CHORUS AUDIO-VISUAL SEARCH
IST - 045480

Second report
Identification of multi-disciplinary key issues for gaps analysis toward EU multimedia search engines roadmap

(D2.2)

Deliverable Type *: PU
Nature of Deliverable **: R
Version: Draft
Created: November 28 2008
Contributing Workpackages: All
Editor: Nozha Boujemaa
Contributors/Author(s): Rolf Bardeli, Nozha Boujemaa, Ramón Compañó, Christoph Doch, Joost Geurts, Henri Gouraud, Alexis Joly, Jussi Karlgren, Paul King, Joachim Koehler, Yiannis Kompatsiaris, Jean-Yves Le Moine, Robert Ortgies, Jean-Charles Point, Boris Rotenberg, Åsa Rudström, Oliver Schreer, Nicu Sebe, Cees Snoek.

** **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: After addressing the state-of-the-art during the first year of Chorus and establishing the existing landscape in multimedia search engines, we have identified and analyzed gaps within European research effort during our second year. In this period we focused on three directions, notably technological issues, user-centred issues and use-cases and socio-economic and legal aspects. These were assessed by two central studies: firstly, a concerted vision of functional breakdown of generic multimedia search engine, and secondly, a representative use-cases descriptions with the related discussion on requirement for technological challenges. Both studies have been carried out in cooperation and consultation with the community at large through EC concertation meetings (multimedia search engines cluster), several meetings with our Think-Tank, presentations in international conferences, and surveys addressed to EU projects coordinators as well as National initiatives coordinators. Based on the obtained feedback we identified two types of gaps, namely core technological gaps that involve research challenges, and "enablers", which are not necessarily technical research challenges, but have impact on innovation progress. *New socio-economic trends are presented as well as emerging legal challenges.*

Keyword List: multimedia, search, research, gap analysis, functional breakdown, use case typology, socio-economic aspects, legal aspects

The **CHORUS Project Consortium** groups the following Organizations:

1	JCP-Consult	JCP	F
2	Institut National de Recherche en Informatique et Automatique	INRIA	F
3	Institut fûr Rundfunktechnik GmbH	IRT GmbH	D
4	Swedish Institute of Computer Science AB	SICS	SE

Information Society Technologies

5	Joint Research Centre	JRC	B
6	Universiteit van Amsterdam	UVA	NL
7	Centre for Research and Technology - Hellas	CERTH	GR
8	Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e. V.	FHG/IAIS	D
9	Thomson R&D France	THO	F
10	France Telecom	FT	F
11	Circom Regional	CR	B
12	Exalead S. A.	Exalead	F
13	Fast Search & Transfer ASA	FAST	NO
14	Philips Electronics Nederland B.V.	PHILIPS	N

Contents

1 INTRODUCTION

After addressing the state-of-the-art during the first year of Chorus and establishing the existing landscape in multimedia search engines, we have addressed the gap analysis during our second year. We have focused our effort on three main directions that have represented our second year three working groups:

- WG1: Technological issues
- WG2: User-centred issues and use-cases
- WG3: socio-economic and legal aspects

When considering the procedure to establishing this gap, we decided to achieve in parallel two central studies:

- a concerted vision of functional breakdown of a generic multimedia search engine
- representative use-case descriptions with related discussion on the requirements for technological challenges

The achievement of these two studies represents the starting point of our gap analysis. The process for fulfilling these two studies was central and fully concerted with our community at large, gathering feedback from EC concertation meetings (multimedia search engines cluster), several meetings with our Think-Tank, presentations in international conferences, and questionnaires addressed to EU projects coordinators as well as National initiatives coordinators.

This process was iterative to enhance progressively and to adjust the outcome to the major players (industries and academia) in the field of search engines. The outcome of these two studies is presented in section (2) and section (3), respectively.

When addressing technological gaps, we have identified very quickly that they are two fold:

- ***core technological*** gaps that involve research challenges
- technological issues that do not necessarily consist on technical research challenges but have impact on innovation progress the that we call "***Enablers***"

These two folds are described in sections 4 and 5, respectively.

Hence in section 4, we have addressed the research challenges that attempt to cover thematically the functional breakdown described in section 2.

On the other hand, the section 5 "Enablers" represents the gaps that are crucial and present a major impact to achieve advances in multimedia search engines. They are some times operational gaps such as "corpora development", etc.

Section 6 points out the emerging trends and challenges regarding the socio-economic and legal aspects. It provides insights on critical issues that need to be addressed and could represent gaps toward the wide deployment of search technologies from the socio-economic and legal view point. This study has been consolidated trough "Sevilla" workshop gathering major players in the non-technical side of search engines.

Section 7 is an annex to the document describing:

- The functional landscape of the EU projects (mapped to our functional breakdown provided in section 2). This provides an overview of the European efforts regarding the activities into each functional box of the generic SE – section 7.1 and 7.2
- A use-case typology through a mind map view. It provides a survey explaining the typology dimensions as well as use-case survey – section 7.3, 7.4 and 7.5
- A survey on market segment – section 7.6
- A glossary defining the terminology used in this document – section 7.7

2 FUNCTIONAL DESCRIPTION OF A GENERIC MULTIMEDIA SEARCH ENGINE

2.1 Purpose

The functional analysis described in this section aims at identifying the major sub-functions comprising a search engine in a fashion as media-independent as possible. It is hoped that this breakdown will clarify some of the technical aspects of this domain, will foster shared vocabulary and understanding of the various functions and facilitate analysis and evaluation of potential research projects in the domain of multimedia search.

It is important to note that what is being presented here is not an architectural diagram of the implementation of a search engine. The various boxes in the diagram represent "functions" that need to be implemented somehow. How and where these functions are implemented belongs to the architectural description for a specific project. There is not necessarily a one to one mapping between the functional description and implementation architecture.

Volumes of digitally encoded documents push towards a two pass solution (index/query)

Search can be achieved in two fashions:

- A *one pass* process amounting to exhaustive examination of all content, looking for the searched feature
- A *two pass* process where a set of features are detected in all documents in a first global pass, and the feature list is examined in a second pass at the request of a user

The first approach assumes that the features sought by the user can be described and identified by the machine, and that there exists a technical process by which such features can be detected in each of the documents which are part of the repository.

There is a long history of research and technical algorithms in this domain ranging from the simple string matching routine of C libraries used in grep to sophisticated object recognition capable of detecting targets within a video stream or spoken keywords within an audio stream.

The volume and variety of data available in digital form has made this one pass solution impracticable for most cases unless one is willing to restrict dramatically one or more of the dimensioning parameters (number of documents to search into, number of potential queries, response time).

The second solution is capable of addressing a much wider spectrum of use-cases, but at the cost of restrictions and uncertainties on the variety of potential queries. Because feature extraction is performed during the first pass one cannot anticipate for all possible features what a user may wish to search for during pass two.

The successes of AltaVista, the first large volume text search engine on the Internet and that of its successor Google have shown that providing adequate or even sub-adequate answers in a very timely fashion (<0,2 sec) was of great value to the users of these systems. The lessons learned in the text domain are likely to be of value although it is clear that each media type is likely to come with its specific problems.

2.2 Multimedia search is implemented through metadata search

A digitally encoded picture is a collection of bits organized into an array of pixels. It may or may not be encoded in a form that achieves lossy or loss-less compression. On this array of bits, algorithmic processes can be run that will extract information such as "color histogram", "texture characteristics" or even "presence of a particular shape". It can also be the case that the author, or later users provide "tags" and "annotations" to be associated with the image. The same is true of audio-speech streams, in which specific processes can extract "spoken words" and "speaker characteristics" (for later differentiation). In audio-music streams, one can detect "music genre, "instrument type" "rhythms", "musical structure", etc. In video streams, beyond analysis of each individual image, one can detect "scene changes", "camera movements", etc. As one can infer, each media type brings with it some specific characteristics that can be extracted, and the lists above are by no means exhaustive and will expand as research progresses. For each of the types of documents above, it is also the case that other information can be known independently of the internal data such as "creation date", "document name and type", "author", "creation process", etc...

By tradition, the "content" of a document (the pixel array, the sample stream, the video stream) is called "the essence" of the document, and information of the latter kind (creation date, name, author, etc.) is called "metadata". By extension, in

the remainder of this document, we will call "metadata" any information that can be directly associated or extracted from the essence. Because of this generalization, it becomes clear that all documents cannot be associated with all types of metadata. Note that the metadata extraction process defined above is also often called "content enrichment" by practitioners of this industrial and research domain.

The reason why we wish to stress the importance of metadata is because of the conjecture that all multimedia searches are implemented through search and comparison of metadata elements.

Beyond the obvious case of searching for an image through a search within the tags and annotations they may each carry, the case of "search by example" can be shown to adhere to this principle. When provided with an image as a "good example" of what the user is searching for, the first thing a search system is likely to do is to process that image and extract from it all sorts of characteristics it knows how to extract. Once this is done, the search engine will compare those characteristics with the same characteristics extracted from the set of images within which the search is performed. For each characteristic, identical or close numerical values provide a clue towards matching images, and the more matching characteristics, the higher overall matching probability.

The process above is very generic, and applies to all sorts of media types, each with its specific set of metadata. Because of this genericity, the remainder of the document calls "Document-metadata" and "Query-metadata" the metadata associated or extracted from the document essence during the first pass (D-metadata) or during the second pass (Q-metadata) described in the first paragraphs of this section.

The analysis above leads to the following overall diagram for a search engine:

2.2.1 Functional breakdown diagram

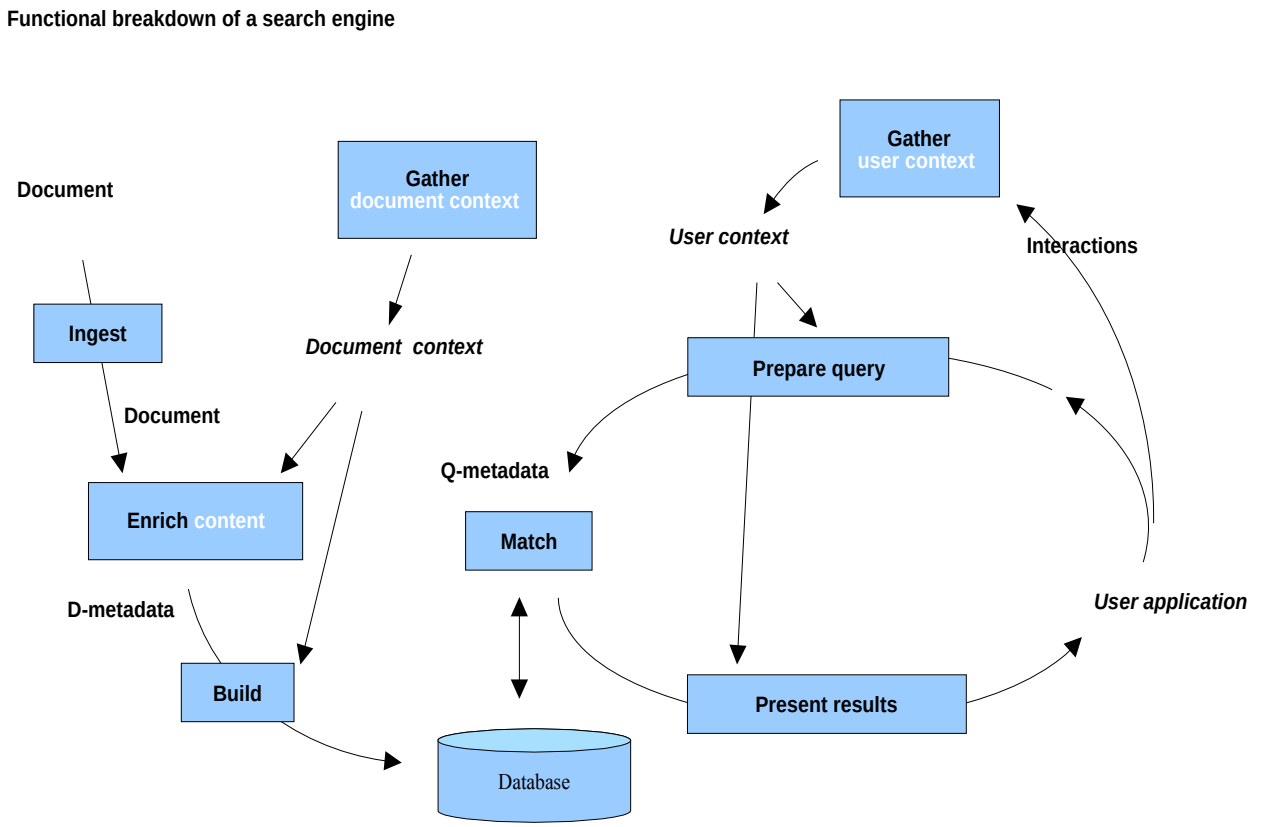


Figure 2.1Functional breakdown of a search engine

This diagram distinguishes pass one (document related) activities on the left part of the diagram from pass two activities (query related) on the right part.
In the middle of the diagram sits the core of the search engine with the database holding the index and the "match" process in charge of matching Q-metadata against D-meta data stored in the index. The remainder of this section discusses in more details the functions performed by each of the boxes of the diagram above.

2.3 The "Ingest" task ranges from trivial for static repositories to very complex for the Internet

Ingest is the task by which documents are discovered and "ingested" into the search engine. This task ranges from trivial when the repository is static (an heritage image database) to very complex when there is potentially no limit to the repository as is the case for Internet wide search engines. Activities associated to this task are not the subject of active research and belong more to software engineering or use-case analysis. The potentially difficult part of the ingest step is the one in charge of maintaining the search engine database up to date in relationship with the evolving document repository (detecting new or content). For this activity, the relationship between the repository and the search engine can be organized in two modes:

- Push mode, in which the content creation application is aware of the search engine and alerts it of any new or updated content
- Pull mode, in which it is the sole responsibility of the search engine to explore the repository and discover new or updated content.

In is clear that the push mode offers better guarantees as to the "freshness" of the database and simplifies greatly the discovery task for the search engine. On the other hand, it relies on the awareness of the content creation applications which is unlikely to happen, especially for applications developed before search engines became popular. Note that this alert task may be also performed by the file system of the OS on which sits the content creation application which may send out alerts whenever a file is created or updated. By monitoring this alert, the search engine can detect content creation or updates more efficiently than by scanning itself the entire file system.

2.4 The "Build" task addresses global issues (ranking), scale, and incremental issues

The "Build" task takes as parameter the D-metadata extracted by the previous step and inserts it into the main search engine database. This seemingly simple step can become substantially difficult when one considers the potential volume of data handled.

The build step performs two main functions:

- It computes any metadata that will be associated with the document, but which relates to the relationship between the document and other documents in the database. Ranking computation is of this nature. Just as the "metadata extraction" step aims not only at computing metadata used for the matching step, but also metadata user for the "present results" steps, the "build" step aims also at contributing to the "present results" step by computing ahead of times global information.
- It inserts all metadata into the database

2.5 "Matching" depends heavily on the collected metadata

As we said earlier, the matching process matches the Q-metadata (and potential supplementary parameters) with the D-metadata stored in the search engine index database. The matching process can take many forms, depending on the specific metadata, and the intent of the user. Individual metadata matching can be exact or approximate. Multiple metadata matching can form a Boolean equation, each term being potentially weighted according to the wishes/preferences of the user. Lets' take two different examples in the following:

For multimedia similarity search, standard metrics (such as L1, L2, etc.) are most often used to filter out a subset of the database. In the context of general purposes search, there is no a-priori knowledge and in that case generic content descriptors are computed which consists on global visual appearance signatures. This mechanism provides approximate similarity search. On the other hand when considering the context of CBCD, it involves specific content description as well specific matching functions. CBCD stands for content-based copy-detection which is useful for automatic detection of illegal video copies for example and hence allow automatic assistance to DRM problems. In this case, precise local descriptors are computed and specific voting functions represent the matching box. They are designed to avoid false alarms.

2.6 "Document context" describes collective metadata

Both the "Enrich Context" and the "Build" functions perform their task by processing a specific document (its essence and its existing metadata), but also by accessing the context in which this document exists. By context, we mean here all that information which relates to a collection of documents rather than a single document. Where a document resides?, Which

collection is it part of?, What is its type or sub-type? (e.g., radio news broadcast, telephone conversation) are information which may be associated with multiple documents. This information may impact significantly the content extraction step through access to dictionaries, ontologies, language models, etc. It is the role of the "document context gathering" function to create and manage these context elements which may be pre-existing (an ontology, a language model), or may be constructed while documents are being ingested (dictionaries, n-tuples, etc.).

2.7 "Enrich content" or metadata production is needed both for the search itself, but also to support user interaction

Content enrichment box represents all about generating new metadata. Metadata used to be manually generated by archivist for the incoming content into the repository. These metadata are human made annotation that are related to a given context, period of time, etc. and hence involve part of the annotator subjectivity. As the increasing of the incoming digital content became far more important than the manual annotations capabilities, in addition to the fact that all multimedia content is not fully annotated or well annotated, the need for automatic metadata generation process was very crucial in the recent years.
We can distinguish several ways to enrich content:

- Computing visual appearance signatures that are most of the time low-level and "signal"-based descriptors. In this regard there are several approaches: global descriptors (for example: including color, shape and texture information for visual content)
- Mono-media content class recognition allowing to generate an overall textual label to a group of multimedia content (e.g., image class recognition: landscape, indoor, outdoor, etc.)
- Mono-media "object" recognition which implies to be able to annotate a subpart of multimedia document

At the end, we have two types of automatic content annotation: on one hand, low-level descriptors which are semantic-less metadata that provide complementary information related to the physical content (sound, color, motion, etc.), on the other hand, mid and high-level textual labels that allow to provide newly added semantic information based most of the time on statistical learning mechanism.
These new approaches for automatic content enrichment make the multimedia content more searchable, enhance the performance (effectiveness) of a search engine from the precision and recall perspectives. In other words, this allows revealing the implicit knowledge and making it rather explicit knowledge. They allow what we call post-annotation of a multimedia document. This have rather important impact as the relevant information that an archivist can annotate at the income of a given content can change over the time. As the manual re-annotation is no longer achievable with the huge amount of digital data collected daily, the automatic content enrichment become an essential functional component for a search engine.

2.8 "Query preparation"

As we said above, the basic shortcoming of the two pass approach is that pass one cannot anticipate for all possible queries that are likely to be presented during pass two.

Users are therefore placed in a situation where they must first discover or guess which of the available metadata (or set of) are available, and how the metadata can be used to serve their ultimate goal. Let’s take the query-by-example discussion above, and assume (a definite restriction for the sake of the argument) that the metadata extracted from the images are limited to "color" and "texture". Users that would choose as an example the photograph of a Ferrari hoping that the search system will find other Ferrari photographs are likely to be disappointed by the results. They may be surprised to find that most of the returned results do not show a Ferrari, but are mostly reddish!

To address this issue, the query interface of the search engine must make visible and meaningful which metadata it can process while staying simple and pleasant. At the same time, the first pass process is "encouraged" to gather as much metadata as possible, focusing on the metadata that is most generic, most discriminatory, and most meaningful to the user.

The query preparation step therefore has two components:

- A Graphical User Interface (GUI) component which implements the physical interface between the system and the user
- A D-metadata extraction component which will extract metadata from the elements supplied by the user.

Both components have access to the user context and can either use it to facilitate the metadata extraction, to "augment" the metadata with contextual elements, or to make the context visible to users so that they may adjust it to the specific query.

2.9 "Results presentation"

The second problem faced by search engines is linked to the potential large amount of results it might return after a query (because of the large initial document repository, because of the broad characteristic of the queried metadata, etc.). To solve this second problem, results returned to users must be organized/sorted/regrouped in a fashion that will be meaningful and helpful to them. By helpful, we mean here both helping them sort through the results and find the "good" one (for instance ranking the results by popularity), or helping them refine their query with additional or new terms that will result in fewer, hopefully more relevant results (clustering by type, by theme, by additional features, etc.). This sorting/grouping can only be done on the basis of part of the metadata collected during the first pass described above.

Again, this functional box is made of two separate components:

- A structuring process in charge of analyzing the D-metadata returned with the results
- A GUI responsible for the actual display and interaction with the user

Of course, it is necessary to look at the two GUI mentioned above in a unified fashion since the goal of the "result presentation" step is to facilitate the task of users in their "query preparation".

2.10 "User context" describes information related to multiple queries, multiple users

Knowledge of the user context creates potential for improving the overall efficiency of the query and ultimately of the user experience. Several scope of context can be taken into account:

- single user, multiple query context such as profiles and preferences (static), geo-localisation (semi static), and relevance feedback (dynamic)
- multiple user context such as recommendations and usage statistics

The obvious example of a query greatly enhanced through knowledge of the context of a user is the geo-localized query ("find restaurants near my current location").
Knowledge of the current location of the user is a fact which is independent from the current query, and is applicable to other queries. The same is true for user profile and preferences (user defined), session activity monitoring (accumulated by the user system), or usage logs activity (accumulated by the search engine system). Gathering such contextual elements, and making them available to the "query processing" step is the task of the "user context gathering" function.

Another example of user context is the context created by the results of the previous query. "Relevance feedback" provides the user with mechanisms facilitating adjustments their query based on the characteristics of the returned results (change the relative weight among several metadata, point at documents that best fit their expectations, choose among suggested parameter adjustments, etc.).

2.11 Discussion

2.11.1 Search Engines and explicit or implicit queries

=> Search Engines are a powerful alternative to databases (explicit queries)

Search engines are mostly known as stand alone applications focused on information access through interactive submission of queries. Born on the Internet (AltaVista/Google) search engines popularity led to their deployment in Corporate environments where they provided access both to unstructured information (documents in file systems, mails) but also to structured information stored into databases. Benefits drawn from the powerful and homogeneous access to information stored in a multiplicity of heterogeneous repositories has been the prime success factor. A good example of the benefits of this approach can be found in the notion of “virtual folders” which extends the traditional notion of folder (the physical, static repository where a collection of files is stored) to that of “the list of files matching a particular query” which is dynamic by nature.

=> Search Engines are now used through implicit queries derived from user interaction

The success of search engines as a means to access unstructured and structured information has extended beyond the stand alone search by query application. There are now interactive applications whose primary goal is not information search, but which require a built-in search functionality. A good example would be a TV SetTop box providing access to a large

collection of programs and films. The user interface of such a SetTop box should allow for easy navigation and search within the Electronic Programs Guide in a fashion compatible with the home and entertainment context in which it is being operated. Early developments and experiments along this idea have shown that search in this context becomes somewhat invisible to the user and is implemented through implicit queries derived from preferences, profiles, and current interaction.

=> Light of heavy reference to user context will span the whole spectrum between explicit and implicit queries

Explicit versus implicit queries is not a binary situation. As has been said earlier in this section, explicit queries can be “enhanced” with information gathered from the user context and thus gain some implicit characteristic.

The impact of this explicit/implicit aspect of queries on the core of the search engine itself is probably negligible. On the other hand, the “Prepare query” and 'Present results” functions are potentially fundamentally impacted. Ultimately, one could say that with a fully implicit query, those two functions are fully integrated into the interactive application under consideration.

The main reason to discuss here this issue (explicit/implicit) is because:

- It opens usage of Search technology to a whole new set of application domains, the most important being today the emerging Digital TV domain
- It puts on Search Engine solutions an engineering, and ultimately a standardization constraint to facilitate their integration into broader applications.

2.11.2 The proposed functional description facilitates distributed or centralized architecture analysis

The functional description above does not make any hypothesis about the centralized/distributed nature of its components. In fact, as described, it is already somewhat distributed across separate functional boxes, each carrying a specific function. Since it is the case that some boxes are accessing data managed by others, the overall distribution must be analyzed with great care, taking into account data access times or possible replications.

Beyond this first level of distribution, it is interesting to analyze whether each box itself can be distributed across multiple machines, possibly across multiple sites.
In a distribution analysis, it is important to distinguish the two levels (machine, site) because of the networking delays involved.

Elsewhere in this document, a full section addresses the issue of the relationship between Peer to Peer and Search. This related deeply to the distribution issue discussed here:

- While some functions can obviously be widely distributed ("Ingest", "content enrichment", "query preparation", "results presentation"), others are centralized by nature ("build/rank") or can only be distributed at the cost of replication ("database").
- Distribution implies network delays. Any such delays involved in the query interaction loop are impacting negatively the overall interaction performance. Solutions to counter this negative aspect have to be proposed.
- Distribution implies resource unavailability. The overall distributed architecture must cater for the fact that nodes may become unreachable. Appropriate redundancy must be proposed.
- Distribution may imply replication, incrementally updating a replicated database

2.12 Conclusion

The analysis above points to the paramount importance of the automatic "content enrichment" or metadata creation step, both because it gathers data/information which will make search feasible, but also it gathers data/information that will help organize results into a manageable and meaningful form.

Of course, the importance of this metadata argument implies both that existing metadata must be preserved through the numerous steps involved during the creation process of documents and content, but also that metadata must be created automatically from the content itself to cater both for the sheer volume of potential documents and for "old" content which is likely to be "metadata poor".

2.12.1 Performance issues drive applicability to use-cases

Performance issues can be found in relationship with all the components discussed above.

It is probably the case that they may be sorted with decreasing importance, but this sorting is likely to vary from one use-case to another.

The two primary driving factors relate each to the two passes described above:

- Pass one must be performed fast enough in relationship with the amount of documents present or appearing in the repository
- Pass two must be fast enough to allow for the trial and error approach induced by the two pass approach.

Beyond these two issues, other performance criteria will play a role in the ultimate user satisfaction:

- *Precision/Recall*: the traditional performance measure of an information retrieval system
- *Relevance*: a measure of the efficiency of the system at finding what is sought
- etc.

The overall analysis above applies independently of the media type of the documents. What will differ from media type to another is the specific technology used to extract metadata (and possibly to search through metadata). Not only the techniques will differ, but also the performance.

Because of the large variation in performance for the metadata creation step, it is expected that use-cases will vary considerably depending on the media type, the volume, and update rate of the repository. It is therefore pointless to describe a specific performance point as a target for a specific technology. On the other hand, measuring the raw performance of a technique (speech to text, face detection, signature computation, etc.) is important both to compare it with other techniques for the same task, and to measure progress achieved over time.

2.12.2 Beyond the functional breakdown, other transversal technical issues must be taken into account

The functional breakdown described above does not address some issues which cannot be described as "functions" but are nonetheless part of the user expectations. Performance is of this kind, and has been addressed above. Most of the issues listed here are quite real, but are not topics for research, but rather for engineering and industrial development. It may nonetheless be the case that some research project, for effective testing requirements, wish to build scale one test-bed, in which case it will have to address problems of this nature.

Other issues of a similar nature are:

- *Scalability*: what is the growth potential of the solution, both in terms of document repository size/update rate and query rate performance is often the limiting factor for security, but that issue is often addressed through architectural and software engineering considerations.
- *Access rights, Intellectual property*: how is document access security handles? Are documents access rights maintained across the search engine? It should be noted that this issue must be taken into account at the early stages of a design as "security as an afterthought" fails almost always!
- *Privacy*: how is the activity of the user kept private? Public search engines accumulate hoards of data about their user's activity. Use of user context to "enhance" queries dramatically increases the privacy issue. Can/should this issue be addressed through technology, regulation, or law?
- *Expandability*: as new document types appear, how easy is it to plug into the system the appropriate modules capable to process such documents?
- *Integrability*: a search engine is not necessarily a stand alone application. It can be part of a production system or of a user application (TV Electronic Guide for instance). This issue addresses the ease by which such integration can be performed. It relates both to architectural issues and to API definitions and their qualities (generality, efficiency, stability - ultimately standardized)
- *Metadata access and ownership*: (covered in a separate section)

2.13A request to the reader

The section above tried to describe multimedia search engine in a media independent fashion. It is believed that most (if not all) current projects currently underway in the domain of multimedia search engines fit in this functional framework.

Readers can contribute to a better analysis and understanding of the problem space by identifying use-cases or projects which do not fit in this architecture, and pointing out where and why. This might reveal novel approaches worth investigating further. In this effort, one should look for large and significant "misfits" rather than detail level deviations that can easily accommodated by the model. One should also remember that the functional breakdown proposed here is not an architecture, and that for a given project, a particular function could be split across several of the modules describing it.

Readers can also identify specific technologies of importance to the search space, and try to "position" them in the diagram, hopefully in s single box, possibly in multiple boxes. For any technology that would not fit, it would be important to understand whether the diagram needs to be altered in a fundamental way or whether it is sufficient to add some new function, or some complexity such as links between the existing functions.

Given a use-case or project not adequately covered by this functional diagram, one should then try to propose a new diagram that would work for the new project while maintaining "compatibility" with the current approach.

3 REPRESENTATIVE USE CASE DESCRIPTIONS AND REQUIREMENTS TO TECHNOLOGICAL CHALLENGES

3.1 Overview

Many search technologies today can be regarded as commodity components due to the fact that they are not designed or evaluated with respect to specific or unique user needs. This means that new services are likely to be built on top of them to address these needs. It has been recommended that target notions with parameters be carefully designed to model the possible use cases for new services (Järvelin and Kekäläinen, 2000). Such a specification could facilitate the quantitative evaluation of a new service. CHORUS has attempted to identify and develop these notions and parameters for new services in multimedia retrieval through the Use Case Typology

Use cases are general, high-level descriptions, or narratives, of how and why an actor interacts with a system. They attempt to find out “Who does what?” and “For what purpose?” (Jacobson et al., 1992; Cockburn, 2002). They are typically used to capture the functional requirements of a system.

Think Tanks were used as a forum by CHORUS to attempt an enumeration of all major use cases for new services in the multimedia search domain. These were analyzed for recurring themes, or attributes. Each attribute was then assigned a simple keyword phrase and placed as a node into a hierarchical typology. A survey was then generated from the typology. This survey allows CHORUS to systemically, thoroughly and automatically construct consistent sets of use cases by polling research projects and initiatives with the survey.

CHORUS has several purposes for collecting use cases:

- determination of functional requirements
- determination of evaluation criteria
- a broad gap analysis of European research topics

In line with the classic reason for collecting use cases, CHORUS could assist project administration by ensuring that resource planning and predicted technical requirements made by project leaders are realistic for the use case generated from the project's survey results.

Secondly, as discussed in the first paragraph, there is a need for better benchmarking and validation efforts for new services in multimedia retrieval. Given the number of constituent components in a search engine and a lack of any standardized architecture, confounding factors abound which complicate efforts to achieve consistent performance measures across research efforts. One way to reduce the complexity of the benchmarking and validation task is to establish a standard set of criteria for each use case defined by CHORUS. A project's use case, as it is determined by their survey results, could then be utilized to determine which criteria are most relevant to their search application. Performance measures among search applications sharing the same use case could then be more easily compared.

Finally, the survey was used in the current effort to conduct a high-level gap analysis of European research efforts in order to discover topical areas which may not be receiving adequate attention as well as those areas that are well researched. Information about research gaps could be useful in funding allocation decisions for new projects as well as determining new research directions for the European research community at large. On the other hand, identifying popular topical areas might suggest areas where increased collaboration is needed, such as a call for standardization or new corpora development.

3.2 Methodology

Five tools were developed in an effort to conduct a systemic and empirical gap analysis of the current field of research in multimedia retrieval:

- a use case typology oriented towards search as a service (*Use Case Typology*)
- a survey based on the previously mentioned typology (*Use Case Survey*)
- a glossary of terms
- a typology for classifying projects into one of five generic use cases (*Market Segment Typology*)
- a survey based on the previously mentioned typology (*Market Segment Survey*)

The first three tools were utilized for data collection, but the final two were shelved after discussions at the 4th Think Tank seemed to reach a consensus that there was little value in classifying research projects into generic use case types.

The *Use Case Typology* was used to design a survey which was deployed on September 19th, 2008 as a web survey. A link to the survey was distributed in a mass email sent by the CHORUS project coordinator to all Projects and Initiatives (national and international) within the purview of CHORUS. The data that was collected was analyzed for gaps and overlaps. The following thirteen projects submitted survey data: VICTORY, VIDI-VIDEO, RUSHES, TRIPOD, DIVAS, PHAROS, VITALAS, MESH, AIM@SHAPE, SALERO, Quaero, AceMedia, and SAPIR.

3.2.1 Use Case Typology

The design of the typology was driven by one important fundamental assumption. Namely, search was conceived as a service. This means that we tried to capture the entire ecosystem of factors that comprise a search engine. These factors are reflected in the names of the six typology categories discussed below. CHORUS not only attempted to capture information about user interaction, but who users are likely to be (market segments), how the system works (technical attributes), and how a system might best be capitalized (revenue sources), to name a few.

As a result of this design decision, it was often difficult for projects to answer many of the questions in the survey due to their specialization. For example, SEA is very focused on P2P protocols and do not invest energy in the development of a full search system. Other times projects had multiple use cases and could answer each question multiple times with different answers depending on the use case they were considering. Projects which reported these difficulties were asked to refrain from filling out the survey.

Characteristics pertaining to the use cases of search engines have been arranged into six somewhat orthogonal categories. The result is a clean, standardized format for project description and reporting. Each category contains a set of attributes and each attribute contains a list of possible values. The hierarchy is illustrated as such:

- Category
    1. Attribute (survey questions)
        1. Value (survey answers)

The six fundamental categories in the *Use Case Typology* are:

- User Interaction
- System Features
- Repository Features
- Indexing Features
- User Demographics
- Socio-Economic Factors

As mentioned above, each category contains a set of related use case attributes about a research project. (When thought of in terms of a survey, these attributes are questions.) For example, the five attributes under Socio-Economic Factors are:

- Revenue Sources
- Trust Management
- Privacy
- Metadata Provenance
- Metadata Licensing

Categories, when designed properly, help to ensure data integrity. They also help to organize the survey questions into groups that facilitate comprehension. If attributes are questions, then values are the possible answers to that question. In summary, the branch for Revenue Sources under the Socio-Economic Factors category looks like this:

1. [Attribute] Revenue Sources
    1. [Value] *Fee*
    2. [Value] Advertisement
    3. [Value] *Licensing*
    4. [Value] *Sponsorship*
    5. [Value] *Subscription*
    6. [Value] *Cross-Selling*

7. [Value] *Not Applicable*
8. [Value] *Other?*

The typology can be reviewed in full (in either list view or a mindmap view) in the Annex at the end of this document.

3.2.2 Use Case Survey

The six fundamental categories in the *Use Case Typology* are used to organize the *Use Case Survey* into six major sections. Each section contains between 3 to 7 questions. To illustrate how typology attributes correspond to questions, the branch for Socio-Economic Factors is listed below, with attribute names in parenthesis at the end of each question:

- What are likely revenue source(s) for a business using your system? (Revenue Sources)
- What strategies are used to increase system transparency and trust so that potentially biasing factors affecting results can be identified? (Trust Management)
- Is personal information maintained by your system? (Privacy)
- Who owns most of the metadata used by your system? (Metadata Provenance)
- Under what conditions do you make available to third parties any metadata you produce? (Metadata Licensing)

The survey can be reviewed in full at the back of this document in the Annex.

3.2.3 Market Segmentation Typology & Survey

During the course of developing the *Use Case Typology*, it was recognized that projects could be classified into one of six generic use case types. These types can be understood to represent six fundamental market categories that exist for multimedia search engines today. The ability to classify projects into generic use cases should reveal which market segments are being served (and ignored) by current European research efforts. The market categories (generic use cases) are identified and defined as follows:

- **Internet Search (IS)** - *Identifying and enabling content across a public network (viz., internet) to be indexed, searched, and displayed to any user.*
- **Enterprise Search (ES)** - *Identifying and enabling specific content across a private network within an institution or company to be indexed, searched, and displayed to authorized users.*
- **Personal Archive (PA)** - *Enabling content to be indexed, searched, and displayed within a small collection of resources organized and maintained by a non-professional for personal purposes; identification of new content by recommendation adds value to the end user only.*
- **Library (LIB)** - *Enabling content to be indexed, searched, and displayed within a large collection of information resources and cultural artifacts organized and maintained by specialists (librarians) for a group of users who are generally granted unrestricted access for minimal or no cost.*
- **Personalized TV (PTV)** - *Enabling multimedia content to be indexed, searched, and displayed within a large collection of resources organized and maintained for a group of users whose access is mediated by commercial relationships; identification of new content by recommendation is commercially important for collection owner.*
- **Surveillance, Detection & Alert (SDA)** - *Identification of a person, object or system process that does not conform to a norm; provision of meta-information (profile or dossier) about people based on feature extraction from image, video or voice data; identification of associative social networks is important.*

Care was taken to normalize the definitions across all six market categories so that they could be analyzed for recurring attribute parameters (feature classes). Five parameters emerged which seemed crucial for differentiating each category. The five parameters are enumerated and defined below with their possible values listed beneath them.

- **Content Acquisition** - *The method of ingesting new content.*
  - Retrieved
  - Submitted
- **Repository Management** - *The level of organization of repository content.*
  - Unorganized
  - Semi-Organized
  - Organized
- **Repository Ownership** - *The applicable licensing model for repository content.*
  - Public
  - Private

- **Repository Access Rights** - *The amount of availability of repository contents to users of the search utility.*
    - Unrestricted
    - Restricted
- **Type of Retrieval** - *The method used for identifying, ranking and returning relevant information resources.*
    - General Recommendation
    - Repository Recommendations
    - Targeted Content
    - Social Networks

Projects can be classified as serving one of the market segments by evaluating what parameters they exhibit and matching these values to the following classification table.

GENERIC USE CASE	KEY ATTRIBUTES
IS (Internet Search)	Content Acquisition = *Retrieved* Repository Management = *Unorganized*
ES (Enterprise Search)	Repository Management = *Semi-Organized* Repository Access Rights = *Restricted*
PA (Personal Archive)	Repository Ownership = *Private Collection* Repository Access Rights = *Unrestricted* Associative Retrieval = *Repository Recommendations*
LIB (Library)	Repository Management = *Organized* Repository Ownership = *Public Content*
PTV (Personal TV)	Repository Management = *Organized* Associative Retrieval = *General Recommendations, Targeted Content*
SDA (Surveillance, Detection & Alert)	Associative Retrieval = *Social Networks*

Table 3.1Use Case Classification Scheme

The final task expended on the generic use cases was to transform them into a simple survey of five questions. Each question corresponds to one of the attribute parameters above. This survey was never deployed due to doubts about the utility of classifying projects. Nevertheless, the survey can be reviewed in the Annex

3.2.4 Glossary

A glossary was compiled to assist survey participants in understanding unfamiliar terms that were used. It should also be consulted by those reading this report so that comprehension of the concepts reported on by project respondents do not diverge from their understanding. The glossary can be reviewed in the Annex at the end of this document.

3.3 Results and Interpretation

Results will be discussed by the six categories outlined in the *Use Case Typology* discussion in the Tools section.

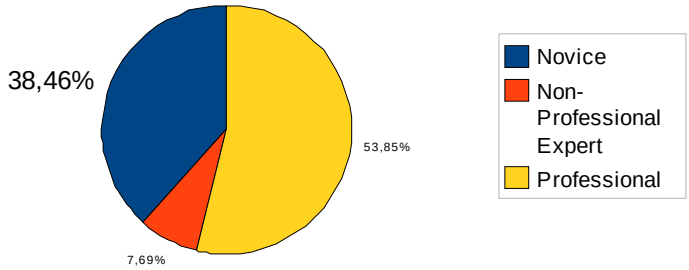
3.3.1 User Demographics

The following questions were asked in this section. Typology node labels are listed in parenthesis at the end of each question.

- *What competence level does your typical user have in regard to your system and knowledge domain? (System & Domain Competence)*
- *How big is your targeted community? (Community Size)*
- *What relationship(s) do your primary users have with the content you provide access to? (User Roles)*

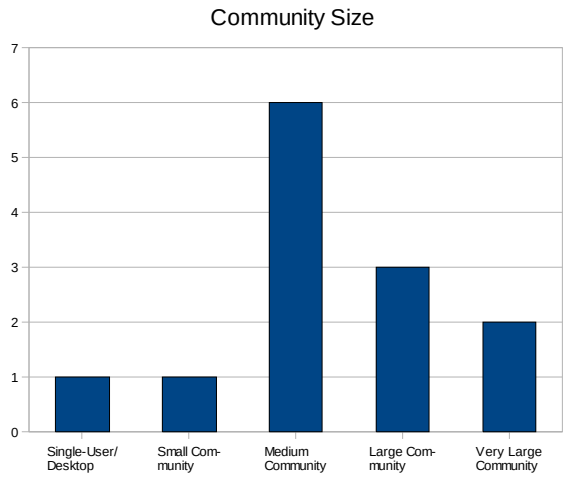
System & Domain Competence. Most research focuses on the professional user, with system design for novices receiving a fair amount of attention (see Fig.1). There seems to be a gap in system design research aimed at the intermediate user. The functional areas where this gap might have the biggest potential impact is in the research drive to improve systems for *query formulation* and *results presentation*. (See the section *Functional description of a generic multimedia search engine*.)

