CHORUS AUDIO-VISUAL SEARCH
IST - 045480

Future of Multimedia Search Engines
Findings by the EU project CHORUS

(July 2009)

Editor: Nozha Boujemaa

Contributors: Nozha Boujemaa, Henri Gouraud, Ramón Compañó, Jussi Karlgren, Pieter van der Linden, Paul King, Nicu Sebe, Joachim Köhler, Alexis Joly, Joost Geurts, Christoph Dosch, Robert Ortgies, Åsa Rudström, Markus Kauber, Jean-Charles Point, Jean-Yves Le Moine

Abstract:

Keyword List: Multimedia search engines, content indexing, metadata, scalability, technology assessment, annotated data collections, search market segments, privacy, user studies, user quality of experience, recommendations,

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
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/IAISD
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.	PHILIP	

Information Society Technologies

Contents

1. INTRODUCTION

CHORUS is an FP6-call6 coordination action on audiovisual search engines. It has set up an information exchange platform for the EU projects, national initiatives and key players in the domain of multimedia search engines (MSE). CHORUS activities aim to bridge the gap between *researchers view* (academia and industry) and the *new services* for end-user (technology consumers: professional and large public content owners) within a market *prospective* in the MSE area. Through its different structures and events: working groups, Think-Tank, "A-V Search" cluster and workshops, CHORUS tackles identifying and deriving *critical issues* through technological aspects together with socio-economic and legal aspects. The CHORUS coordination action spotlights the convergence of a number of broad categories of information access-based activities to spotlight trends and challenges for future development projects.

Search and information access is a base technology for most human intellectual activities. Search technology as it is understood today is a huge and growing business. We have only seen the start of this business area. The general purpose search engine as made popular by the major web service providers is certainly the most spectacular aspect of the technology field. But there are numerous untapped other opportunities in this field, for specialised and general search tools alike. The premise of this vision section is that new types of tools for access to digital information will be necessary and that there is a possibility to influence the direction of development in the near future through informed research efforts directed towards a common goal.

Figure 1- **Chorus meetings**

In the past few years, the CHORUS coordination action has through extensive and continued contact with research projects, both European and national, as well as through conferences and think tanks, formulated a number of tenets and trends which are likely to influence the near future of research and development in the general area of retrieval of image, video, audio and other non-textual objects. Most importantly, development of new technology alone cannot provide the basis for take-up and societal impact of new services. Identifying crucial information needs, not served by technology today, is the natural path to successful introduction of technology. This document intends to spotlight several such crucial convergences of need and potential technology, and to provide directions and to identify the major trends and tendencies which strategic research efforts best are advised to work within. We base our premise on a number of observations.

Firstly, and most obviously, from the advent and increasing rate of arrival of more data through increased network connectivity, lowered publication thresholds and digital production systems, and database uplinks and retrospective digitalization of heritage and previously inaccessible yet valuable material from archives, museums, private collections and corporate data warehouses. The dematerialization of previously physical information services such as music and movie distribution is part of this trend. The tools in place today are vectored towards material currently available in the collections they are deployed with: when material of different editorial levels and different standards are mixed, new requirements are placed on access technology.

Secondly, through the increased pace of social innovation with new tasks, new patterns and situations of usage, and the arrival of new categories of users. User-generated content, reflecting the above-mentioned lowered publication threshold on the one hand and a rapidly changing media landscape on the other, poses new challenges for access, both for the general public, for the editorial efforts to make sense of the new information world, and for the future archivists trying to understand the times we currently live in. This calls for new approaches to interface, index, retrieve, and display multimedia content in ways which encourage and empower user engagement and channel it to uses which will benefit the society of users the most: when a critical mass of users approach information and communication technology with confidence, to take part and enjoy what is on offer, but also, increasingly today, to contribute and share in the common information space this will be for the benefit of all.

Thirdly, the obvious technology challenge of handling the arrival of new content types, and new content sources: the largest growth of data is on audio-visual material, not on text; the largest growth is on non-English material. Text retrieval technology has been afforded by the major service providers and the technology to provide basic retrieval is well-known and available at low or no cost - but extensions to handling new languages and new media are challenging questions to address.

Fourthly, except for the major web search engines, search has not been the primary factor for the emergence of new services or products on the Internet. Examples such as Flickr, YouTube, Facebook, MySpace, LinkedIn and other similar popular services are based on organisation of content, and require satisfactory search components to be useful and gain acceptance; however, effective search mechanisms are not the competitive advantage of the services.

