National Initiative	THESEUS	N/A	Extraction/Indexing	Audiovisual	Feature extraction	Indexed Content (Audiovisual)	Unspecified	Unspecified	Content Provider	
National Initiative	THESEUS	N/A	Retrieval (Search)	Audiovisual	Query by Example	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Semantic navigation and interaction.
National Initiative	THESEUS	N/A	Retrieval (Search)	Controlled Metadata	Query by Controlled Metadata	Retrieval System (General)	Unspecified	Unspecified	End User (Vague)	Semantic navigation and interaction.
National Initiative	THESEUS	N/A	Classification Systems (General)	Controlled Metadata	Semantic Classification (General)	Controlled Vocabulary Development	Unspecified	Unspecified	Content Provider	Ontology design, mapping, management, evolution; reasoning.
National Initiative	THESEUS	N/A	Extraction/Indexing	Audiovisual	Machine Learning		Unspecified	Unspecified	Content Provider	Statistical machine learning.
National Initiative	THESEUS	N/A	Classification Systems (General)	Controlled Metadata	Semantic Classification (General)	Standards Development	Unspecified	Unspecified	Content Provider	Metadata standards and standardization.
National Initiative	THESEUS	N/A	Personalization	Audiovisual	Unspecified	Unspecified	Unspecified	Unspecified	Content Provider	User adaptation and personalization
National Initiative	THESEUS	N/A	Classification Systems (General)	Controlled Metadata	Semantic Classification (General)	Ontology/Taxonomy Manager/Editor	Unspecified	Unspecified	Content Provider	Visual ontology editing framework
National Initiative	THESEUS	N/A	Extraction/Indexing	N/A	Semantic Classification (General)	Indexing Manager/Editor (General)	Unspecified	Unspecified	Content Provider	Visualization techniques for semantic annotation

National Initiative	THESEUS	N/A	Retrieval (General)	N/A	GUI Development	Vague	Unspecified	Unspecified	End User (Vague)	Modular GUI framework for the visualization of semantic information in Web.
National Initiative	THESEUS	N/A	Retrieval (General)	N/A	GUI Development	GUI	Unspecified	Unspecified	Content Provider	Visual ontology editing framework.
National Initiative	THESEUS	N/A	Extraction/Indexing	N/A	GUI Development	GUI	Unspecified	Unspecified	Content Provider	Visualization techniques for semantic annotation.
National Initiative	THESEUS	N/A	Vague	Unspecified	Vague	Vague	Unspecified	Unspecified	Unspecified	PROCESSUS: Integration of semantically enriched process-chains in and across industrial companies.
National Initiative	THESEUS	N/A	Vague	Unspecified	Semantic Classification (General)	Vague	Unspecified	Unspecified	End User (Vague)	ALEXANDRIA: Publishing platform for user generated content; combines semantics and community recommendation.
National Initiative	THESEUS	N/A	Unspecified	Image	Vague	Unspecified	Unspecified	Unspecified	Content Provider	MEDICO: Scalable semantical analysis of diagnostic images in medicine.
National Initiative	THESEUS	N/A	Unspecified	Audiovisual	Vague	Unspecified	Unspecified	Unspecified	Content Provider	CONTENTUS: Process chain for providing semantic access to AV-archives as a part of safeguarding the national cultural heritage.

Class	Project	UC	Primary Goal	Secondary Goal
Project	DIVAS	1	Segment location in audiovisual database of user relevant information.	
Project	DIVAS	2	Deliver new audiovisual information with respect to the interests and needs of a user in almost real time.	
Project	DIVAS	3	Content description and machine-assisted indexing of audiovisual corpora.	
Project	DIVAS	4	Locate audiovisual content (particularly original source video) corresponding to textual or visual material.	
Project	DIVAS	5	Identify whether a video sequence has been correctly viewed by users.	
Project	DIVAS	6	Retrieve audiovisual content using an extract of it.	Rapidly browse an extended database to identify and locate the originating audiovisual document.
Project	DIVAS	7	Search for audiovisual duplicates.	
Project	DIVAS	8	Efficiently search and tag unknown files within private music archives and within compressed domains.	Automated delivery of updates of new releases of songs or video clips.
Project	RUSHES	1	Find content by searching with keywords and semantic concepts.	
Project	RUSHES	2	Find content by visual search (similarity search)	
Project	RUSHES	3	Find media of a certain location, date and time.	
Project	SAPIR	1	Search for photos in social network's photo collections.	
Project	SAPIR	2	Search for receipts by multimedia queries on a social network.	
Project	SAPIR	3	Discover and purchase new music on basis of a recorded audio excerpt.	
Project	SAPIR	4	Retrieve information in situ on buildings or objects depicted on the tourist's photos.	
Project	SAPIR	5	Extract information related to films that occur on TV, and search for other films or video clips based on extracted features.	
Project	SAPIR	6	Utilize other people's closeness to particular events to retrieve a photo.	
Project	TRIPOD	1	Augment existing photo captions.	
Project	TRIPOD	2	Create rich photo captions using geo-metadata.	

Project	TRIPOD	3	Recommend postcards to send using a user's preferences.	
Project	VICTORY	1	Search for the design of automotive parts.	
Project	VICTORY	2	Vague	
Project	VICTORY	3	Vague	
Project	VICTORY	4	Share customized figures among gamers.	
Project	VICTORY	5	Provide access to large repositories of maintenance, installation and support documents.	
Project	VICTORY	6	Provide information about objects of interest to travelers.	
Project	VITALAS	1,1	Personalize users access to content.	
Project	VITALAS	1,2	Retrieve pictures that are relevant to a user's interests.	
Project	VITALAS	2,1	Automatically label visual concepts.	
Project	VITALAS	2,2	Provide a graphical interactive and efficient overview of a collection of pictures (~2000 items).	
Project	VITALAS	2,3	Provide an interactive navigation based on proximity criteria with user feedback.	
Project	VITALAS	2,4	Retrieve pictures from textual query using conceptual links.	
Project	VITALAS	2,5	Retrieve pictures of a well-known person.	
Project	VITALAS	3,1	Provide a cross modal similarity search to find videos extracts.	
Project	VITALAS	3,2	Establish audio and visual classification.	
Project	VITALAS	3,3	Navigation of video content by concepts.	
Project	VITALAS	4,1	Create an interactive structured map of a video program to provide a general structured overview.	
Project	VITALAS	4,2	Find content about a subject that is similar to a news program.	
Project	VITALAS	4,3	Localisation of audio or graphic temporal "markers" associated with a program.	
National Initiative	iAD	N/A	Research next generation precision, analytics and scale in information access.	Build international networks to identify and execute on global disruption opportunities.
National Initiative	IM2	N/A	Understand, present, and retrieve information from multimodal recordings of face-to-face meetings or lectures	Development of new meeting browsers.
National Initiative	MultimediaN	N/A	Intuitive and easy multimedia retrieval (search and browse) using semantic concepts.	Enable P2P video indexing and metadata exchange.
National Initiative	MundoAV	N/A	Provide professionals, companies and institutions a complete, comprehensive, timely and authoritative referent of the audiovisual sector.	
National Initiative	QUAERO	N/A		
National Initiative	THESEUS	N/A		