Figure 3.1: System & Domain Competence



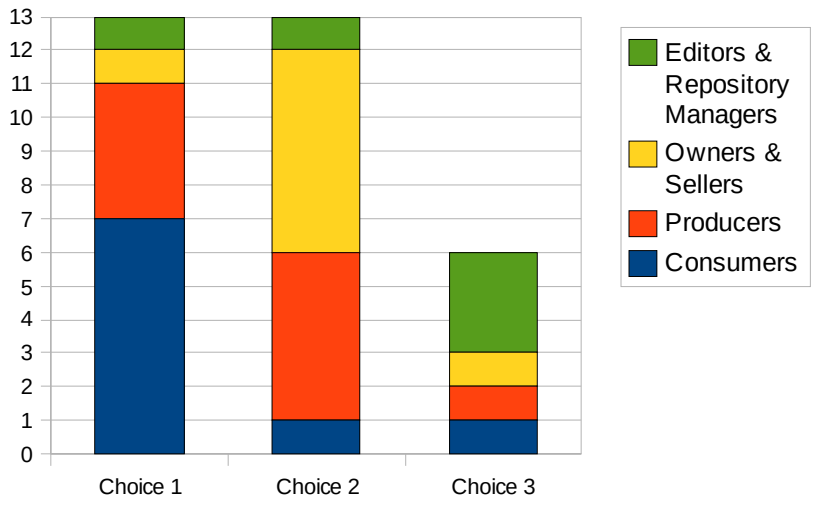
Community Size. Search applications aimed at the individual user/desktop and small community groups are only being investigated by one project each (see Fig. 2). Activity in the area of the desktop comes from P2P research in the form of client-side applications.

Figure 3.2 Community Size



User Roles. Among the four types of users, systems aimed at consumers and producers receive the most attention whereas owners and sellers are largely targeted as a second choice (see Fig. 3). Editors and repository managers, on the other hand, seem neglected. The negative impact of this oversight might be felt most saliently for the following functional components: *ingest, enrich content, gather document context,* and *build.* However, it is not clear whether the needs of editors and repository managers for these functional components are different in any important way from those of producers. So it is difficult to draw conclusions. Future versions of the survey should attempt a better disambiguation of the needs of these groups.

Figure 3.3 : User Roles



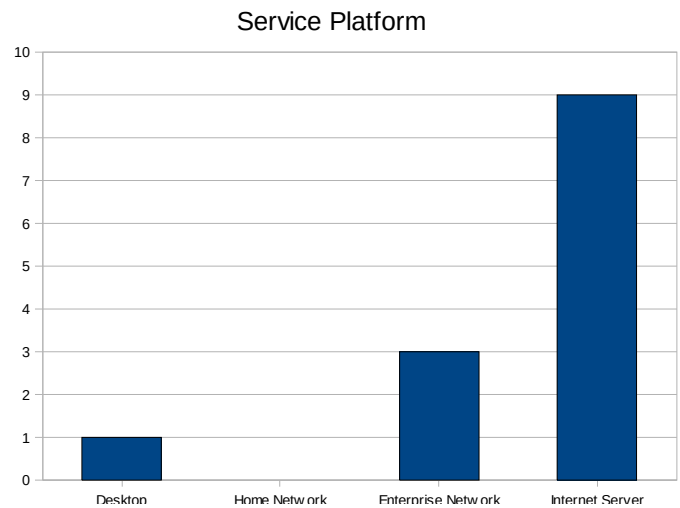
3.3.2 System Features

The following questions were asked in this section.

- *Where is your service designed to be hosted? (Service Platform)*
- *Which device(s) have you explicitly designed your search service for? (Device)*
- *How does your system primarily display search results? (Results Composition)*
- *How do you measure the quality of results your system provides? (Quality of Results)*
- *What kinds of contextual data does your system use? (Contextualization)*
- *What kind of semantic tools are being used by your service? (Semantic Technologies)*

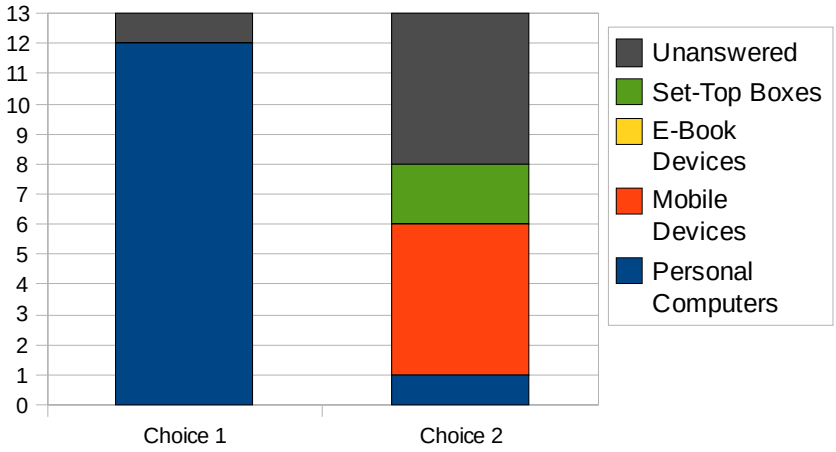
Service Platform. Responses reveal that a majority (69%) of research is focused on internet applications, with enterprise applications receiving 23% (see Fig.). Only two projects (AceMedia and SAPIR) report any research activity into desktop applications and none report anything for home network applications. These findings correspond with those of *Community Size*, which suggested that most research is focused on medium to large communities.

Figure 3.4 Service Platform



Device. The primary user device targeted by all research projects is the personal computer (see Fig 5). Secondary targets are mobile devices and set-top boxes. However, set-top boxes are targeted by only two projects, and only as secondary choices. No one reported on research efforts that targeted e-books. It is not clear whether this device has unique research requirements, but it should be followed by CHORUS.

Figure 3.5 Device



Results Composition. Research into user interface methodologies shows that eight of eleven projects report working on ranking methodologies (Single Ranked List) whereas just three report on efforts that are involved with faceted views (see Table 2). There seems to be no activity in clustering (Cluster Views) or the development of innovative visualization (Graphs) whatsoever. It is not possible to derive any conclusions, however. This question should be reformulated to ask about research in these areas, rather than who uses these methods in their systems. Furthermore, the values/answers should be reworked to reflect the work of Datta et. al. as discussed in the section 4.3 (Organization and Navigation). Namely, they should be the following: *relevance ordered (ranked)*, *clustered*, *hierarchical*. The *composite* view could be captured by allowing respondents to make multiple selections. In addition, the *faceted* and *graphical* views should be included for their unique contributions.

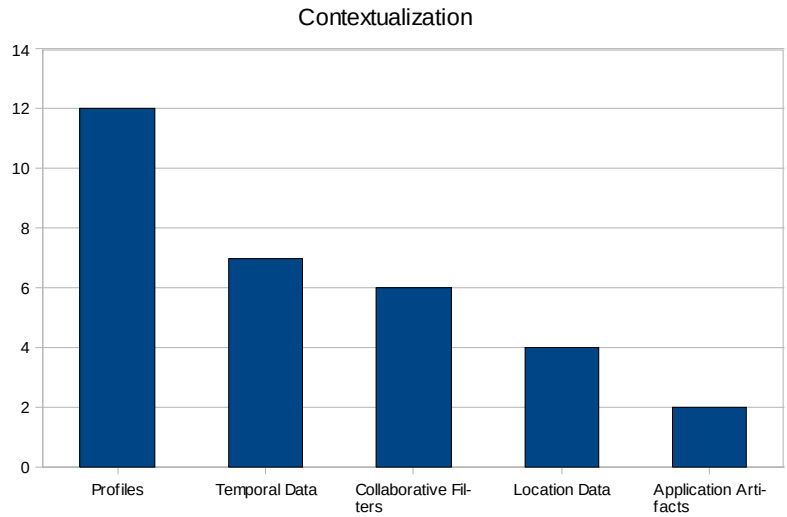
Results Presentation	Projects
Single Ranked List	10
Faceted View	3
Cluster View	0
Graphs	0

Table 3.2 Results Composition

Quality of Results. Nothing useful can be determined from this question. In fact, future versions of the survey should probably remove this data point since it is more relevant to benchmarking and validation, which is a later stage informed by the use case data collected from this survey. As discussed in the Overview section, use case data can be used to determine appropriate criteria for measuring the performance of search applications, but it is not the role of use cases to gather the metrics currently being used by the projects themselves.

Contextualization. Research into contextualization methodologies are overwhelmingly focused on profiles (see Fig. 6). Furthermore, the popular use of collaborative filtering and location data probably comes from profiles as well. This means that profiling techniques are a very popular area of research. In fact, it is difficult to imagine a contextualization effort that does not use data likely related to profile data in some way.

Figure 3.6 Contextualization



Semantic Technologies. Most projects reported on research efforts into all the semantic technologies included in the survey (ontologies, semantic language specifications, reasoning engines, semantic query language, ontology development tools). Furthermore, each technology was reported to be investigated by six or seven projects. There does not seem to be a gap.

3.3.3 User Interaction

The following questions were asked in this section.

- *What is the primary retrieval strategy of your typical user? (Retrieval Strategy)*
- *What is the goal for most of your users? (Goal of Interaction)*
- *How do users formulate a typical query on your system? (Query Formulation)*
- *What types of queries does your system primarily use? (Query Type)*

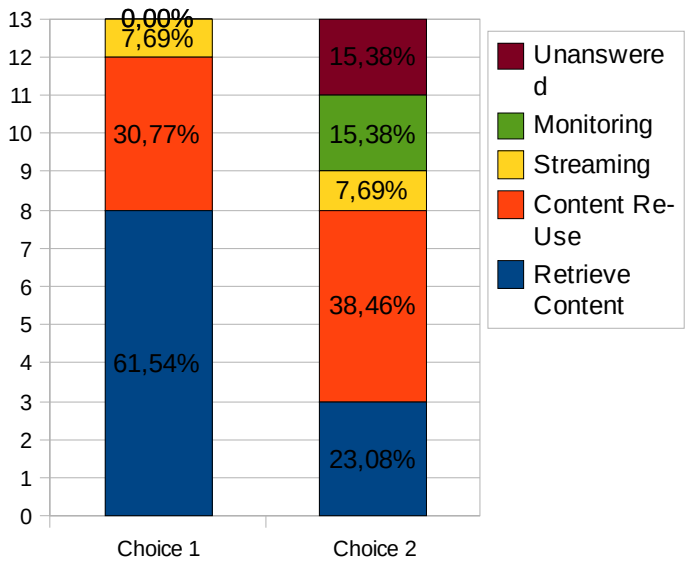
Retrieval Strategy. A majority of projects are involved with systems primarily using Known-Item Search (62%) (see Table 3). Browsing is the primary retrieval strategy for 23% of projects and recommendation systems comprise 8% of the projects surveyed. This question did not allow multiple selections, so there is undoubtedly much overlap, viz., most of the projects probably use a combination of methods. Therefore it is not advisable to make any determinations from this data. Future versions of the survey should ask respondents to make multiple selections to this question.

Retrieval Strategy	Projects
Known-Item Search	8 (62%)
Browse	3 (23%)
Recommendation	1 (8%)
Unanswered	1 (8%0

Table 3.3 Retrieval Strategy

Goal of Interaction. A majority of research projects are interested in the typical content retrieval and re-use scenarios (93% and 61% as first and second choices, respectively) (see Fig. 7). Monitoring and content streaming are receiving little attention.

Figure 3.7 Goal of Interaction



Query Formulation. Nothing interesting can be discovered from this question. It should possibly be dropped from future versions of the survey.

Query Type. All projects reported that the primary type of query used in their system is an explicit query and none reported using implicit queries. This would be an alarming trend if, as it seems to suggest, no one was using implicit queries. However, the survey question did not allow multiple choices, so projects were forced to select only one option. Consequently, it is not advisable to draw any conclusions from this result. Future versions of the survey should reformulate this question altogether and ask what kind of implicit query strategies, if any, are being investigated.

3.3.4 Repository Features

The following questions were asked in this section.

- *What size of document repository are you targeting with your system? (Repository Size)*
- *What is the likely cull rate (removal of obsolete index entries) per year for targeted repository documents that you index? (Repository Churn Rate)*
- *What is the likely growth rate per year for targeted repository documents that you index? (Repository Growth Rate)*

Repository Size. Almost all projects are working with collection sizes that are open, or innumerable (see Table 4). That is, they can never be complete. This corroborates the apparent trend revealed in the *Service Platform* question under the "System Features" section where 69% of projects report that their systems were designed for internet servers. Most corporate and institutional repositories are not innumerable. Therefore, our results from the *Repository Size* question seem to support the prior tentative conclusion that enterprise search is not receiving much attention.

Repository Size	Projects
Innumerable	10
1 TB	1
50 GB	1

Table 3.4 Repository Size

Repository Churn & Growth Rates. These two questions don't reveal anything interesting for a gap analysis. However, this data would be useful in the determination of functional requirements, the first purpose for collecting use cases as discussed in the Overview section.

3.3.5 Socio-Economic Factors

The following questions were asked in this section.

- *What are likely revenue source(s) for a business using your system? (Revenue Sources)*

- *What strategies are used to increase system transparency and trust so that potentially biasing factors affecting results can be identified? (Trust Management)*
- *Is personal information maintained by your system? (Privacy)*
- *Who owns most of the metadata used by your system? (Metadata Provenance)*
- *Under what conditions do you make available to third parties any metadata you produce? (Metadata Licensing)*

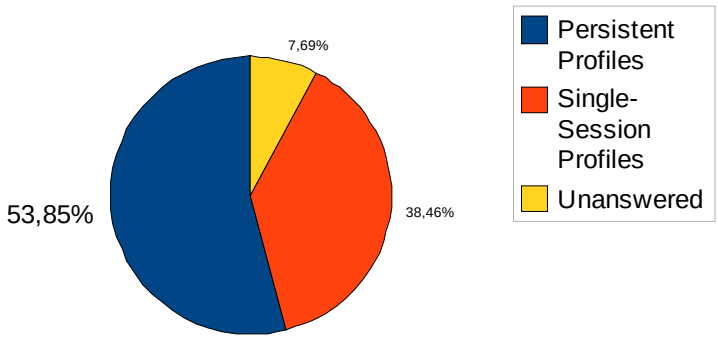
Revenue Sources. The data collected is inconclusive. Most projects do not have business plans, so this question is probably a measure informed by unqualified speculation. Furthermore, it does not contribute to understanding functional requirements or benchmarking and validation criteria. Therefore, future versions of the survey should probably remove this question.

Trust Management. This question suffers from the same drawback as the previous one. It deals with search as a complete service within a competitive, commercial environment. Management of trust among users is a customer service problem. In other words, it is a commerce issue. As research efforts, most projects do not approach search as a commercial actor. Rather, they focus on solving some technical aspect of search. Consequently, it is not surprising that most projects report this question as not applicable to them (see Table 5). It is interesting to note that several projects do, nevertheless, use some form of trust management strategy. In particular, PHAROS and Victory both report using reputation systems. However, it is suspected that these projects were reporting on the reputation system as a technical problem.

Trust Strategy	Projects
Not Applicable	8
Provenance Information	3
Reputation System	2
Sponsors Clearly Identified	1

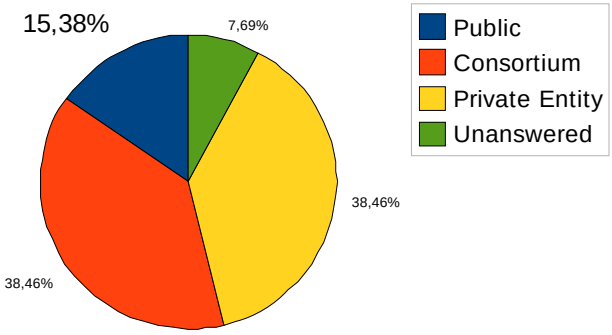
Table 3.5 Trust Management

Figure 3.8 Privacy



Privacy. A majority of projects (54%) are using persistent profiles in their systems and 38% report using single-session profiles, for an overall 92% project reporting rate for profile usage (see Fig. 8). Given the importance of content enrichment with metadata external to the content itself to overcome the semantic gap (as discussed in the *System Features* section), this is a good trend. However, there will be repercussions within the social-economic and legal spheres.

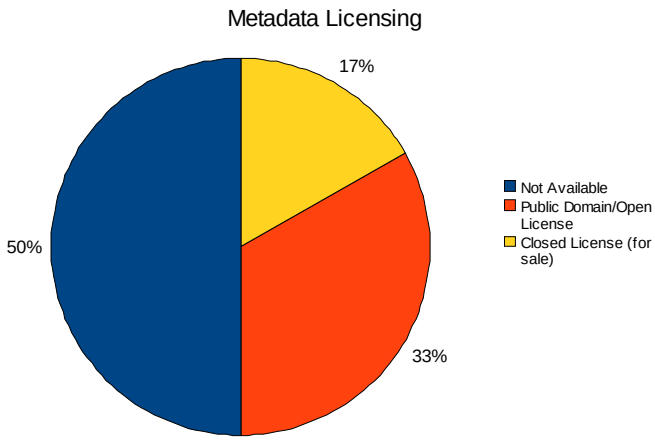
Figure 3.9 Metadata Provenance



Metadata Provenance. A large amount of metadata used by projects is privately owned (38%) whereas publicly available metadata is used in only 15% of projects (see Fig. 9).

Metadata Licensing. We see a similar trend in the data in this question. Many projects (38%) do not make their metadata available at all and another 13% have applied a closed license to it, which means it is for sale (see Fig. 10). In other words, a majority of metadata used by projects is not in the public sphere. Only 25% of projects have released their metadata into the public domain. It is difficult to know whether there is a significant trend in this data. It is probably not important that all projects release metadata, since the focus of research is on the advancement of the art of the technical *process* of metadata generation. The consequent index (of metadata) is of secondary interest insofar as it determines the quality of the process that was used to create it. Metadata can always be recreated.

Figure 3.10 Metadata Licensing



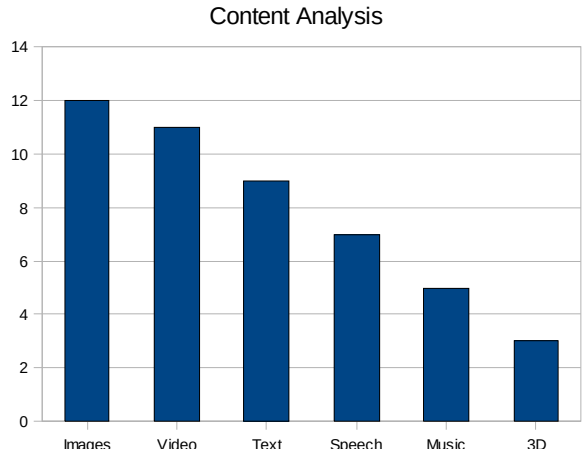
3.3.6 Indexing Features

The following questions were asked in this section.

- *What kinds of content features are primarily processed to build your index or to differentiate your index from others? (Content Analysis)*
- *How quickly can new content be annotated, published and made available? (Indexing Timeliness)*
- *What kind of index does your service create and utilize? (Index Scheme)*
- *How is most of your metadata generated? (Metadata Generation)*

Content Analysis. As expected, most projects are using image and video processing techniques. In addition, text processing is used by a little more than 2/3 of the projects (see Fig. 11). This corroborates, to some extent, the findings of the *Privacy* question which revealed that 12 of the 13 projects reported working with user profiles. In addition, results from the *Contextualization* question indicate that all but one of the projects were utilizing user profiles. Again, it is important that research continues in this area for the purpose of closing the semantic gap inherent in multimedia retrieval.

Figure 3.11Content Analysis



It should be noted that the rather exotic research topics of music and 3D processing are also represented in project research with 5 and 3 projects reporting activity in each, respectively. This seems to represent an overall good mix of content analysis research.

Indexing Timeliness. This question does not seem to reveal anything notable for a gap analysis. All projects report the capability to regenerate their index on an hourly or daily basis. Since we do not known the size of repository that the projects are reporting these numbers against, it is impossible to determine whether their methods are satisfactory for commercial purposes. This question should be considered a candidate for removal from future versions of the survey if it does not contribute to understanding the functional requirements of the system.

Index Scheme. Not surprisingly, almost all projects are using a centralized index scheme (see Fig. 12). However, three projects (VITALAS, Victory and SAPIR) report using distributed indexes. Distributed search remains an intriguing area of research today. It is good that some activity has been reported in the use of decentralized indexes and this should be ensured in the future.

Figure 3.12 Index Scheme

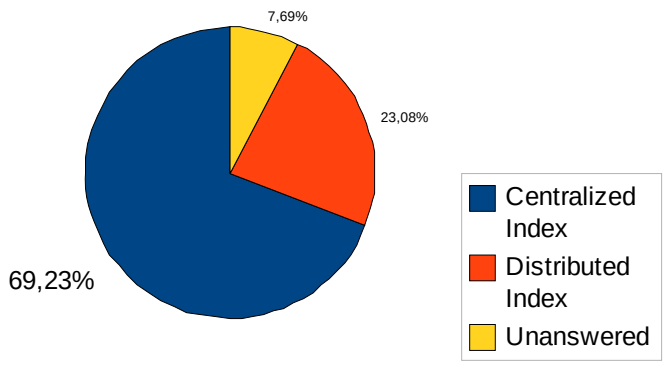
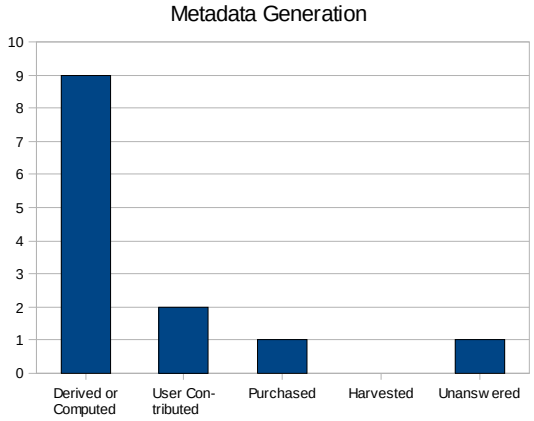


Figure 3.13 Metadata Generation



Metadata Generation. Nothing surprising is revealed. Almost all projects generate their metadata using computational, or algorithmic, means, rather than simply copying what exists in the collection (harvesting) (see Fig. 13). Again, we must remember that the community surveyed were researchers, not commercial entities. Metadata creation is an important problem in search which they are investigating. If the surveyed group were commercial actors we would expect to see harvesting and purchasing as the most popular means of generating metadata, since the primary goal would be to build complete and useful indexes, not to investigate some new method of deriving metadata.

3.4 Discussion and Future Work

3.4.1 Major Findings

3.4.1.1 Functional View

As we have seen from the *Results and Interpretation* section, there are various implications of the findings as they apply to the functional view of search engines, resulting in the most direct impact on technical progress. Namely, the apparent neglect of intermediate users in system design would most likely negatively impact the pace of innovation that applies to the *Query Formulation* and *Results Presentation* components. Additionally, the possible neglect of editors and repository managers would likely result in a negative impact within the *Ingest, Enrich Content, Gather Document Context,* and the *Build* components of search engines. In order to safeguard a fully informed European research agenda with wide coverage of all functional search components, CHORUS should ensure that these user groups receive attention.

3.4.1.2 Enterprise Search

Several data points corroborate the finding that enterprise search may be a neglected area (see Community Size, Service Platform, and Repository Size). This is further supported by the fact that CHORUS Think Tank discussions tend to focus on internet mediated search. In fact, there seems to be an agenda of attempting to engineer a European competitor to Google within the internet search arena. This is unfortunate due to the fact that enterprise search is a very important and lucrative market sector and it is an area where Google does not dominate. Additionally, the contributions of enterprise search technologies to internet search may become much more important as the semantic web emerges.

3.4.1.3 Market Segments

Multimedia retrieval is a crucial differentiating factor for the generic use case of *Personalized TV* (see PTV in the section *Market Segmentation Typology & Survey*), which seems poised to become an important market segment in the near future. Consequently, CHORUS should hope to see the set-top box as a more common end user device within European research projects (see *Device*).

Monitoring and streaming applications also seem to be neglected (see *Goal of Interaction*). These application areas are closely aligned with the generic use cases of *Personalized TV* and *Surveillance, Detection and Alert* (see PTV and SDA in the section *Market Segmentation Typology & Survey*). CHORUS should ensure that these topics are sufficiently invested in future projects.

3.4.1.4 Semantic Gap

There seems to be much work happening with Profiles (see *Privacy*) and Collaborative Filters (see *Contextualization*). Due to the significant difficulties that arise from the semantic gap inherent in multimedia retrieval, this is a rather good trend. Content-based processing techniques of multimedia are very limited in their ability to generate conceptual annotations and

this will be the case for quite some time. Although we cannot expect to close the semantic gap using content-based processing any time soon, this does not mean that multimedia retrieval cannot be made to perform as well as content-based processing of text documents. The most promising strategy for closing the semantic gap in a reasonable amount of time seems to come from content enrichment of metadata using data from sources external to the content, such as profile data. Recommendation and reputation systems both exhibit this approach. We should continue encouraging significant research efforts that investigate profile techniques for contextualization.

3.4.1.5 SOCIO-ECONOMIC AND LEGAL FACTORS

The use of profiles, although good for closing the semantic gap, is a challenging issue for policymakers. Most projects are investigating the use of profile data (see *Privacy* and *Contextualization*) and collaborative filters (see *Contextualization*). Technically, research in this area needs to continue if we want effective multimedia search engines and intelligent human-computer interaction. However, this is a socially and legally challenging issue. The development of profiles, or dossiers, about people has its precedent in several recent and rather dark historical times.

3.4.1.6 METADATA

The results are interesting from the *Metadata Provenance* question. They reveal that only 15% of metadata used by projects comes from the public sphere compared to 38% that comes from a private entity. There is a compelling case that can be made for ensuring that metadata, by and large, remain in the public sphere due to preservation and lifecycle issues (see section 5.4, *Metadata Lifecycle and Preservation*). However, an equally compelling case can be made that metadata creation and enrichment lies in the domain of value added services within the commercial sector and, as such, should be regarded as a revenue generator. No matter where we may fall on this issue, a large number of projects (38%) report using consortium metadata, which may balance the field and ensure that a diversity of approaches are pursued.

However, it is probably important to follow up with the 38% of projects that reported they did not make their metadata available in the *Metadata Licensing* question in order to find out why and what kind of metadata they generate.

3.4.2 Survey Design

We have attempted to model research projects utilizing a use case model that is designed to capture the functional requirements of a service. Considering search as a service in the design of the *Use Case Typology* and *Survey* resulted in somewhat of a misalignment between how we are attempting to view and classify research projects and how they actually operate. In particular, search as a service tends to embody the commercial point of view. That is, it presumes search to be a completed system, deployed in the market and servicing an actual user base with a business model driving its development. On the other hand, projects are research oriented. They are more focused on some small aspect, or technical problem, of search to be solved. As such, they tend to represent commodity components in the sense that they are best deployed as part of a new service that could be aimed at any arbitrary use case. This has two results. First, some projects reported that many of the questions do not apply to them. Secondly, since the work of many projects can be applied within multiple use case scenarios, some asked that they be allowed to respond with multiple answers to many of the questions. Projects that reported either of these difficulties, however, were asked not to respond to the survey. In general, when applying the *Use Case Survey,* as it is currently designed, to research projects, we should expect missing and inconsistently collected data.

The market segment survey should be merged with the use case survey so that broad market segments can be analyzed. At the very least, this could help corroborate some of the findings suggested by the use case survey.

Finally, questions should be classified more carefully by one of the three purposes for collecting use cases outlined in the Overview section. Namely, (1) determination of evaluation criteria, (2) determination of functional requirements, and (3) a gap analysis of research topics. These questions should then be evaluated by stakeholders and revised so that future surveys address more specific needs in order to achieve a maximum productive impact for the considerable effort involved.

REFERENCES

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.

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

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.

4 RESEARCH AND TECHNOLOGICAL GAPS

The functional description given in chapter 2 above involves several crucial research issues. They center around the application of information access to known information access needs of users and current as well as coming applications designed to address those needs. The functional description does not make statements as to what information is used for indexing and querying the system; it does not presume any specific interaction model; nor does it specify a usage context or task. All of these factors are determined by the application the system is put to use in. Chapter 3 describes use cases and attendant usage scenarios as understood by current research projects in the field. Use cases do not specify the functionality in the system assumed in the scenarios. Linking the space of possible use cases given by chapter 3 to the functional description we find the crucial challenges given by the specifics of the knowledge representation which must be designed to be habitable for query preparation, result presentation, and indexing functions alike. Improving one without the others - while possibly a fruitful research task - is pointless from an application perspective. Figure 4.1 shows the interplay of index, query, and result presentation in the context of the functional description from chapter 2.

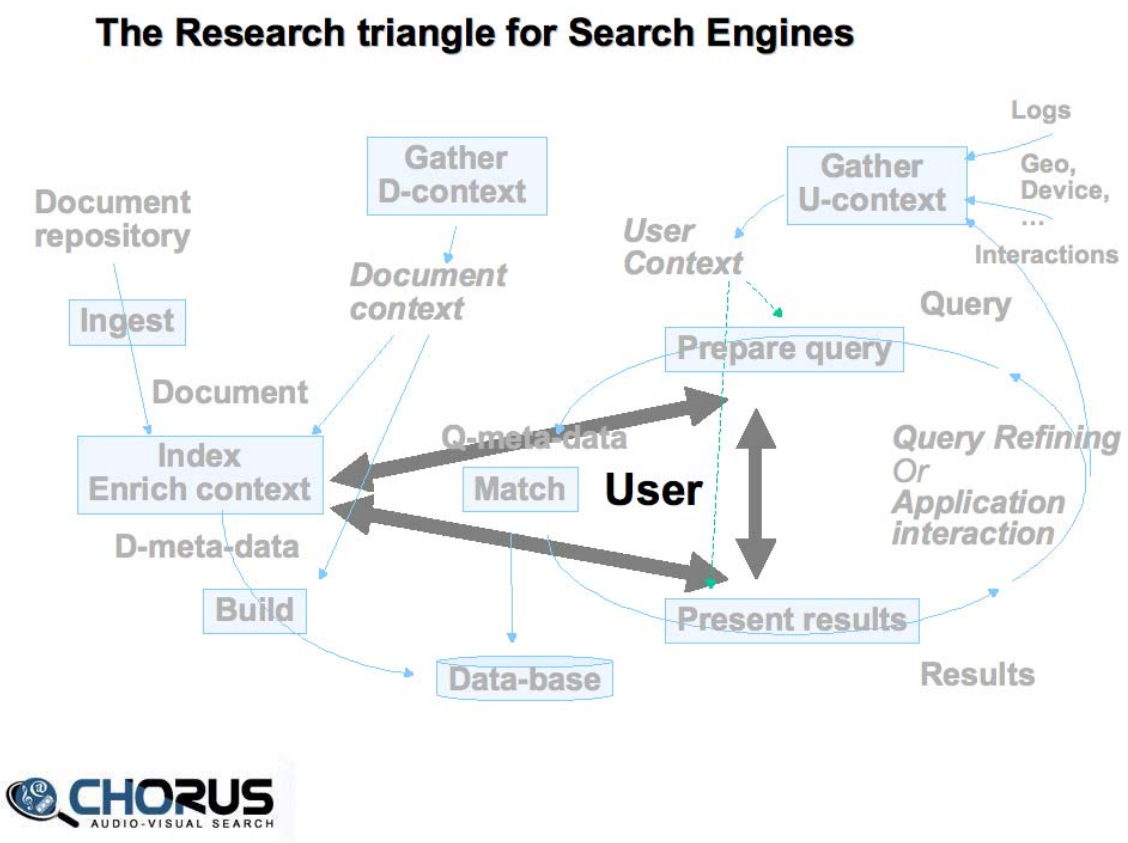


Figure 4.1 Users interact with the information access system at various functional points

The following sections will address, in turn, techniques for extraction salient information descriptors from the various data documents arriving for treatment in the system (section 4.1); techniques for formulating information needs in terms the system can use for preparing a query appropriate for the index in play (section 4.2); and techniques for presenting the found results in a manner appropriate for the information need at hand (section 4.3). Interaction models and service designs based on envisioned use cases and usage scenarios can be instantiated variously, based on the functional description from chapter 2.

It should be noted that the query preparation function from the functional description should not be equated with a user interface for entering queries; neither should the result presentation necessarily be understood as a ranked list of documents. While these are typical interaction modes for ad-hoc text retrieval systems such as have been made familiar by commercial web search engines in recent years, they are by all means not the only possible interaction model with an information access system using an underlying search engine. The possibilities of different interaction models are made much more obvious once systems for non-text content have been made available for real tasks: the topic oriented search associated with text retrieval systems is not alone among usage scenarios anymore.

4.1 Advanced content enrichment methods

Human informational output, whether collated in edited and well-defined documents or implicitly present in data streams, data bases or collections, contain large amounts of potential information that information access systems must process and manage to make them visible, retrievable, and useful. What the object of interest is can be understood in several ways, whether a text (in a digital library or in a web page, e.g.), a multimedial animation or presentation (embedded in the output of some proprietary tool), a video clip or audio capture (in any of a multitude of formats). These broadly different understandings of what constitutes a multimedia document require anyone working with multimedia information access to define the notion of syntax and semantics of multimedia documents. The syntax of a multimedia document is an aggregation of several elements from different data types that provide rich information and enhanced experience. Standards for such definitions are being accepted broadly in the field. Visible and audible data types are text, images, graphics, video, and audio; structural elements may not be visible by themselves but are used to determine the spatial and temporal organization of the other data types; interactive elements provide a way for the user to interact with the content.

However, while many of the challenges of text retrieval and multimedial retrieval are similar, there are important differences. Unlike text documents, multimedia documents do not necessarily contain symbols that users can use directly to express their information need. This problem has roots in two different aspects. The first one is the richness of the searched information itself: visual information can communicate a wide variety of messages and emotions; audio content can also communicate feelings and emotions; structure also gives a different organization and usability (or user experience) to communication. In other words, multimedia documents give more freedom to the semantic interpretation of the communicated message. The second aspect is the communication gap between users and systems at the actual time an information need arises: while the representation the computational systems work with may be consistent and reliable, it may not be responsive enough to the situational dynamics the human user is grappling with, including factors related to the expression of ideas, emotions and feelings.

Several techniques have been developed and researched to empower users with new tools to express their query to achieve a better mapping between what users can express, what the system can extract from multimedia, and what the system can successfully match. In the following the advanced content enrichment methods are discussed for several different types of media. Every media type requires specific algorithms which have to be analyzed and extended to deliver reliable indexing and retrieval methods. Finally all indexing methods should be combined to a full multimodal approach to extract the maximum amount of information for a multimedia retrieval task.

4.1.1 Speech

The reliable indexing and retrieval of speech content is an important component in a generic multimedia search engine. Upcoming services like the Google Election Video gadget, which uses speech recognition to index the political speeches of the 2008 US campaign shows that these kind of analysis is of increasing interest. A summary of the current status will be followed by current trends and long-term perspectives in speech based information retrieval.

During the last two SIGIR workshops on Searching Spontaneous Conversational Speech (SSCS),
held in 2007 and 2008, the existing speech retrieval technology and applications were presented and demonstrated. Although achieving high word recognition rates with Large Vocabulary Continuous Speech Recognizers (LVCSR) are always the main goal during the indexing process several aspects has been investigated. For broadcast news task containing read and non-spontaneous speech the word error rate is in an area of 15 to 10% for a speaker independent task. However, for other domains like indexing lectures or meeting the error rates reaches 50%. Any degradation of the speech signal due to different speaking styles, dialects and background noises decreases the recognition performance. For a retrieval task the pure recognition rate is only one aspect. To optimize the retrieval rate is has been shown that a lattice based decoding improves the recall and precision numbers of a speech based retrieval system. The lattice output containing graph based recognition alternatives have to be integrated in a common retrieval environment including standard based indexing structures. It has been shown that the usage of lattices improves the retrieval rates significantly independent from the pure recognition rate. A major drawback of systems which apply large vocabulary continuous speech recognition for spoken document retrieval on heterogeneous data is their vocabulary dependence: the recognition component needs prior lexical knowledge, i.e. the set of recognizable words is limited. Vocabulary independent approaches try to overcome this by estimating speech transcripts on the subword level. Spoken term queries are broken down to a subword representation, and a pattern matching algorithm locates this representation in the subword transcript. Current work includes lattice based phoneme, syllable and word indexing, as well as hybrid word-subword approaches. Scalability issues arise with the complex retrieval on subword units and lattices. For universal search engine the aspect of multiple languages and multilinguality is of increasing interest.

Besides speech retrieval technology preprocessing algorithms to segment the speech signal in homogeneous segments have been investigated and developed. Speaker and sentence segmentation are important techniques to segment the speech signal.

4.1.1.1 TRENDS/PERSPECTIVES

- short term
    - Hybrid speech retrieval systems: The combination of word and sub-word (i.e. syllables, phones) unit retrieval systems will overcome the out-of-vocabulary problem with respect to performance issues.
    - Lattice integration: The usage of lattice speech decoding outputs shows improvements regarding the retrieval rates in comparison to single best output. Here the integration into existing indexing systems must be investigated and improved. Also the usage of posterior probabilities derived from the lattice output is under investigations.
    - Robust feature extraction: Background noise and complex multimedia recordings (e.g. double talk, mix of speech and music) requires robust methods to extract and transform the speech signals. Here further research is needed to achieve human like robustness.
    - 10% error rate for broadcast news: Speech recognition systems with 10% word error rate in broadcast news will be developed for several languages.

- long term
    - Vocabulary independent retrieval systems: To fully overcome the vocabulary dependency of a speech retrieval system, investigations in the area of sub-word unit modelling, language model adaptation and flexible indexing structures has to be carried out.
    - Multilingual system: Starting form systems for multiple languages further research work on multilingual speech retrieval systems has to be done. Especially for less investigated languages (e.g. dialects of Afghanistan) more intelligent approaches than collect and annotate huge speech corpora has to be investigated
    - Spontaneous speech: The aspect of spontaneous speech (different speaking styles, different types of language models) must be investigated
    - Adaptation: The domain adaptation is important to achieve better retrieval results for new and changing domains.
    - 10% error rates for meeting: The meeting domain covers many of the defined challenges. As long term goal the speech decoding system should generate less than 10% error for the difficult meeting domain (i.e. background noise, double talk, open vocabulary, …)
    - Blind source separation: The separation of audio sources is still unsolved but important for practical tasks
    - Audio event detection, audio scene analysis: The automatic segmentation, indexing and description of complex audio scenes including audio events is long term goal for audio scene analysis

4.1.1.2 CHALLENGES

- A great challenge is to overcome the strong **lexicon dependency** of today's speaker recognition systems. I.e., it is imperative to improve speech recognition in contexts where no adequate language model is present.
- This is especially true where the **variability of speech** is very high, e.g., in natural language processing.
- Most speech retrieval systems are tuned for a single target language. Thus, **multilinguality** is a topic still to be addressed in the implementation of more widely applicable systems.
- Realistic environments are often characterised by **very complex audio sounds**. In such situations, speech recognition accuracy tends to drop considerably.
- Finding methods for **unsupervised adaptation** of speech models is a key step in dealing with varying speakers and acoustic situations.

4.1.1.3 PROMISING DIRECTIONS

- language model adaptation, tight integration of speech decoder output and indexer
- audio modelling (anchor models)
- phonetic or syllable based approaches, new forms of context dependency

4.1.2 Music

Finding music of interest in today's large media repositories is a strong need for many users. Music is also present in a large number of media such as television shows and radio broadcasts and has to be handled by search engines dealing with these types of media.
After a short overview of the current status of music information retrieval, future challenges for the development of musical information retrieval systems are identified from current and upcoming trends.

Digital representations of music come in two flavours: symbolic representations and signal-based representations. The first kind of representation is closer to musical scores and allows direct access to abstract concepts like notes and pitch but

usually lacks information concerning the interpretation of the music. The analysis of signal-based representations, on the other hand, has to start with more basic tasks such as note onset detection.
A problem that is widely regarded as solved is the task of audio identification, also known as fingerprinting. Here, given a short excerpt of an audio recording of music, the task is to identify it, using a large database of recordings, by giving, for example, the name of the song, the composer or interpret. Often, the exact time position in a recording from the database is to be found. This can be done reliably even for very large databases and in the presence of noise, lossy compression, and a number of other signal degradations.
A more challenging task extending the scope of audio identification is audio matching. Here, the goal lies in resolving queries by example in situations where more severe alterations of the music are present. For example, it is desirable to be in a position to identify different versions of a song, involving musical alterations such as nonlinear temporal distortions or changes in instrumentation. Here, research is still in its infancy.
Synchronization of different music representations such as score and audio allow gaining high-level information not from the audio signal alone but by using available additional information. This approach works very well in controlled audio situations such as classical music performances with a small number of instruments.
Symbolic music representations avoid a number of the problems inherent in signal-based representations and allow access to a number of concepts from musical semantics, such as notes, instruments, and voices, which are hard to extract from other representations. Thus, the problems attacked in the analysis of symbolic music representations usually reside on a higher semantic level. A lot of research has been focused on finding similarity measures for melodies. These could be a cornerstone of melody retrieval systems that are closer to the human perception of musical similarity.

4.1.2.1 TRENDS, PERSPECTIVES

Now that extracting and applying low-level audio features have become familiar techniques for researchers and system designers, a main trend is to move music information retrieval techniques closer to a general audience by addressing topics that are more closely related to musical semantics. Research in audio matching tries to yield robust recognition of music subject to typical musical variations. Tasks like the following aim at supporting users in organizing their music collections:

- genre classification
- artist recognition
- mood detection
- automatic playlist generation
- music recommendation

These techniques are also of particular interest in the context of modern music distribution involving online stores.
More basic tasks find application in retrieval engines as well as in analysis tools for music scientists:

- music segmentation
- music summarization
- rhythm extraction
- melody extraction
- instrument identification

A growing trend in audio analysis is source separation. This research field aims at developing algorithms for the identification and extraction of the sources constituting an audio scene. For mixtures of a small number of audio sources, current techniques already give good separation results. Extracting audio sources from complex mixtures, however, is still out of reach of these methods.
A common approach to bridge the semantic gap between low-level audio features and musical semantics is involving communities into the annotation process. Thus, the enthusiasm of music lovers can be exploited to gain high quality annotations.