Fifthly, situated and tailored information applications, with the advent of ubiquitous information technology, making use of location and other contextual factors, will enable new services which adapt their interaction model to specific usage rather than the most general. This includes personalisation technologies which based on user behaviour or its relative similarity to behaviour of others, allow systems to recommend choices where none yet have been made by the user and defaults where interaction is limited; position-aware services which tailor their responses to the location of the user; ubiquitous and ambient computing initiatives where search is likely quite often to be embedded in other services and performed by the system after inferring user needs. Users will not necessarily recognise the search as a specific set of actions performed by them.

These trends provide an opportunity for service providers in Europe: the field is open for new entrants to provide services for specific cases, for new media, for new service contexts. Content availability, scalability, and service reliability are currently the most crucial bottlenecks for establishing take-up, rather than technology differentiation. From a European perspective, the industrial base for search engines lacks actors. While European search technology research is of internationally established quality and some corporations deliver services based both on repackaging existing technology and developing in-house technology, the large consumer services are mostly trans-Atlantic. Providing the right starting points for growth in European service providers is one of the challenges identified by CHORUS. The opportunities outlined above give cause for optimism if the technology gaps can be addressed in an effective and goal-directed manner. This document is intended to provide a guide for that purpose.

2. DEFINITION OF MULTIMEDIA SEARCH ENGINE

CHORUS elaborated and agreed to a common **functional breakdown** model of a generic search engine[1] regardless of the application domain or business sector. It presents the benefit of shared projects' description and vocabulary across industry and academia. Search is about *making best use of available meta-data*[2] to provide the user with *meaningful information* in spite of the fact that the user's request is possibly poorly formulated and typically *unanticipated.* Keeping the *"user in the loop"*, maximize its efficiency.

Searching for information within a digital document requires that the information sought be first transformed into a "content signature".

[1] See D2.2 deliverable on "Gap Analysis" for more details

[2] Metadata are defined as being all information besides the raw content that make the content searchable: manual annotations, automatically generated low-level signatures, automatically generated semantic labels, device generated media context …

Doing so for textual documents is easy - the "signature" is simply the ASCII representation of words (although associating words with “meaning” is somewhat harder!!). Performing the same task for images or sounds is much more difficult, and has been an active research topic ever since the emergence of digitized multimedia content: images, videos and sounds.

The information explosion on the Web has increased the complexity of information search in a fundamental novel way: Direct *exploration* of digital document to locate the sought after a "signature" (traditional code) is not possible anymore because of the volumes involved. It is therefore necessary to split the search process in two phases:

- A first off-line phase (batch oriented), during which "content signatures" are collected and organized for optimal access. These signatures now fall under the general definition of metadata (the human and machine generated metadata.).
- A second interactive phase during which a user formulates queries and examines results over the signature matching results performed by the search engine.

These two phases decomposition turns the already complex problem of meta-data extraction into an even harder one: phase one meta-data collection cannot anticipate for all possible future queries!

Recognizing this fundamental issue and translating its consequences into the components of a search engine is the goal of the functional description proposed here. Its goal is to create a shared understanding and vocabulary, to focus efforts where the pain and difficulty resides, and to facilitate synergies and convergence detections.

The diagram below describes a generic search engine through the various "functions" that must be implemented is one form or another. This description is not architecture, and each function must not necessarily be implemented as one module. Moreover, some aspects such as scalability, security and other software engineering issues are not addressed by this functional decomposition while they should be addresses explicitly in any actual architecture design or implementation.

In the proposed functional design, the two phases mentioned above appear to the left for the document oriented phase one, and to the right for the query interaction of phase two. Phase one result in the production of Document meta-data organized into a structured search oriented database, while phase two alternates between Query meta-data preparation and Results presentation. Both phases meet at the central Matching function whose key and complex role ranges from simple lookup of words into an inverted index to fuzzy matching of large signatures.

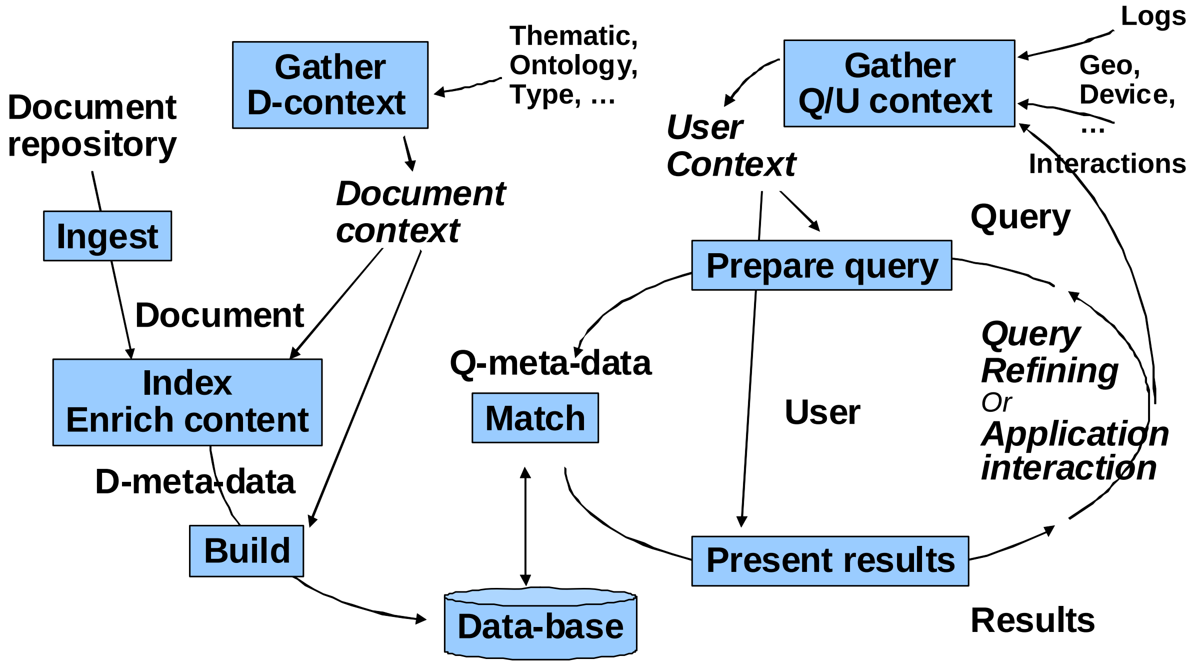


Figure 2 **Functional breakdown of a search engine**

Phase one can be decomposed into several functions. Although a single box view of the document processing activity might seem sufficient, it is necessary and productive to decompose into several functions, each driven by one of the major characteristics of the problem space.
- Ingest deals with the discovery of documents to be included into the search space. The characteristics of this function are driven by the volume of documents, the rate of evolution of this volume, and by the expected "freshness" of the resulting service. This is where "real-time" search may happen if it is an objective, which may result in a push ingest model where the application world pushes new information towards the search service. Conflicting constraints between freshness, shared interfaces and computing power will result into varied solutions adapted to various markets needs.

- Index deals with extracting meta-data form one document. The characteristics of this function is driven by the nature of the processed document (text, image, sound, video, 3D, other, ...) and by the rate at which new documents appear. The ultimate goal is to extract as much information as possible from the document, exploiting any contextual information available. Multi pass indexing is indeed possible, existing meta-data being an integral part of the document context.
- Build deals with assembling the meta-data produced by the previous function into a usable database. This function deals with the processing which requires a global view of the system. For instance, ranking a new document relative to others, or clustering documents together cannot be performed without such a global view. The characteristics of this function are driven by the nature of the computation performed and its possible centralized nature,
- Document context deals with any in formation shared between the documents of the searchable corpus (or subsets of). Notions such as language, origin, collections, are typically needed by the Index function. Automatic detection of the context is an essential aspect of this function, and when not possible, may result into manually applied restrictions to the ingest function which may in turn have an impact on the which use-case may or may not be implementable.