4.1.2.2 CHALLENGES

A high number of low-level audio features for music description have been developed and many of them have been standardized in the MPEG-7 standard (Moving Picture Experts Group). There is, however, no comparable process for the extraction of high-level features.This is one of the main challenges in signal-based music processing: to move from low-level signal descriptions to high-level features corresponding to musical semantics. Advances in this context will show effect in many areas such as musical similarity, music retrieval, and summarization.
The main problem in achieving this goal lies in the high complexity of music stemming from the complex sounds generated by musical instruments and the human voice as well as the interaction of several of them. This makes extracting semantic information from audio recordings of music very difficult and very desirable features such as instrument recognition and music transcription are currently out of reach in the setting of general polyphonic music.
Another group of difficult problems arises in situations where tasks are hard to define properly. For example genre classification and mood detection have a strong subjective component which makes it hard to define ground truth data and exact problem definitions for algorithm development. Other tasks, such as melody extraction from audio or symbolic representations, come with similar problems and are usually hard problems even for human listeners.

4.1.2.3 PROMISING DIRECTIONS

- Moving from **controlled audio situations to more complex situations**: the main problem in achieving higher-level music information retrieval tasks is that of extending methods working well in controlled audio situations to more complex situations, e.g., finding a way to apply methods coping well with small numbers of instruments to strongly polyphonic situations with more or more complex instruments. One way to achieve this is to strengthen research in source separation techniques. A robust segmentation of audio scenes into its sound sources would enable a large number of existing analysis techniques to work in more complex situations.
- **Robust extraction of high-level information**: the robust extraction of high-level information from the low-level features that are the cornerstone of contemporary musical information retrieval systems has to be bolstered, ideally in cooperation with music scientists.
- **Learning algorithms for time-series data**: another main blocking stone in algorithmic music analysis is the fact that current state-of-the-art machine learning techniques are not very well adapted to time series-based data such as music. Usually, great effort has to be spent in order to find representations of music signals suitable for the application of machine learning algorithms. More effort should be put into the development of machine learning algorithms tailored for time-series data.

4.1.3 Image

Content-based image retrieval (CBIR), as we see it today, is any technology that in principle helps to organize digital picture archives by their visual content. By this definition, anything ranging from an image similarity function to a robust image annotation engine falls under the purview of content-based image retrieval. This characterization of content-based image retrieval as a field of study places it at a unique juncture within the scientific community. While we witness continued effort in solving the fundamental open problem of robust image understanding, we also see people from different fields, such as, computer vision, machine learning, information retrieval, human-computer interaction, database systems, Web and data mining, information theory, statistics, and psychology contributing and becoming part of the content-based image retrieval community. Moreover, a lateral bridging of gaps between some of these research communities is being gradually brought about as a by-product of such contributions, the impact of which can potentially go beyond content-based image retrieval.

Despite the effort made in the early years of image retrieval research, there is not yet a universally acceptable algorithmic means of characterizing human vision, more specifically in the context of interpreting images. Hence, it is not surprising to see continued effort in this direction, either building up on prior work or exploring novel directions. Considerations for successful deployment of content-based image retrieval in the real world are reflected by the research focus in this area. By the nature of its task, the content-based image retrieval technology boils down to two intrinsic problems: (a) how to mathematically describe an image, and (b) how to assess the similarity between a pair of images based on their abstracted descriptions. The first issue arises because the original representation of an image which is an array of pixel values, corresponds poorly to our visual response, let alone semantic understanding of the image. From the design perspective, the extraction of signatures and the calculation of image similarity cannot be cleanly separated. The formulation of signatures determines to a large extent the realm for definitions of similarity measures. On the other hand, intuitions are often the early motivating factors for designing similarity measures in a certain way, which in turn puts requirements on the construction of signatures. Significant effort in the recent years has been on developing a large diversity of image signatures. Advances have been made in both the derivation of new features (e.g., SIFT) and the construction of signatures based on these features, with the latter type of progress being more pronounced. The richness in the mathematical formulation of signatures grows alongside the invention of new methods for measuring similarity. In terms of methodology development, a strong trend which has emerged in recent years is the employment of statistical and machine learning techniques in various aspects of the content-based image retrieval technology. Automatic learning, mainly clustering and classification, is used to form either fixed or adaptive signatures, to tune similarity measures, and even to serve as the technical core of certain searching schemes, for example, relevance feedback.

4.1.3.1 TRENDS/PERSPECTIVES

- short term
    - semantic search and concept-based detection in complex background: dealing with cluttered background and increasingly complex concepts.
    - interactive search and agent interfaces: Agents are present in learning environments, games, and customer service applications. They can mitigate complex tasks, bring expertise to the user, and provide more natural interaction. Creating natural behaviors and supporting speaking and gesturing agent displays are important user interface requirements. Research issues include what the agents can and should do, how and when they should do it (e.g., implicit versus explicit tasking, activity, and reporting), and by what means should they carry out communications (e.g., text, audio, video). Other important issues include how do we instruct agents to change their future behavior and who is responsible when things go wrong.
    - Efficient (visual) feature extraction: extracting efficient and robust features form a dense grid on the image (e.g, SURF).

  - New learning models: collaborating with the artificial intelligence and learning research community for new paradigms and models of which neuro-based learning is only one candidate. Learning methods have great potential for synergistically combining multiple media at different levels of abstraction.
  - Statistical models for semantic access: extensive effort toward probabilist models that can benefit from training data and are robust to missing information.
  - Automatic annotation: methods that use the concept detectors to automatically annotate the images.

- long term
  - Experiential computing: Focus on methods for allowing the user to explore and gain insights in media collections. On a fundamental level, the notion of user satisfaction is inherently emotional. Affective computing is fascinating because it focusses on understanding the user's emotional state and intelligently reacting to it. It can also be beneficial toward measuring user satisfaction in the retrieval process.
  - Human-centered models: t main idea is to satisfy the users and allow them to make queries in their own terminology.
  - Emergent semantics: study the potential to learn the goals of the user in an interactive way.
  - Multi-user environments: discovering more effective means of human-human computer-mediated interaction is increasingly important as our world becomes more wired or wirelessly connected. Very important here is the query model which should benefit from the collaboration environment.
  - Neuroscience models: Use the results from neuroscience and cognitive psychology to discover and, in some cases, validate abstract functional architectures of the human mind. Even the relatively abstract models available from today's measurement techniques (e.g., low fidelity measures of gross neuroanatomy via indirect measurement of neural activity such as cortical blood flow) promise to provide us with new insight and inspire innovative processing architectures and machine learning strategies.

4.1.3.2 CHALLENGES

- Semantic gap: this is still an unsolved problem. The next evolution of systems would need to understand the semantics of a query, not simply the low level underlying computational features. More sophisticated user models and interaction are needed.
- Image segmentation: integration of the user context and needs. Image segmentation will only be solved in a restricted domain.
- Fusion with other modalities: this challenge refers also to video retrieval. The idea is that the multiple modalities will not only bring extra information but can also help the disambiguation of each other.
- Interpretation: this refers to the user model and context. Interpretation should be done in a particular context and for a specific user.

4.1.3.3 PROMISING DIRECTIONS

- New machine learning algorithms based on human perception and cognition
- High-performance computing
- Folksonomies
- Evaluation with emphasis on representative test sets and user patterns

4.1.4 Video

The ease with which video can be captured has lead to a proliferation of video collections in all parts of society. Getting content-based semantic access to such collections is a difficult task, requiring techniques from image processing, computer vision, machine learning, knowledge engineering, and human computer interaction.

The current methods in video search rely on the automatic detection of semantic concepts, subsequently these methods use the shot-based confidence values associated to large sets of concept detectors for video retrieval. A basic concept detector exploits (visual) features related to color, texture, shape, and motion, and supervised learning using support vector machines. The basic concept detectors are typically extended in several ways. These extensions include fusion of multiple features, fusion of multiple classifiers, and modeling of relationships between multiple concept detectors using graphical models, data mining techniques, and ontologies. Research in automatic detection of semantic concepts in video has now reached the point where over a hundred, and soon more than thousand, concept detectors can be learned in a generic fashion, albeit with mixed performance. Current research directions in retrieval explore how to select relevant concept detectors automatically given a user query, by relying on query topics, example images, and again ontologies. Despite the potential of concept-based video retrieval, however, automatic methods will not solve all search problems. Thus, eventually user involvement is essential. To aid the user in concept-based video search, several advanced visualization techniques have been proposed recently that aid the user by displaying and browsing video retrieval results. In addition, active learning and relevance feedback are active areas of research for interactive video search. It should be noted that while concept-based video retrieval is a promising technology, performance is still far from perfect; the state-of-the-art typically obtains reasonable precision, but low recall in a relatively narrow domain. Hence, robustness of concept detection methods needs to be improved further to cater for broad-domain applicability. Note the inherent connection between video retrieval and image retrieval: they both share most of the same problems and the current trends in both cases rely on improving the feature extraction and the classification methods.

4.1.4.1 TRENDS/PERSPECTIVES

- short term
  - Incremental machine learning improvements: the extension of the classical support vector machine approaches towards multi-kernel variants.
  - Incremental computer vision improvements: use more robust and efficient features such as keypoint/codebook features.
  - (Inter)active learning: develop learning methods that use the minimal user feedback and that despite this can cope with difficult problems.

- long term
  - Inclusion of temporal dimension: most of the current video retrieval methods are reduced to image based analysis but ignoring the temporal dimension. The new trend is to extract features that can capture the dynamics of the video.
  - Concept localization: most of the current concept detection methods are not able to precisely localize the detected object. The new trend is to concentrate on the concept localization which opens the way for multiple concept detection.
  - Temporal concepts: the development of models that can capture the dynamic concepts (e.g., a running car).
  - Dynamic concept interactions: the use of the co-relations between the concepts in a dynamic setting.
  - Include common-sense: some of the problems are too complex to be solved in general but they can be significantly be simplified if common sense constraints are used.
  - video mining/event detection: detection of the most important broadcasted video, most important event of the year, etc.

4.1.4.2 CHALLENGES

- Broad-domain applicability: there is a need for algorithms that are not only scalable to the available data but can easily be generalized to other applications.
- Lack of training data: develop methods that can learn in the presence of scarce training data.
- Computational efficiency: the development of real-time analysis that can efficiently allow semantic search.
- Robust performance: keep high robustness of the algorithms even if difficult conditions (e.g., scarce straining data) occur.

4.1.4.3 PROMISING DIRECTIONS

- Leveraging social tagged media as substitute for training data
- Transfer learning
- High-performance computing

4.1.5 3D Search

The approaches in content-based image retrieval (CBIR) and content-based video retrieval (CBVR) rely on semantic concepts, which are derived from visual features in an image or key frame representing a video shot. The next step beyond analysis of visual features is the analysis of properties describing the motion of the camera or the 3D structure of a scene. The term 3D search is dedicated to 3D properties represented in a video sequence. Hence, the retrieval of 3D objects or 3D models is not considered here, although this field of research becomes relevant even in the case of handling and accessing large repositories of 3D models. The progress in computer vision provides a large variety of methods to analyze the camera motion and the 3D structure of a scene under specific conditions. The exploitation of these techniques for derivation of novel semantic description of the scene content has become of increasing interest in the past years. After a short summary of the SoA, the current trends and long-term perspectives in 3D scene analysis for video information retrieval are presented.

In MPEG-7 some features related to the motion of the camera have been already described such as camera motion, global motion, motion trajectory and parametric motion. Due to the rapid development of robust techniques in the field of self-calibration by using structure-from-motion techniques, a complete 3D description of the camera path and a sparse description of the 3D structure of the scene are available. However, assumptions on the motion of the camera and the motion in the scene are still made in order to allow a robust estimation of the 3D structure of the scene. Based on a descent 3D model derived from video sequence first approaches are available in order to categorize natural scenes in flat or diverse structured scenes. The combination of 3D scene structure with visual 2D features is already starting and is exploited for the detection of urban and rural scenes. Furthermore, 3D reconstruction has been progressing in the past decade, categorized in shape from depth and shape from silhouette methods by using multiple views. In addition to self calibration approaches, properties of projective invariants are exploited in order to classify video information according to its 3D structure. Beside the analysis of 3D structure from video, the availability of associated 3D information in form of depth maps seems realistic in the near future due to the advances in the field of 3D TV and 3D cinema. It is recognized that the movie producers will provide associated 3D information or at least stereo content for future productions. Standardization activities for video and

depth information are on the way. Hence, the exploitation of this information for search and retrieval becomes increasingly relevant. It has to be stated that the advances in computer vision are not sufficiently exploited for information retrieval (IR).

4.1.5.1 TRENDS/PERSPECTIVES

- short term
  - Exploitation and improvement of existing self-calibration methods towards video analysis in more general scenarios
  - Derivation of a robust estimation of sparse 3D models from video sequences based on results of self-calibration methods
  - Temporal consistency of 3D camera motion and 3D scene structure
  - Combination of depth and motion analysis depending on the dynamic properties of the scene i.e. moving camera/motion in the scene or both
  - Classification of 3D structures and 3D camera motion and transformation into semantic concepts
- long term
  - Combination of visual information and structural information derived from 3D scene structure
  - Usage of associated 3D information such as depth information form other sensors or generalized disparity maps for search and retrieval
  - Transfer of sparse 3D models to more complex ones
  - Extension of 3D camera motion and 3D scene analysis towards unconstrained scenarios with a mixture of motions by the camera and within the scene

4.1.5.2 CHALLENGES

- Research progress in the two independent communities Computer Vision and Information Retrieval (IR) needs to be brought together
- The whole set of 3D properties of a video scene i.e. motion of the camera and 3D structure of a scene needs to be embedded in novel semantic concepts
- Unified syntax for 3D data/depth is required to allow standardized processing and analysis
- Bridging the gap between 3D data/depth and semantics. How can 3D scene properties be described with meaningful semantics?
- Lack of training data and ground truth needs to be sorted out. Performance evaluation of new approaches requires serious ground truth data
- Computational efficiency becomes again relevant due to the complex analysis algorithms

4.1.5.3 PROMISING DIRECTIONS

- User-centred research
- Exploitation of projective geometry
- Combination of structure-from-motion, motion estimation and multiple view geometry
- Clustering of 3D data/depth

4.1.6 Multimodal Analysis: Limitations and Challenges

Current approaches of extracting information from semantic-multimedia content are based on a scenario that involves several processing steps with an associated loss of information. In this setting the most relevant points of information loss are:

- Data and annotations: the traditional semantic-multimedia retrieval setting is defined by some training data and the corresponding annotations. This is a simple and traditional model that has been explored in different areas and the lack of training data has many times been pointed out as the overriding issue. However, simple annotations are an oversimplification of the problem domain: information semantics are more complex than simple binary annotations of examples. The large amount of weakly-labeled data available from different sources, e.g., Wikipedia, Flickr, or news sites, limits the usability of the traditional setting.
- Keyword vocabulary: several social-media applications allow uncontrolled vocabularies for content labeling which affords users more creativity and expressivity. This risks both lowered quality since no guarantee of the benevolent intent or competence of other users can be had and also wider semantic gaps, since no guarantee of identical or comparable perspectives between users can be had. In a controlled vocabulary such guarantees are inherent in the system, but at a cost: training users to understand the controlled vocabulary can be a threshold for infrequent or inexperienced users.
- Low-level representations of information: computer vision, audio processing and natural language processing all apply a compressed representation of information to enable information extraction algorithms. In all cases, there is a significant loss in the semantics preserved in these low-level representations of data. Thus, richer low-level features that preserve the most relevant information are required.
- Machine learning: techniques that allow computers to learn a given task have not yet achieved the level of human understanding and perception required by semantic multimedia retrieval. Both similarity functions and learning algorithms are tied to the low-level representation of information, thus suffering the same type of limitations.
- Lack of bootstrapping mechanism: Many of the retrieval modules are trained for a specific domain. When these modules should be applied in new domain, which are not covered during the training/learning phase the

performance degrades significantly. Therefore fast adaptation and bootstrapping mechanism are needed to achieve reliable retrieval performance for the new domain without collecting and annotating huge amount of reference data

- Context information to reduce the domain space: For large domains (i.e. internet search) the variability of data is huge. It is well known that the retrieval modules show better performance, if the application domain is more limited. Semi-automatic information about the current context and the domain can improve the retrieval rates. Therefore the detection of the context using a specific information type will reduce the domain variability for a retrieval task using further multimodal retrieval methods. To achieve this domain detection and limitation automatic methods are required to identify and classify specific domains.

As can be inferred from the above list, semantic-multimedia information retrieval combines many different types of expertise and with that it also inherits the different limitations of each area.

The limitations pointed out above are currently addressed in research and the following research challenges have been identified:

- Users and Social Media. In recent years multimedia retrieval research has focused on the ambitious task of extracting meaningful information from semantic multimedia. Large efforts were allocated to this task, and little effort has been dedicated to the understanding of real user problems (Forsyth 2001). These concerns have again been voiced recently at a talk at the British Computer Society by van Rijsbergen (2007). Fully automatic analysis algorithms have been pursued feverously by the community. However, a number of social-media applications have successfully inserted the user in the loop, giving evidence that semi-automatic methods are adequate in these scenarios. Social-media applications like Flickr, YouTube7, del.icio.us8, IMDB9, Wikipedia and others have strategically put users in the information processing loop where they are constantly providing valuable feedback. This new setting contrasts greatly with the classic multimedia retrieval model and motivates users to cooperate as a large community. Understanding the possibilities of the new problem setting allows scientists to work on solutions that can help users and bring more success to the area of semantic multimedia retrieval.
- Large-Scale Data Resources. While the traditional problem setting favors supervised methods that model labeled data, the new problem setting makes available large amounts of unlabeled data that create a demand for a new breed of unsupervised algorithms. The objective of these algorithms should be the deduction of a knowledge base concerning the way users perceive and interact with semantic-multimedia information.
- Weakly-Labeled Data Resources. Another challenge derives from the community aspect of the new setting. A problem is the strong dependence that today's algorithms have on the quality of annotations. Multimedia is currently available with different comments, tags, links and other information that multiple users assign to a given document. This community effect provides algorithms with many different relevance judgments that should be exploited.
- Cross-Media Information. The new multimedia retrieval scenarios combine many different types of information sources with different semantics. Many other information sources are available in multimedia, (e.g., authorship, location, event), capturing device characteristics (e.g., lenses depth of field). New algorithms must cope with the multitude of information sources and with the increased complexity and heterogeneity that they exhibit.

4.2 Query preparation: Establishing an information need

4.2.1 Information Seeking Strategies

How users approach the formulation of their information need can be given various different analyses. The think tank processes organised by the CHORUS project and the questionnaire study presented in chapter 3 both are congruent with previous research on user behaviour. A foundational analysis of information seeking behaviour is made by Belkin et al (1995), drawing distinctions between browsing or scanning a collection on the one hand, and searching for a item which is known or predicted to exist on the other; between specifying the information need in some terminology and recognising appropriate items as they appear on the other. These distinctions, and other similar ones, are reflected in the use case analysis given in the preceding section 3 - Cf.Q10 of the questionnaire, which indicates that the project prototypes envision more than known-item retrieval as an access method and Q11 which shows that the projects intend their technology to be useful for more than retrieving single known items. The think tanks have at various occasions stressed the need for distinguishing between lean-forward interaction, where the user takes the initiative to actions for obtaining information, and lean-backward interaction, where the user is served interaction opportunistically by the system, not based on user requests but on modelling user actions. These various distinctions serve as a functional basis for the design of most information access systems in play.

The interaction between user and system can be designed to conform to various usage scenarios: browsing, searching, exploring, formulating recommendations, receiving recommendations - and naturally, this is not an exhaustive list. The interaction can be based on system-initiated recommendation rather than user-initiated search, on likeness to examples

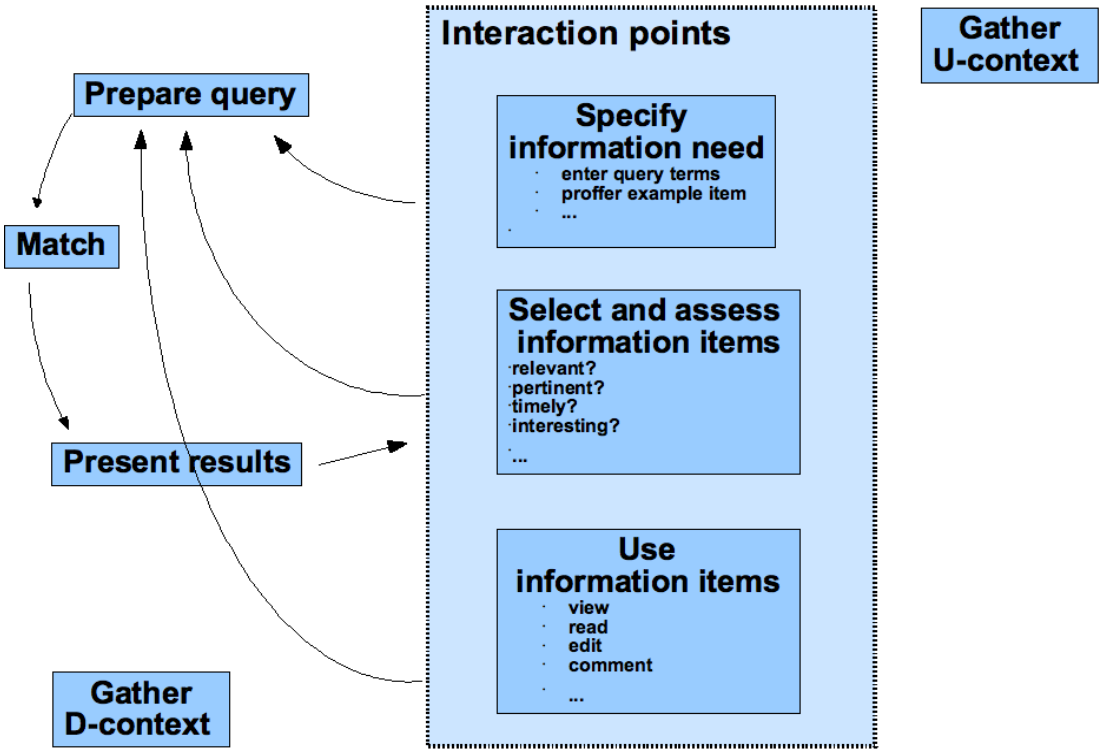
rather than goal-directed matching to queries, on satisfying rather than optimising needs of the users. Referring back to section 3 on use cases, we can see how different envisioned use cases and attached scenarios predict different search strategies, and implicitly have ramifications for appropriate session design and for the attendant formulation of information needs.

Typical information access systems provide three main points of interaction for the information searching process:

1. specifying information need,
2. examining sets of information items to find the appropriate ones for the information need in question, and
3. using the information item in question to fulfil some known information need.

Figure 4.2 shows the interplay of user-specified information need (corresponding to prototypical query formulation) with feedback from the result presentation and information usage stages: many system architectures which instantiate the functional description of chapter 2 in various system designs will design the interaction model in other ways, e.g. by using observed data on user behaviour or historical data from categories of users to prepare a query without overt user supervision. Section 4.2 below will return to this issue. A central and well-established functionality in this respect is that of relevance feedback, where users are allowed to manipulate system suggestions for queries or elaborations to user-formulated queries. This is a technique which makes explicit the interaction between system index and query, and which has been found quite valuable in certain settings.

Figure 4.2 User-specified information



In a typical text information retrieval setting the first step involves the user entering some search terms, the second scrolling through a ranked list of items selecting some subset for further perusal; the third retrieving the text item and reading it. [Oard 2001, e.g.] A ranked list of retrieved items is typically composed of summary representations of the information items: title, date, and possibly a brief extract of the item content. Both the task of specifying an information need and selecting among items, given the summary information, is non-trivial in the general case even for a knowledgeable user. It involves establishing what the information system contains, what the user needs, and what the system competence is to specify, match, and provide for the need. Figure 4.2 above illustrates some of the dependencies between the actions the system performs and further retrieval from the information items the system knows about.

The most typical search engine is built to cater for searching rather than browsing, and is based on user specified information needs rather than prompted recognition of relevant items. However, this is not a necessity in the general case. Given a set of information items previously in use, they can be used to extract an information need from the user actions. Examples can be extracting content descriptors from items viewed by a user, or tracking usage of items, relating similar

usage patterns to each others. Given a representation of user need, this can be entered into the information access system as a query or presented to the user for refinement, depending on system and representation setup.

In fact, today, most information access is not performed by search in specific information access systems, but by perusing ones document collections, by scanning over the content of folders and directories, or even by happenstance. This is partially reflected in the results from the project questionnaire, which shows both recognition and specification as query formulation methods for the prototype systems.

4.2.2 Multimedia accelerates move to different usage models

This movement from user-specified to system inferred information need is hastened by the move from text to multimedia for several reasons: multimedia information access brings with it a whole series of alternative formulations of information access.

First, as has been discussed in preceding sections, multimedial information items are not as simply segmentable into meaningful micro-items as are texts. In the case of texts, words provide simple cues to text content. No words are available in the case of e.g. images. Enabling indexing and search is a greater challenge for non-textual items from reasons of knowledge representation. Several projects use existing proxies for information such as user-contributed comments or tags; others use contextual data mined from user behavior or meta-data harvested from extraneous knowledge sources; some perform content analysis of various kinds. The discrepancy between indexing vocabulary and user formulation of information need is in the general case larger in a multi-medial usage context than in a word-based text retrieval context. The translation from the information need which motivates users to use the system to the system-internal representation is a non-trivial step in itself, adding to the complexity of query formulation. (Cf. section 2.7 above, and Q9 of the questionnaire).

Secondly, multimedial information sources are not as topically focused as the collections of textual data that are targets for text retrieval systems. To provide useful systems, it is necessary to better understand a broader spectrum of usage rationale. For usage which is more directed towards less urgent entertainment rather than fulfilling timely information needs, user needs is less obviously modellable in terms of concise search queries. Further, while some video material may by topically analysable: newscasts, instruction clips, or lectures, e.g. others are intended to provoke a sensation or provide momentary enjoyment - finding common content features over a set of such materials is not obviously possible before the fact.

4.2.3 Information access in general is moving from retrieval of known items to other contexts

Information access, currently most encountered in the form of retrieval systems geared towards explicit search of document collections is merging into more specialised contexts. These contexts may not be natural hosts for information retrieval in the prototypical sense. While the questionnaire results show that the current project prototypes all are built to accept explicit rather than implicit queries (Q13), the information access inherent in some specialised task or usage context may be more naturally realised implicitly rather than explicitly; users may not even necessarily be aware of the need formulation, matching, and retrieval processes underlying the functionality the system serves them. Referring back to section 2.7 we see how the functional architecture will receive queries however they have been formulated: the back-end information retrieval system does not need to be rebuilt if the query entry has been designed in a suitably modular manner.

4.2.4 Lowered publication threshold

The lowered publication threshold and true, if not completely symmetric, bi-directional communication technology allows users to contribute content. Data sources will be greater in number and more heterogeneous than before. There will be more personal and user contributed data, some intended for public use, but also information repositories intended for personal information usage and management. The needs for information access in a structured repository are much different from those in a collection where no quality control and no editing is performed.

Several recent projects work with the question of how to provide a framework which

1. encourages and motivates users to contribute to shared information systems and
2. provides guidance, quality assurance, and a shared semantic space to contributions,

better to build a common body of knowledge. Much of the user-generated content will be used the same way that professional content is, but the models for data quality, persistence, archival character, intellectual rights issues differ importantly.

In a specialised task or usage situation, contextual factors will be an important facet of formalising the information need of the user. The context itself may carry crucial information which must be made explicit to the system and utilised in the matching process to serve the appropriate information to users, rather than something which is considered to be noise and must be abstracted away from. It is worth noting that while the functional description given above in section 2 is primarily built with to address the capture of topical information, it in no way limited to extracting, indexing, querying, and matching based on topical data only. The model straightforwardly handles non-topical characteristics of the information items at hand. Features such as positional information, sourcing, pricing, stylistic characteristics, granularity or other not primarily

topical information may be as salient as topical content for some information access scenarios. The questionnaire gives an indication that many projects in fact make use of various contextual information: positional, personal, platform data etc. (Cf Q8 of the questionnaire.)

4.2.5 Character of representation: transparency, maintenance, enrichment and refinement

The representation scheme used as a surrogate for the information items in the index is intended to be maximally functional with respect to the information needs the system is intended to cater to. For some system setups, this representation needs to have sustainable and archival qualities over long periods of time, or over heterogeneous user groups; for others it can be allowed to change and evolve dynamically. In either case, usage information will be a crucial factor in assessing the success of the representational scheme. A system will need to communicate its capabilities to its clients; it will need to be aware of the potential inherent in usage as a sign-off for the qualities of its representational scheme; it can in some cases use that information to further enrich its index.

4.2.6 Challenges for future systems with respect to establishing information needs of users

Establishing information needs of users for multimedia systems must be sensitive to all the specific challenges in play: content representation, contextual factors, use case and session design, rationale for use, and further lowered publication thresholds. Most systems being designed, developed, and deployed today have variously informed approaches in this respect. Some are very specific to the perceived and understood needs of the targeted user community; others have modeled themselves on previously established practice without attempting to make new advances in this respect. Given the use case analysis of section 3 above and the functional description of section 2 above, future projects should be encouraged to make explicit their assumptions with respect to use cases, thus setting several of the parameters necessary for further system design.

The first crucial question for achieving future impact is understanding why systems with excellent user experience qualities fail, whereas systems with less accomplished design features can go on to become successful services even for broader audiences. To do research and development projects must be able to share research results; for this, general models for interaction with information acccess must be developed. The CHORUS project provides the functional description of chapter 2 and the use case analysis of chapter 3 to enable such leveraging of previous and parallel results.

The second, more technical challenge is how to achieve a knowledge representation which cuts across cultural divides and use cases to enable general information access performance; how to frame mid-level semantic descriptions that empower users to provide annotations and other resources to extend and become useful.

The third, more behavioural challenge, is how to study information access in non-topical and non-task settings, how to build interaction models based on those studies, and how to build sustainable evaluation schemes based on non-topical concerns.

REFERENCES

Nicholas J Belkin, Colleen Cool, A Stein, and U Thiel. 1995. Cases, scripts and information-seeking strategies: On the design of interactive information retrieval systems. Expert Systems with Applications, 9 (3).

Peter Ingwersen, Kalervo Järvelin. 2005. The Turn: Integration of Information Seeking and Retrieval in Context. Berlin / Heidelberg: Springer.

Douglas W. Oard. 2001. ``Evaluating Interactive Cross-Language Information Retrieval: Document Selection'' In: Cross-Language Information Retrieval and Evaluation. Carol Peters (ed.) Lecture Notes in Computer Science Berlin / Heidelberg: Springer. Pages 57-71

Andorra Conference Report (Deliverable 4.x of the CHORUS Project)

4.3 Organisation and navigation in result content

Presentation of search results is perhaps one of the most important factors in the acceptance and popularity of any retrieval system. Datta et al [2008] characterize common visualization schemes for image search as follows.

- **Relevance-Ordered**. The most popular way to present search results is relevance ordered, as adopted by Google and Yahoo! for their image search engines. Results are ordered by some numeric measure of relevance to the query.
- **Time-Ordered.** In time-ordered search, multimedia documents are shown in a chronological ordering rather than by relevance. Google's Picasa system for personal collections provides an option to visualize a chronological timeline using pictures.

- **Clustered.** Clustering of images by their metadata or visual content has been an active research topic for several years. Clustering of search results, besides being an intuitive and desirable form of presentation, has also been used to improve retrieval performance.
- **Hierarchical.** If metadata associated with images can be arranged in tree order (e.g., WordNet topical hierarchies), it can be a very useful aid in visualization. Hierarchical visualization of search results is desirable for archives, especially for educational purposes.
- **Composite**. Combining consists of mixing two or more of the preceding forms of visualization scheme, and is used especially for personalized systems. Hierarchical clustering and visualization of concept graphs are examples of composite visualizations.

Study of organizations which maintain image management and retrieval systems has provided useful insights into system design, querying, and visualization. The final verdict of acceptance/ rejection for any visualization scheme comes from end-users. While simple, intuitive interfaces such as grid-based displays have become acceptable to most search engine users, advanced visualization techniques could still be in the making. It becomes critical for visualization designers to ensure that the added complexity does not become an overkill.

In order to design interfaces for retrieval systems, it helps to understand factors like how people manage their digital collections or frame their queries for visual art images. Several user studies on various ways of arranging images for browsing purposes were conducted, and the observation is that both visual-feature-based and concept-based arrangements have their own merits and demerits.

Thinking beyond the typical grid-based arrangement of top matching images, spiral and concentric visualization of retrieval results have been explored by several researchers. For personal collections, innovative arrangements of query results based on visual content, time-stamps, and efficient use of screen space add new dimensions to the browsing experience.

Portable devices such as personal digital assistants (PDAs) and vehicle communication and control systems are becoming very popular as client-side systems for querying and accessing remote multimedia databases. Portable-device users are often constrained in the way they can formulate their query and interact with a remote image server. There are inherent scrolling and browsing constraints which can constrict user feedback. Moreover, there are bandwidth limitations which need to be taken into consideration when designing retrieval systems for such devices. Some additional factors which become important here are the size and the color depth of display. Personalization of search for small displays by modeling interaction from the gathered usage data is an important research direction. Also, equally important is the research of developing image attention models for adapting images based on user attention for these small displays as well as finding efficient ways of browsing large images interactively, such as those encountered in pathology or remote sensing, using small displays over a communication channel.

Image transcoding techniques, which aim at adapting multimedia (image and video) content to the capabilities of the client device, have been studied extensively in the last years. This class of methods known as semantic transcoding aims at designing intelligent transcoding systems which can adapt semantically to user requirements. For achieving this, classes of relevance are constructed and transcoding systems are programmed differently for different classes.

The abovementioned issues apply mostly to the case of image-based systems but in the case of video or multimedia documents there are several other important aspects to be considered. First, how is the video content itself visualized? Second, how is the structure of the video collection presented and how can users navigate this structure?

The basic means for visualizing video content is of course a generic video player, but for a video search system we need different techniques especially when the video items are not small coherent clips, but are programs of half an hour or more. Watching the entire video is not an option, therefore many people have developed techniques to reduce the length of the video and keep the relevant information only. Extensive overviews of methods for summarization are presented in [Money and Agius 2007] and [Truong and Venkatesh 2007].

The most common way of visualizing the content is by a set of static key frames, typically 1-3 per shot. Most of the existing browsers can be mapped to specific values for the following dimensions in the design space for key frame based video browsing systems: the layeredness, the spatial versus temporal presentation, and the temporal orientation.

Simple storyboard methods are actually very effective. Rather than taking the shot as basis, several researchers visualize the story units in the video as a collage of key frames. This can be done by exploiting named entity extraction on the automatic speech recognition results to map the key frames to the associated geographic location on a map and combine this with various other visualizations to give the user an understanding of the context of the query result [Christel et al 2002]. As an alternative, a very elaborate video browsing environment based on summaries is presented by [Haubold and Kender 2007]. It includes mechanisms for showing the timeline, all faces appearing in the video, speaker changes, and visual segmentation cues, which in their application are for example slide changes in a video lecture.

Almost all systems take a user query as the basis for visualization. As the query yields a ranked list, a linear set of key frames are presented to the user. Commonly, the result is paginated and each page is displayed in a 2-dimensional grid. In the grid the linear ranking of results is mapped to left-right top-down reading order. To go through the results the user just flips through the various pages. In [Hauptman et al 2006] an extreme approach is followed where the pages are dynamically presented in a very rapid fashion to the user which only has controls to indicate whether the results are correct or not. Rather than assuming reading order, an alternative is to employ the 2-dimensional nature of the grid by computing optimal data-driven coordinate axes. Thus in every page of results, clusters of visually related key frames appear. Another option is to use a true 2-dimensional display, but not to rely on a grid. Instead, key frames can be placed on the screen in such a way that the dissimilarity among the features of the key frames are optimally preserved using various projection methods. In this case, the data is organized in a hierarchy to facilitate browsing at the overview level, viewing one representative frame per cluster, as well as more detailed viewing of individual clusters.

A disadvantage of the query-based methods is that users have to go through the result list and have to switch back to the query screen when they cannot find more relevant results. To break this iterative query-view approach the systems should give options to depart from the initial result presented and to delve into local browsing opportunities. A solution is to present the timeline of the video on the horizontal axis [Rautiainen et al 2004] and to use a full grid where the vertical dimension shows similar shots for every key frame in the timeline. The browser presented by [Snoek et al 2007] does not use a full grid but shows the ranked results as a linear list using the other dimension for showing the timeline of the video the current shot is part of. One can then take this a step further and give multiple different ranked lists starting at the current shot to explore, including dimensions for visually similar shots, but also for shots that are similar given their concept scores [de Rooij 2008]. As such, new results are found by browsing along different linear lists, instead of entering new queries. As an alternative to using linear lists it is possible to visualize relations between shots with a hyperbolic tree, allowing for hierarchical browsing [Luo et al 2007].

It is important to stress that visualizations give the user insight in the video collections and the results of the query, but that the main purpose of visualization is to provide the means for effective user interaction.

4.3.1 Learning from the User

An actively pursued direction in image retrieval is to engage humans in the searching process, that is, to include a human in the loop. Although in the very early days of content-based image retrieval, several systems were designed with detailed user-preference specifications, the philosophy of engaging users in recent work has evolved toward more interactive and iterative schemes by leveraging learning techniques. As a result, the overhead for users, in specifying what they are looking for at the beginning of a search, is much reduced. For video retrieval the user performs two types of interactions with the system: selection of relevant/irrelevant results and commands to navigate the collection of results.

Good overviews of generic methods for learning from user interaction are presented in [Zhou and Huang 2003, Huang et al 2008]. We make a general distinction between methods based on relevance feedback and active learning. In relevance feedback the system uses the current sets of relevant and irrelevant results to update the model it has built for the particular query. The system then presents the next set of results and the process reiterates. Active learning does not only optimize the model, it also computes which elements from the unlabeled data pool are most informative. Hence, the system picks these for presentation to the user.

Relevance feedback provides a compromise between a fully automated, unsupervised system and one based on subjective user needs. While query refinement is an attractive proposal when considering a very diverse user base, there is also the question of how well the feedback can be utilized for refinement. Whereas a user would prefer shorter feedback sessions, there is an issue as to how much feedback is enough for the system to learn the user needs. One issue which has been largely ignored in past research on relevance feedback is that the user's needs might evolve over the feedback steps, making weaker the assumption of a fixed target. New approaches such as [Jaimes et al. 2004] and [Fang et al 2005] have started incorporating this aspect of the user's mind in the relevance feedback process.

While the progress in relevance feedback research is evident, the issue remains that, we do not see many real world implementations of the relevance feedback technology, either in the image or text retrieval domains. This is potentially due to the feedback process that the users must go through, which tests the users' patience. New ideas such as memory retrieval, which actually provide the user with benefits in the feedback process, may possibly be a key to popularizing relevance feedback. The future of this field clearly lies in its practical applicability, focusing on how the user can be spared the greatest amount of effort in conveying the desired semantics. The breaking-point and utility derived out of this process, at which users runs out of patience and at which they are satisfied with the response, respectively, must be studied for better system design.

Chen et al [2005] were among the first to apply active learning in concept-based video retrieval. They select the elements closest to the current boundary between the relevant and non-relevant items as these are the most uncertain elements. This

is then combined with relevance feedback to improve precision in the top-ranked list. Their results indicate that this approach often leads to a focus on a small part of the information space. They therefore propose to perform the active learning for different modalities and combine them in the next phase. Important here is to note that different semantic concepts require different elements to be judged by the user, and different sampling strategies. The system in [Luan et al 2007] takes such an adaptive sampling strategy to obtain a balance between seeking high precision and high recall. They start off by presenting elements far away from the decision boundary, elements most likely to be relevant to the query. If the user feedback indicates that those elements are indeed relevant they continue sampling this part of the information space. When the number of elements labeled as relevant by the user becomes too small they start selecting elements close to the boundary to update the current model, hoping to increase recall by finding new areas to explore. Another option is to choose elements close to the boundary but in an information space based on distances to a number of selected prototypes. By doing this and using a reasonable number of prototypes, a distance preserving projection to 2-dimensional can be used as a good approximation to the original space. This has the advantage that the decision boundary and the selected examples can be shown to the user in a manner leading to intuitive feedback.

Although interaction was recognized early as a key element, the above interaction methods only started to unlock the potential that intelligent interaction has for video retrieval. A thorough understanding of the relation between the space spanned by different concepts, the proper sampling strategy, intuitive displays and effective navigation of this space is urged for.

As query-by-concept is close to what we are used to in traditional search engines, it is the query method of choice for video retrieval whenever the quality of the detectors is sufficiently high. Automatic concept detection methods are getting to the point that for a given information need in many cases they will select the same concept a human would. The methods are not yet capable of employing common sense, associating the information need to objectively spoken a not directly related concept. For example, how to teach a system to select the “political leader” concept when searching for “people shaking hands”? Furthermore, how a system can decide on the semantic and visual coverage, while simultaneously taking the quality of available detectors into account is an open problem.

In any case, interaction will remain a part of any practical video search system. So rather than considering the interaction as a method to counteract any errors of the automatic system we should incorporate the interaction right into the design of the system. To that end we fortunately see a large body of work in the machine learning community on optimal active learning based techniques. These should be adapted to deal with the large volumes of data in video and the varying quality of some of the concept detector results. In addition, visualizing the query results in an intuitive way is as important in the design of any video retrieval system. We should develop methods which truly integrate active learning with visualization and do so while conforming to principles from the field of human-computer interaction.

REFERENCES

R. Datta, D. Joshi, J. Li, and J. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” ACM Computing Surveys, vol. 40, no. 65, pp. 1–60, 2008.

A. G. Money and H. Agius, “Video summarisation: A conceptual framework and survey of the state of the art,” Journal of Visual Communication and Image Representation, vol. 19, no. 2, pp. 121–143, 2008.

B. Truong and S. Venkatesh, “Video abstraction: A systematic review and classification,” ACM Transactions on Multimedia Computing, Communications and Applications, vol. 3, no. 1, 2007.

M. Christel, A. Hauptmann, H. Wactlar, and T. Ng, “Collages as dynamic summaries for news video,” ACM Multimedia, pp. 561–569, 2002.

A. Haubold and J. Kender, “VAST MM: multimedia browser for presentation video,” in ACM International Conference on Image and Video Retrieval, pp. 41–48, 2007.

A. Hauptmann, W.-H. Lin, R. Yan, J. Yang, and M.-Y. Chen, “Extreme video retrieval: Joint maximization of human and computer performance,” in ACM Multimedia, pp. 385–394, 2006.

M. Rautiainen, T. Ojala, and T. Seppanen, “Cluster-temporal browsing of large news video databases,” in IEEE International Conference on Multimedia & Expo, 2004.

C. G. M. Snoek, M. Worring, A. W. M. Smeulders, and B. Freiburg, “The role of visual content and style for concert video indexing,” in IEEE International Conference on Multimedia & Expo, pp. 252–255, 2007.

O. de Rooij, C. G. M. Snoek, and M. Worring, “Balancing thread based navigation for targeted video search,” in ACM International Conference on Image and Video Retrieval, pp. 485–494, 2008.

H. Luo, J. Fan, J. Yang, W. Ribarsky, and S. Satoh, “Analyzing large-scale news video databases to support knowledge visualization and intuitive retrieval,” in IEEE Symposium on Visual Analytics Science and Technology, pp. 107–114, 2007.

X. Zhou and T. Huang, “Relevance feedback in image retrieval: A comprehensive review,” Multimedia Systems, vol. 8, no. 6, pp. 536–544, 2003.

T. Huang, C. Dagli, S. Rajaram, E. Chang, M. Mandel, G. Poliner, and D. Ellis, “Active learning for interactive multimedia retrieval,” Proceedings of the IEEE, vol. 96, no. 4, pp. 648–667, 2008.

A. Jaimes, K. Omura, T. Nagamine, and K. Hirata K, “Memory cues for meeting video retrieval,” In ACM Workshop on Continuous Archival and Retrieval of Personal Experiences with ACM Multimedia, 2004

Y. Fang, D. Geman, and N. Boujemaa, N., “An interactive system for mental face retrieval. In ACM International Workshop on Multimedia Information Retrieval with ACM Multimedia, 2005.

M.-Y. Chen, M. Christel, A. Hauptmann, and H. Wactlar, “Putting active learning into multimedia applications: Dynamic definition and refinement of concept classifiers,” in ACM Multimedia, pp. 902–911, 2005.

H.-B. Luan, S.-Y. Neo, H.-K. Goh, Y.-D. Zhang, S.-X. Lin, and T.-S. Chua, “Segregated feedback with performance-based adaptive sampling for interactive news video retrieval,” in ACM Multimedia, pp. 293–296, 2007.

4.3.2 Data Organization

In this section we will deal with data organization and presentation by discussing and comparing faceted categorization with ontology based classification. More specifically, an analysis of existing approaches that adopt semantic web representation will take place revealing in that way the potential of ontologies in the area of content representation, data presentation as well as user interface interaction and navigation.

4.3.2.1 FACETED CLASSIFICATION

A significant refinement of the hierarchical scheme has been demonstrated using faceted categories. Faceted classification is a method that groups sets of hierarchical metadata into categories. Yee et. al. [13] report that users express a strong preference for faceted classification over Google's relevance-ordered scheme even when the prototype interface for the faceted system was an order of magnitude slower than the baseline. Although faceted classification has historically been performed manually by trained librarians, Hearst et. al. [12] have made significant progress in automating this task. Hierarchical faceted classification is now common in many e-commerce sites such as Buy.com, eBay.com and Shopping.com.

The DARE method suggested by Frakes et. al. recommends building a thesaurus for each facet [14]. Indeed, currently this is common practice in the industry. However, a significant drawback to this approach is that there are no formalized relationships between facets. Since these relationships are not encoded, they cannot be displayed, merely implied. Moreover, the meanings of the facets themselves are open to interpretation. Users typically guess what is meant using cues such as the name of the facet and it's context. However, users make mistakes. Ontologies offer a solution. By anchoring faceted categories to an ontology, relationships between facets can be formalized and expressed through the interface, enabling more sophisticated browsing capabilities.

4.3.2.2 ONTOLOGY-BASED DATA ORGANIZATION

In order to offer meaningful and efficient presentation of data, a proper organization of the content has to take place. This means that content representation techniques adopted for the structuring and the organization of a digital library is essential. Although faceted categorization of content has been shown to be satisfactory, recent research demonstrates that Semantic Web technologies could be even more effective in terms of describing and linking parts of the content.

The Semantic Web essentially advocates “crossing the chasm” from unstructured keyword-based models to richer logic-based annotations that would eventually provide a basis for reasoning. This entails that the logical model of a document becomes a set of logical assertions (annotations) about its contents (and perhaps also about its physical structure, its relationships with other documents and other meta-information). In addition, the form of the queries becomes a logic expression with an arbitrary level of complexity in its structure. [1]

The Semantic Web uses ontologies as a means for a formal and explicit representation of content. In the sense coined by the field of Artificial Intelligence, an ontology is “an explicit and formal specification of a conceptualization of a particular domain of interest” [2]. Thus, ontologies provide a way of capturing a shared understanding of a domain that can be used both by humans and machines to support information exchange and integration and to facilitate interoperability between engines. Problems caused by structural and semantic heterogeneity of different models can be avoided. Furthermore, ontologies are an effective means for making implicit system design decisions and underlying assumptions explicit. This makes it easier to reason about the intended meaning of the information.

An ontology consists of a list of terms and relationships between these terms. The terms denote important concepts (classes of objects). The relationships include hierarchical class relations which specifies a class C to be a subclass of another class C if every object in C is also included in C. It is important to note that ontologies not only model content, but also add to the content modelling expressiveness and reasoning capabilities. Ontology rules provide a way to define behaviour in relation to a system model.

4.3.2.3 ONTOLOGY-BASED DATA PRESENTATION

Ontology structured content allows data presentation based on the hierarchical structure of concepts. This means that a number of different classification schemas could be achieved. At first glance, such a categorization would not reveal the power of ontologies, as a similar presentation could be possible with hierarchical tree-structured facets. However by taking a closer look, ontologies have the ability to present multiple browsing hierarchies to the end user. Ontologies generally contain a richer set of relations between concepts and can exploit any subset of those relations for presentation at the user interface. Classical hierarchies that are driven by taxonomies or thesauri, on the other hand, have fewer or less well-defined relationships. Taxonomies are limited to superordinate and subordinate relations, whereas the semantics of the related terms encoded by thesauri are vague. As such, these tools are capable of offering only a single hierarchical view into a collection. The different hierarchical views within an ontology can be triggered by various contextualization methods, including user profiles, location and time data.

Ontologies are not only capable of improving upon the hierarchical visualization scheme, but the clustering scheme as well (if it is based on metadata) by using the technique discussed above. That is, multiple types of clusters can be triggered by exploiting subsets of concept relations within an ontology. Furthermore, the unique visualization cues inherent in clustering means concept relations may be used to enhance data representation in interesting and possibly more intuitive ways, such as highlighting certain classes of concepts in various colours or textures.

In addition, as discussed in the faceted classification section, ontology-based semantics offer the additional value of explicitly representing the relationships between concepts, which are otherwise ambiguous and must rely on guesses by the user. This should facilitate a better browsing and search experience.

A final, and rather interesting, application of ontologies is in their use as cross-domain references. That is, links with other external ontologies could be established in order to provide navigation paths between domains that may be of interest. It should be noted that the use of ontologies for the annotation of Web resources is not limited to categorization processes, as those that have been extensively used in Web catalogues like the popular Yahoo! [3].

4.3.2.4 ONTOLOGY-BASED RESULT SET REFINEMENT

Although the presentation of results may not gain significant advantages from semantic representation, further querying and filtering capabilities, as well as search and browsing options, could be integrated in a search engine interface to take advantage of the full potential of ontologies. Ontologies can be exploited to support retrieval in several ways.

One of the possible approaches is the employment of an interface in which users type terms and use ontologies later for the expansion of the query. Another work is described in [5] where the interface supports concept filtering and browsing while query expansion is also realized with the suggestion of semantically related recommendations. Based on the existing state of the art another more complicated approach is suggested in [4]. In this case the query formulation process is intentionally iterative, since search proceeds through guided navigation of the knowledge stored in ontologies. The relationships among terms in the ontologies are used to build the user interface, and the system gets actively involved in suggesting alternative paths or possibilities that end up in a query comprised by the navigation and selections of the user.

As far as the navigation is concerned, a number of tactics have been applied to different search engines. As stated in [7], the user prefers to control the search procedure by using certain search strategies and tactics that are present in the interface. Tactics are defined by Bates as “one or a handful of moves made to further a search”. As strategies are considered concrete plans for search that may combine several tactics and have a far-reaching scope. A summary of these tactics that can be effectively supported by ontologies are presented below:

- SUPER: To move upward hierarchically to a broader concept
- SUB: To move downward hierarchically to a narrower concept
- RELATE: To move sideways hierarchically to a coordinate concept
- CONTRARY: To search for a term logically opposite
- SELECT: To break complex search queries down into sub-problems
- PARALLEL: To make the query broader including synonyms or otherwise conceptually parallel terms.
- PINPOINT: To make the query narrower by reducing the number of parallel terms
- SPECIFY: To search on terms that are as specific as the information desired.
- EXHAUST: To include more that one concepts in the query.

The SUPER and SUB tactics are straightforward navigations that allow browsing the hierarchy of the ontology, and have been applied yet in a number of systems like in [8]. The RELATE tactic could be applied with the aid of ontologies and broaden queries by using concepts with the same "father". The CONTRARY tactic could be a complicated task for ontologies as antonyms are not often explicitly stated. The PARALLEL tactic exhibits similar difficulties, as synonyms are in general not inferable from the ontology. PINPOINT operates on the same information, but in this case allowing the user to reduce some terms that are marked by the system as parallel. The SELECT technique can be implemented by using a child browser window for each branch, thus maintaining the link with the original one for an eventual re-joining. SPECIFY and EXHAUST indicate the need for "and"-like and "or"-like semantics for query formulations including several concepts, so that more specificity or flexibility in selecting search results is provided. Some of these tactics have already been integrated in the interface of several tools [9, 10, 11].

4.3.2.5 SUMMARY OF CONTRIBUTIONS OF ONTOLOGIES

Considering the above, it is derived that ontologies can be used to support a range of well-known user interface search tactics. This means that an ontology based user interface could be very efficient in terms of navigation and query formulation as the hierarchical structure of it, as well as the relations between the concepts, can be exploited by the aforementioned tactics. In addition, existing ontology-based approaches that support browsing, recommendation, hierarchical and cluster presentation reveal the potential of ontologies in terms of content representation, data presentation and navigation. Finally, as information needs to be properly represented through a number of high level concepts, ontologies appear to efficiently support the query formulation and the information retrieval process. Comparison with classical hierarchies driven by taxonomies or thesauri revealed the superiority of ontologies since taxonomies are limited to superordinate and subordinate relations, whereas thesauri are unable of providing semantic relations between the involved terms. In conclusion, recent research activities have provided some significant work, however further research is required to deal with usability and technical issues, including scalability and performance as well as data presentation.

REFERENCES

[1]Elena Garcia and Minguel-Angel Sicilia, "User Interface Tactics in Ontology-based Information Seeking", PsychNology Journal, Vol 1, Num 3, pp 242-255, 2003

[2]"Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce", D. Fensel, 2001, Springer

[3]Labrou, Y. & Finin, T. (1999) Yahoo! as an Ontology: Using Yahoo! Categories to Describe Documents. Proceedings of the Eighth International Conference on Information Knowledge Management, 180–187

[4]Andreasen, T., Fischer-Nilsson, J. & Erdman-Thomsen, H. (2000). Ontology-based Querying. In: Larsem, H.L. et al. (eds.) Flexible Query Answering Systems, Recent Advances, Physica-Verlag, Springer, 15–26.

[5]S. Vrochidis, C. Doulaverakis, A. Gounaris, E. Nidelkou, L. Makris and I. Kompatsiaris, "A Hybrid Ontology and Visual-based Retrieval Model for Cultural Heritage Multimedia Collections", 2nd International Conference on Metadata and Semantics Research, 11-12 Oct., Corfu, Greece, 2007

[6]Baader, F., Calvanese, D., McGuinness, D., Nardi, D. & Patel-Schneider, P. (editors). (2003) The Description Logic Handbook. Theory, Implementation and Applications. Cambridge.

[7]Bates, M.J. (1990). Where Should the Person Stop and the Information Search Interface Start?. Information Processing & Management 26, 575-591.

[8]Papazoglou, M.P., Porpoer, H.A. & Yang, J. (2001) Landscaping the Information Space of Large Multi-Database Networks. Data & Knowledge Engineering, 36(3), 251–281.

[9]Sicilia, M.A., Garcia, E., Aedo, I., & Diaz, P. (2003). A literature-based approach to annotation and browsing of Web resources. Information Research Journal 8(2).

[10]Garcia, E. & Sicilia, M.A. (2003). Designing Ontology-Based Interactive Information Retrieval Interfaces. Proceedings of the Workshop on Human Computer Interface for Semantic Web and Web Applications, Springer Lecture Notes in Computer Science 2889, 152–165.

[11]Garcia, E., Sicilia, M. A., Diaz, P. & Aedo, I. (2003). An Interactive Ontology-Based Query Formulation Approach for Exploratory Styles of Interaction. Proceedings of the 10th International Conference on Human - Computer Interaction (HCI 2003). Lawrence Erlbaum Associates.

[12]M. Hearst, "Clustering versus Faceted Categories for Information Exploration," Communications of the ACM, 49 (4), April 2006.

[12]P. Yee, K. Swearingen, K. Li, and M. Hearst, "Faceted Metadata for Image Search and Browsing," in the proceedings of ACM CHI 2003.

[14]Frakes, W., Prieto-Diaz, R. and Fox C. DARE: domain analysis and reuse environment. In Annals of Software Engineering, (5) 125-141, W. Frakes (Ed.) Baltzer Science Publishers, September 1998.

4.4 Scalability

Today's scalability issues already put brake on growth of multi-media search engines. The searchable space created by the massive amounts of existing video and multimedia files greatly exceeds the area searched by today's major engines. And unfortunately, this will become more and more critical in the next decade: **the amount of raw data is indeed still growing exponentially** and most recent **content enrichment techniques produce more and more heavy features**, even on relatively small datasets. **Consistent breakthroughs are therefore urgent** if we don't want to be lost in data space in ten years.

On the other side, solving scalability issues will also widely **benefit the quality of multimedia search engines**. It is indeed well known that text-based search engines became popular only when the number of managed documents became sufficiently high to improve the precision of the results (often regardless recall performances). It is today unfortunate, that most multimedia search engines using content-based technologies do not work with enough contents to really show the power of such technologies. Consistent breakthroughs regarding scalability issues will therefore also benefit to bridge quality gaps.

Of course, **hardware progress** has an important role to play in the increasing capacities of today's and tomorrow's search engines. However, even if some specialists still predict some good days for Moore's law, up to 2017, hardware solutions will not be sufficient to solve all scalability issues. It is first important to remind that exponentially improved hardware does not necessarily imply exponentially improved software performance to go with it. There are indeed problems when exponential increases in processing power are matched or exceeded by exponential increases in data amount and/or in algorithm complexity as the problem size increases. Furthermore, not all aspects of computing technology develop in capacities and speed according to Moore's law. Random Access Memory (RAM) speeds and hard drive seek times improve at best a few percentage points each year. So, despite hardware progress, the key lies rather in software solutions.

P2P technologies will be another crucial key for solving scalability issues and their use has to be generalized as much as possible. Sharing costs between peers has indeed the major advantage to reduce the amazing amount of unexploited computing resources all over the world. However, P2P technologies themselves do not reduce the intrinsic complexity of the used algorithms. Furthermore, not all technologies and algorithms are easily distributable and the corresponding research challenges have first to be solved before deploying P2P solutions.

Unfortunately, the big challenge of multimedia search engines scalability does not rely on a single bottleneck on which we could invest all efforts. To process the massive scale of new data created every day with more and more complex technologies, **scalability considerations must be taken into account at all stages of the indexing and retrieval workflow**, from content analysis to social tagging to search results organisation. In the following, we explore six main challenges for the next decade.

4.4.1 Breaking algorithms complexity

Complexity remains the keyword when speaking about scalability. Whatever the efficiency of the implementation and the use of powerful hardware and distributed architectures, the ability of an algorithm to scale-up is strongly related to its time complexity and space complexity. Nowadays, efficient multimedia search engines rely **on various high level tasks** such as content-based search, navigation, knowledge discovery, personalization, collaborative filtering or social tagging. They involve complex algorithms such as similarity search, clustering or machine learning, on heterogeneous data, and with heterogeneous metrics. A large part of these algorithms still **have quadratic and even cubic complexities** so that their use in the large scale is not affordable if no fundamental research is performed to reduce their complexities. Ideally, to support efficiently massive growths of data, a scalable search engine should not exceed near-linear complexities for background or batch processes such as indexing and sub-linear complexities for online services such as searching or interactive learning. Furthermore, **besides the amount of input data** and generated features, complexity may need to be reduced for other growing quantities such as number of **users**, number of **information sources,** number of **data attributes** or features **dimension**. A complete study would be required to identify all essential complexity gaps; we give here a short list of some identified complexity reduction open problems, from content description to result content organization:

- **Complexity of content-based features extraction based on large vocabularies:** Representing multimedia contents by sets of words belonging to some content-based vocabularies is an attractive model that allows benefiting from a lot of research results obtained in textual domain. However, computing the content-based words of a single document when using large vocabularies can be an intensive task and the complexity is usually linear in the number of documents and in the size of the vocabulary. Reducing the time complexity of processing a whole database is thus required.
- **Learning complexity of kernel-based machine learning techniques**: In recent years, there has been a lot of interest on using kernels in various machine learning problems, with the support vector machine being the most prominent example. Many of these kernel methods are formulated as quadratic programming with cubic training time complexity and quadratic space complexity.
- **Time complexity of recent clustering algorithms**: Many classical data clustering approaches have now reduced complexities but their performance and applicability often force particular choices of data representation and similarity measure which are problematic for a wide range of modern knowledge discovery and data mining processes. On the

other side, recent algorithms solving these issues, such as shared-neighbour, spectral or kernel-based clustering still suffers from high complexities (usually quadratic).

- **High dimensional similarity joins complexity:** Combining two datasets based on some similarity predicate into one set such that the new set contains pairs of similar objects of the two original sets is an important but very expensive primitive for many multimedia databases mining approaches. Compared to the traditional point-at-a-time approach that computes similar neighbours for all data points one by one, set oriented similarity joins can accelerate the computation dramatically. For now, it does not exist any similarity join algorithm for high dimensional data with proved reduced complexity.
- **Complexity of algorithms combining multiple search results:** Due to the diversity of indexed attributes and used approaches in multimedia information retrieval, generic meta-search engines become more and more essential. Merging and organizing all the results retrieved by heterogeneous information retrieval systems is therefore an important process, for which several algorithms have been proposed. Their complexity for increasing numbers of information sources is however not well studied.
- **Curse of dimensionality**: High-dimensional problems are still difficult to solve for many algorithms for which the complexity increases exponentially with the dimension. This is proved by providing exponentially large lower bounds that hold for all algorithms. Examples of problems suffering from the curse of dimension include nearest neighbour search, optimal recovery of functions, global optimization, partial differential equations and many machine learning problems. Breaking the curse of dimension is thus still a big challenge, particularly for multimedia content descriptions based on large vocabularies (textual or content-based).

4.4.2 Generalizing multidimensional indexes and similarity search structures

Despite the recent development of efficient approximate similarity search structures for high dimensional data, their use in multimedia retrieval systems remains sporadic. Several issues can explain this lack of generalization:

The first problem is that there is a huge literature related to multidimensional indexing and similarity search techniques but **very few comparative experiments** that would allow an easy choice for eventual consumers of these technologies. There is a therefore a need of introducing unique data sets accepted by the whole community for verifying all NN solutions.

The second problem is that there is **no complete enough software** providing all the required functionalities (despite existing theoretical solutions to solve them). A generic similarity search structure should deal with a sufficiently large number of data types, similarity measures and query types. It should allow dynamic and batch indexing and should work on disk or in memory.

Finally, they are **still some fundamental open problems** that prevent the use of similarity search structures in a lot of multimedia indexing and retrieval scenarios. We list here some of them:

- Algorithms with both **linear space complexity** and **logarithmic search complexity.**
- Nearest neighbours for **sparse vectors** in Euclidean space.
- Automatic and/or adaptive **search query parameters** selection.
- Optimisation issues **for heterogeneous queries.**
- **New dynamic aspects:** object descriptions are changing and sometimes even similarity function is modified with time.
- **Relational nearest neighbours**: using graph structure of underlying domain.
- **Probabilistic analysis** for specific domains: introduce reasonable input distributions and solve nearest neighbours for them.
- Low complexity **Reverse nearest neighbours** algorithms.

4.4.3 Large scale evaluations and analysis

Conducting retrieval experiments on common datasets has always been a central part of the Information Retrieval discipline and **test collections** consisting of documents, queries or topics, and relevance assessments for those topics have been around since the earliest days. These tended to be **very small in size** and not universally available but there has always been reasonable comparability among researchers in terms of empirical research work. About 15 years we saw a scale-up on the size of the collections and more importantly a concerted effort at providing relevance assessments on these collections, e.g. with the introduction of the TREC exercises. However **the size of the collections** used today for **multimedia retrieval evaluation tasks** remains **several orders of magnitudes too small** to analyze the effect of **substantial datasets growth**. For a broad range of techniques including retrieval, filtering, summarisation, browsing, classification, clustering, automatic linking, and others, it is therefore unclear whether they are **consistently effective in the large scale**, for **different users**, and under **different search contexts**. Setting up large scale multimedia retrieval evaluations and **analysing the effects of growing datasets** on existing algorithms is therefore an important challenge for the next few years.

4.4.4 Development of technology aware algorithms

Throughout the history of computing space/time trade-offs have always been strongly influenced by the shifting relative costs of CPU cycles versus storage space and speed. For instance, it now makes sense in many cases to trade space for time, such as by precomputing heavy indexes and storing them in ways that facilitate rapid access, at the cost of using more disk and memory space: space is getting cheaper relative to time. Another example is the arrival of new storage supports, such as solid state drive, that will also strongly influenced the optimum trade-offs. Since such changes in relative costs and growths will occur again in the future, it is important to **develop algorithms intelligent regarding computing technologies**.

Automatically estimating optimum trade-offs, empirically or theoretically, **can provide several orders of magnitude gains in performance** and makes technologies **widely more durable**. Very few algorithms used in multimedia search engines perform such optimization for now, even most similarity search and indexing structures that are yet strongly dependent of such considerations.

Another important challenge will be to generalize the parallelization of basic algorithms such as distance computations, convolutions or sorts. It is indeed noticeable that the recent multiplication of multi-core CPUs does not necessarily reflect a similar increase in practical computing power, due to the unparallelised nature of most applications. On the other side, **most time spent by complex systems is usually due to numerous calls of basic operations**. Parallelizing them on several cores will provide important gain factors.

4.4.5 Rationalization of indexing workflows

Most recent low level content analysis methods produce **very large amounts of features**, usually **larger than the content itself** if we want to have a complete description of all media with enough global and local attributes. Storing all this information in the large scale should not be an issue and is even recommended to avoid costly re-computing of lost information. On the other side, **directly post-processing** these features with higher levels content enrichment or structuring techniques will probably **not be affordable** for huge multimedia documents datasets. **Indexing workflows should therefore be rationalized** according to algorithms complexity and granularity of user needs. A typical example is the case of personal pictures collections exported to online photo-sharing Web sites. All complex tasks, such as categorization and learning have to be done locally on relatively small datasets and only their results will be indexed in the large scale.

To allow such strategies, it will be fundamental **to standardize the underlying content descriptions and vocabularies**. Doing so will provide interoperability and let the multimedia community focus ongoing research on a well-defined set of semantics.

4.5 Effects of Network Architecture and P2P issues

Peer to peer technology, distributed and federated architectures, overlay networks and other related technologies show great promise in their application to consumer oriented multi-media information access. Today, web search is almost exclusively under the control of centralized search engines. Only lately have various projects started building and operating P2P web search networks, and so far these endeavors are fairly small in scale.

The purpose on this section is to investigae challenges that must be met and that bottlenecks must be addressed by research and engineering efforts in the near future. The section is based on the results of the first workshop on peer to peer architectures for multimedia retrieval (1P2P4mm), organized by Chorus, that took place in Vico Equense, Italy, on June 2008, co-located with the Infoscale 2008 conference,

A typical P2P search system can be seen as consisting of two parts — the underlying distributed system and its mechanisms for transfer and delivery, and the search tool on top. There is little doubt today that the P2P model for distribution is a very powerful solution for distributing content around the web. While BitTorrent and similar systems clearly announce their P2P nature, peer-based solutions are also used behind the scenes by many applications, with Skype as a large and notable example. A recent and very important example is also the *Proactive network Provider Participation for P2P* or P4P inititative, a joint effort for internet service providers that are investigating methods for optimizing peer-to-peer connections. The incitement behind this initiative is to reduce costs and speed up download time by e.g. preferring local connections.

4.5.1 Competing with centralized solutions

There are a number of advantages of using P2P as compared to centralized solutions:

- A P2P network provides robustness: although performance may go down, it is very difficult or impossible to shut down all peers.
- With P2P infrastructure costs are shared (no huge startup costs which are a barrier for entry, e.g. to compete with Google and its huge infrastructure of servers)

However, to be able to compete with centralized solutions, P2P networks have to deal with a chicken-egg problem: they necessarily start with a small number of users but have to be competitive (e.g. index space and crawler speed for a complete index) to the incumbent services from the first day, otherwise they will not attract new users. Beyond all technological challenges, the most crucial success factor is reaching the critical mass as soon as possible.

P2P technology provides an intriguing base for innovation in the search engine market, by overcoming the market entry barrier of huge infrastructure investments. Thus P2P search could have a chance to challenge the increasing concentration

in search, and to strengthen the plurality of information. On the other hand the required technology poses huge technological challenges in order to be competitive in terms of reliability and speed.
A substantial amount of users is not only necessary for the popularity and/or commercial success of P2P networks. In order for research to be able to experiment and test hypotheses and solutions, a large test bed is needed (further discussed below).
However, for P2P to really be competitive to centralized solutions, interoperability is key. The P2P paradigm must be extended from cooperation between independent users to cooperation between independent developers. HTML/HTTP, RSS or even Gnutella have been successful on a global scale, because they allowed interaction and exchange between independently developed applications. So perhaps the development of a common internet protocol which allows different P2P initiatives and parties to work seamless together and to share resources and information, while having a creative competition of different client software, with different user interfaces, different index structures and ranking algorithms and different feature sets could be an alternative approach to the attempt of having all partners working together at a single solution. This new protocol could add functions for providing a joint discovery service for peers of different P2P applications, building and accessing a common distributed storage, basic search functions, common crawler functions for independent text/multimedia, common standards for exchanging attention data, and standardization of user profiles for incorporating unified social network functionality across the different applications.
While such massive development efforts still remain to be seen, it should be noted that the new Flash 10 beta version includes P2P technology. If P4P became an adopted standard by providers this would be another possibility for such a unification of P2P applications to become a reality. In this context, it could also be noted that the European Broadcasting Union (EBU) is currently performing trials with P2P distribution of content.

4.5.2 Content –based search and hybrid approaches

The main issue for search in distributed systems is that not only the data but also the indices used for search are distributed. Although each peer may be queried for its content for simple data, indices are often centralized. Hybrid approaches become especially important when search is based on (semantic) descriptions of the content being searched.
An advantage for P2P over centralized solutions is that the peers in the network, besides providing sharing and downloading capabilities, may also contribute in creating index information (e.g. do low-level feature extraction from audiovisual objects). However, an inherent feature of P2P networks is that storage space and computing power will vary considerably among peers. Including mobile clients emphasizes this issue.
More research is needed on content based search and hybrid approaches. One example from the projects under CHORUS' realm is the IST VICTORY project that develops search mechanisms that are both content and context based, facilitating the formulation of queries and enabling search by example using 3D/2D objects, sketches, etc. VICTORY proposes a hybrid approach combining P2P and Grid technology to efficiently utilize the computational power in the network. In addition, peers can take on different responsibilities in their role as servers (serving as highly capable super peers or more computationally week edge peers).

4.5.3 Benchmarking and Distributed large test collections

An important issue for research in any area – P2P mm search included – is benchmarking. The goal of benchmarking is to assess the quality of the benchmarked system and to allow comparison to other, similar systems. The existing benchmarks for evaluation of search performance (mainly accuracy) were designed for and are utilized in local repositories of media. Such metrics cannot be directly applied in the P2P environment because of its nature. On the other hand, since P2P environments are characterized by a significant delay in communications, search time may be more worth measuring. Other benchmarking issues for P2P are how to create a benchmarking environment for the P2P overlays; what search functionality (what services) to select for benchmarking; and how to make measurements comparable.
Related to benchmarking is also the importance of large depositories that algorithms may be tested and compared on. One initiative in this area is the CoPhIR test collection,(Content-based Photo Image Retrieval, http://cophir.isti.cnr.it/) CoPhIR is part of the SAPIR project and the target collection is 100 million images.

4.5.4 P2P as a "political" statement

There are a number of reasons for preferring P2P network solutions over centralized solutions that are based on what may be referred to as "political" grounds. The first is in line with the currently very strong movement towards sustainable solutions. In the promoters' perspective, our highly computerized societies where huge numbers of 7/7-24 hours running machines are heavily underexploited by their owners, are evidently wasting a substantial amount of valuable computational resource. P2P techniques are promising as a way to make this resource available for re-use.
P2P could also be promoted for reasons of info-diversity and privacy. 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".
But P2P is not only an interesting and valuable distributed network technology. The peers represent users, who have made an active choice to connect to the network. Becoming a member of the network not only implies getting access to the services offered by the network. By joining users also agree to contribute by letting their computers be used as servers by other users.
Although the participation and agreement may be implicit and even not known to the user (as is the case for Skype), the contributing peer is an interesting concept worth investigating further. This is part of the web 2.0 revolution, an ongoing

movement that could be described either as “altruistic” (users providing content for the common good) or as “exhibitionist” (getting your “15 minutes of fame” through the web). Regardless of description, the fact that the Internet is a non-centralized structure gives support for the applicability of P2P based functionality. The data is originally highly distributed, residing on millions of sites (with more and more individuals contributing, e.g., through their blogs).
The fact that not only professional providers produce content is also important. The advent of more and more user generated content creates new challenges e.g. for generating index information and metadata.
Continuing to view peers as users, peer-to-peer networks can be used as infrastructures for peer-centric information access. First, P2P allows cooperation between resources with equal capabilities for organizing, representing, accessing and distributing information. Second, they increase uncensored sharing of audiovisual information in a scalable way. Moreover, P2P networks may keep track of the context in which the collection of objects managed by a peer was organized, e.g., the history of the local queries or the list of the most significant features of the objects. Finally, P2P networks permit to scale the sharing of the evidence gathered through implicit feedback when the amount of multimedia data grows.
Audiovisual content also carries problems with rights management. While thumbnails or similar information may be distributed over the network, the object itself and its rights could be kept locally on the owner’s peer.
Currently, most user-generated audiovisual content is made available using centralized servers, notably YouTube and Flickr. Another solution would be the establishment of a peer-to-peer network where every content producer is also a provider. Continuing today’s trend for user-generated content, a significant growth can be expected for the volume of multimedia content owned by each peer. An index can then become necessary even for the content of a single peer and, to remain autonomous, the peer may prefer to have a local copy of its index. It can be expected that end-user software for content-based indexing of images and video will be released in the near future by several editors, using various image or video descriptors and multidimensional index structures.

4.5.5 Concluding remarks on P2P solutions

In conclusion, the problem of achieving a critical mass of peers and competing with the dominating centralized commercial search solutions seems to give P2P search a bleak future. Close to all issues where P2P technology might have something to offer are mixed with reasons for preferring a centralized solution instead. While waiting for P2P technology to be included in internet protocols, a few non-technical and more political suggestions for how to make P2P take off anyway can be made.
However, more research is needed before P2P and other distributed network solutions can be taken into use or ruled out. In particular, more focus on hybrid solutions and on the semantic parts of search is needed..

5 ENABLERS

5.1 Corpora development

The development and availability of data corpora to develop, enhance, and evaluate multimedia search engines and their underlying modules are an important aspect. Due to the fact that most modules of a multimedia search engine are based on statistical classifiers and pattern recognition technology the demand for training corpora is very high. These modules learn and model the pattern of a multimedia search query. The following two examples will demonstrate this fact. First, training a state-of-the-art speech recognition system for the purpose of audio search requires a training corpus of at least 50 hours of speech recordings. Second, for the training of a video classification system, about 1000 hours of video clips per class are needed in order to create sufficiently robust models. In both cases, the effort of collecting and annotating the training corpora is huge and requires financial and human resources and the right knowledge to perform the process of collecting and annotating in an efficient way. If the multimedia search engine consists of several audio visual indexing and retrieval modules that are combined to perform a flexible multimedia search application it is important to have access to training corpora to optimize the complex multimedia search engine. It can be stated that the availability of such huge training corpora is a key issue to providing high-performance search engines.
Multimedia corpora are also used for internal testing and optimization of the media indexing and retrieval modules to achieve the best performance for a given task. Official benchmarking activities make use of well defined and general agreed corpora to compare different systems with each other or against given requirements. Examples for this benchmarking are TrecVid or NIST speaker recognition task.
The amount of effort necessary to build corpora with high quality annotations and metadata of significant size and quality is enormous. This leads to the situation that this task is often underestimated in research proposals. The creation of such corpora is usually not in the focus of research groups and often requires non-scientific work. Availability alone does not lead to better and new research approaches and so the motivation to build large corpora is not high enough. Further, the financial cost for corpus development is substantial and underestimated in many research proposals. This leads to the situation that the application scenarios of multimedia search engine projects are not well evaluated and the progress can only be demonstrated on official benchmarking corpora. It will be important for the future to plan and provide more resources and effort for the creation of multimedia corpora including annotations and controlled metadata descriptions.
To support the process of corpora development, standardized methods and tools have to be established and provided for research and development (R&D) purposes. The process of corpora development should be documented in detail so that new partners can follow the instructions easily. The guidelines for corpus development have to be created by experienced

partners who are active in this field. Further, the tools for corpus development, like annotation tools or video recorders and players should be made available to different partners without complex licensing restrictions. The usage of semi-automatic annotation tools to create proofed reference annotations is an important approach to increase the efficiency and decrease the effort in corpus development. The availability of open source corpora development tools would be helpful to decrease the entrance barrier to new partners in this area.
The design of a corpus depends heavily on the application scenario or research topic which is in the focus of a specific project, application or product. For example the scenario of video concept retrieval requires an annotation of the whole video clips into predefined tasks. The effort of this manual classification is lower than the effort of a more detailed annotation on segment level. Especially to perform a high quality segmentation of audio-visual recordings requires huge effort. Although it will be difficult to develop corpora of general usage it should be investigated for which additional research work and applications the new developed corpora can be foreseen before starting the work.
Corpora distinguish each other regarding size and homogeneity of the material. To provide meaningful corpora these two criteria should be investigated and defined clearly. For many retrieval tasks and module development huge amount of samples are required to train the classifiers robustly. There is always a trade-off between size and level of annotation. It will be a challenge to create detailed reference annotations for large-size corpora.
The homogeneity and the domain dependency of a corpus have to be analysed carefully. To solve a specific retrieval task a domain dependent and homogeneous data set will lead to optimized indexing and retrieval modules. This can lead to new research approaches and results. However, the generality of such new modules can be limited and only be valid for specific domains.
The research community must have an open access to the available corpora. A technical and legal infrastructure to distribute the corpora must be created and maintained. Therefore specialized data organisations (e.g. NIST, ELRA, LDC) has to extend there portfolio of corpora. Investigations on alternative access methods (e.g. web services) to get data from the corpora should be carried out.
Finally it will be important to made huge amount of well design corpora for multimedia search technologies available. Especially multimodal corpora containing all different kinds of media together are the basis for advances in this area.

The following list summarizes the most important issues for the corpus development in the area of future multimedia search engines:

- Defining standard process and guidelines for corpus development
- Providing open source tools to support the corpus development process
- Controlled standards for annotation formats and processes
- Development of large multimodal corpora for general multimedia retrieval tasks
- Consideration of real world applications and task before starting the corpus definition (application centric approach)
- Clear definition of corpus development tasks in future research projects
- Creation of a reliable and open distribution infrastructure
- Providing significant resources and budgets for corpus development (i.e. targeted corpus development projects)
-

All these requirements will lead to provide better and large scale corpora which are the key enabler to achieve progress in the area of multimedia search

5.2 Multimedia search engines assessment

As multimedia search engines research domain is very active, the technological know-how acquired a critical mass and the multiple research results became mature. In this context, having a reference evaluation framework has a great importance not for competitiveness objectives but to provide **landmarks** on technologies frontiers and performances. It should also provide guidelines for search engines "technology consumers" for the most appropriate technology regarding a given professional or personal usage and need.

5.2.1 Performance assessment: for which purpose?

We have noticed that the setting-up of early benchmark initiatives was technology driven as the research field was young and the benchmarks were more dedicated to assess the emerging multimedia content indexing and search technology performances. Hence, benchmarking campaigns were focusing on evaluating the effectiveness and the efficiency of indexing methods such as multimedia content signatures generation (automatic content enrichment). Many performance measures (the most known are precision and recall) based on ground truth data base were computed. Most of the time evaluation databases were whether toy/lab databases so that the ground-truth is relatively easy to generate. But the impact of such benchmark was obviously affecting and dedicated to the academia community. As the technology became mature and the proof of concept of such multimedia content search established, the end-users, the industries and content owners express a growing interest to use such technologies. At that moment, the community became aware of the so called "semantic gap", since the experiments of search engine usage (in-situ) were not so successful in the early years. That was not very surprising as most of the multimedia search engines consider the query by example paradigm (the earliest multimedia search). This paradigm was of coarse not an obvious one for an end user. There were multiple tentative to face the semantic gap which represent now one of the major challenge in content search technologies. That derives diversity in

indexing and search technologies that give birth to multiple benchmarking initiatives and tasks: image annotation, object recognition, high level features detection, interactive search … In the D2.1 chorus document, we related the different existing benchmarking campaigns that runs such tasks. The most recent discussions conducted within the community point out the crucial role of users needs and requirements to make tasks definition more realistic. The impact of such discussion is to make benchmarking campaign useful not only for academia but also provide guidelines for end-users to be able to choose the most appropriate technology for a given usage or scenario. In this regard, the performance assessment measure should not only be dedicated to technology and system but user-centred measure with a set of pre-defined use-case. The usability, utility and acceptance measures are not so common in the performance assessment of a search engine neither in the existing benchmark campaign neither in the evaluation framework of the ongoing European or national initiatives in search engine domain. The early European projects were also technology driven but we observe that the real use-cases and scenarios are getting important in the definition of the target technologies that need to be specified and developed; which represent a very recent good sign of the evolution of the community toward more impact of the developed technologies.

5.2.2 Recommendation for benchmarking framework:

As we have already pointed out in the previous Chorus deliverable (D2.1), existing benchmark initiatives are rather fragmented with overlaps and also with missing pieces. To be able to achieve a significant progress in the benchmarking objectives and achievements, the community needs to invest a lot of efforts with regards to several directions. It would very helpful if the European commission provide the necessary support for such European efforts and initiative that should target the international community of multimedia search engine and hence will insure the leading position of Europe in this domain. In the following, we list some directions which will help achieving significant progress and will impact benchmarking campaigns providing benefit to the academia as well as to end-users:

1. Analysis of existing benchmarks initiatives with regards to:
    a. Task definition review & assessment of all components of retrieval systems (including evaluation metrics).
    b. Investigation of performance evaluation. Particular attention should be paid to the fairness of comparison within and across benchmarks.
    c. Review of existing data collections toward re-use of content and recommendations for ground truth methodologies.

When analysing the ***existing state of the art*** in the benchmarking domain, a major outcome should be the **identification of gaps and communalities:** technical/research approaches never evaluated in any initiative; and on the other hand tasks that have different titles but share the underlying technologies: evaluating the same technological challenges in redundant ways (different data collections with the related ground truth generation…).

2. Advance with the search engine and content description communities beyond the existing evaluation tasks and foster **task incubation**. Benchmarking process should be a living and dynamic process that has the sufficient flexible mechanisms to allow the emergence of new tasks but also the death of tasks that are considered as solved.
   The scalability issue, mentioned in sections 4.4 for benchmarking in the context of large data collections and also mentioned in section 5.4 for P2P context (distributed collection), is one of the most obvious example of starting incubation effort.
   Therefore, such community effort should be able to:
    a. Investigate **new technical opportunities** that **foster research innovations**.
    b. Address **user-driven task** definitions, which represent a lack in most of the existing benchmark initiatives. In this way, stakeholders should be involved to be able to identify the best techniques to address real application needs so that the evaluated tasks meet the end users' expectations.