- Phase two is decomposed into several functions combined into a single interaction loop:
- **Query preparation** deals with the difficult task of capturing the user intent, and transforming it into processable query meta-data. This preparation may take as input explicit query terms, parameters supplied by the user, contextual information extracted from the device (GPS location), the user profile and query history, or collective information extracted from logs. In this task, it is important that the system reveal to the user as much as possible of its capabilities, so that she can adjust her search task to maximize the effectiveness of available tools rather than insist on using a non existing one!
- **Results presentation** deals with the task of returning to the user the potentially large amount of information matching his query. The goal of this function is to maximize the effectiveness of the system for its user. This is achieved by trying first to present to the user results which are most likely to fit his intent, but also by revealing to the user as much as possible of the structure and general characteristics of the result set. As it is very likely that the first set of results may not satisfy the user intent, it is important again for the system to reveal as much as possible about its capabilities, this time in the context of a first query. The combined effectiveness of the two steps of query preparation and results presentation should maximize the user effectiveness in a positive feedback interaction loop.
- **Query and User context** deals with extracting and collecting any contextual information that will contribute to the query preparation function. In such, it is not directly part of the main interaction loop, but its impact is potentially large, in particular in the case of implicit querying where the query results almost entirely from context data.
- The balance between these three functions is greatly impacted by the explicit or implicit nature of the search task.

Given the interactive nature of phase two, its User Interface aspect is indeed essential in a fashion that goes well beyond the purely graphical aspects of such an interface. Keeping the user in the loop, and allowing him to build a mental and predictable image of what the search system can do best is a critical aspect of search systems research and design, and is one of the strong recommendation of Chorus resulting from the functional analysis task.

The functional analysis reveals also the key role played by meta-data in addressing the search issue. The creation of document meta-data is serving multiple purposes, and is the core technical challenge of search engines technology:

- Document meta-data is the only mechanism through which content becomes accessible. It must therefore anticipate as much as possible the future user requests while acknowledging the illusory aspect of this goal.
- Similarly, translating the user intent into query meta-data is of a comparable illusory nature.
- In this fundamentally difficult situation, one of the key roles of meta-data, especially in the explicit search context, is to reveal to the user the effective capabilities of the search system and to maximise his/her ability to use them towards achieving his/her goal.
- To that extent, it is fair to say that meta-data represents basic or intermediate level information that contributes to reducing the width of the "semantic gap" without ever totally closing or bridging it.

The functional analysis above is a potentially powerful tool for analyzing the feasibility and business applicability of specific use cases, and to contribute to the technical Gap Analysis. For a given use case to be implementable, all the functions of the proposed diagram must have an implementation compatible with the overall scope of the use-case. Computing the expected meta-data must be first doable and second in a span of time compatible with volume and change rate of documents on one side, and with the interactive nature of queries on the other side. Contradictory aspects of the system, such as freshness (real time) and completeness (exhaustive coverage), must also be balanced in accordance with the use-case requirements.

The functional description above does not address several aspects of search engines which are transversal to the system must be covered at implementation time such as scalability and security. One particular aspect is the possible distributed nature of proposed architectures often envisioned to resolve the scalability issue. Real distribution3 implies network latencies and possible failures, which may hinder fundamentally some functions which require global knowledge such as the build function (ranking for instance).

3. MARKET SEGMENTS AND VISION:

3.1 INTRODUCTION

Search, as given in the previous section, is a base technology for very various services and activities spanning over most human intellectual activities. In this report, we have chosen to distinguish six broad categories of information access-based activities to spotlight trends and challenges for future development projects to address. The categories we have chosen are neither meant to be exhaustive nor entirely distinct from each other – the salient characteristics we have chosen to model are one conceivable analysis – based on our perspective, depending on how content and repositories are managed, a new service or application can be understood by the following fields of activity:

- Web Search
- Personalized TV
- Enterprise Search
- Public Archives and Digital Asset Management
- Personal Archive Search
- Monitoring, Detection & Alert

The attributes that define the market categories are enumerated and defined in the following table.

Attributes	Definition	Values
Content Management	*The level of organization of repository content.*	Unorganized, Semi-Organized, Organized
Content Ownership	*The applicable licensing model for repository content.*	Public, Private
Repository Access Rights	*The availability of repository content to*	Unrestricted, Restricted

[3] Note: as opposed to distribution or replication of a task across multiple processes residing in the same machine, or on the same high performance LAN

	users of the search utility.	
Revenue Model	*The primary type of income stream for achieving sustainable, long-term profitability of the search service.*	Direct Revenue, Subsidized Revenue, Content Licensing

Content Management: Organized content refers to well-structured content that is managed professionally (i.e., by a librarian). Unorganized content refers to unstructured content that requires additional processing and is not under the purview of professional management. Semi-organized content refers to documents that have some structure, but where the structure requires interpretation and normalization or it may refer to a collection of decentralized repositories that may or may not be professionally managed.