1. Analysis and evaluation of the **operational aspects of the benchmarking workflow**. This should allow to specify and to develop software architectures to harmonize and support a **joint framework** for the existing initiatives. This should investigate till the hardware infrastructure to support the baseline platform and data distribution for the newly identified tasks. Such software platform will allow to foster **more evaluation fairness** between existing benchmark initiatives and to produce a common harmonized workflow for benchmarks.

2. Strong international collaborations:
    c. **Monitor the impact** of benchmarking on research and innovation that imply to monitor the impact on the scientific community and the end users.
    d. Strong communication and coordination with the diverse international initiatives as European researchers are deeply involved in such international campaigns.

The approache of perfomance assessment will promote the uptake of innovative content technologies by stakeholders and reduce the gap between technology production and its usability. Such effort will have impact on community to support innovation and to facilitate the technology uptake by the stakeholders throughout Europe and at the international level.

5.3 Essence and metadata preservation from end to end

Technological key issues can be derived by considering the "state we like reach in the future" according to the following chapters of the CHORUS "Intermediate Vision Document (D3.2)" :
"2.1.1 Search awareness during production and distribution of media" [1]
"2.2.2 Coordinating source-to-sink (end-to-end) systems that preserves

a.) meta data
b.) essence quality for better automatic generation of meta data and even better user experience" (avoid information loss caused by cascaded decompression and compression (i. e. transcoding), high transfer rates of Networks needed, even on consumer side)" [1]

5.3.1 The advantage of preserving essence and metadata end-to-end for search

As already stated in the CHORUS "Intermediate Vision Document" [1], today's production and distribution technology does not necessarily care about the fact, that, eventually, the content (content =essence + metadata) has to be searchable after having being produced and distributed. Nevertheless, this need for searchability is well recognised by all owners of large (audiovisual) archives, who are eagerly awaiting solutions to enhance their content by additional metadata. One simple example is the date and time information a videowas shot by the camera. This information, which is often available in the raw material (the rushes) and which is highly desired for search, is normally not preserved from recording over postproduction to distribution. The loss of these machine-created metadata is due to current technologies usually used in subsequence for video production, postproduction and distribution (including compression, decompression and recompression at each of these stages).

For example, a miniDV based video camcorder (several 100 million devices have been sold worldwide) records the recording date (along with other data that is useful for search). When copying this recording on a DVD, the needed conversion from miniDV format to DVD causes the loss of the recording date and the loss of all other metadata generated by the miniDV camera such as shutter time or location (GPS) data (provided, of course, that the camcorder is equipped with this technical option). That kind of information is neither signalled to the television receiver/terminal in the broadcast mode nor when streaming the content over the internet. The reason for this is that distribution formats had primarily been developed to minimize data-rate and to respect other constraints of the network and the media.

In a future world of broadband connections, according to our vision, it will be more important that distributed material becomes searchable rather than any saving of data rate. As transmission bandwith/data-rate will not be a limiting factor at long term, it will simply not be economic to reduce the data rate at the expense of losing available and associated metadata which allows for the searchablity of essence. **We are lacking a technology which would preserve associated information (i. .e. metadata) throughout the whole value chain of networked media.**

This relates first of all to metadata but applies, in longer term, also to the data of the audiovisual essence. The secondary aim should be to preserve essence quality instead of lowering data rate to an amount that the loss of information would lead to a increased error-rate for object recognition and thus allow for distributed production and search processes. Further research should demonstrate to what extent object recognition technologies can be improved by using material which was processed by lossless, data compression technologies, which by themselves might be subject for future research.

Keeping essence objects separate instead of mixing them together during production is assessed to advance object recognition as well. For instance, keeping voice tracks separate instead of mixing them with back-ground noise or music would help in automatic object and speech recognition for audiovisual search and annotation, but could increase the necessary bandwidth and storage. Composition of separately transmitted or streamed audio-visual objects is the basic concept of the MPEG-4 system specification which, however, due to its complexity has not really find its entry yet into today's audiovisual industry.

Undoubtedly, the new technology we are seeking for to perform the sustainability of essence and metadata has still to take into account bandwidth (or transmission) cost, especially when providing the data to a mobile or portable terminal. Therefore the new technology should allow operating a search engine at different entry points in the production, postproduction and distribution chain (including the interaction channel): at the player/browser/consumer device, at the (mobile) network provider side, at the archive and in postproduction and production.

It is to be noted that the associated data (i. e. the metadata) need not only be maintained in the direction from production to consumption. Of at least equal importance is to preserve the metadata provided by the consumers/prosumers when uploading or exchanging their audiovisual data ("essence"). The consumers/prosumers may accept to provide the (automatically available or manually created) metadata in addition to their audiovisual data in order to help others to find their material via a search engine. This certainly involves issues on privacy and ownership of the metadata. Such issues

should also be covered by appropriate (new) technology in order to encourage even sceptical consumers to contribute and enrich their data with metadata which are so desirable to improve search.

The technology should allow all producers of metadata - professionals as well as consumers/prosumers - to determine whether they would like to retain the generated metadata or provide these data to a (general or selected) audience. Additionally a technical solution would be desirable which would allow the owner of the metadata to revoke the (use of the) generated metadata by search engines in case they had been provided unintentionally, under false assumptions or under a different legal scheme.

Future research is needed to generate knowledge that would allow the development, specification and standardization such "search-aware production and distribution technology". This research would have to respect given technologies and given platforms including professional and end-user devices for production, post-production, distribution and presentation.

Examples on the consumer side are: A networked television receiver that is equipped with a camcorder input and which will be used for postproduction and distribution (and not just for consumption); or a mobile phone that could capture and post-process video items whilst supporting this action by "search-aware production and distribution technology". It is essential to establish user requirements for this specification for both, the professional and the consumer/prosumer domain.

"Search-aware production and distribution technology" could be decisive for consumers when buying new devices, such as cameras, because the self-generated audiovisual content can be found more easily (on the device, in the personal archive, or in the public web).

5.3.2 Already started activities to preserve metadata

There are already ongoing activities within the industry to address parts of the technological key issues of a "search-aware production and distribution technology" at different levels for different sectors cited bellow.

One recent example in the photographic sector was documented in a press release entitled "Metadata Working Group Introduces First Specification for Interoperability and Preservation of Metadata in Digital Photography" [4]. The intent of the Metadata Working Group is to publish technical specifications that describe how to effectively store metadata into digital media files. These royalty-free specifications will be made available to manufacturers and service providers so that they may create products that store metadata in a consistent way, and that allow consumers to maintain control over their valuable information.

That means search engines can make use of this metadata in a standardized form. The specification [5] is, of course, sector-specific but the principle is desirable for other sectors all well such as the audio and video sector if not the whole multimedia sector.

In the multimedia research community there is substantial support that a media asset gains value by the inclusion of information about how or when it was captured or used, and how it has been manipulated and organized [2]. Such information can be exploited by search engines. However, as described in Section 5.3.1 above, today's audiovisual technology (even in the professional production and distribution sector) do not preserve such information established during the production and/or the distribution process. Like in some other sectors, one simple example is again the date and time of the recording, which is almost never preserved by today's video broadcast or IP streaming production technology, but is typically added manually during archival.

The total workflows need to be scrutinized. Some information is already preserved from production to the consumer (end-to-end) and could be used already for search. For example the teletext transcript that is typically available for television broadcast, is a high quality resource that may be used to annotate video at hardly no additional costs. In addition to this, most of the information gathered during the production process cannot be inferred from the essence itself but is nevertheless relevant for search and should be preserved by means of appropriate technology. A more exotic example are the names of the actors in a television series. This type of information is available in the script during production, but usually disappears in the post-production process, and is sometimes reconstructed manually for the electronic program guide. Nevertheless, for search and retrieval (both in a personal and a professional audio visual archive) this type of information is invaluable. In fact, the script provides even more information, such as actions, motivations, and discourse structures.

Metadata is undoubtedly a valuable commodity not only for search but also for every automatic or semi-automatic process performed during the media life-cycle. In fact, looking at the approaches of life-long recording, as in the project "Mylifebits" performed by Microsoft, they seem absolutely essential for making the material accessible and exploitable. As the access to the material will vary over time, the different views on the essence (e.g. interpretation, use, maintenance, generation) need to be kept so that additional support services can be provided.

Thus, the more information is available about a media item in the form of metadata, the larger its economic value will become. Production of metadata is very costly. Their preservation is therefore highly desirable. The general problem is still that researchers and developers are lacking the corresponding awareness. Producers, manufacturers, archivists etc should coordinate in order to achieve preservation of metadata. In the research sector, a first step in this direction is achieved through the proposed model of canonical processes of media production [3]. This model is an approach to address the issues of metadata capture, preservation and exchange, and to improve interoperability of (semantically-rich) multimedia systems. The basic idea is, though, to keep generated metadata (where metadata not only describes content, but also the media asset's context, e.g. its use over time, its changes, its different discourse roles, etc).

5.3.3 Activities to preserve essence data

As stated in Section 5.3.1 above, it may not only be beneficial to preserved from metadata throughout the multimedia value chain. It might as well be of importance to preserve the essence quality itself from extensive processing such as high-rate data compression. Uncompressed audio and video material (or material which just underwent loss-less compression) would not show coding artifacts that, when present, might render object recognition more problematic.

In today's broadcast production, post-production and distribution sector it is absolutely common to use so-called "non lossless video data compression formats" like MPEG-2 or H.264 (MPEG-4 Part 10) and losing information of the essence, because of the non lossless nature of such codec implementations. Especially when cascading a series of non-lossless video data compression formats, loss of information becomes visible, depending on the length of the cascading chain and the different formats and parameters that are involved. From video production, post-production and distribution over the air or internet the cascading chain typically consist of seven transcoding and re-encoding steps with typically some five different compression formats. One approach to lower the resulting transcoding losses in such a given production, post-production and distribution chain is that broadcasters are aiming to choose a low compression ratio with a very a high headroom whereever it is economical in this chain to avoid visible loss of information (so called artifacts) in order to satisfy the consumer with better picture quality. They simply spend a headroom of up to six times of the actually needed distribution bandwidth to keep transcoding and reencoding losses low. But there is no evidence that this precaution that lowers visible artifacts is good enough for future automatic object recognition, annotation and search, because of very complex artifacts caused by so many different compression formats (wavelet vs. DCT based, intraframe vs. interframe) involved in this chain e.g. DNxHD, DV, MJPEG, MPEG-2, H.264.

For user generated content it is even more important than in the broadcast sector to avoid additional losses due to transcodings, because of higher economic constraints regarding storage and network. Spending a headroom of six times of the actually needed bandwith to lower transcoding and re-encoding losses seems not to be feasible, as broadcasters do. A first attempt to address this technical issue is provided, for example, by one internet video host [6], where untranscoded raw material of user generated content can be uploaded, stored and accessed if the user agrees to, aside the rendered and transcoded post-produced product which is for streaming.

REFERENCES

[1] R. Ortgies, R. Neudel, C. Dosch et. al., (2007) CHORUS Deliverable 3.2 Vision Document, intermediate version http://www.ist-chorus.org/documents/Chorus_Del3.2_Final_Nov2007.pdf
[2] Nack, F. & Putz, W. (2004) Saying What It means: Semi-automated (News) Media Annotation. Multimedia Tools and Applications, 22, pp. 263 - 302,
[3] Hardman, L., Obrenovic, Z., Nack, F., Kerherve, B., and Piersol, K. (2008) Canonical Processes of Semantically Annotated Media Production. To appear in the Special Issue on Canonical Processes, Multimedia Systems Journal, http://www.springerlink.com/content/100377/?Content+Status=Accepted
[4] Pressrelease of the Metadata Working Group: (COLOGNE, Germany — Sept. 24, 2008) „Metadata Working Group Introduces First Specification for Interoperability and Preservation of Metadata in Digital Photography. Adobe, Apple, Canon, Microsoft, Nokia and Sony form Metadata Working Group. "
[5] Metadata Working Group (September 2008) GUIDELINES FOR HANDLING IMAGE METADATA Version 1.0 http://www.metadataworkinggroup.com/press/pdf/photokina_pr_2008_09_24.pdf
[6] http://www.vimeo.com

5.4 User needs and requirements

Information retrieval systems are interactive systems, intended to be used by some human to achieve some goal. Not only is it important for the system to function properly and efficiently, it also has to meet the users' expectations. At the most basic level, a search result is relevant only if it is considered to be so by the person who did the searching.

Users interact with search engines in many different ways, adding, annotating and retrieving information. The user population is heterogeneous, ranging from professionals to just anyone trying to find some information. In addition, the

goals and tasks of the user may be very different and the context in which these goals are to be achieved vary. Thus, from the user point of view, an information retrieval service can take on many different faces. As discussed in section three, new services may be built by assembling components of search technologies.

The functional breakdown diagram of section two, reproduced in the beginning of section four, shows where in the search process user interaction takes place. Section two also makes clear that the user context may be taken into account by the search components themselves.

5.4.1 User involvement and user centered design

User centered design and human-computer interaction (HCI) was discussed quite extensively in deliverable D2.1. In this section, user-related issues are briefly reviewed.

More and more voices are raised to stress the importance of involving the user in the design and development of new services and products. There is little 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.

User centered design refers to the 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. The user involvement can take place at different stages of the development process and takes on different forms. Users may be studied (observed, interviewed), be asked for their opinion and/or feedback (focus groups, user evaluation) or be directly involved as participants of the design team (participatory design).

The first and probably most crucial success factor is to understand the context in which the search engine will be put to use. Within the field of HCI, ethnographically inspired, observational methods are often applied to assess information about the connection between the task to be solved and the surrounding work processes.

Once a system has been implemented to some extent, user evaluation of prototypes and large scale beta testing are common methods to get in touch with the user and collect feedback. By using rapid prototyping, users may be involved in testing at a much earlier stage, as exemplified by the new trend of *Agile* systems development.

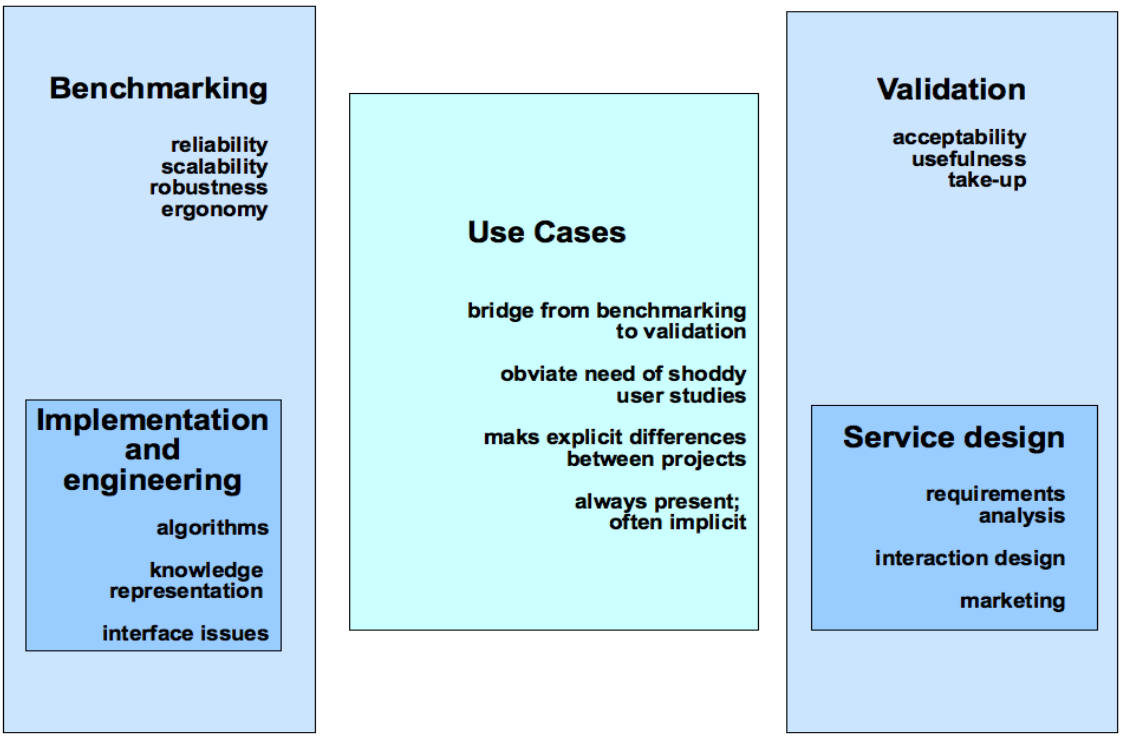
Usability testing is also important. Here, the focus is on the graphical user interface, and the goal is to answer questions such as: Does the flow of information through the system follow a natural and understandable path for the user? Does the user understand what menu choices and buttons do?

5.4.2 Use cases as a tool for assessing user needs

As discussed above, to create really useful technology there is a need for a thorough understanding of user issues and a likewise thorough knowledge and mastering of an arsenal of techniques and methods for assessing user needs, studying user behaviour, evaluating user satisfaction and performance. This is the object of study of the field of human-machine interaction and interaction design. Hitherto, multi-media information access projects relatively seldom have identified interaction as a pressing issue – technology and system factors have overridden those concerns. Simultaneously, the human-machine interaction research field rarely uses multi-media information access as domain of study.

In deliverable 2.1, CHORUS suggested use cases as a solution for information retrieval research to take user requirements and expectations into account without having to perform extensive HCI research. Use cases also bring other benefits, e.g. in relation to benchmarking and validation as summarized in Figure 4.3.

Figure 5.3 Benefits of Use Cases and their relation to Benchmarking and Validation



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. As discussed in section 3.1, the multimedia information retrieval field has a need for improved benchmarking and validation methods. The collection and analysis of both existing use cases and cases perceived as generic by industry lays the foundation for the establishment of a standard set of criteria for use cases; an effort that both benchmarking and validation will benefit from. Classifying use cases along these criteria greatly simplifies comparisons between projects. The typology of use cases and the framework for describing them may also serve as a source of inspiration for future projects. Finally, use cases not only describe tasks to be accomplished but also bring in the context of use that is so crucial in understanding user needs and in designing really useful services.

In conclusion, generalizable results and guidelines on interaction in multi-media information access need appropriate methodology and craft from the interaction field. Unfortunately, such collaboration between fields is rarely a possibility in research projects. Carefully designed use cases may however allow projects to take user issues into account in a structured way, as a better alternative to omitting user studies completely or performing them amateurishly when interaction expertise is not available.

5.5 Trends affecting research in MM search

5.5.1 User Generated Content

Another trend that will affect search is that of user generated content. Web 2.0 has enabled users to contribute with all kinds of material to the body of content available on the internet. Although a lot of this material is mostly personal and/or of very low quality, home technology and software has improved to a point where this content may become interesting also for professional actors.

Although it could be argued that user generated content is like any other content and thus does not put specific demands on search engines, there are at least three issues that have an effect on search and thus on research on search. The first issue is that user generated content currently is stored and indexed in a very different setting from professional content. The table below lists some examples of differences between “YouTube-style” and professional content.

	Professional	YouTube-style
Metadata	highly structured factional based on ontologies	Unstructured emotional based on folksonomies
Search type	Keyword, directed	Browsing, associative
Rights management	Strict, controlled	Anarchistic
Quality	High, controlled	Very varying, uncontrolled
…		

Table 5.1 Differences between "YouTube-style" and professional content

A second issue is that of scale. If the current trend of ordinary people's tendency to store audiovisual data on the interenet is extrapolated, there is no arguing that the repositories will be not only unstructured but huge. In his keynote speech on "Challenges in searching social media" at the CIKM 2008 Workshop on Search in Social Media (held in conjunction with ACM 17th Conference on Information and Knowledge Management end October 2008), Andrew Tomkins, Chief Scientist of Search, Yahoo!, noted that if every second user stores ten images per day, this amounts to 10 exabytes of storage per year.

The third 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.

5.5.2 Mobile and set-top-box search

A third trend that has been touched upon in other places of this document is that the computer no longer is the only device upon which search will take place. Mobile phones raise new challenges in terms of small screens and other limitations in interaction, limited computing power and battery lifetime concerns, as well as accessibility and bandwidth. Another type of device that will come into play is the set-top box (as discussed elsewhere in this document). Here, search will be very different from the users' point of view, and the user community will be even more heterogeneous than for internet search.

6 SOCIO/ECONOMIC & LEGAL ASPECTS

6.1 Introduction

This chapter serves different (interrelated) purposes. The first one is to *update and present new socio-economic trends* with regard to the previous reporting period. We see little value repeating issues where there has been no significant change with regard to the past reporting period. An example for this is the problematic of copyright and trademark infringement, whose source tension has practically not change and which has already been extensively analysed both from a legal and socio-economic point of view in the past reporting period. Rather trying to be exhaustive covering all aspects, we will focus on new issues or changing issues and discusses these. An example in case is the consequences of an increasing personalization of services.

The second objective is to pinpoint to *emerging challenges*. Most notably this is the case for privacy concerns. Personal data is proliferating in the internet, particularly in social networks, and search engines are increasingly more performing to link these data, as people search engines do.[1] It has, therefore, attracted the attention of the legislator and many stakeholders over the past year.

The third objective is to report on the *results on the workshop* the IPTS organized on September 2008, in which close to 40 experts (see Annex 7.8) discussed during two days the status and prospects of the search engine landscape, the socio-economic challenges, and some policy options. With the aim trying to quantify the perception for sensitive and conflicting issues between stakeholders, the IPTS interrogated the workshop participants anonymously. The results of this survey are presented in Annex II. Although the sample is not large enough to be representative, we see its value in quantifying in the degree of consensus of stakeholders in broad lines and in the pondering of different challenges. Therefore, in the body of this report we do refer to this workshop when discussing the individual issues. Marks of the type (Q.X.Y) refer to the questions in the survey (Annex 7.9).

Finally, we present and analyse *policy options* in the search domain, beyond the pure research and development policy. We conclude presenting some forward-looking considerations on placing the citizen's relationship with search engines and a broader base, taking a novel view on *electronic Identity, based on user's autonomy*.

6.2 Encompassing Trends

Web search engines providers have been a major source of innovation for internet applications in the past and are confident that they will remain also important drivers in the future of the internet (Q1.6). Future applications and services will be determined by the consolidation/emergence of novel technological capabilities responding to explicit or latent user demands. Given the pace of speed of innovations introduced (a couple of new tools or services practically every month) experts have huge difficulties making assumptions on upcoming applications in the mid-term based on evidence of grounded information. Thus, statements on prospects of web-based applications and novel services expected by the year 2015 are more speculative in the search domain than in other ICT areas, in view of the workshop participants. In spite of this difficulty, there are some long-term user demands, which are likely to shape the search domain. These include user's demand for a higher degree of personalization, for more participation and social interactivity, for more naturalness and greater accessibility (e.g. mobility)

6.2.1 Personalization

Technologically speaking, search engine providers compete amongst themselves from their competitors thanks to the efficiency of their crawlers, the richness of their index, quality of their algorithm to determine the relevance of search hits, and the speed to deliver the results to the users. One consequence of search engine personalisation, however, is the enrichment of the latter process of defining relevance by means of another component: a database containing the user's personal information (IP-address, web history etc.). 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.

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,

[1] Examples: www.spock.com or www.pipl.com

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 in the end-user's interest. 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 a notable imbalance 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 (Machill 2004: 325), personalising may be one way to provide the end-user with more relevant hits (Teevan 2005).

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 patterns are commercially relevant information on 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. Therefore 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 is increasingly extending 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 often many times a day. In the future, search may introduce *"prospective" elements*. Search engines would match a user's record against new information passing through their matching engine. In sum, personalisation of search appears to converge into benefits for both the commercial players and for the end-user.

The concept to personalise search is not new (e.g. with Hotbot thinking about it as far back as 1996) (Gasser, 2006: 204), but now technological advances as regards storage, processing power, and artificial intelligence (Olsen 2006), has driven more personalisation in recent years. Two major approaches are commonly used, 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, Yahoo! Search Builder, and Google's 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 (Sherman 2006; Hafner 2006; Bradley2006). 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 is to re-rank results automatically provided by search engines, or to show different users different results based on their past behaviour. A prominent example of this approach is A9, an Amazon service which uses the Google index.

Consequences of personalization

The question arises whether the increased use of personal data by search engine operators in the course of their attempts to personalise search might have a negative impact for the user beyond privacy concerns (e.g. with respect to freedom of information, media pluralism or other). There is a need to analyse two possible (and inter-related) claims. First, search personalisation relies to a large extent on user data collection. User awareness of such user data collection might produce chilling effects by *restricting "curious searches", searches that are politically or culturally sensitive*. Second, personalisation tends to create strong ties to or affinities with a particular (personalised) search engine, making it less likely that users switch to competing engines (*"stickiness"*-argument).

As regards the first argument, it is clear that increased collection of data will limit my autonomy in that it may have a chilling effect. If I know that my data is being logged, I may become more careful to enter certain key words in the search box. Some courts have indeed recognised the importance of allowing for anonymous surfing, and held that there should be a right to navigate anonymously in the interest of freedom of expression. In the US, for instance, the Delaware Supreme Court has protected the identity of a blogger in the case of Doe vs. Cahill, finding that the government had failed to meet the strict standards required by the First Amendment (right to freedom of expression) to unmask an anonymous critic. However, users might well desire some personalisation, depending on the topic they are searching for, and the capacity in which they are searching. No personalisation can be achieved without some form of data collection. Personalisation of search thus requires us to add the identity-dimension to the search realm. Personalisation necessitates us to acknowledge not only the fact that all users have a multiplicity of (sometimes conflicting) identities, and that it makes no real sense to work with one single, monolithic, personalised profile per user. In addition, it poses the question to what extent and when one may identify the user who made a given search. That is, when may one engage in the process of linking certain data to certain individuals or identities (Microsoft 2006). In sum, personalisation cannot be achieved without some form of data collection. Personalization inherently raises issues of identification that have a direct connection with data protection issues. Therefore, many experts point to need to grant the users some form of control as to the identity or capacity in which they are searching and being recorded.

The second concern is probably the most important one. Personalization fits in the general strategy of search engines to build a fixed customer base, thereby fortifying and expanding their market share. Indeed, there are clear signs of market concentration in the search engine sector, as a result of the concentration of advertising revenues flowing from search. But switching costs are virtually non-existent. Users may easily use different search engines, and switch between them, depending on their degree of satisfaction with each of them. Search engines have thus been very actively looking for ways to tie users into their platform, and one very powerful way would be through search engine personalisation. If it is true that personalisation will yield better and more relevant search results depending on the number of searches effectuated on a given search engine, then users have would have fewer reason to switch search engine: changing search engine would mean that they would need to start from scratch, re-build their history and user profile. As a result, personalisation has the effect to get users 'stick' to search engines.

This last issue cannot be seen independent from the other two concerns raised above. In the case of a lack competition between search engines the above arguments for building in more user control, or checks and balances ensuring a sufficient degree of internal media pluralism become more compelling. One option could be to ensure some form of *'standardisation or interoperability'* with a view to maintaining a healthy degree of competition between media players. These standards or rules would enable the user to *actually 'port' his/her data or user profile onto other platforms*. In other words, more interoperability would, in part, enhance user control over, or autonomy with regard to, the search engine they are using.

6.2.2 Naturalness

Search engines' 'holy grail' is to become more sophisticated in their ability to understand meaning. With emerging tools, users no longer have to write down their queries with keywords and can ask questions in natural languages. Getting closer to the 'semantic web' paradigm is a formidable challenge of the future, which would set the base for disruptive applications. The workshop panel was confident that natural language processing and other 'semantic' elements will be available by 2015 (Q1.1). These elements will enable to improve significantly the 'search experience. This vision is highly positive, but the devil is in the detail and a differentiated view is necessary. For instance, effective face detection is practically around the corner, yet face recognition requires semantic attributes that are not yet feasible today. The recognition of more complex images and video may be mid-term (5-10 years) while searching an audio excerpt in a complex (noisy) environment is likely only in the long run (next 20 years). Also, while technology is progressing steadily it is difficult to judge if the user's perception of the satisfactory performance of audio-visual search technology will be similar to today's text-search (Q1.2). User expectations co-evolve with technological progress.

Another promising outlook that will shape the way forward is to make use of semantic capabilities for *contextual search*. Location-dependent or geo search is a well-identified example of contextual search. Less under discussion, but an equally interesting tool of the future, *chronologically dependent search* whereby a user is able to 'query in time' (e.g. 'when did I say something'). Even more attractive would be the possibility to query local and chronological (i.e. when and did I say something in that place).

In addition to stand-alone search tools, many experts consider that a number of valuable applications are at sight whereby 'search engines' are one (important) module embedded into a more complex system. One example of an attractive applications, are services for *task-tailored structuring* of information on demand. Similar tools are currently developed for 'large clients' (e.g. enterprises) only. In the future these may be available for small entities, or even on individual demands. Another example, are systems for *online personal data management*.

In times of information overflow, a competitive advantage of search engines is to render a more *intuitive and enriching visualization* of results. Search engine providers are being expressed in new ways, from automated clustering and categorization to actual answers to questions. New methods of presentation such as clustering, tag clouds, graphical scales that widen or narrow searches based on parameters, and automated categorization make it easier to navigate results. Search results are more accurate and automatically summarized, with relevance determined by individual preferences.

One potential drawback of too much personalization of search engines is that it leads to too 'expected' results. Possibly the relevance of results might be enhanced by introducing features that emulate human behaviour, like curiosity. It is well-known that people have a tendency to hang out with people who are like them.[2] This is so as much in the real world as in the online world. While it is true that the Internet lowers the cost of finding people who are like us, and directing people toward people who think like us, personalisation of search may have a very powerful potential to do just the contrary. It would achieve this by enabling *serendipity*. Serendipity is the situation in which one is pleasantly surprised, or makes desirable discoveries by accident. To be sure, even within homogenous groups of people with shared beliefs there are great variations in terms of interests. The shared interests might thus provide a better platform for expanding one's own interests. Put differently, I might be more able and willing to accept new ideas and recommendations that come from someone who has something in common with me; someone with whom I share a common set of values. This has also commercial value,

[2] this phenomenon is called 'homophily' (Vedantam 2006)

trying to direct people searching one product to get interested in other products. (When we go into a bookstore and get attracted by literature we were not looking for in the first place. In a more rudimentary way, serendipity in online bookstores is emulated by suggesting books that have been bought by people having searched similar ones). Users would see value in search engines suggesting unexpected results, information and knowledge. And, upcoming personalised search engines might adapt their personalisation strategies so to introduce some unexpected elements beyond the mere reiteration of common, shared information. Technologically speaking this is not easy: such a search tool must deliver not only unexpected but also relevant results, making the algorithm particularly difficult as it has to comply with partially opposing requirements.

6.2.3 'Social search & computing'

Social computing attitudes which are now affecting all aspects of the evolution of the Internet are also contributing to the evolution of the search sector, providing opportunities for innovative solutions to enter the market and re-establish a plurality of choices for the users. In particular the social computing trend is expected to lead to a variety of solutions answering the needs of different audiences, allowing for more diversification in the market.

According to Sherman (2007), social search can be grouped into the following 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 (Rollyo, Google Custom Search, Eurexter,), collaborative harvesters (Digg, Netscape, Reddit) and social Q&A sites (Yahoo Answers, Answerbag). In view of many experts, social search will expand, because *humans still perform better than search engines at doing some search tasks*. The increasing popularity of social search is that it mitigates some of the problems of algorithmic search. Often, social search comes closer than the huge crawler based indexes to give relevant results. This is particularly true for organizing non-text content like images and movies until more performing audio-visual search tools will be available. One example is the video search engine Sproose. Sproose uses the same database as the popular Blinkx search tool, and has added a “digg-like” interface that allows registered members to vote on retrieved video results. Search engines that exploit the 'wisdom of the crowds' are proliferating (examples include human-powered search directory Mahalo, the people search engine Pipl, social bookmarks driven search engine Nsyght, etc.) In the future search engines will need to make extensive use of social networking phenomenon, where search engines will focus more strongly on the presentation of results based on *recommendations of social networks' users*. Similar user-driven approaches will be possible by making user-related searches become available, thus helping to increase the *quality of the search results*

The rise of Web 2.0, poses two challenges to search engine providers. First how to find best information in Web 2.0 applications and, second how to make best use of the Web 2.0 potentialities and their communities for rendering the search engines better performing. With regard to the former, there is proliferation of information formats and sources which is usually addressed by (mostly user-unfriendly) applications. Increasingly these solutions no longer meet the needs of users, entranced by the flexibility and collaborative dynamics offered by user generated tools, like blogs, bookmark and content sharing tagging, etc. However, *user-generated content* will be increasing and will *inevitably influence search products*. Web 2.0 features are not limited to search on PCs; participatory features will penetrate also the mobile domain. Apart of satisfying the 'commodity' of making phone calls, smart phones will be more and more targeted for search related activities like finding directions and will embed web 2.0 features. For instance allowing not only searching for nearby restaurants, but also facilitate the reading and writing reviews on the restaurant. The basis for this is already set. Many new generation mobile phones have GPS included and some of them have cameras geo-tagging capabilities. Once many handheld devices will be deployed, this will opens a whole new window of opportunity of location-based web services, like sharing information, finding friends or likeminded people in the area and searching for local information. How these services will be supported (e.g through subscription fee, direct ads placement, affiliate industry partnerships, etc.) may be service specific and needs further investigation. *Contextual advertising* seems an obvious option, but the willingness of users to receive advertising in a specific location *is not uncontested*. Concluding it appears that some lessons from the web search business model may be applied to the mobile world, but certainly not transposed one-to-one.

6.3 Market and businesses

6.3.1 Web search

There is little doubt that the today's web search market is shaped by a quasi-monopolistic position. Google is the main player in Europe, holding over 90% of the search engine market in most EU Member States). Over the last year the market leader Google has strengthened its position and there are no signs that the situation is turning in the short term. On the contrary, Yahoo's delicate financial situation and the need to redirect its strategy, runs the risk to weaken significantly this particular player; (possibly having to go alliances or sold to Microsoft), diminishing further the number of independent actors. The prevailing search engine business model relies almost exclusively on advertising (on search engines sites or affiliated). Other non-advertising business options, like subscription-fee service, enterprise search, pay for inclusion, consulting, etc. a very minor role. With regard to advertising, Google follows two routes, either through their own web sites or through their network sites. While in 2003 revenues where (nearly) equally split, by 2007 advertising in Google web

sites accounted for double the revenues of the Google network sites. The reason for this trend can be attributed to a deliberate change of the company's strategic objectives: providing a diversification of products, form a search tool provider, to personalized services, browsers, etc.

The first question that arises is if the web search market is a *natural oligopoly*. Many experts consider that this is indeed the nature of the web search engines landscape given the advertising business model. Online advertising needs to attract very large audiences which can be reached only by few powerful general-purpose search engines. In addition, there is some room for specialised (e.g. thematic, vertical) search engines, interesting for specific advertisers to reach targeted audiences, but these are niche markets. This view of a natural oligopoly was largely supported by the workshop attendants (Q2.2). The industry landscape might differ in the case of an alternative business model to advertising, but any viable (e.g. subscription-fee service) alternative is not at sight.

What are the implications and risks of such a natural oligopoly? One risk could be the *abuse of dominant position*, whose ruling is covered by competition law. More generally, the question arises whether there current structure gives confidence in a sane and robust competition amongst web search engines. The panel opinion was mixed (see also Q2.1). Supporters of a *robust competition* argue that competition amongst players is largely based on innovation, for instance offering new and useful services to users. In addition, users have the choice to select the search engine that suits, as there is no technological lock-in and the switching costs are virtually non existent. Critical voices, state that the dominance of the market leader is so immense (over 90% in most European countries) that there are no real alternatives for advertisers to reach large audiences. In addition, they say that Google is no longer only a search tool provider, but a provider of diverse products (email, personalized services, browsers, storage, etc), that creates dependency making a change difficult. This 'stickness' argument will be further discussed later.

While there are diverging view on the degree of competition in such an oligopoly, there are far less when on the dynamics. Most experts believe that there is and will be *opportunities* space left *for newcomers* (Q2.3). They consider that the entry barrier is not insurmountable for newcomers and the route to success is through innovation and specialization. Nevertheless penetrating into the market is far from easy: it requires resources and expertise. Entrants would need considerable technical competence and innovative ideas in order to provide search results of high quality to convince and displace the network actors (i.e. users, advertisers).In a subsequent step, the necessity of reaching economy of scale and networking effects limits the emergence of strong competitors.

From a policy point of view, an interesting point arises whether and how much the lack of compatibility, de-facto standards would lead to a lock-in situation and network effects in the long run. For instance favourable 'default' settings contribute to determine user's choices (i.e. the power of default) and leading to 'user stickiness' are an example of measures to make it more difficult for newcomers, although –in principle– users have the choice to move to other service providers.

6.3.2 Web search and the media industry

The role of the media and citizens has traditionally been very distinct. In the realm of the internet this is no longer true. User-generated content is proliferating and users are becoming 'prosumers'. The blurring of frontiers affects also industrial players. Search engine providers offer news syndication and other services that were traditionally provided by media companies. The workshop participants discussed the changing relationships in the 'triangle' search engines, media/advertisers and users.

There was consensus that the search engine industry has already profoundly changed the advertising market. While the adverting market is stagnating or shrinking in traditional media in TV and newpapers [9], adverting on the internet is still growing. Search engines providers play a key role, as they reach a large audience and they can direct ads.

The structure of markets is changing quicker than years before. The rapid change of the economic status of people (by loosing their jobs), the increasing mobility of people, the disintegration of family structures, amongst others, are the origin of a broader segmentation of the target audiences. As this fragmentation is progressing, the advertiser's prime objective are novel ways (ideally 'real-time') to identify and approach the specific target audience. Search engines have been disruptive and successful in this respect. The second objective of advertisers is to turn the brand-customer relationship from a monologue (one way advertising) into a two-way 'dialogue' where the customer gets engaged, listens, and acts. To get into this dialogue, users must perceive advertising as an enriching user experience and for this to happen *advertising will get context dependent*. From a marketing point of view to target the audience best, the brand would be able to follow the 'customer's journey'. Search engine providers might potentially contribute to this *behavioural marketing* by profiling users. But profiling individuals for this purpose without their consent is legally not permitted and ethically unacceptable. More importantly, however, is that for personalized advertising, individual profiling is not even necessary. Knowing the patters of particular groups might suffice to generate customized services and to support the business of the web search industry Workshop participants see a number of upcoming trends. Workshop experts foresee that in the future *network TV and multimedia web will merge* and advertising will be going hand by hand with programming (i.e. for product placements). Marketing agencies will learn how to use social networks, product placements, and other non-intrusive methods to sell their products. A particular battle has started over the control of or alliances with popular social network sites. Examples include

MySpace, owned by News Corporation (Fox Interactive Media), Microsoft having a minority stake of Facebook [10], or Orkut owned by Google. The player's success to provide customised services in this evolving web 2.0 environment for clients will influence the advertising share, and thus shift the *balance of power of from 'traditional' toward the new media players.* Albeit recognizing a change in power, the panel's did not consider that search engine providers will sideline completely traditional media players, but rather converge into a different equilibrium (Q1.5).

The view on role and the power of users with respect to search engines and advertisers is controversial. Many see tension in the fact that search engines offer a service to the citizens and to society but they make their business out of selling advertising space. They consider –that under the prevailing business model– the advertisers are the sole customers of the web search industry Opponents, on their side, consider this a short-sighted position: search engines actually do make their profit by balancing the interests of both users and advertisers, as much as newspapers do. In the opinion of search engine operators, both advertisers and users are de-facto customers of search engine providers, as ¬for the business to function– both side need to be satisfied.
Supporters of the first view, argue that the goal of multimedia web search is not different to network television, namely to channel as many eyeballs as possible to advertisers. In other words, their business is not selling content but attracting audience. Consequently, this conflict of interests is the origin of a number of tensions amongst stakeholders leading to several concerns. One is that web search engines –due to their nature– have the possibility to intervene on page rankings and could make misuse of this for instance with the motivation to compel potential advertisers into subscribing to the search engine advertising programmes (Q3.7). Representatives of search engine providers ruled out this option completely. Other participants pointed out that most of the major search engines had abandoned their paid inclusion programs, precisely because they undermined the legitimacy of search results, a legitimacy that is needed to engender trust in the public.

6.3.3 Mobile search

Mobile search is key for the development of a thriving mobile content market. Just as on other digital platforms characterised by the proliferation of content and services, mobile search is the natural interface between user and mobile content. Indeed, given the personal nature of mobile devices, mobile search may well become the most important user interface. Yet up to now the technology was only slow to take-up. Analysts recently reported that only 30 percent of mobile users access the Internet on their mobile devices, though 75 percent of those who do also conduct searches [11].[3]

Nevertheless, a good indicator for the perceived importance of this market is the current 'clash between titans'. A number of telecoms players, device manufacturers, software companies, search engines, directory service and content providers all want (part of) the mobile search cake, as the market is picking up speed. In fact, the number of subscribers may be still modest in absolute terms, but the growth rates are impressive. The GSM Association reported 32 million mobile broadband connections in March 2008, up from only 3 million broadband connections in March 2007![4] More importantly, mobile Internet and search is an expanding market. Its main attractions are local information, like weather, maps or directions. It is estimated that in 2008 there are 2.5 billion mobile users world-wide, of whom roughly 40% have 2.5 G and 10% 3G technology. As the number of mobile subscribers is still increasing world-wide (particularly in highly populated, less developed countries where the penetration rate is modest) and the share of 2.5G mobiles or higher will be increasing, the number of subscribers will also increase. Today, 489 million people have access to mobile Internet (not necessarily broadband), and this may double by 2011. As the vast majority of the mobile Internet users will also be searching, they become potential customers for search engine providers. In fact, eMarketer considers that the mobile search advertisement market 221 m$ in 2008 and will increase to 2,361 m$ by 2011.[5]

These estimations are similar to Juniper Research forecast for the period 2008-2013, which expects direct revenue growth from nearly $1.5 billion in 2008 to $4.8 billion in 2013. The annual growth rate for total revenues from mobile search related business is about 27% in average for this five year period. The strongest market growth is expected for (mobile) general web and local search, according to Juniper. Regardless the accuracy of these estimations, the prospects is very promising.

For this to come about, however, a number of challenges, which arise from the fact that mobility imposes specific requirements with regard to user interaction, retrieving data and displays, need to be resolved. Another issue is adapting or creating content suitable for mobile devices. Content search particularly will have added value when content is adapted to the user and combined with other technologies, e.g. location-based services. This will render the search experience more personalised.

The large majority of the workshop participants, subscribe to the optimistic view that is a only question of time (e.g. time to get to more adequate handheld devices, supporting infrastructures, appropriate contents, etc) that internet on the move will

---

4 www.gsmworld.com
5 www.eMarketer.com

ramp up and thus also mobile search. There was even a large base that 'search on the move' ('mobile search') will overtake desktop search (Q1.3) by 2015. One argument for this optimism, is, mobile technologies provide a rapidly growing user base for the participatory web. This is particularly true for those regions of the world where internet access via fixed line is not available or expensive, due to lack of infrastructures or geographical challenges. Here, mobile technology can often be the only way for people to get internet access. Another reason is that smart phones and specifically designed web applications are starting to enable performing common online tasks on a mobile phone. This emerging trend leads to increased web use by existing users and consequently to more user-generated content.

A major argument is that future wireless grids (e.g. mobile communications, Wifi, sensor networks, RFID-networks), together with added-value enablers (e.g. GPS) will enable new applications. With this outlook, location-based services requiring 'search-solutions' are likely to become increasingly important. (Q1.4). With mobile internet services and the deployment of higher-speed mobile networks, users can connect and access the intranet and information systems of their organisation from almost any location at any time, through always-on connectivity. The next step in this technological revolution is to connect inanimate objects and things to communication networks. RFID allow for the accurate identification of objects and the forwarding of this information to a database stored on the internet or on a remote server. In this manner, data and information processing capabilities can be associated with any kind of object. This means that not only people, but also things will become connected and contactable. In such a truly ubiquitous network – anytime, anywhere, by anyone and anything - users will ask search solutions to tracked information as well as various documents and objects.

Mobile devices offer a very promising platform for all kinds of electronic communications services, given the very high penetration rate. More mobile content is now available than ever before in form of music, ringtones, images, text, games, etc. Mobile browsing is also being improved through combination of 3G and better devices, and some analysts predict that mobile entertainment sales are going to rise above $38 billion by 2011. [12]

From a business point of view mobile-search is still in its infancy, the market structure is far from being consolidated. In this buoyant phase, the adequacy and viability business models needs still to be tested. Although the winning model is unclear, the panel considers that the telecom providers' favoured walled-garden business model is likely not to prevail. (Q2.7). A positive regard of unconsolidated and promising area with a very dynamic actor landscape, is to see it as opportunity for Europe. Some workshop participants would consider it beneficial to support European actors, particularly small but innovative players, as this may lead to a vibrant industry in Europe and set the basis for a same competition in the domain.
Four main mobile search strands can be identified:

- Text based local search is the returning of answers in text format (e.g. 4Info, Google SMS, TellMe SMS, CitySearch, etc.);
- Mobile directory services and voice search where the answers to user queries are provided automatically or through human agents (e.g. GOOG 411, TellMe, etc.);
- WAP local search where the application is based on the search engine provider's server (Yahoo OneSearch, Yell, etc.);
- Via a downloadable application for mobile (Yahoo Go, Ask Mobile, etc.) that includes the search functionality.[13]

Most market players believe that search advertising has most chances of generating revenues, and the leading business model appears to be evolving from caller pays to advertising supported mobile search.[14] Analysts predict that the US mobile search advertising revenues amount to $33 million in 2007, $102 million in 2008, and $1.4 billion by 2012.[15] The same trend can be spotted as regards global mobile search from $1.5 billion by 2011 to over $11 billion by 2008. [16] However, mobile search technology will need to transform further in order to reach its full potential. Mobile search technology has to overcome a number of inherent problems. First of all, we use mobile search not only in order to save time, but foremost in order to find geo-specific information about businesses, news, weather, sports. The geographic dimension is key to success. For instance, the types of information and/or business searched for will help determine the size of the radius around the user in which to locate relevant businesses. Likewise, the famous page-rank algorithm is not nearly as useful on mobile as it is in web search since people look for very specific answers, not general information about something. In other words, a different approach is required. It is not sufficient to jam existing search onto the mobile device.

Second, in order to implement this new search paradigm a number of technical limitations will need to be overcome. (a) Limits in bandwidth make it difficult to perform speedy searches or to run various searches in parallel. It is thus imperative to serve very relevant results from the first time. (b) The small size of the screen moreover leaves very little room for multiple results, while the small size of the keys leads to more incorrectly spelled words and thus more need for more personalisation and recommendations. Again, accuracy and efficiency in retrieving the right information with very little information are key to success. (c) The query method has to be very flexible since searches are performed in varying circumstances. Voice searches (with the help of speech recognition technology) may be more useful in some situations, but in other situations text will prevail as a query method.

6.3.4 Enterprise search

Search engine providers investigate various technological and functional options to improve the relevance and performance of search to provide information and knowledge. Several technological developments can actually be identified. They correspond to users demand on the one hand (single point access to information inside the organization, web information and commercial information; suggestions, backtrack…), and supplier solutions on the other hand (scalable technologies with existing search engines, modular structure, integration with information software…). As a consequence, these dynamics should be held as substitute perspectives shaping various techno-economic trends; each one supporting different industrial paths of development and new structuring of the marketplace and value chain.

In enterprise search, suppliers sell to clients (on-shelf) products and specific appliances (implementation, maintenance, add-on, tuning of services…) tailored to the companies needs. Enterprise search aim at increasing the productivity and reduce information overload by providing employees, partners and customers with the ability to find relevant content in a wide range of repositories and formats. Often, enterprise search solutions cause profound changes in the management of information system, in the efficiency of the firm and the way it implement competitive strategies, with noticeable changes also for the employees and management.

Enterprise search differs from web search, in several ways. To be valuable to users, enterprise search solutions need to deliver results that are relevant and in context, for further analysis and evaluation. Here, business environments call for specificity in the designing of search engines functionalities. For instance, computer-generated clustering is a useful tool in business environments, as users –unlike web search– cannot judge the relevance based upon popularity (e.g. number of links to other sites) or social tagging (i.e. reputation-based concepts relying upon the wisdom of the mass). Rather, employees often do not ask for something specific, but aiming to discover new material they could not explicitly specify at the first place. There, some enterprises search solutions include an automatic categorization, using a kind of 'guided navigation'. Another important difference between enterprise search and web search is the access to the companies content. While web search scans the entire 'public' internet and points to sites that everybody can access, this is not necessary the case in enterprise search, where a security structure is in place the use and access of content may be restricted.

Given the technological specificities and the market specificities, the experts during the Seville workshop unanimously agreed to the view that enterprise search domain to be very distinct market from the web search. There was ample consensus that the market for tailored search solutions will not disappear (e.g. replace by performing web search engines) in the mid-term (Q2.5). This does not mean, however, current (small) that the current market structure will remain the same. It is possible enterprise search developers might be absorbed by (larger) enterprise resource providers (Q2.6) and signs of a consolidation effect are already visible, due to the market relevance.

The reason for the 'appetite' of ERP to acquire technology providers is that the market for enterprise search is both booming and fragmented. Booming, because it was growing 39% to $1.4 billion in 2006, according to consultancy firm IDC. And fragmented, as the enterprise search solutions marketplace is not monolithic in its requirements. Products, prices and features vary widely depending on the client's request for configuration (some systems cost several hundreds of thousands of Euros). The marketplace is populated with numerous companies (Vivisimo, Omnifind, Exalead, Autonomy, Fast, Coveo, Endeca, etc.) of different sizes; mostly mid-sized for software companies. The diversity of demands on search technologies has been little attractive for very large vendors and has offered the possibilities for small, but dynamic, players to focus on distinct niches. This seems to be changing slowing, with large software companies seriously announcing products in the enterprise search market. During the past reporting period, Microsoft bought FAST for $1,2 billion, Intel and SAP invested $15 millions into enterprise-search company Endeca, and Google developed an offering that is adjusted to business market.

6.4 Social Aspects

6.4.1 Search Engine Bias and Media Pluralism

In Europe, it is commonly accepted that media information and services (i.e. ultimately culture) are intrinsically different in nature from 'non-cultural products'. A non-functioning media market would have adverse effects not only for the media players themselves but more importantly for society at large. Therefore pluralism of the media is recognised in Europe as deserving special regulatory attention in the interest of freedom of expression and freedom of information. Media pluralism regulation seeks to further two types of values that are generally thought to benefit society and individuals: democracy and autonomy. At times, these values underpinning much of media regulation may be endangered. For instance, in a context of high barriers to entry media outlets, and the viewers' attention is a scarce commodity. The risk exists thus that information provided by a particular media player may be biased, and induces people to have only a partial overview over a given issue, thus affecting both individual autonomy and democracy. These problems are usually remedied by two types of measures. 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.

Due to the lack of clear metrics for assessing media pluralism, correcting for perceived market failures is highly complex, and intervention needs to be carried out with caution. This is especially so in fast-paced technology markets. A number of commentators are currently analysing to what extent existing media pluralism regulations ought to be applied to search engines. In recent years, search engines have acquired a prominent role in granting users widespread access to information, and in giving the various advertisers targeted eyeballs. Technologists argue that technology itself has no values embedded in it, while others have argued that this is not the case [17]. These commentators have argued consistently that there is a clear bias in the search sphere [18]. Search engines appear to be biased by not being able to index the whole web and therefore defining criteria to crawl and display results. They appear not to be reliable in their selection criteria by systematically favouring certain sites (popular sites with many linkages to them, American sites). In general, this bias is caused by the algorithms (computer logic routines) that produce the search results, rather than to deliberate human intervention. However this has implications for which sites may be systematically displayed and which systematically neglected. For example, American sites, by having been online longer and supporting a larger population, are more likely to be linked to. Niche sites with few links to them have difficulty to emerge.
That some search engine bias exists is rarely disputed, what is contested is if the observed bias is unintentional and due to (some) unavoidable or unmanageable cause (see also survey). This would regard bias as a conflict of norms rather than a deviation from a pure objectivity, which is not the case. Given that there is a lack in transparency in the search algorithm, this offers room for hypotheses suggesting that the bias is deliberate (e.g. Search engines favour certain websites because of advertiser pressure, Search engines favour certain websites because of governmental pressure or regulation to promote or restrict access, etc.)

Another source of problems is that the customers of search engines are the advertisers, not the users. Assuming this hypothesis (see previously on opposing views claiming that also users are clients), there is no economic incentive for generating and sharing better content, just “good enough” content; in fact some incentive exists to drive more people to associated advertisements. Consequently, countries that have no appeal to the advertisement industry end up being under-represented.

Another issues due to economic pressure is that web site operators and content providers are known to do anything to get as high as possible on the most prominent search engines' rankings. Thus, there is a whole industry that devotes itself to search engine optimisation [19]. Examples of proposed internal pluralism measures are increased transparency and various labelling and signalling measures by trusted third parties; one much discussed example of an external pluralism measure is public investment in alternative search engines.

Some experts say that it is unavoidable that search engines contribute to editorial judgments. They argue, for instance, there is a kind of snowball effect in which the first hits get more user attention and thus more links or "user votes", and are thus more likely to appear at the top of the page. Algorithms are also said to favour older pages that had more time to accumulate links and readers [21,22]. Search engine providers argue that their in ranking algorithm have to take into account countermeasures to fight against malcontents, fraudsters and spammers. Two dynamics limit the bias of search engines. First, low switching costs between search engines in a competitive market environment mean that if a search engine's bias degrades the relevance of search results, users will explore and shift to alternatives. Second, in this view the shift toward personalisation of search results moots the search engine bias since it breaks the above described snowball effect, and it caters for minority interest [23].

In sum, there is consensus that web search engines are not neutral with respect to page rankings and that it influences pluralism. Whether search engines influences is positive or negative media pluralism remains controversial (Q3.4).On one side, there is evidence of adverse effects. For instance, Machill [24] shows the risk for 'circular investigations' amongst media professionals, and by this lowering quality of content. In addition, many experts subscribe view that the advertising market moves from content to intent [20]. And this –in turn– threatens pluralism and diversity, for lack of incentive to do original research. On the other side, search engines contribute positively to the expansion of news, by offering access to local news, by finding thematic news or minority views, etc. Generally speaking, however, there was no strong sentiment that the current situation was such threat to pluralism that would requires immediate active intervention by the regulatory authorities. The major argument being that it cannot be claimed (or proofed) that there is a systematic bias or denial of information (underrepresentation of niche voices or minority voices) or that current player can be accused to have an editorial agenda and to rank search results accordingly.

6.4.2 Access to knowledge and opinion making

The discussion on the impact of search on society started from the consequences of the quasi monopoly that characterises today's market. It was mentioned that large search engines show little interest in those markets, particularly small and poor countries (e.g. minority languages), where is little business to gain (advertising), which results in the fact that sites and content from these minorities are underrepresented in page-ranks. Some of the panellists argued that search should be regarded as an elementary service, but no consensus on the implication for stakeholders of such definition could be reached. There was no agreement possible regarding the definition of the quality of service a search engine should provide.

Some experts mentioned that the evolution of search service market is partially also the reason for to the problem of *exclusivity of access to data* in some domains. The case was made for public funded content and databases, such as in the case of satellite images. To their opinion, the access of this public information is hampered by the cost model used by public authorities. Access to public content might be enhanced by changing the predominant cost-recovery model (of investments into creating content) used by public institutions towards a marginal cost model for distribution.

The panel also discussed if and how web search results do shape opinions, such as by distorting *news trustworthiness* through selective news syndication. Some argued that the information diversity (in analogy to bio-diversity) should be protected, and measures should be designed to prevent both the distortion of information and the interruption of services. Others argued, however, that one has to distinguish between the diversity of the Web and the search engine interface to that diversity. Search engines do actually help searchers find “narrow” content that otherwise would go unnoticed.

Because of the non-disclosure of the search algorithms there is empirical lack of evidence of potential intentional misrepresentation of information, the behaviour of crawlers, filters and ranking systems. The survey highlighted that neutrality and veracity of search results are considered as a major social concern. Particularly delicate issues are the manipulation by manual selection between algorithms, the verification of the 'truth' of information and the guarantee of a continued openness and freedom of the net. Therefore some experts that policy should support Public Broadcasters to enter the Search sector since they are already working in the information domain and are publicly funded; this way stressing the 'public role' of the search engines' service.

6.4.3 Search engines as a public service

The survey results indicate that there is a common understanding impression that web search engines do perform a kind of public service (enabling the information society) and should be treated as such by stakeholders (Q3.2). For some, to consider search engines as a fundamental cornerstone for citizens' information, knowledge, civil rights and liberty would call for direct actions by governments. They argued that the European Commission and national governments should fund the development of a 'European Search Engine' to grant the European Union information independence from commercial operators with decision centres in the US. (Following a risk / benefit analysis of independence, e.g. what would be the costs / implications for Europe if Google would shut down their services for 24 hours). This would prevent Europe to suffer the consequences of possible unexpected failures of the existing system. Opponents, consider that geo-political argument does makes little sense in globally networked economy. Moreover, basing arguments of on the dichotomy European vs. Non-European is difficult to sustainable, given that all major US-owned search engine providers have companies and research units in Europe.

Although is it undisputed that search engines functions in the 'public's interest' and it key for of economy and society, the views remain polarized between supporters and opponents of a public European Search engine. The reason lies in unconceivable understanding whether the search engines a primary basic service to EU citizens (in the traditional sense), or if search engine infrastructure should be considered as a critical infrastructure (implying and posing specific requirements to Europe to safeguard its vital interests).

Irrespective of the exact governmental implications, the question arises if Europe would be capable to create a powerful alternative at all. Given the industrial search engine landscape, it appears rather unlikely that the market would coagulate into a powerful European search engine industry by its own. One reason is the fragmentation of expertise in Europe, notwithstanding the important and useful European initiatives supporting the development of future architectures and technologies. Another reason is the lack the human resources to create a competitive search industry. The battle is on promising talents is fierce, even more so since US search engine companies are establishing research units in Europe. Anyhow, it is felt that Europe still maintains a strong pool of expertise both in industry and academia. Clustering this expertise may give birth to new companies and technologies, for instance Europe has several 'alternative' search tools covering the social web.

As a matter of example, Europe has considerable expertise in P2P-search. This technology is investigated as a potential solution once distributed systems reach their technological limits. Europe as excellent researchers and research programmes has boosted the academic activity. Whether Europe is able to turn these research results into innovation setting the basis for a vibrant market in the future is not clear (Q2.8). A business model that is viable for P2P environments is still to be tested in field. Supporting a research and industry suitable framework cluster to let the market come up with innovative models via cooperation and joint effort may be a way forward.

6.5 Privacy

Search engines invoke the image of people being able to gain knowledge about other people's private lives using search engines [25]. It is important, however, to bear in mind that responses to data profiling by search engines may take many facets: law, technology, social norms and market. One example of a technological response to surveillance and data-

profiling by search engines by search engines is TrackMeNot, a browser extension which produces a lot of 'noise' and obfuscates the actual web searches in a cloud of false leads [26].

Search engine logging raises legal regulatory issues of information privacy, regarding 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 (Article 2(a)). Such personal data should be processed fairly and lawfully (Article 6(1)(a)); they are to be collected for specified and legitimate purposes (Article 6(1)(b)). In addition, the data processing in question needs to be relevant (Article 6(1)(d)) and not excessive in relation to the purpose for which the data have been collected (Article 6(1)(c), Articles 7-8). Finally, the data need to be kept accurate and up-to-date, when necessary with the help of data subjects (Article 7(a)), and ought to be stored no longer than necessary for attainment of the objective; they may be disclosed only with the consent of the data subject (Article 7(a)). The first question that arises is 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' 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 (Sullivan 2006). But 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.

It is important to note, however, 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 would have the tools for getting 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 [27]. 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.

The search engines' specific importance in the current data protection debate derives not that much from the fact that they gather huge amounts of data and hence constitute a good place for hackers, third parties, or governments to mine data ('security argument'). Indeed, there are other players that are also (sometimes even more) compelling for these purposes. Examples of such players are the Internet Service Providers, banks, web sites visited by the user. Even the information stored on the users' computer in the browser's cache or on the hard drive may be more useful [28]. Users leave numerous traces during the browsing and searching experience and enforcing user privacy is goes beyond simply ensuring that the major search engines delete the users' personal search history.

The specificity of search engines is the specificity that they direct people. Search engines are like the pointers of the information age, they move the various users in the desired direction. In doing so, they affect user autonomy. Search engines are masters in determining where the users will leave digital traces. Users pass through the search engine on their way to the web site content where they will leave various traces, and the search engine thus could determine with which content a given IP address will be associated, and which web site will be able to place a uniquely identifying cookie on the users' computer.

Increasing concerns on data protection and privacy with respect to search engines, caused considerable activity on regulators side over the past year, both at national and at European level. Most notably, the advisory body on data protection and privacy, set up under Article 29 of the Data Protection Directive 95/46/EC.

Opinion of Article 29 Working Party

On 4th April 2008, the Article 29 Working Party, an independent European advisory body on data protection and privacy, adopted an Opinion on data protection issues related to search engines.[29] The objective of the Opinion was to strike a balance between the legitimate business needs of the search engine providers and the protection of the personal data of internet users.

In their opinion, the Working Party states the responsibilities for search engine providers as controllers of user data resulting from the Data Protection Directive (95/46/EC). They consider that –as providers of content data (i.e. the index of search results)– European data protection law also applies to search engines in specific situations, for example if they offer a caching service or specialise in building profiles of individuals.

This Opinion addresses the kinds of data processed in the provision of search services, the legal framework, purposes/grounds for legitimate processing, the obligation to inform data subjects, and the rights of data subjects.
A key conclusion is that the Data Protection Directive applies to the processing of personal data by search engines, even when their headquarters are outside the EEA, and that the onus is on search engines in this position to clarify their role in the EEA and the scope of their responsibilities under the Directive. The Data Retention Directive (2006/24/EC) is clearly highlighted as not applicable to search engine providers.

The Working Party considers that personal data must only be processed for legitimate purposes. Search engine providers must delete or irreversibly anonymise personal data once they no longer serve the specified and legitimate purpose they were collected for and be capable of justifying retention and the longevity of cookies deployed at all times. The consent of the user must be sought for all planned cross-relation of user data and for user profile enrichment exercises. Website editor opt-outs must be respected by search engines and requests from users to update/refresh caches must be complied with immediately.

More specifically, the Working Party sets out the specific recommendations. The WP considers that personal data registered by search engines must be erased as soon as possible, and after a 6-month period at the latest. In any event, Directive 2006/24/EC relating to the storage of traffic data does not apply to search engines; they do not have thus any legal obligation to store information concerning users traffic data, unlike Internet access providers for example. The WP recommends that Internet users be also clearly informed of their rights, in accordance with Directive 95/46/EC relating to the protection of personal data: information on the purposes of the data process, the terms of exercise of their right of access, modification and erasure. Lastly, Internet users must give their consent to the use of their data for consumer profiling purposes in particular.

Following the WP's Opinion, feedback was provided by several actors, including the market leader Google. On 8th September 2008, Google responded to the Article 29 Working Party Opinion on data protection issues related to search engines, and as a tangible result, Google announced to cut the retention time of data to 9 months. On 2nd October 2008 announced that the Working Party would organise hearings with three search engine operators. In addition, Article 29 Working Party announced also that they are preparing an opinion on on-line social networks (SNS). This announcement is important also for search engine operators, as they make extensive use of data collected by on-line social networks sites and some search engines are also owners of (or have a stake in) popular SNS. Similarly, mobile search is also starting to raise privacy concerns.

Data protection and mobile search

The success of mobile search depends on personalisation and user profiling. Generally speaking, user data is necessary for improving (i.e. personalising) search results, adapting the user interface (e.g. look and feel), and raise advertising relevance with a view to generating more income. In the mobile context, the need for user data is even greater. It is necessary not only in order to serve geographically relevant information, but also in order to overcome the technical limitations inherent in mobile search.

Data protection will therefore likely to play determinant role in the success of mobile search. Data protection and privacy laws have come to the fore as key determinants of the success of the search business models. User profiling and the logging of user search queries are certainly very sensitive issues given the personal nature of user queries.

This debate is even more important in the mobile search era. First, mobile search raises issues of 'locational' privacy; an issue which is not well-understood and still being investigated, for instance, in relation to RFID. Second, there is an intimate connection between a user and his/her mobile device that is used. The data protection debate in mobile search is thus much more sensitive, and profiling and recommendations more likely to give the feeling of intrusion that triggers privacy debates. Data protection law are likely to influence the deployment of mobile search and the marketing of mobile search services. The recent controversies surrounding the US Child Online Protection Act, and AOL's release of user queries, or the exchange of opinions between the Art.29 Working Party and Google, makes us belief that in the future there will be an extensive discussion on the appropriateness of various legal instruments (EU data protection, e-privacy and data retention directive) and their application in the mobile search sector.

An upcoming question is to what extent the existing debate can be transplanted to the mobile search field, and to what extent there are differences between the two sectors in terms of data protection. Mobile presents two riddles that increase the need for careful data protection scrutiny. First, there is an intimate connection between the user and the mobile device used for the searches, potentially making profiling and personalisation much more precise (and hence informational self-determination much more important). Second, mobile inherently includes locational aspects and thus enriches the user profiles with a key aspect that users may not want to be recorded.

Results from the workshop

The survey shows that more than half of the workshop participants consider that privacy has irrevocably been eroded (Q4.1). The perception that the 'genie is out of the bottle' is in line with a large fraction of the society.[31] One explanation is that the notion of privacy is changing. Privacy as a fundamental value is not longer a static, monolithic entity, but multimodal, dynamic, context dependent, both in space and time. For instance, an image or a sentence in a given moment can be perfectly appropriate, but sometimes if released later would harm the person's right to privacy. Therefore, come experts see a need for a methodology (or tool) to deal privacy in a fluid, a kind of personal privacy right management system.

Whether profiling by search engines (e.g. cookies, log-files, IP-addresses, etc.) is fundamentally different to tracking via other digital footprints (e.g. credit card records, cell phone calls, ATM machine use, etc) (Q4.4) needs a deeper analysis. One difference is that information is stored in distributed systems posing a particular set of responsibilities and legal uncertainties. Some experts claimed that users should be enabled to see who owns what about them and that in distributed systems ('in the cloud'). Another difference is the type of data collected. And here, there seems to be an issue on consent and IP statistics. IP information could be regarded as personal data once these statistics can lead to a unique identifiable person. The SE-providers argue against extending the perspective of the privacy directives including this type of information. According to data protection authorities this request does not solve the problem, as providers do have alternatives to get the info they want.

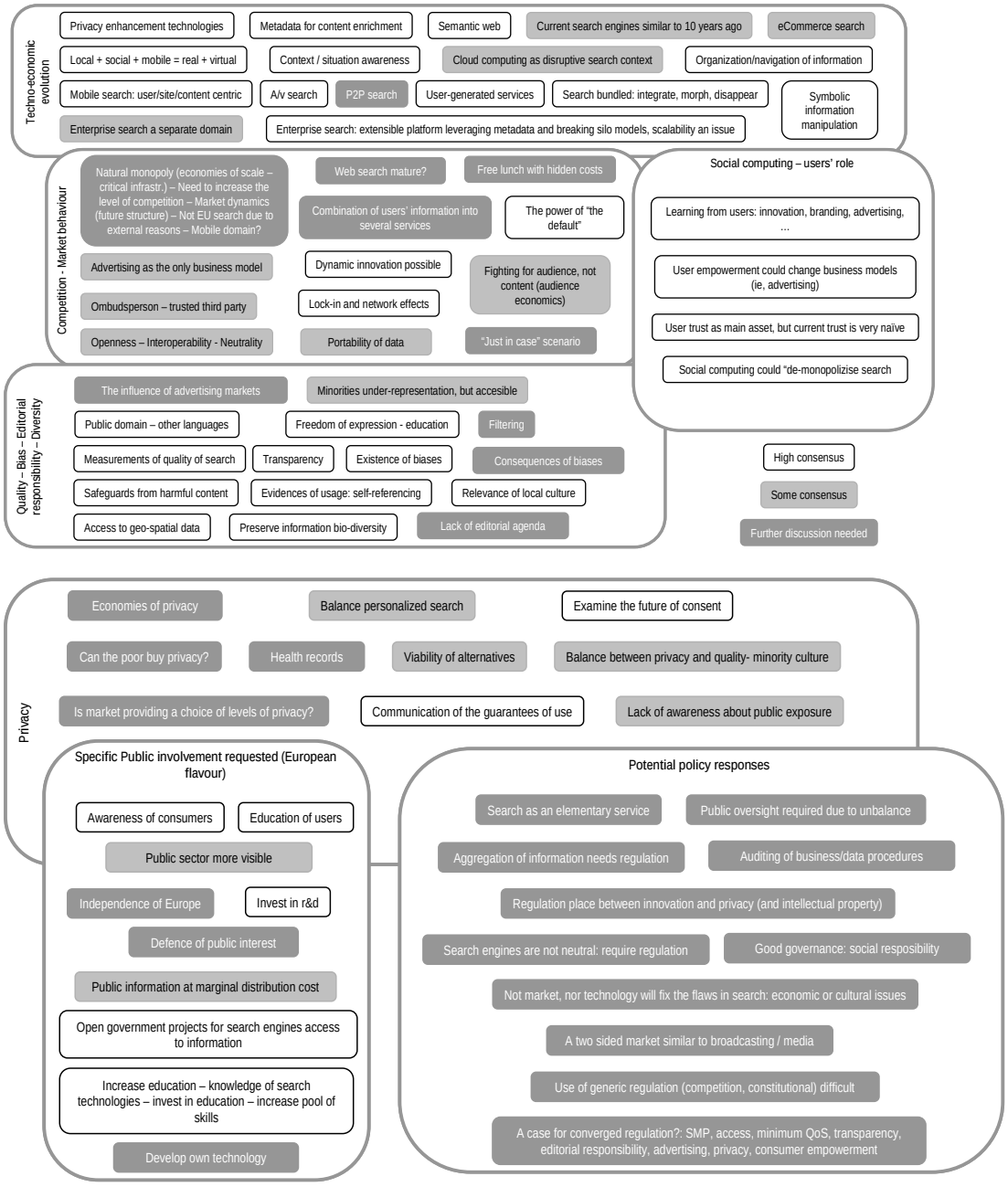
There was consensus amongst the panellist that *consent-based approaches are a good measure to empower people* and sensibilise users of their doings. Adequate use of it should be strengthened and more used. Rather than a binary and static concept, consent shall be regarded as a flexible and dynamic one. Flexible, in the sense that it may allow for different degrees of consent (e.g. by granting the possibility to the user to choose the most appropriate level of privacy), and dynamic in the sense of option to change (e.g. for instance by allowing a testing time with minimum data collection). Limitations to consent-based approaches may be of practical reasons nature. Assuming that consent has be awarded for a specific purpose, how do deal with the case that purpose gets (even slightly) modified? In the case that consent has been granted for two different purposes separately, under which circumstances operators be allowed to aggregate this information? Similarly, how to deal with revocation in a user-friendly and practical way? More importantly, however, is the concern that relying (too much or mainly) on consent would shift the burden of being informed too much on the user and industry may take this as motivation for not assuming their full responsibility.

Search engines do already make efforts to make 'understandable' their privacy policies, thought many panellist considered that there is ample room to improve them communicate them better and to *enhance transparency* with respect to the use of personal data collected from users' searches and use of services. Many panellists called for policies that make users more aware of the consequences, for instance, for a default opt-in mechanisms (rather than opt-out). Some suggested *awareness campaign* like to "tobacco kills" on the cigarette pack: web banners could flag that user's actions are going to be recorded for commercial purposes. Other participants emphasised the role of education and the importance of reaching people already outside the educational system. These *education measures* should ensure that users are in the position to make informed choices. Appropriate sensibilization can turn people's sensibility to privacy also into an opportunity. This is the case when society will acknowledge the search engines efforts to set the highest privacy standards. In other words, *assuring high privacy will become a competitive asset*.

In sum, search players store enormous amount of information about users, and they may provide useful services making best use of it. Major *risks* arise from the *aggregation of content.* This gets especially problematic when the markets drive into a monopolistic equilibrium, as it seems to appear. In this case the regulator needs to investigate the necessity to protect the interest of citizens. In this respect, some expert argued that the rationale for EU intervention should concentrate on the security of data to counter act the risk of data leakage. Others plead for a *European Watchdog* that –if with sufficient privileges– could monitor abuse and misuse of personal data by SE providers. Others suggested an ombudsman-like approach, whereby individuals have the possibility to put forward complaints. Although such an organization would not have legislative of executive power, it would contribute to solve problems (mainly) by mediation. The possible nature, governance and roles of this entity were discussed at some length, and different solutions proposed. This is further described in the section on policy options.

6.6 Policy Options and Outlook

The ultimate objective of this piece is to make recommendations for policy and decision makers with regard to audio-visual search. A starting point toward this aim is to recall the degree of consensus and divergence for the different issues discussed previously. Figure 1 and Figure 2 underneath tries 'visualize' different aspects. They are clustered around following thematic areas: 'techno-economic evolution', 'competition & market behaviour', 'quality/biases/diversity/editorial responsibility', 'social computing and users' role' and 'privacy'. The areas of high consensus are underlined in white, while themes with a low degree of consensus put in dark-grey boxes. These Figures were presented and validated by the workshop participants.



The panel discussed some policy options in order to make some recommendations. These recommendations were summarized in the workshop minutes which were approved by the participants. Options include both regulative and non-regulative measures, like awareness initiatives for users, trust-building measures for stakeholders, or research policies. While trust-building measures or research policies enjoyed large consensus, options involving regulation were far more disputed amongst panel members. Regulative measures, both a governmental and government/industry, were discussed at length without reaching consensus. Therefore the range of viable options will be presented without prioritising them:

Non interventionist approaches:

1. 'Laissez faire'

Some panellist considered that a more liberal attitude would be the most adequate. They were confident that the potential risks discussed during the workshop can and will be overcome without the need to implement any regulation. Potential concerns, like consumer protection, privacy protection or unfair competition are already covered under general rules. No specific regulation would be needed as solutions would naturally arise from a combination of user empowerment, technological developments, and/or market dynamics.

The optimism that the market will be taking care of itself, particularly on the user's privacy and personal data protection aspect, relies on the consideration that user's trust in the search engines is vital for their business. Therefore it is in the industries' own interest that user's sensibilities and privacy are correctly treated. In addition, trust and privacy-assurance becomes a differentiating factor in a competitive environment.

Supporters of a 'laisser-faire' approach were concerned that governmental intervention would increase the industry's burden to comply with (unnecessary) regulation. They were also concerned that more intervention could become detrimental for innovation.

Regulation and co-regulation:

2. Encouraging privacy certification process

The panel was positive in supporting measures that entail the certification of privacy enhancing products. The underlying concept relies upon the certification of IT products and IT-based services privacy compliance with European data protection regulations. This is done by introducing a transparent and revisable procedure supervised by independent authorities. The European Privacy Seal is awarded for IT-products and IT-based services that have proven privacy compliance in a two-step certification procedure. First, an evaluation is performed by specialized IT experts and then this evaluation report is checked by an independent certification body (more information at www.european-privacy-seal.eu). Privacy seal have been successfully introduced in Germany and are currently being tested at European level.
The seal offers positive incentives for market players to implement privacy-enhancing measures. A win-win situation is pursuit envisioned for all stakeholders: the user benefits from a certified quality product or service, the manufacturer profits from market advantages, the privacy protection authorities benefit from a relief in control tasks. Finally, IT-industry gains confidence and acceptance, and society, from an improvement of basic rights protection.

3. Creating a 'third-party' agency

Another policy option discussed was the creation of an independent 'third-party' agency with legislative mandate in auditing and controlling the processes of search engine and other providers of services storing, managing and processing large amounts of data (e.g. providers of social network sites). There was a long discussion on the need and ways to increase public oversight at European level. In particular, there were diverging views on nature, mandate of such an agency including the relationship with other organizations (e.g. national data protection agencies), the required instruments (e.g. current regulation in place), etc.
In this context the establishment of an independent mediator party (e.g. ombudsman-like) was also discussed. In contrast to the aforementioned 'third-party', such an organization would be a recognized body, with a clear mandate to mediate but would not enjoy any legislative power. The panel felt that –if any– the organization to be established should have powers.

4. Adapting current directives and creating specific ones

Current EU regulation, like the electronic communications directive, the audio-visual Medias services directive, or the electronic commerce directive, does not cover adequately or are not applicable to search engines. Many panellists consider that search engines are not adequately or sufficiently covered by current directives and –as search as an essential service– a regulation specifically tailored to them would be beneficial as to assuring the correct functioning of the industry in the long run. This specific regulation shall set minimum standards in terms of quality of service, transparency, and protection of privacy. As a matter of example it should cover the following aspects: It should oblige search engine providers to disclose whenever they correlate or link personal data. It should oblige also to disclose when personal data are sold or provided to third parties. Further, it should impose strict data retention times (possibly below 6 months) and ways for users to get access to their personal data.

The panel considered that new regulation should be targeted to solve specific issues (e.g. privacy) and not go so far to infringe more generic or sector regulation. For instance, the panel felt no strong need to regulate the editorial responsibility of search engines. Similarly, the rules of competition laws seem to adequately cover this industry.

Awareness and trust-building measures:

5. Raising user awareness

There was large consensus amongst the panellist that users must have a basic understanding of the functioning and operation of search engines. One measure to raise consumers' awareness could be through particular campaigns. More effectively at the long-term, however, would be to increase the IT literacy in general. More education in this respect would be desirable, including training programmes to increase the knowledge/skills in search technologies, the basics of data protection, the appropriate use of privacy enhancing technologies etc. Some panellist, suggested that such IT training modules should be offered to pupils at schools as part of their obligatory educational programmes (e.g. a kind of expansion of the IT 'driving licence')

6. Establishment of a permanent stakeholders platform

The panel considered that trust is one (probably the most) important unifying element amongst all stakeholders. It would be therefore advisable to establish a permanent platform that include representatives from industry, governments and civil society to discuss, propose, and implement measures that increase the trust in search-based services. One action of this platform could be to establish a kind of a European 'observatory of search', whereby a number of relevant issues are monitored and regularly updated. This monitoring might comprise cartography of actors, a summary of complaints and

pitfalls, and other relevant issues. Another action may be to discuss the need and viability of a European Code of conduct in this domain, and if viable to define it.

Research policy:

7. Fostering research

Some participants mentioned that the European research landscape in search suffers from fragmentation and the fact that – with the exception of very few exceptions– the industry is consists of small and medium enterprises (SME). Given this situation, some participants recommended to pursuit a R&D strategy that focuses on particular technological challenges in the search domain and that facilitates the involvement of European SME. These activities shall be aimed to create clusters search-related technologies. Research policy shall take into account that search technology will be shifting from a stand-alone application towards one (highly important) 'module' embedded into more complex high-value applications. The challenge is identify high-value applications (that have a search component – but not only) and to gather expertise to develop them. If expertise should not be available within the Member States, Europe should consider setting-up strategic research collaborations with abroad (i.e. in Asia)

In addition to search technologies, the panel considered it worthwhile to strengthen the research efforts of privacy-enhancing technologies.

Other supportive initiatives:

8. Supporting 'public' search engines

The panel discussed whether governments should fund the creation of 'independent' search engines. The views were much polarised.

9. The government as an actor

Governments are holders of massive public information. Some experts considered that –beyond their tasks of legislating– they have also an important and role to play as an actor in the market dynamics by enabling to make better accessible this content. Governments may consider reviewing of their policy, by changing from recovery cost model towards a marginal cost model for charging the access to this content.

Outlook: Autonomy as a basic principle for electronic identity

It is increasingly recognized that the *principle of autonomy* lies at the centre of data protection, rather the principle of secrecy. 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 user's 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 [32]. 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 no longer 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.

This chapter has highlighted that personalisation, the ensuing 'stickiness' of search engines, and the privacy concerns may turn out to become detrimental for the competition in the search engine market and the whole development of innovative electronic services. To prevent such a potential clash in the interest, we might need to focus on increasing user autonomy in three respects. First, we will need to ensure a degree of autonomy as regards the use of a given search engine. This means that decision makers will need to ensure *sufficient interoperability to avoid lock-in situations and fight stickiness*. Second, we will need to focus more on the users' *autonomy as regards their multiple (and sometimes conflicting) digital identities* when carrying out searches. A closer focus on digital identity is paramount to create one of the cornerstones of media pluralism. Clear data protection rules in relation to user identities may prevent the stifling of other fundamental liberties such as the right to freedom of information. Third, decision makers should seek to *foster user autonomy* through appropriate. This would include policy initiatives such as imposing greater transparency and raising awareness as to the specific data collection activities of search engines.

The digital environment requires us to think about the conception of data protection. Data protection in a digital environment is no longer about secrecy, but foremost about autonomy. Loss of autonomy is no longer connected solely to identifying information of particular individuals, but may take place even in a context where it is not possible to ascertain users' real world identities. Likewise, data protection is no longer just about particular individuals and intrusions into their

private sphere. It is a systemic problem in need of a systemic solution. Search personalisation, then, provides a perfect example of how this systemic problem may affect other fundamental values too.