Content Ownership: Public content is generally not restricted by licensing terms whereas private content is.

Repository Access Rights: Unrestricted access refers to repositories that can be accessed by anyone. Restricted access refers to repositories that require authentication for viewing content or it's metadata.

Revenue Model: Direct revenue refers to fees or subscriptions generated from the purchase or leasing of the search engine. This is the simplest business model. Subsidized revenue refers to income generated from secondary commercial relationships involving advertising, cross-selling or sponsorship. Advertising is revenue generated by a paid announcement or product promotion appearing in the search service interface. Cross-selling is 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.) Sponsorship is revenue generated from fees paid for granting the right to associate another organization's name, products or services with the search service or company. Content licensing refers to revenue that is generated from the licensing or selling of content within the repository.

The table below gives an overview of all the above market segments together with their attributes defined. Each market is characterized by how its four major attributes are expressed together. Each market is described in more detail in the sections that follow.

MARKETS	*Content Management*	*Repository Ownership*	*Repository Access*	*Revenue Model*
Web Search	Unorganized	Public	Unrestricted	Subsidized Revenue
Personalized TV	Semi-organized	Private	Unrestricted, Restricted	Subsidized Revenue, Content Licensing
Enterprise Search	Semi-Organized	Private	Restricted	Direct Revenue
Public Archives and Digital Asset Management	Organized	Public	Unrestricted	Direct Revenue
Personal Archive Search	Semi-organized	Private	Unrestricted	Direct Revenue, Content Licensing
Monitoring, Detection & Alert	Organized	Private	Restricted	Direct Revenue

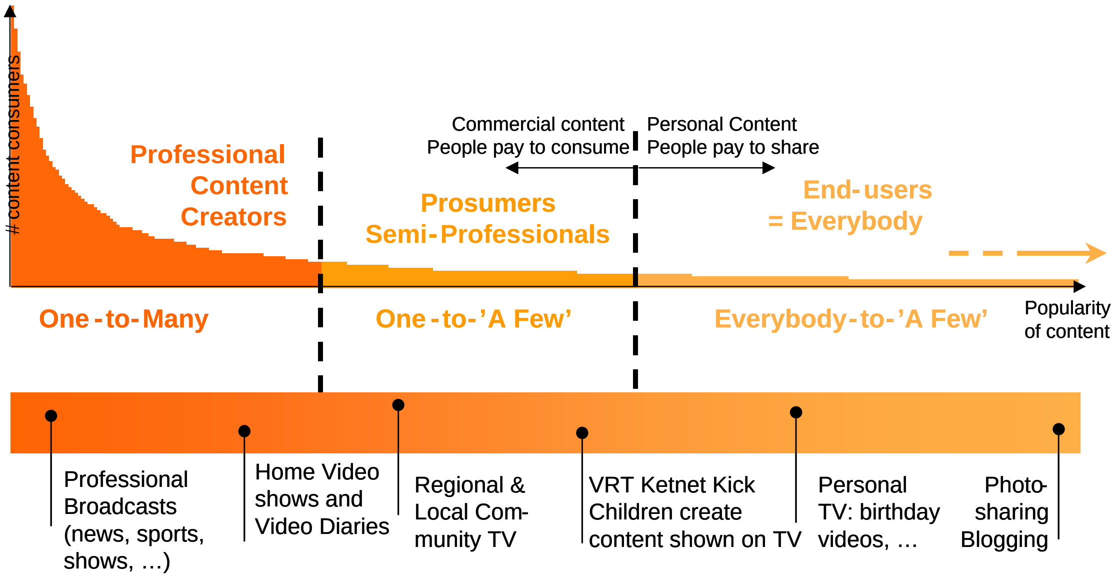

# content consumers
Professional Content Creators
Commercial content People pay to consume
Personal Content People pay to share
Prosumers Semi-Professionals
End-users = Everybody
One-to-Many
One-to-'A Few'
Everybody-to-'A Few'
Popularity of content
Professional Broadcasts (news, sports, shows, ...)
Home Video shows and Video Diaries
Regional & Local Community TV
VRT Ketnet Kick Children create content shown on TV
Personal TV: birthday videos, ...
Photo-sharing Blogging