REFERENCES

[1] 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, 321-347
[2] 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
[3] Gasser, U.: Regulating Search Engines: Taking Stock and Looking Ahead. In: Yale Journal of Law and Technology, Vol.9, 2006, pp.124ff; available at http://ssrn.com/abstract=908996
[4] 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
[5] Sherman, C.: Google Launches Custom Search Engine Service, October 24, 2006; available at http://searchenginewatch.com/showPage.html?page=3623765
[6] 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
[7] Bradley, P.: Your Search, Your Way, September 19, 2006, available at http://searchenginewatch.com/showPage.html?page=3623434
[8] Microsoft Research: The Identity Metasystem: Towards a Privacy-Compliant Solution to the Challenges of Digital Identity, October 2006; available at http://www.identityblog.com/wp-content/resources/Identity_Metasystem_EU_Privacy.pdf
[9] The prices for television advertising are falling. In 2008, prices reach their lowest level since 1987., see 'TV advertising declines further in the UK' by Tim Bradshaw, Financial Times, 11th November 2008 see www.ft.com/cms/s/0/a523c27c-af7d-11dd-a4bf-000077b07658.html
[10] On 24 October 2006 Microsoft purchased a 1.6% share of Facebook for $246 million. www.microsoft.com/Presspass/press/2007/oct07/10-24FacebookPR.mspx
[11] See iCrossing Study Finds Most Mobile Internet Users Connect to Search, April 25th, 2007, at http://news.icrossing.com/press_releases.php?press_release=icrossing_mobile_search.
[12] Mobile TV Revenues to "skyrocket" says Informa, December 20, 2006, *at* http://www.mobilemarketingmagazine.co.uk/2006/12/mobile_tv_reven.html.
[13] Web search giants have presence in all market segments while the others specialise. Analysts predict an increasing integration of those four streams of mobile search. *See* Greg Sterling, Segmenting Local Mobile Search: The Major Players & Mobile Search Types, August 27, 2007, *at* http://searchengineland.com/070827-110648.php.
[14] OVUM Study, Mobile Search Comes of Age: Opportunities & Challenges, June 2007, p.3, *at* http://www.m-e-f.org/fileadmin/user/Suhail/Search_and_Discovery/MEF_Ovum_mobile_search_white_paper.pdf.
[15] Tameka Kee, Kelsey: Ad-Supported Directory Assistance Key To Future Mobile Search Boom, September 11, 2007, *at* http://publications.mediapost.com/index.cfm?fuseaction=Articles.showArticle&art_aid=67178.
[16] Mike Slocombe, 'Mobile Search Business to thit $2.4 billion by 2011', Digital Lifestyles, 17th July 2007
[17] Introna, L.; H. Nissenbaum: Shaping the Web: Why the Politics of Search Engines Matters. In: The Information Society, Vol. 16, No. 3, 2000, pp.169-186
[18] Van Couvering, Elizabeth 'Gatekeeping the web: cultural and economic origins of the biases of search engine results' unpublished
[19] van Eijk, N.: Search Engines: Seek and Ye Shall Find? The Position of Search Engines in Law. In: IRIS Plus, Vol.2, 2006; available at http://www.obs.coe.int/oea_publ/iris/iris_plus/iplus2_2006.pdf.en
[20] Battelle, J.: The search. How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture. London [Nicholas Brealey Publishing] 2005
[21] The Economist: Egalitarian Engines, November 17th, 2005
[22] Fortunato, S. et al.: The egalitarian effect of search engines. In: Proceedings of the National. Academies of Science (USA), Vol,103(34), 12684-12689, 2006; available at arxiv.org/abs/cs.CY/0511005
[23] Goldman, E.: Search Engine Bias and the Demise of Search Engine Utopianism. In: Yale Journal of Law and Technology, Vol.9, 2006, pp.188ff.
[24] Marcel Machill, Markus Beiler and Martin Zenker "Journalistische Recherche im Internet. Bestandsaufnahme journalistischer Arbeitsweisen in Print, Fernsehen, Radio und Online." Band 60 der Schriftenreihe Medienforschung der Landesanstalt für Medien NRW (ISBN 978-3-89158-480-4) Berlin: Vistas Publishers 2008, 412 pages.
English (shorter) version: Marcel Machill and Markus Beiler "The importance of the internet for journalistic research. Multi-method study on the research performed by journalists working for daily newspapers, radio, television and online." JOURNALISM STUDIES Vol. 8 (2008)
[25] 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
[26] Howe, D.C.; H. Nissenbaum: 'TrackMeNot' 2008, Information about the browser extension TrackMeNot (where it is also downloadable) is available at http://mrl.nyu.edu/~dhowe/trackmenot (last visited 3rd November 2008)

[27] 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.249 ff.
[28] Sullivan, D. Private Searches Versus Personally Identifiable Searches, 2006; available at http://blog.searchenginewatch.com/blog/060123-074811
[29] 'Opinion 1/2008 on data protection issues related to search engines', Article 29 Data Protection Working Party, Brussels 4 April 2008, available at http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2008/wp148_en.pdf
[30] "Response to the Article 29 Working Party Opinion on data protection issues related to search engines", 8 September 2008, Google, available at http://64.233.179.110/blog_resources/google_ogb_article29_response.pdf
[31] This is confirmed by a online survey with 5.200 youngsters carried out in July – August 2008. "Survey on eID", Wainer Lusoli, Ramón Compañó and Ioannis Maghiros, IPTS report forthcoming.
[32] Solove, D.: The Digital Person. Technology and Privacy in the Information Age. New York, NY [NYU Press] 2004
[33] 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

7 ANNEX

7.1 Functional landscape of the EU research projects

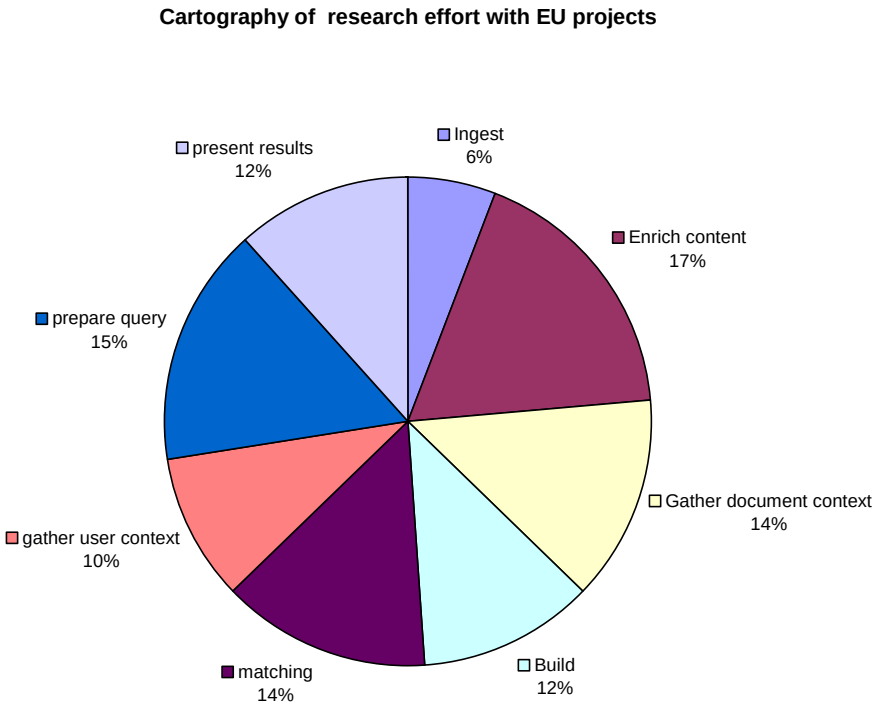


Figure 7.1 The graph maps research effort of the EU projects to the components of the functional breakdown

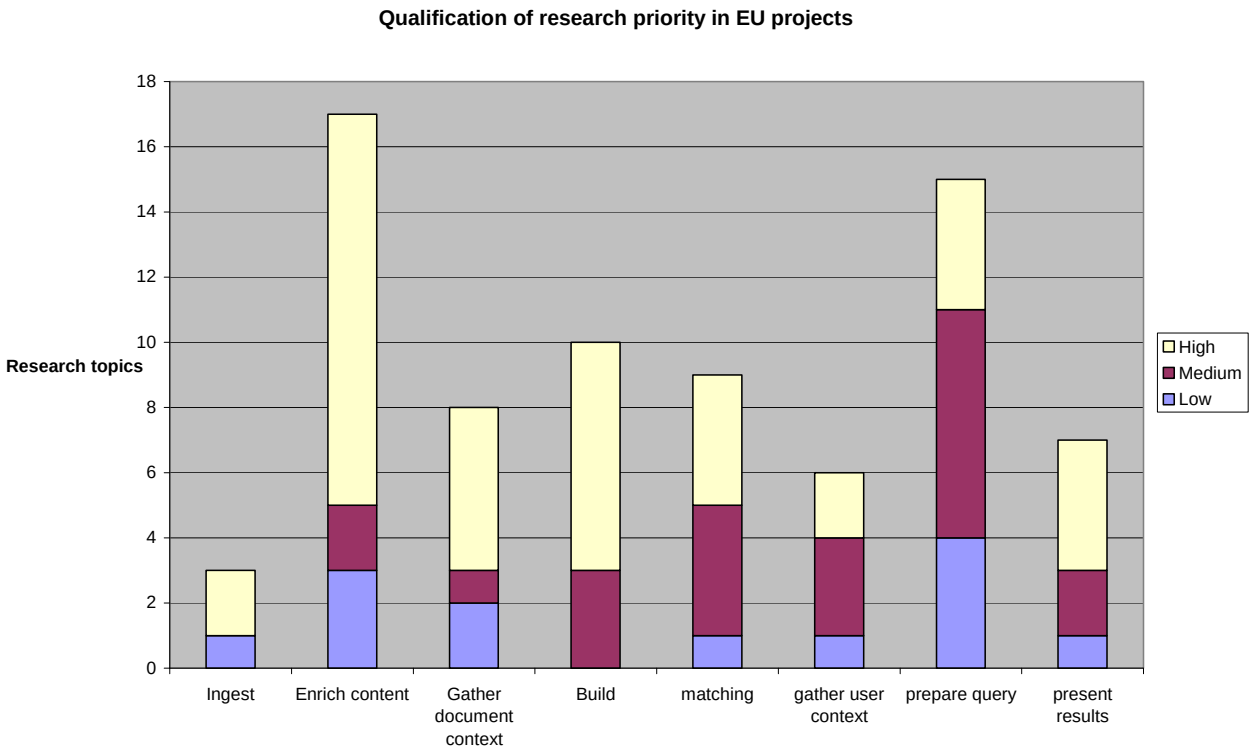


Figure 7.2 The graph above represents a qualification of research topics within a component of the functional breakdown.

7.2 Survey results functional breakdown

Function	Project	Research Topic	SoA Status	Progress Description
Ingest	TRIPOD	Gathering documents from digital libaries	Low	Working
	DIVAS	Fingerprint creation form compressed files	High	Succeded for several file formats
	AIMSHAPE	Certification of 3D Shapes	High	Tools for the analysis & synthesis of 3D shapes
Enrich context	VITALAS	Text data mining	Low	Improved text disambiguation tool
	VITALAS	Content-based Audio/Speech indexing	Medium	Generic low-level audio description methods for indexing and structuring audio content.
	VITALAS	Content-based Audio/Speech indexing	High	Topic detection of an audio segment based on automatic segment extraction and automatic transcription
	VITALAS	Content-based Image and Video indexing	High	Low and mid level visual content description methods for large scale video mining and structuring
	VITALAS	Content-based Image and Video indexing	High	Generic visual object recognition methods for automatic annotation
	RUSHES	Content-based Video indexing	High	Development of novel semantic description based on analysis of 3D scene structure and 3D camera motion
	PHAROS	Annotation process	High	Definition of a distributed and configurable multi-annotator and multi-modal annotation framework
	MESH	Multimodal indexing	High	Use simultaneously visual, audio (speech) and text sources to extract meaning
	AIMSHAPE	Geometry based annotation	High	Common shape ontology, tools for the automatic extraction of metadata
	VIDIVIDEO	video raw data indexing	High	good in TRECvid competition
	VICTORY	Annotation propagation	Low	Tool for automatic annotation of non-annotated 3D objects
	VICTORY	3D content-based indexing	High	Low and mid level visual content description methods for distributed indexing
	VICTORY	3D content-based indexing	High	Integration of different search methodologies
	SALERO	Semantic annotation	High	protopy satge
	MESH	Speech recognition	Medium	two-phase speaker independent speech transcription
	MESH	Video indexing	High	automatic concept detection from video

	SALERO	automatic extraction of semantic information from unstructured data	Low	
Gather Document Context	VITALAS	Content-based Image and Video indexing	High	Relevant visual thesaurus construction, such as automatic video structuring patterns extraction
	VITALAS	Cross-media indexing and retrieval	High	Tools for automatically generating vocabularies for cross-media retrieval, based on cross-media machine learning.
	TRIPOD	Building spatial ontology	High	Ongoing
	RUSHES	Onthology definition	Low	Automatic generation of user-centric onthology
	QUAERO	Gathering and fusionning multiple types of data from sources.	High	Capability to align scripts, new agency alerts etc, with video and audio content.
	PHAROS	training of the annotators	Medium	Annotators might be specifically trained on the desired collection of resources, before starting the actual indexing, to improve the performances
	AIMSHAPE	Geometry based 3D shape indexing	High	Common shape ontology, tools for the automatic extraction of metadata
	VICTORY	3D content-based search and retrieval in mobile devices	Low	Creation of a 3D object search engine for mobile devices
Build	VITALAS	Search scalability issue	High	Scalable cross-media document representation that can handle 1000 – 3000 Concepts
	VITALAS	Content-based Audio/Speech indexing	Medium	Large scale audio indexing for up to 10,000 hours of audio data
	VITALAS	Search scalability issue	High	Generic similarity search structures for large scale high dimensional features
	RUSHES	Metadata modelling	Medium	Customized metadata framework
	PHAROS	Indexing process	High	Definition of a distributed, possibly incremental, dynamic indexing process, automatically aware of right management
	MESH	Semantic indexing	Medium	Refer all metadata to domain knowledge structures for increased precision
	VIDIVIDEO	video raw data indexing	High	good in TRECvid competition
	VICTORY	P2P Network	High	Hybrid architecture allowing for sharing, downloading, searching and visualization of 3D objects and accompanied information (2D, sketch, video, text)
	VICTORY	Search scalability issue	High	Generic similarity search structures for large scale high dimensional features

	VITALAS	Content-based Image and Video indexing	High	Similarity search on ten thousand hours of video and ten millions of images
Matching	TRIPOD	Promote topic diversity in search results	Medium	Ongoing
	RUSHES	Content-based Image and Video indexing	Medium	Textual search based on semantic annotation of unedited raw video
	PHAROS	Query execution	Medium	Queries are performed by triggering different query systems
	MESH	semantic-based personalized query engine	Medium	process the query into the Knowledge Base taking into account user preferences for ranking
	AIMSHAPE	Semantic and Geometric search engines	High	similarity search (global and partial)
	VIDIVIDEO	concept based machine learning	High	TRECvid superior performance
	VICTORY	3D Content-based object retrieval	High	Similarity search on thousands of 3D objects using as query text, 3D object, 2D image, sketch
	VICTORY	Text, 2D, Sketch to 3D object retrieval	Low	Innovative methods for content-based cross modal 3D object retrieval
	VICTORY	3D Content-based search and retrieval	High	Innovative methods for 3D object retrieval (view-based, 3D information, graph-based)
Gather User Context	RUSHES	Collective annotation and socially derived profiling	Medium	New semantic taxonomies derived from social profiling in the conext of user scenarios
	TRIPOD	Gather user use data for a recommender system	Medium	Ongoing
	PHAROS	Social context	High	Extraction of social annotations and evaluations for improving the search results
	MESH	Distributed user context	Medium	Keep user profiles both at server and client-side, with different granularities
	MESH	Dynamic user contexts	High	Use profile acquisition at different temporal levels (short-term & stable) simultaneously
	VITALAS	Access rights	Low	Secure system focused on digital content mixed search engine, with secure data transference
Prepare query	VITALAS	Interactive and user adapted cross-media retrieval	Medium	New relevance feedback models based on cross-media document descriptions

VITALAS	Interactive and user adapted cross-media retrieval	Medium	Enhance query selection by analysis of user context, such as interaction logs
VITALAS	Personalization	Low	Personalized query interface focused on mixed digital content.
RUSHES	Interactive and user adapted cross-media retrieval	Medium	New relevance feedback models based on cross-media document descriptionson cross-media document descriptions
PHAROS	multi-modal queries	High	Queries can be formulated by mean of different pieces, each addressing different media characteristics. Queries can be formulated using different media (e.g., similarty queries on images)
MESH	Semantic processing	Medium	transform the query into a semantic query by mapping to elements in the defined knowledge structures
DIVAS	FIngerprint matching based	High	FIngerprint matching based on movie events (independent of file characteristics and coding technologies)
AIMSHAPE	Semantic search of shapes	High	The user is able to formulate graphical queries alleviating much of the typical burdens associated with dealing with query languages. The graphical representation is processed by the Search Engine and the query is translated to the appropriate format (nR
VIDIVIDEO	ontology and user interfaces	Medium	
VICTORY	Access rights	Low	Development of Identity management Unit
VICTORY	Access rights	Low	Development of novel blind 3D object watermarking methods
VICTORY	Personalization	Low	Methods for personalization of the retrieved results based on user's interests
VICTORY	Interactive and user adapted cross-media retrieval – Relevance feedback	Medium	Methods for achieving better retrieval accuracy using machine learning algorithms (per person/per session)
VICTORY	Interactive and user adapted cross-media retrieval	Medium	Enhance query selection by analysis of user context, such as interaction logs
VICTORY	Access rights	High	Development of Digital Right management Unit

Present results	VITALAS	Personalization	Low	Personalized result interface focused on mixed digital content.
	VITALAS	Visualization for navigation and search	High	Clustering visualization and navigation techniques for browsing in large scale audio visual repositories.
	RUSHES	AV timeline representation	High	Visual rendering of audio and video temporal structures
	PHAROS	Visual-effective representation of results	High	Results are shown comprising a user-friendly interface for their annotations. E.g., video results are displayed together with timelines and highlighted temporal segments where contents are found. Faceted search and refinements can be applied
	MESH	hypermedia browsing	Medium	present results with semantic links between video documents to improve browsing
	VIDIVIDEO	information visualisation	High	won prices on demo in conferences
	VICTORY	Visualisation and shared sessions	Medium	Methods and tools for visualization of 3D objects in mobile devices and shared sessions between multiple PCs and mobile devices

Table 7.1 Overview of submitted surveys

7.3 Use Case Typology (mind map view)

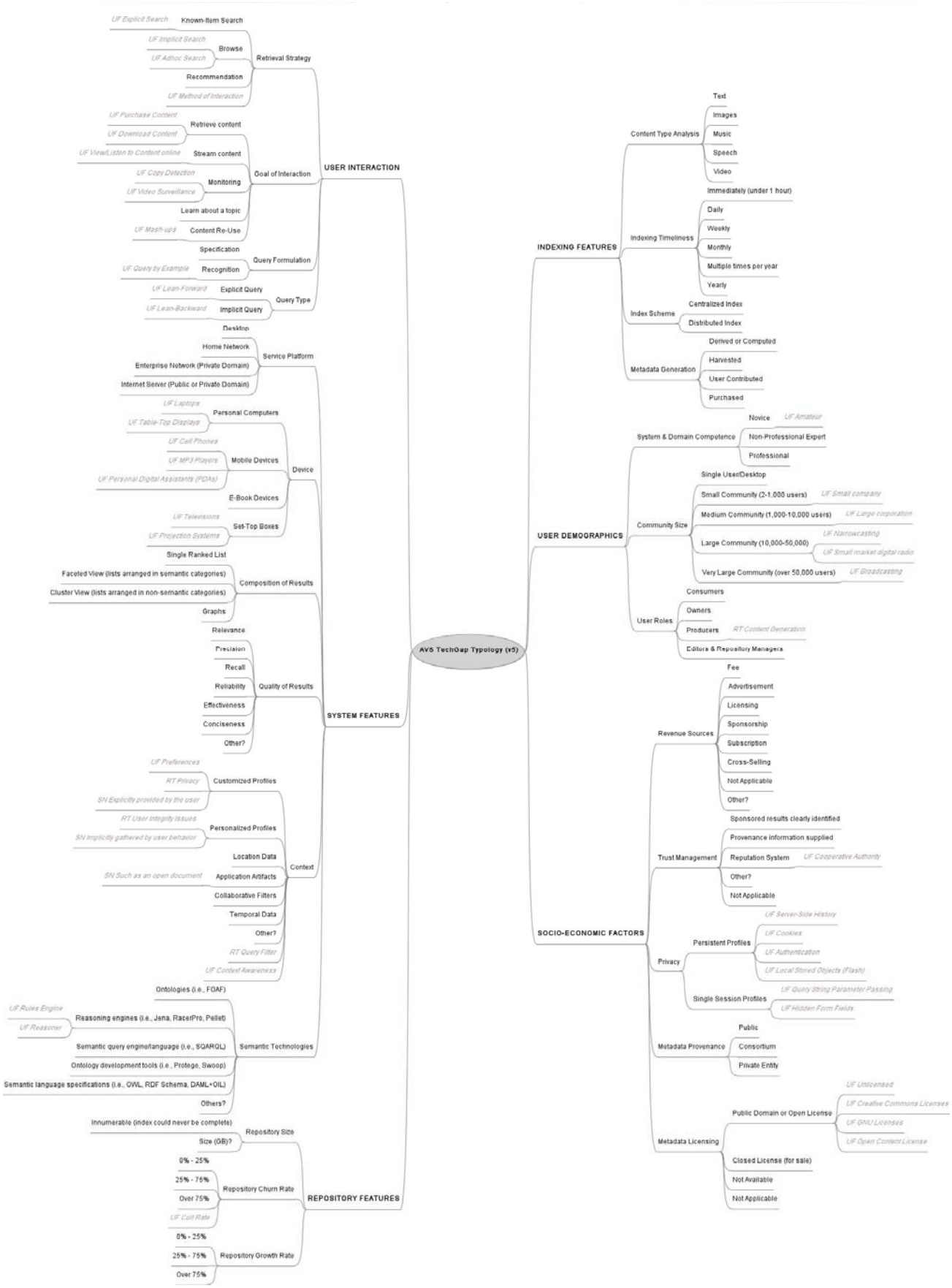


Figure 7.3 Use Case Typology (mind map view)

7.4 Use Case Typology (list view)

- **USER DEMOGRAPHICS**
    1. System & Domain Competence
        1. Novice (UF Amateur)
        2. Non-Professional Expert
        3. Professional
    2. Community Size
        1. Single User/Desktop
        2. Small Community (2-1,000 users) (UF Small company)
        3. Medium Community (1,000-10,000 users) (UF Large corporation)
        4. Large Community (10,000-50,000) (UF Narrowcasting, Small market digital radio)
        5. Very Large Community (over 50,000 users) (UF Broadcasting)
    3. User Roles
        1. Consumers
        2. Owners
        3. Producers (RT Content Generation)
        4. Editors & Repository Managers
- **SYSTEM FEATURES**
    1. Service Platform
        1. Desktop
        2. Home Network
        3. Enterprise Network (Private Domain)
        4. Internet Server (Public or Private Domain)
    2. Device
        1. Personal Computers (UF Laptops, Table-Top Displays)
        2. Mobile Devices (UF Cell Phones MP3 Players , Personal Digital Assistants (PDAs))
        3. E-Book Devices
        4. Set-Top Boxes (UF Televisions, Projection Systems)
    3. Composition of Results
        1. Single Ranked List
        2. Faceted View (lists arranged in semantic categories)
        3. Cluster View (lists arranged in non-semantic categories)
        4. Graphs
    4. Quality of Results
        1. Relevance
        2. Precision
        3. Recall
        4. Reliability
        5. Effectiveness
        6. Conciseness
        7. Other?
    5. Context (RT Query Filter; UF Context Awareness)
        1. Customized Profiles (UF Preferences; RT Privacy)
        2. Personalized Profiles (RT User Integrity Issues)
        3. Location Data
        4. Application Artifacts
        5. Collaborative Filters
        6. Temporal Data
        7. Other?
    6. Semantic Technologies
        1. Ontologies (i.e., FOAF)
        2. Reasoning engines (i.e., Jena, RacerPro, Pellet) (UF Rules Engine, Reasoner)
        3. Semantic query engine/language (i.e., SQARQL)
        4. Ontology development tools (i.e., Protege, Swoop)
        5. Semantic language specifications (i.e., OWL, RDF Schema, DAML+OIL)
        6. Others?
- **USER INTERACTION**
    1. Retrieval Strategy
        1. Known-Item Search (UF Explicit Search)
        2. Browse (UF Implicit Search, Adhoc Search)
        3. Recommendation (UF Method of Interaction)
    2. Goal of Interaction

    1. Retrieve content (UF Purchase Content, Download Content)
    2. Stream content (UF View/Listen to Content online)
    3. Monitoring (UF Copy Detection, Video Surveillance)
    4. Learn about a topic
    5. Content Re-Use (UF Mash-ups)
  3. Query Formulation
    1. Specification
    2. Recognition (UF Query by Example)
  4. Query Type
    1. Explicit Query (UF Lean-Forward)
    2. Implicit Query (UF Lean-Backward)

- **REPOSITORY FEATURES**
  1. Repository Size
    1. Innumerable (index could never be complete)
    2. Size (GB)?
  2. Repository Churn Rate (UF Cull Rate)
    1. 0% - 25%
    2. 25% - 75%
    3. Over 75%
  3. Repository Growth Rate
    1. 0% - 25%
    2. 25% - 75%
    3. Over 75%

- **SOCIO-ECONOMIC FACTORS**
  2. Revenue Sources
    1. Fee
    2. Advertisement
    3. Licensing
    4. Sponsorship
    5. Subscription
    6. Cross-Selling
    7. Not Applicable
    8. Other?
  3. Trust Management
    1. Sponsored results clearly identified
    2. Provenance information supplied
    3. Reputation System (UF Cooperative Authority)
    4. Other?
    5. Not Applicable
  4. Privacy
    1. Persistent Profiles (UF Server-Side History, Cookies, UF Authentication, Local Stored Objects (Flash))
    2. Single Session Profiles (UF Query String Parameter Passing, Hidden Form Fields)
  5. Metadata Provenance
    1. Public
    2. Consortium
    3. Private Entity
  6. Metadata Licensing
    1. Public Domain or Open License (UF Unlicensed, Creative Commons Licenses, GNU Licenses, Open Content License)
    2. Closed License (for sale)
    3. Not Available
    4. Not Applicable

- **INDEXING FEATURES**
  1. Content Type Analysis
    1. Text
    2. Images
    3. Music
    4. Speech
    5. Video
  2. Indexing Timeliness
    1. Immediately (under 1 hour)
    2. Daily

    3. Weekly
    4. Monthly
    5. Multiple times per yeay
    6. Yearly
3. Index Scheme
    1. Centralized Index
    2. Distributed Index
4. Metadata Generation
    1. Derived or Computed
    2. Harvested
    3. User Contributed
    4. Purchased

7.5 Use Case Survey

The term within brackets that precedes each question correlates to a node in the *Use Case Typology*. These were omitted on the final survey, but are included in this document for clarity. Parenthetical comments following the questions indicate how choices were constrained.

- **USER DEMOGRAPHICS**
    - [*SYSTEM & DOMAIN COMPETENCE*] What competence level does your typical user have in regard to your system and knowledge domain? (1 choice)
        - Novices
        - Non-Professional Experts
        - Professionals
    - [*COMMUNITY SIZE*] How big is your targeted community? (1 choice)
        - Single User/Desktop
        - Small Community (2-1,000 users) (e.g., small company)
        - Medium Community (1,000-10,000 users) (e.g., large corporation)
        - Large Community (10,000-50,000) (e.g., internet "narrowcasting", small market digital radio)
        - Very Large Community (over 50,000 users) (e.g., broadcasting)
    - [*USER ROLES*] What relationship(s) do your primary users have with the content you provide access to? (Rank 2)
        - Consumers (do not create content)
        - Producers (create content)
        - Owners & Sellers (original content distributors)
        - Editors & Repository Managers
- **SYSTEM FEATURES**
    - [*SERVICE PLATFORM*] Where is your service designed to be hosted? (1 choice)
        - Desktop
        - Home Network
        - Enterprise Network (Private Domain)
        - Internet Server (Public or Private Domain)
    - [*DEVICE*] Which devices have you explicitly designed your search service for? (Rank 2.)
        - Personal Computers (including laptops, tabletop displays)
        - Mobile Devices (cell phones, PDAs, MP3 players)
        - E-Book Devices (using e-ink)
        - Set-Top Boxes (television, projection systems)
    - [*COMPOSITION OF RESULTS*] How does your system primarily display search results? (1 choice.)
        - Single Ranked List
        - Faceted View (lists arranged in semantic categories)
        - Cluster View (lists arranged in non-semantic categories)
        - Graphs
    - [*QUALITY OF RESULTS*] How do you measure the quality of results your system provides? (Rank all that apply.)
        - Relevance - Numerical score is assigned to a search result representing how well the information need of the user that issued the search query was met
        - Precision - The number of relevant resources retrieved by a search divided by the total number of resources retrieved by that search.
        - Recall - The number of relevant resources retrieved by a search divided by the total number of existing relevant resources.
        - Reliability - The extent to which the search service yields consistent, stable, and uniform results over time.
        - Effectiveness - Measure by which a user can accomplish their task.

        - Conciseness - Economy in communicating search service results that is achieved by expressing a great deal in just a few words
        - Other?
    - [*CONTEXT*] What kinds of contextual data does your system use? (Rank 3)
        - Customized profiles - personal details explicitly provided by the user (i.e., ratings, preferences)
        - Personalized profiles - personal details are derived from the analysis of user behavior (i.e., items purchased, pages viewed)
        - Location data
        - Application artifacts (i.e., open documents)
        - Collaborative filters (profiles from other similar users)
        - Temporal data
        - Other?
    - [*SEMANTIC TECHNOLOGIES*] What kind of semantic tools are being used by your service? (Choose all that apply.)
        - Ontologies (i.e., FOAF)
        - Reasoning engines (i.e., Jena, RacerPro, Pellet)
        - Semantic query engine/language (i.e., SQARQL)
        - Ontology development tools (i.e., Protege, Swoop)
        - Semantic language specifications (i.e., OWL, RDF Schema, DAML+OIL)
        - Others?
- **USER INTERACTION**
    - [*RETRIEVAL STRATEGY*] What is the primary retrieval strategy of a typical user after they are familiar with your system? (1 choice.)
        - Known-item search (identification of a specific resource)
        - Browse (exploration of a collection for discovery purposes)
        - Recommendation-based "search"
    - [*GOAL OF INTERACTION*] What is the goal for most of your users? (Rank 3.)
        - Retrieve content (download, view or listen online, purchase)
        - Learn about a topic (not retrieval of content)
        - Monitoring - Identification of anomalies within a continuous data stream using feature extraction techniques for image, video or voice data (e.g., video surveillance, copy detection)
        - Content Re-Use - Re-packaging and re-distribution of content with added value (i.e., mash-ups)
    - [*QUERY FORMULATION*] How do most users formulate typical queries on your system? (1 choice.)
        - Specification - Query with known metadata about the content
        - Recognition - Metadata for the desired content is unknown, so metadata from related content is used (viz., query by example)
    - [*QUERY TYPE*] What types of queries does your system primarily use? (Rank 3)
        - Explicit queries (formulated by the user)
        - Implicit queries (formulated by the system from profile, history, and other usage data)
- **REPOSITORY FEATURES**
    - [*REPOSITORY SIZE*] What size of document repository are you targeting? (1 choice)
        - Innumerable (index could never be complete)
        - Size (GB): _____
    - [*REPOSITORY CHURN RATE*] What is the likely cull rate (removal of obsolete index entries) per year for targeted repository documents that you index? (1 choice)
        - 0% - 25%
        - 25% - 75%
        - Over 75%
    - [*REPOSITORY GROWTH RATE*] What is the likely growth rate per year for targeted repository documents that you index? (1 choice)
        - 0% - 25%
        - 25% - 75%
        - Over 75%
- **SOCIO-ECONOMIC FACTORS**
    - [*METADATA PROVENANCE*] Who owns most of the metadata used by your service? (1 choice)
        - Public
        - Consortium
        - Private Entity
    - [*METADATA LICENSING*] Under what conditions do you make available to third parties the metadata you produce? (1 choice)
        - Public domain or open license (copy and modification rights do not require fiscal payment)
        - Closed license (for sale)
        - Not available

        - Not applicable (we do not produce our own metadata)
    - [*PRIVACY*] Is personal information maintained on your system? (1 choice.)
        - Yes (Persistent Profiles)
        - No (Single Session Profiles)
    - [*TRUST MANAGEMENT*] What strategies are used to increase system transparency so that potentially biasing factors affecting results can be identified? (Choose all that apply.)
        - Sponsored results are clearly identified
        - Provenance information is provided
        - Reputation system
        - Other: _____
        - Not applicable
    - [*REVENUE SOURCES*] What is the most likely revenue source for a business using your system? (Rank 3 choices.)
        - Fee
        - Advertisement
        - Licensing
        - Sponsorship
        - Subscription
        - Cross-Selling
        - Not Applicable
- **INDEXING FEATURES**
    - [*CONTENT TYPE ANALYSIS*] What kinds of content features are primarily processed to build your index or to significantly differentiate your index from others? (Rank all that apply)
        - Text (i.e., Natural Language Processing)
        - Images (i.e., edge or color detection)
        - Music (i.e., music-to-text annotation)
        - Speech (i.e., speech-to-text annotation)
        - Video (i.e., motion processing)
        - Other: _____
    - [*INDEXING TIMELINESS*] How quickly would new content generally be annotated, published and made available? (1 choice)
        - Immediately (under 1 hour)
        - Daily
        - Weekly
        - Monthly
        - Multiple times per year
        - Yearly
    - [*INDEX SCHEME*] What kind of index does your service create and utilize? (1 choice.)
        - Centralized Index
        - Distributed Index
    - [*METADATA GENERATION*] How is most of your metadata generated? (Rank 2)
        - Derived or computed (metadata is inferred from content)
        - Harvested (associated metadata is well defined for the content, so it is copied with minimal inferential processing)
        - User Contributed (i.e., includes collaborative filtering based on user preferences)
        - Purchased

7.6 Market Segment Survey

- [*TYPE OF RETRIEVAL*] Which types of retrieval are most important for your service? (Rank 2)
    1. General Recommendations - Documents or resources similar to the item(s) of interest (queried resources) are presented or suggested.
    2. Repository Recommendations - New content is suggested for adding to a repository, library or corpus.
    3. Targeted Content - Personalized content, such as advertisements or public service announcements, are displayed.
    4. Social Networks - The item of interest (queried resource) is a living entity whose interdependent relationships are analyzed to reveal and display other closely related entities; used in scientific fields such as epidemiology, biology, economics, and law enforcement (terrorist networks).
- [*REPOSITORY MANAGEMENT*] How is the content in your targeted repository organized? (1 choice)
    1. Unorganized - Large variety of disparate content that, for the most part, does not reside in managed collections.
    2. Semi-Organized - Homogeneous content resides in various collections that are managed but without consistent criteria across the collections (isolated silos); or, content resides in one or more collections that may have previously been cataloged for other purposes or are not managed by an information specialist

(i.e., librarian).
3. Organized - Homogeneous content from one or more collections that are professionally managed by an information specialist who catalogs all new content.

- [*REPOSITORY ACCESS RIGHTS*] What access rights do end users have to the content in the targeted repository? (Note: Do not confuse with Repository Ownership.) (1 choice)
  1. Unrestricted Access - Users can view most content items.
  2. Restricted Access - User permissions to view some content is often restricted due to security clearance issues or pending remittance of payment
  3. Not relevant
- [*REPOSITORY OWNERSHIP*] Who owns most of the content in the targeted repository? (1 choice)
  1. Public Content (unlicensed or openly licensed and accessible by the general public)
  2. Private Content (closed license or not accessible to the general public)
- [*CONTENT ACQUISITION*] How does your system acquire most new content? (1 choice)
  1. Retrieved Content (i.e., web crawler)
  2. Submitted Content (i.e., by a user or librarian)

7.7 Glossary

Advertisement - Revenue generated by a paid announcement or product promotion appearing in the search service interface.

Browse - Retrieval characterized by exploratory behavior within a content collection for the purposes of discovery and research.

Centralized Index - A metadata catalog that is stored on a single host.

Closed License - Copy and modification rights are exclusively held by a closed entity which requires fiscal payment for the transfer of those rights.

Cluster View - *Items are listed within unnamed (or loosely named) categories.*

Conciseness - *Economy in communicating search service results that is achieved by expressing a great deal in just a few words.*

Consumers - *The end point in the supply chain for the information provided by the service.*

Content-based image retrieval (CBIR) - *The application of computer vision to the image retrieval problem*

Content Churn Rate - *The rate at which existing content is considered obsolete and removed. (Measured as a percentage of the total number of documents replaced per year.)*

Content Re-Use - *The re-packaging and re-distribution of content with added value (i.e., mash-ups).*

Content Type Analysis - *Content features that are processed for the purpose of generating metadata for an index.*

Controlled Vocabulary - *A controlled indexing language formally organized so that a priori relationships between concepts are made explicit (i.e., thesauri)*

Cross-Selling - *Revenue generated from selling an additional product or service to the user (i.e., the music search engine Seeqpod sells concert tickets to users who have searched for an artist who will be playing in the user's geographical area.)*

Customized Profiles - *The service tailors information for each user based on personal details or characteristics that are explicitly provided (i.e., ratings, preferences).*

Derived or Computed - *Metadata that is inferred from it's associated content.*

Desktop - *The service exists as a locally executed program on a computer.*

Distributed Index - A metadata catalog that is stored among many hosts.

E-Book Device - *A low-power electronic computing device that uses electronic paper.*

Editors & Repository Managers - *A content provider actor who adds value to an information resource by submitting changes in format, grammar, semantics, sequence and/or context to individual documents or by managing the document or its collection.*

Effectiveness - *Measure by which a user can accomplish their task.*

Enterprise Network - *The service is hosted on a private (i.e., business) network.*

Explicit Query - *A query that is explicitly formulated by the user (i.e., "Lean-Forward Query").*

Faceted View - *Items are listed within named categories (groups, facets).*

Fee - *Revenue generated from a fixed charge paid by a downstream user or distributor of the search service.*

Graphs - *Rather than listing items that are arranged in graphical categories.*

Harvested - *Associated metadata is well defined for the content so it is copied with minimal inferential processing.*

Home Network - *The service is hosted on a small, domestic network.*

Implicit Query - *A query that is not actively formulated by the user but constructed by the system from the user's behavior and profiles (i.e., "Lean-Backward Query").*

Indexing Timeliness - *The speed which new content is annotated and published by the search service. (NOTE: Critical for the monetization of content. For example, a clip of a goal is worth a lot in the 10 minutes (up to maybe 10 hours) following the goal, but is valueless a week later.)*

Innumerable - *Size of the repository is unbounded, incalculable or not important.*

Internet Server - *The service is provided through the primary, public, global network. (Public or private domain.)*

Known-Item Search - *Retrieval characterized as the identification of a specific resource.*

Large Community - 10,000-50,000 users (i.e., narrowcasting, small market digital radio).

Learn About a Topic - *The primary goal is not to retrieve an item, but to learn about a topic.*

Licensing - *Revenue generated as royalty paid for the transfer of intellectual and proprietary knowledge to a third party.*

Medium Community - *1,000-10,000 users; (i.e., large corporation).*

Metadata - *Information that serves as a document (resource) surrogate and is used for matching to a query during a search session. Metadata about control, management, and usage is considered out of scope.*

Metadata Licensing - *The conditions under which metadata created by the project is made available to third parties.*

Mobile Device - *A compact computing device designed for travel. System interaction is limited by a primitive input device, a miniature keyboard, or a touchscreen and output is printed to a very small display screen (i.e., cell phones, PDAs, mp3 players).*

Monitoring - *Identification of anomalies within a continuous data stream using feature extraction techniques for image, video or voice data.*

Non-Professional Expert - *Domain experts who regularly provide content but do so as an avocation. Any proceeds generated from their activities are mainly used to offset expenses for the service and are not a significant source of financial income (i.e., bloggers).*

Novice - *An end user who lacks extensive experience with the type of service provided and who does not expect advanced capabilities. They would typically use the service to accomplish personal goals.*

Ontologies - *A representation of a set of concepts within a domain and the relationships between those concepts that is used to reason about the properties of that domain, and may be used to define the domain.*

Open License - *Copy and modification rights do not require fiscal payment (i.e., Wikipedia, Creative Commons Licenses, GNU Licenses, Open Content Licenses)*

Owner - *An entity possessing legal rights to the information resources provided by the servic*e.

Peer-to-peer (p2p) - computer network architecture that uses diverse connectivity between participants in a network and the cumulative bandwidth of network participants rather than conventional centralized resources.

Persistent Profiles - *Information used to differentiate users and maintain data related to the user during navigation and across multiple visits.*

Personal Computer - *A consumer desktop, laptop or tabletop (display) computer.*

Personalized Profiles - *The service tailors information and/or the user interface for each user based on personal details or characteristics that are implicitly provided (i.e., items purchased or pages viewed).*

Precision - *The measure of exactness or fidelity of a search service; The number of relevant resources retrieved by a search divided by the total number of resources retrieved by that search.*

Privacy - *The amount and nature of information that the search service maintains about its users.*

Producers - *A content provider actor who overseas the management of creating the information resources provided by the service.*

Professionals - *An end user who possesses considerable experience with the type of service provided and who is knowledgeable and expects advanced capabilities from the service. They would typically use the service for business purposes.*

Public Domain - *Unlicensed.*

Recall - *The measure of comprehensiveness of a search service; The number of relevant resources retrieved by a search divided by the total number of existing relevant resources (which should have been retrieved).*

Recognition - *Metadata for the desired content is unknown, so metadata from related content is used (viz., query by example).*

Recommendation - *(Recommendation system) A content-based filtering technique that rates a collection of items by comparing data between user profiles and metadata about the information item. (RT: Reputation System)*

Relevance - *Numerical score assigned to a search result that represents how well the information need of the user that issued the search query was met.*

Reliability - *The extent to which the search service yields consistent, stable, and uniform results over time.*

Repository Growth Rate - *The rate at which the size of the overall data repository varies over time. (Measured as an average percentage increase of the total dataset size per year.)*

Repository Size - *Size of the document repository that is targeted by the search service.*

Reputation System - *A collaborative-based filtering technique that rates a collection of people using the opinions they have about one another and then promotes information items associated with those people. (RT: Recommendation)*

Retrieval of Content - *The primary goal is to find and retrieve an item (viz., downloaded, viewed online, purchased).*

Retrieve Content - *To purchase or download content.*

Revenue Sources - *Primary source of revenue stream for achieving sustainable, long-term profitability.*

Set-Top Box - *Television, projection systems.*

Single Ranked List - *Items are listed without categorization.*

Single Session Profiles - *Information used to differentiate users during navigation for a single visit. Data related to the user is not maintained after the session ends.*

Small Community - *2-1,000 users (i.e., small company).*

Specification - *Query with known metadata about the content.*

Sponsorship - *Revenue generated from fees paid for granting the right to associate another organization's name, products or services with the search service or company.*

Subscription - *Revenue generated from a licensing agreement in which a buyer purchases the search service for a specified period of time.*

System & Domain Competence - *Expertise of the typical user in regards to the knowledge domain and the search system.*

Trust Management - *The credibility of the search service in regards to providing non-biased results.*

User Roles - *The relationship of a typical user to the information they receive from the service.*

Very Large Community - *Over 50,000 users (i.e., broadcast community).*

7.8 Socio-Economic Workshop

The Institute for Prospective for Technological Studies organized a workshop on 29th and 30th September 2008. The list of participants, in alphabetical order, and the agenda of the workshop is given underneath.

7.8.1 Participant List

Name	Organisation
Loretta Anania	European Commission DG INFSO D2, Brussels
Rolf Bardeli	Fraunhofer, St Augustin
Francesco Barbarani	Fox Interactive Media - My Space.com, Milano
Nozha Boujemaa	INRIA – IMEDIA, Rocquencourt
Markus Bylund	Swedish Institute of Computer Science, Stockholm
Cécile Chamaret	Ecole polytechnique- CNRS, Paris
Leonardo Cervera Navas	European Commission, DG Markt, Brussels
Mihai Datcu	DLR, Cologne
Christoph Dosch	IRT, Munich
Christoph Glauser	Institute for Applied Argumentation Research IFAAR, Berne
José Luis Gómez	UNED, Madrid
Henri Gouraud	JCP Consult, Rennes
Gregory Grefenstette	Exalead S.A., Paris
Lucas Introna	Lancaster University Management School, Lancaster
Jussi Karlgren	Swedish Institute of Computer Science, Stockholm
Markus Kauber	JCP Consult, Rennes
Joachim Klerx	Austrian Research Centers GmbH, Vienna
Per Koch	Pandia.com, Oslo
Susanne Koch	Pandia.com, Oslo
Nicklas Lundblad	Google Sweden, Stockholm
Marcel Machill	University of Leipzig
Sjoera Nas	Dutch Data Protection Office, Den Haag
Javier Oliete	Ovilgy One Worldwide, Barcelona
Robert Ortgies	IRT, Munich
Fabrizio Porrino	European Commission, DG INFSO D2, Brussels
Wolfgang Sander-Beuermann	Suchmachinen Verein, Hannover
Nicu Sebe	University of Amsterdam
Eric Oluf Svee	Swedish Institute of Computer Science, Stockholm
Jan Schallaböck	Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein, Kiel
Hendrik Speck	University of Applied Sciences Kaiserslautern
Elizabeth van Couvering	London School of Economics
Pieter van der Linden	Thomson, Paris
Nico van Eijk	University of Amsterdam
David Broster	IPTS
Margherita Bacigalupo	IPTS
Ramón Compañó	IPTS
Claudio Feijóo	IPTS
Ioannis Maghiros	IPTS
Wainer Lusoli	IPTS
Marc van Lieshout	IPTS

7.8.2 Agenda of the workshop

Day 1 – Monday, 29th September 2008

Opening

09:20 – 09:15 Welcome to the IPTS
by Dave Broster (IPTS, European Commission),
09:15 – 09:30 Objectives of the Workshop
by Ramón Compañó (IPTS), Loretta Anania (DG Information Society)
09:30 – 10:00 Roundtable Presentation

Techno-economic trends (Chair R. Compañó)

10:00 – 10:10 Are there any disruptive and groundbreaking applications in sight?
by Nozha Boujemaa, Inria – Imedia, France
10:10 – 10:20 Issues arising from the current techno-economic model of web search
by Hendrik Speck, Univ. Kaiserslautern, Germany
10:20 – 10:30 From web to mobile search: barriers and opportunities
by Christoph Glauser, Institut für angewandte Argumentenforschung Switzerland
10:30 – 10:40 The future of Enterprise Search
by Gregory Grefenstette, Exalead, France
10:40 – 10:50 Trends in online advertising
by Francesco Barbarani, Fox Interactive Media - MySpace.com & Italian Internet Advertising Bureau, Italy
10:50 – 11:10 Coffee Break
11:10 – 13:10 Discussion *(Chair Ioannis Maghiros, IPTS)*
Introduction to discussion presenting results from the Questionnaire
13:10 – 13:30 Session Conclusion
by Ioannis Maghiros

Search engines and Society (Chair L. Anania)

14:30 – 14:40 Search engines and the Citizen: Possibilities, interests and risks
by Lucas Introna, Univ. Lancaster, United Kingdom
14:40 – 14:50 How Social Web will impact Web search.
by Susanne Koch, Pandia.com, Norway
14:50 – 15:00 Gatekeeping the web: cultural and economic origins of the biases of search engine results
by Elizabeth Van Couvering, London School of Economics, UK
15:00 – 15:10 How to overcome the 'monoculture' of the search engine landscape in Europe
by Wolfgang Sander-Beuermann, Univ. Hannover and Suchmachinen Verein, Germany
15:10 – 15:30 Coffee Break
15:30 – 15:40 The implications of search engines for journalism and media policy
by M. Machill, Univ. Leipzig, Germany
15:40 – 15:50 Legal gaps and options to reduce them
by Nico van Eijk, Institute for Information Law, Univ. Amsterdam, The Netherlands
16:00 – 17:30 Discussion *(Chair: Ramón Compañó, IPTS)*
Introduction to discussion presenting results from the Questionnaire
17:30 – 18:00 Session Conclusion
by Ramón Compañó

Day 2 – Tuesday, 30th september 2008

09:00 – 09:10 Welcome and coffee
09:10 – 09:20 Summary of findings of the first day
by Claudio Feijóo (IPTS), Fabrizio Porrino (DG Information Society)
09:20 – 09:40 Discussion and validation of main conclusions

Privacy and Search engines (Chair Ramón Compañó)

09:40 – 09:50 Tug of war between value for users and need for privacy: Where are the limits?
by Per Koch, Pandia.com, Norway
09:50 – 10:00 How much profiling is needed for targeted advertising
by Javier Oliete, Ogilvy Worldwide, Spain
10:00 – 10:10 Transparency and meaningful user choice in privacy design
by Nicklas Lundblad, Google, Sweden
10:10 – 10:20 How can civil society contribute to privacy challenges?
By Joachim Klerx, ARC-sys, Austria
10:20 – 10:40 Coffee Break
10:40 – 10:50 Self-regulation, co-regulation and regulation: what is the right mix?
by Jan Schallaböck, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein, Germany
10:50 – 11:00 Gaps and Options for protecting the privacy of citizens across the

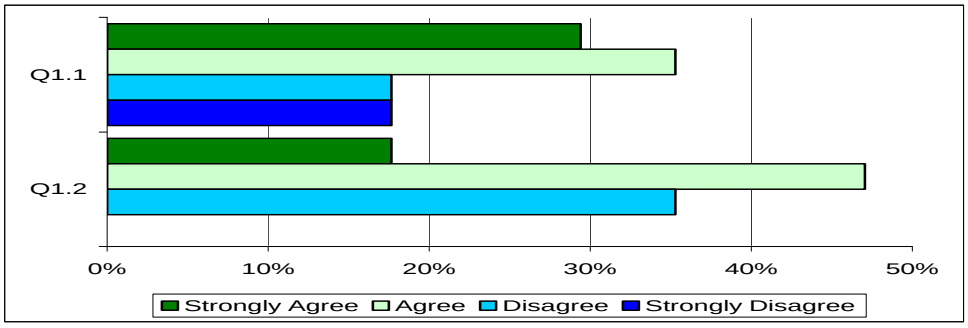
European Union
by Sjoera Nas, Dutch Data Protection Office, The Netherlands
11:00 – 13:00 Discussion *(Chair: Ioannis Maghiros, IPTS)*
Introduction to discussion presenting results from the Questionnaire
13:00 – 13:30 Session Conclusion
by Ioannis Maghiros
13h30-14h30 Lunch
14:30 – 15:00 Second Round Questionnaire
by participants
15:00 – 15:30 Comparison of attitude change of participants to questionnaire before and after the workshop
by Wainer Lusoli, IPTS
15:00 – 15:30 Workshop Conclusions &Wrap-Up
by Ramón Compañó (IPTS), Loretta Anania (DG Information Society)
16:00 Close

7.9 Survey Results Socio-Economic Workshop

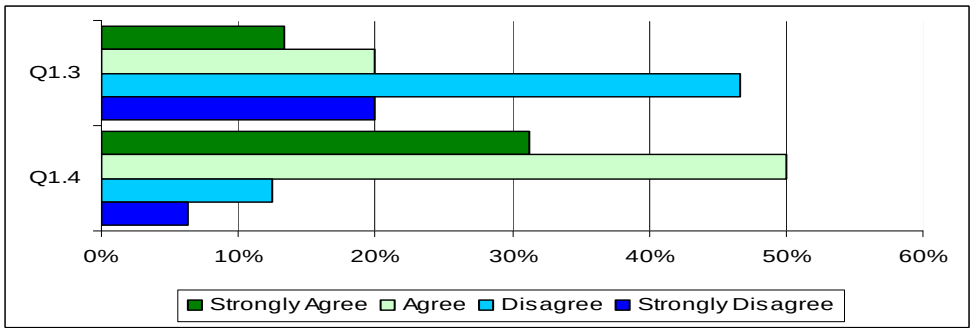
A workshop on socio-economic matters took place on 29th and 30th September 2008. The participants to this workshop were asked to fulfil a questionnaire prepared by the IPTS, which was completed by 28 people. After the event, the panel was asked to revise their original responses and to fill again the same questionnaire, of which 16 out of the previous 28 did. The second round can be considered as a fair reflection of the views of the panel after having exchanged arguments with respect of each of the questions. Therefore, –although not being a Delphi exercise in the strict sense– it gives an indication of the level of consensus for each of the questions.

The survey results are reported by the sequence of discussion in the workshop them (trends, techno-economic aspecs, social aspects, privacy, policy options). Their numbering (QX.Y) is the same than in the body of the report

1. Trends

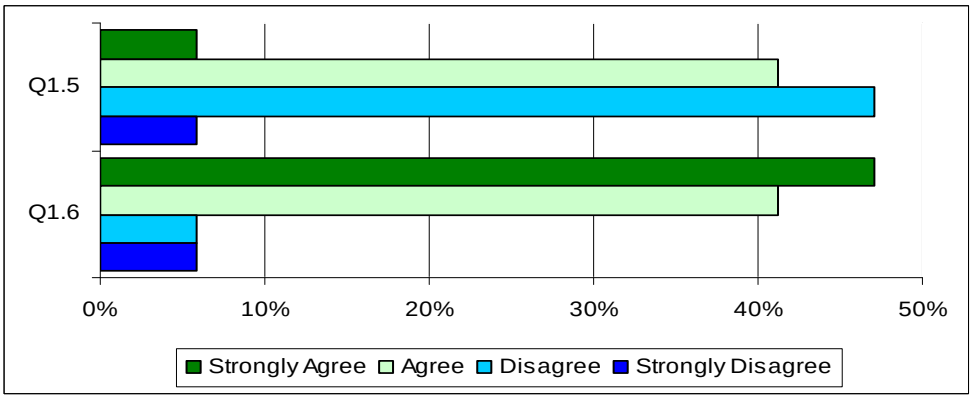


Q1.1 By 2013: Semantic search will be available
Q1.2 By 2013: Audio-visual search technology will become as performing as today's text-search



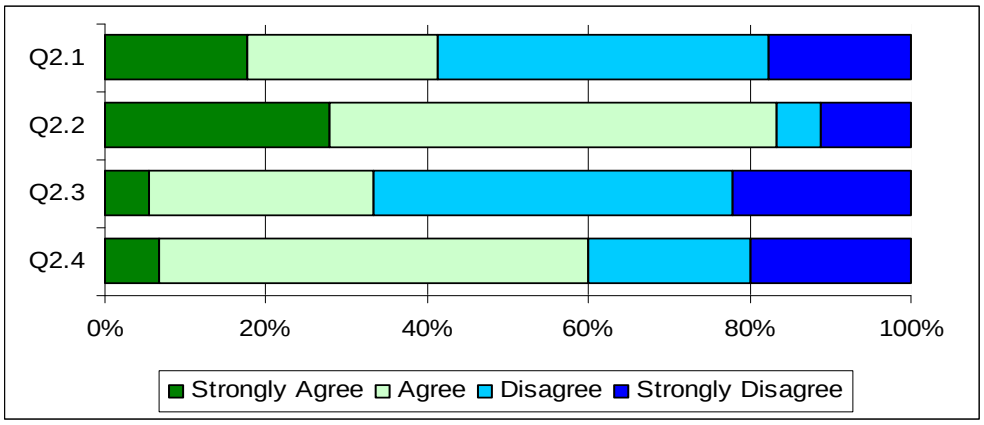
Q1.3 By 2013: Mobile search will overtake desktop search
Q1.4 By 2013: Location-based search will become a 'killer' application enabled by wireless grids (e.g. mobile communications, Wifi, sensor networks, RFID-networks) and GPS.

Q1.5 By 2013: Search engines providers will have sidelined traditional media players (e.g. by reducing their advertising

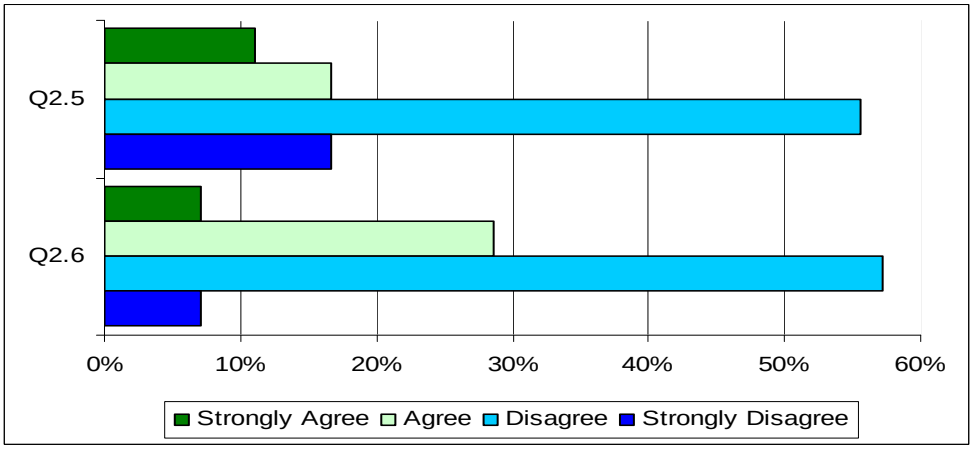


share, by providing customised services, etc.)
Q1.6 Until 2013: Search engines providers will remain important drivers of innovation for internet applications.

2. Techno-economic aspects

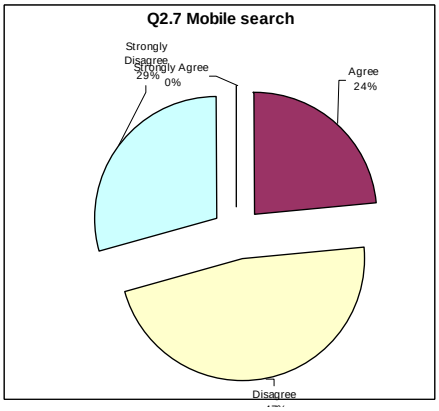


Q2.1 Web search: there is robust competition amongst search engines (e.g. low switching costs for users, innovation based competition, etc.)
Q2.2 Web search: is a natural oligopoly market (e.g. few dominant search engines and possibly several smaller thematic search engines)
Q2.3 Web search: there will be no space left for newcomers (e.g. the entry barrier is insurmountable)

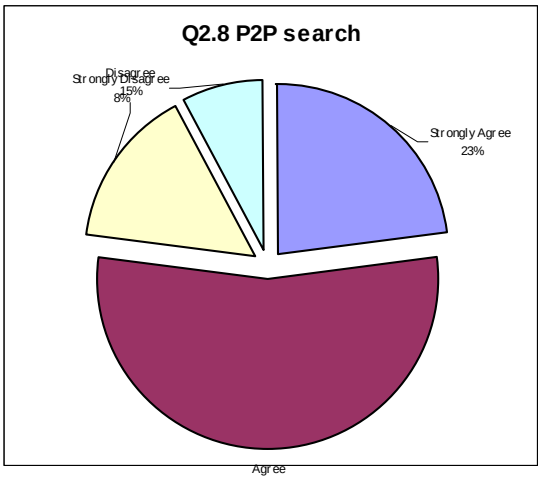


Q2.4 Web search: advertising is the only viable business model

Q2.5 Enterprise search: tailored solutions will disappear (e.g. future web search engines will provide similar good job)
Q2.6 Enterprise search: most of today's companies will be acquired by web search engine providers

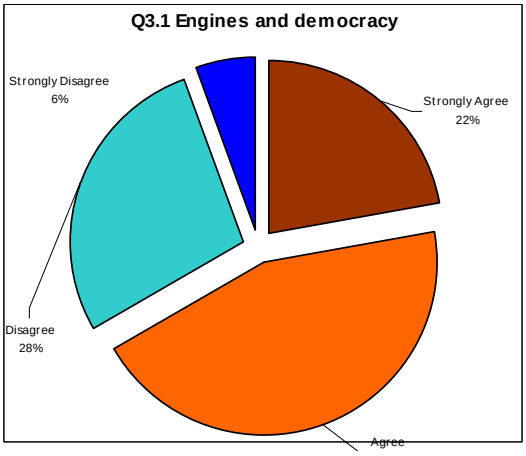


Q2.7 Mobile Search: telecom providers' favoured walled-garden business model will prevail

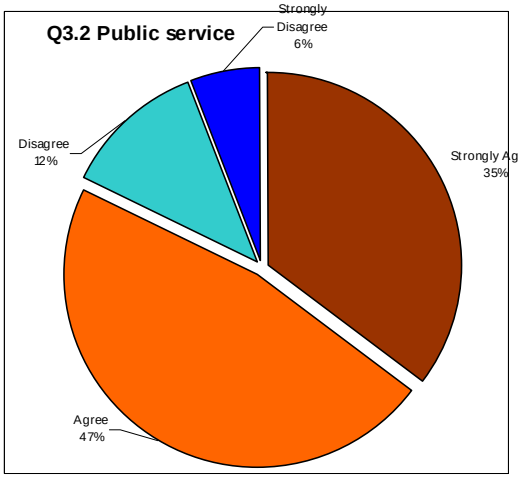


Q2.8 P2P search: there is no viable business model in sight

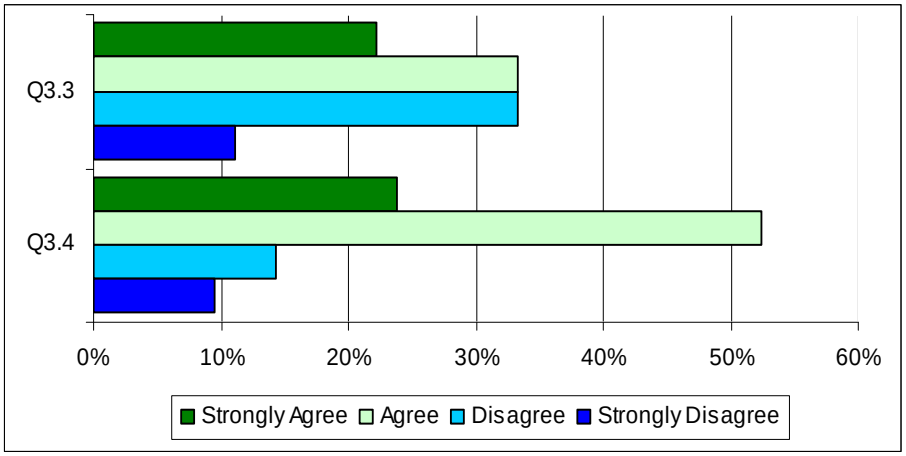
3. Search Engines and Society



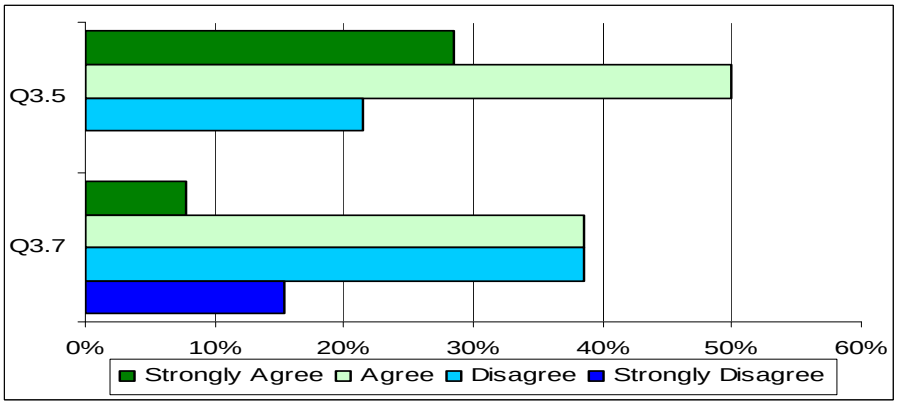
Q3.1 Web search engines are an unstoppable democratic force for civil rights and liberty (e.g. by enabling citizens to find information)



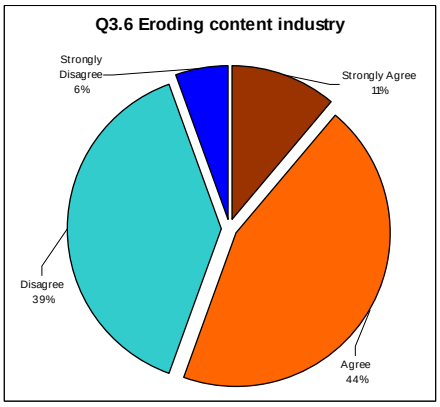
Q3.2 Web search engines are a public service (enabling the information society) and should be treated as such by stakeholders



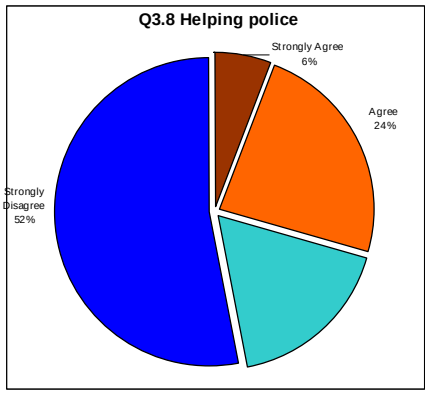
Q3.3 Web search engines manipulate public opinion (e.g. by distorting news trustworthiness through selective news syndication, etc)
Q3.4 Web search engines enhance media pluralism (e.g. by offering local news, by finding thematic news or minority views, etc.)



Q3.5 Web search engines manipulate page rankings
Q3.7 Web search engines compel potential advertisers into subscribing to the search engine advertising programmes (e.g. by 'bullying' techniques, intentional low ranking of non-subscribed potential advertisers, etc.)

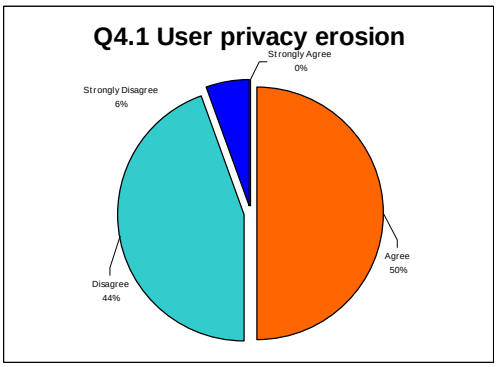


Q3.6 Web search engines erode the professional content industry markets (e.g. providing news syndication, video sharing platforms, etc.)

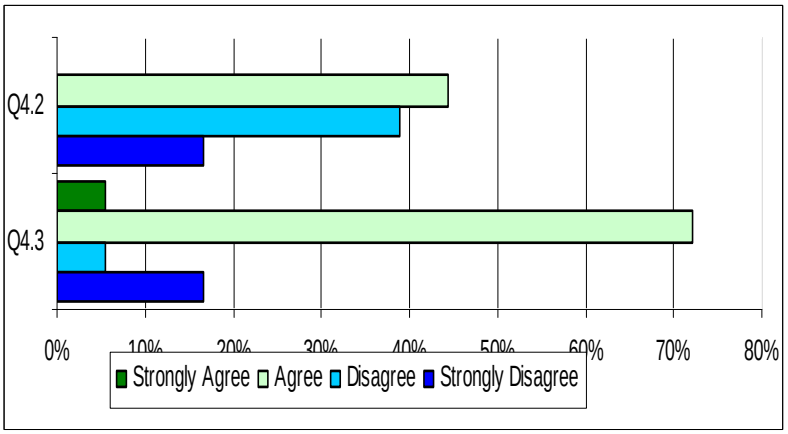


Q3.8 Web search engines should actively help the police in identifying criminals, paedophiles, racists, etc.

4. Search Engines and Privacy

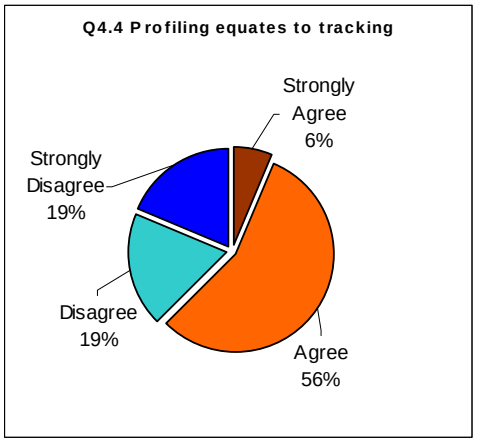


Q4.1 Privacy has already been irrevocably eroded

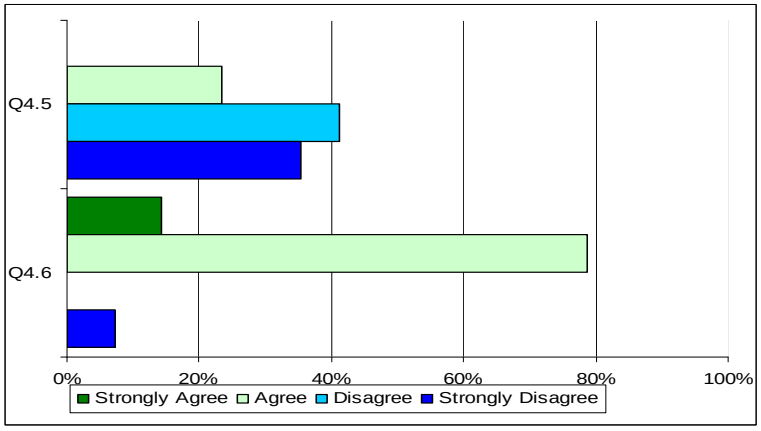


Q4.2 Profiling of small groups (or individuals) is indispensable to generate customized services

Q4.3 Users are willing to exchange personal data for customized services

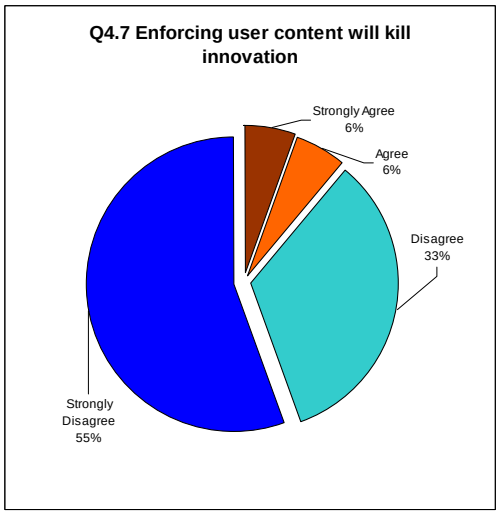


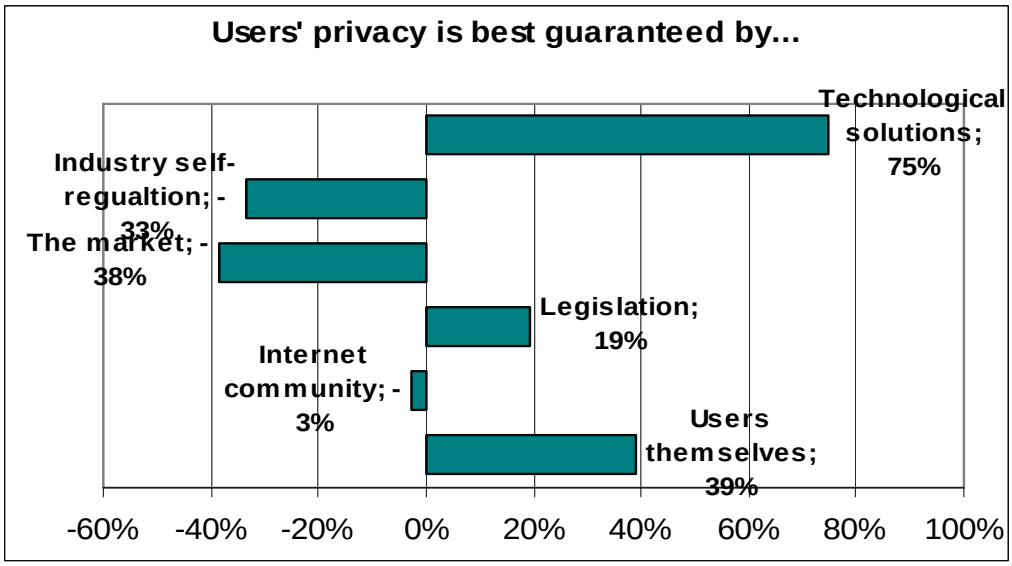
Q4.4 Profiling by search engines (e.g. cookies, log-files, IP-addresses, etc.) is fundamentally not different to tracking via other digital footprints (e.g. credit card records, cell phone calls, ATM machine use, etc)
Q4.5 Privacy by design (e.g. privacy-enhancing, transparency-enhancing technologies) is not viable (e.g. hacking, costs,



etc)
Q4.6 Users should be given the option to 'lease' their privacy rights in exchange for transparent benefit (This should be made in a regulated in a transparent and controllable way, for instance, rights could not be resold)

Q4.7 Approaches requiring user consent for processing their data will kill innovation (e.g. as opt-in options difficult to implement, as use of data for the development of future applications unclear, etc)





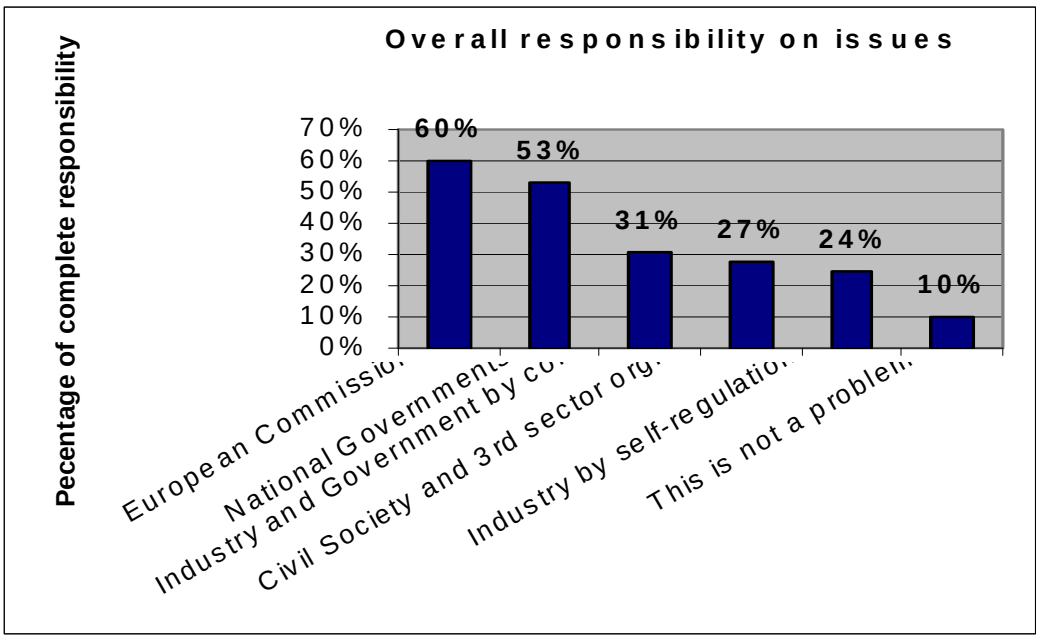
5. Ensuring Privacy

User's privacy will be best guaranteed by...
Q5.1 ... technological solutions (obscuration, anonymisation, encryption techniques, etc.)
Q5.2 … by industry self-regulation (e.g. codes of conduct)
Q5.3 … by the market (e.g. privacy has a fair price; privacy-enabling services will emerge)
Q5.4 … by legislation (e.g. criminal law, data protection, data retention directives, etc.)
Q5.5 … by the internet community (e.g. hacktivism, provision of P2P search, etc.)
Q5.6 … by the users themselves (e.g. controlling own profiles in social networks, etc.)
(up to three responses could be tagged – therefore sum does not to add-up to 100%)

6. Responsibilities [up to three responses can be tagged]

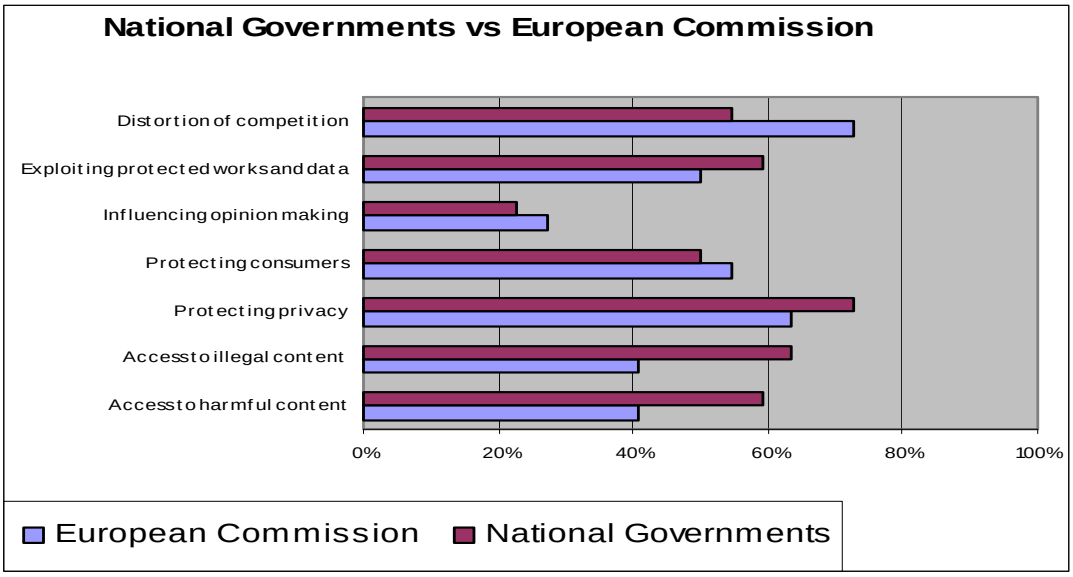
Who is responsible amongst the following (European Commission, National Governments, Industry by self-regulation, 'Industry and Government by co-regulation', 'Civil Society and 3rd sector organizations' or 'this is not a problem') is responsible for
Q6.1 Access to harmful content (e.g. child pornography, racism)
Q6.2 Access to illegal content
Q6.3 Protecting privacy (e.g. data processing for non authorized purposes)
Q6.4 Protecting consumers (e.g. ranking methods of sponsored links)
Q6.5 Influencing opinion making

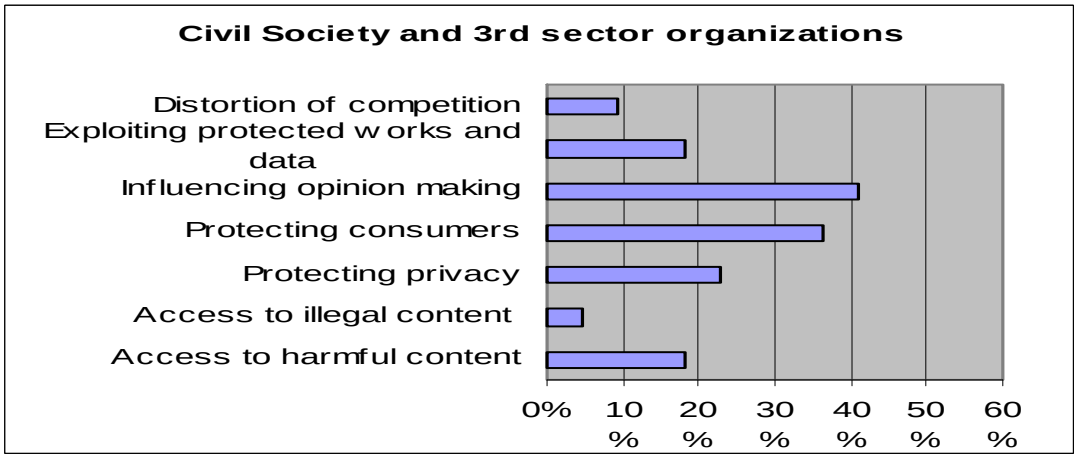
Q6.6 Exploiting protected works and



data

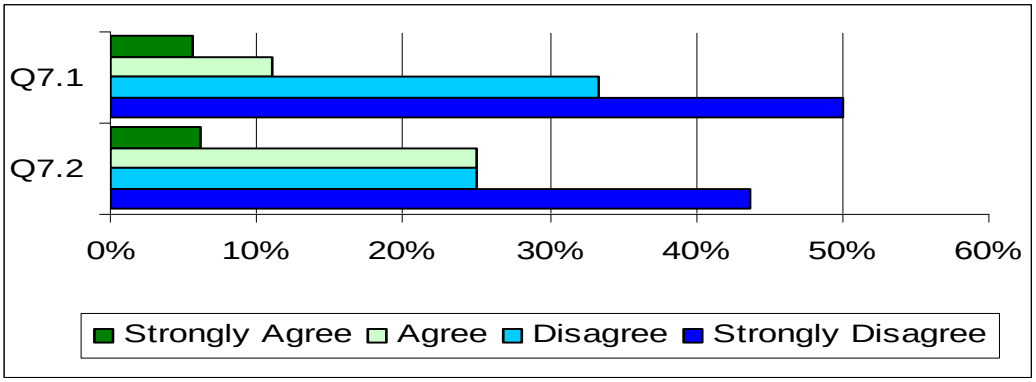
Q6.7 Distortion of competition (e.g. abuse of dominant position)



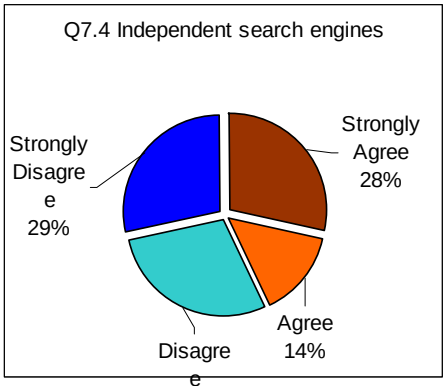


7. Policy Options

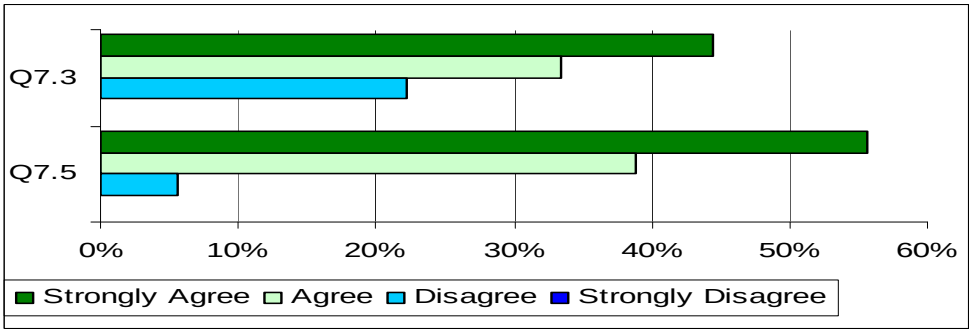
Q7.1 Potential concerns, like consumer protection, privacy protection or unfair competition are already covered under general rules. Search engines do not need any specific regulation (as in the physical world, bookshops, libraries or kiosks have been selecting and facilitating access to content for decades without any special rules).



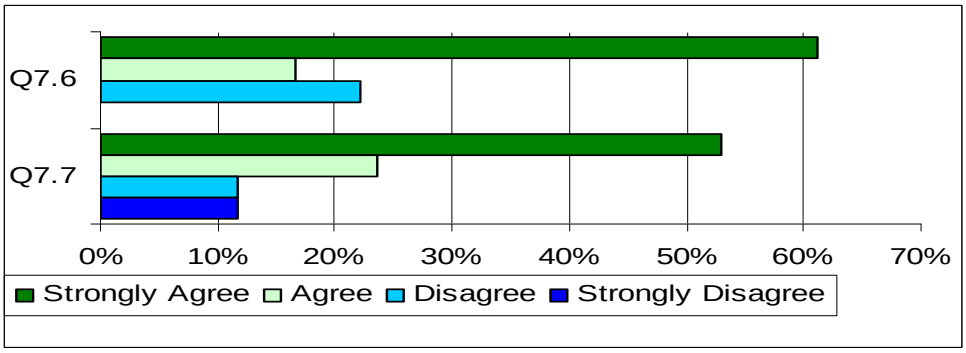
Q7.2 Sector EU directives, like electronic communications directive, audio-visual media services directive, e-commerce directive, do not fit the search engines industry and there is no need to take them into account in future updates.



Q7.4 Governments should fund 'independent' search engines



Q7.3 Governments should oblige search engine providers to disclose when they carry out data correlation (e.g. browsing behaviour from search engines with user profiles from social network sites)
Q7.5 Governments should oblige search engine providers to disclose when they sell marketing data



Q7.6 Search engines should retain personal data no longer than 6 months

Q7.7 Governments should oblige browsers to have privacy-enhancing technologies built in

- End of document -