Figure 3 Market is shifting from mainstream business to individual and fragmented tastes where citizens evolve from a passive media consumer of mainstream content towards an active role in the media chain

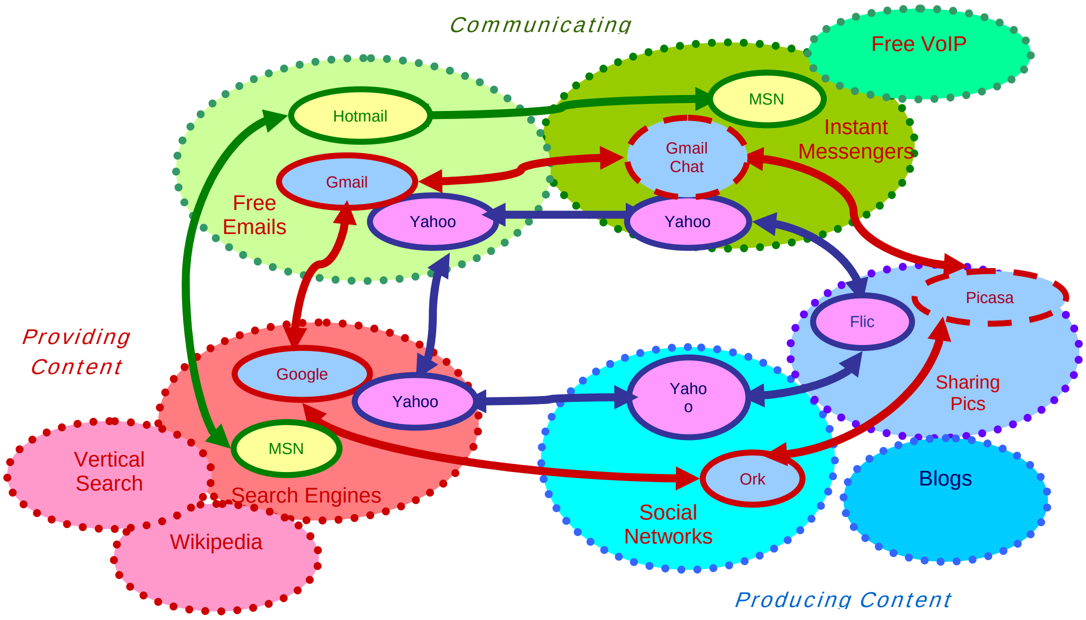
3.2 Web Search

The web search market involves the identification and indexing of large scale content across numerous information sources on the Internet or other publicly available network, and the subsequent provision of public access tools to the content in question. The Web Search paradigm, familiar to any Internet user, is dominated by a small number of major players, currently lead in the global arena by Google, Yahoo! and Live Search (Microsoft) which all use an advertising based business model to monetize search, supplemented by licensing agreements. There is a wide-spread concern that the dominance of these global actors will irreversibly make it impossible for smaller players to enter the search market. However, the history of web search services is short in industrial terms, and there are many competitors to the major services, such as Exalead, Ask, AllTheWeb, Clusty, and Lycos. These services provide broad and generic web search for general purpose search.

On the other hand a search service can target some specific set of resources, selected by type of resource, by source, or by topical area to yield a search service for specific needs, e.g. for professionals in some business area or enthusiasts with some specific interest to provide a vertical search service or cater to the long tail. Examples of companies in this field include Business.com for business owners and entrepreneurs, GlobalSpec.com for engineers, and SearchMedica for healthcare professionals. It is quite likely that this sort of specialised service will become more prevalent, as business models to target special interests grow in acuity: the quality a specialised search service can provide to professional societies and interest groups can be heightened through tuned algorithms as well as through editorial contributions.

While the technology for building a search engine is well-established and several industrial-strength state-of-the-art search engines are available for free download under open source licensing schemes this is not enough to ensure more competition in the market. Coverage, response speed, reliability, scalability, and consistency are the most important competitive factors to gain market share, which in turn determines advertising revenue. These competitive factors are all determined by the availability of large and efficient computing resources: servers, local architectures that allow robust and scalable handling of large numbers of transactions. Obtaining and maintaining such infrastructure is demanding in terms of investment. Specialised search services can elect to build their own indexing and search technology - which will be necessary in some cases, to cater for specific needs of the target group - or to build on existing commodised technologies.

To achieve the open and creative environment where new business ideas can work to integrate new forms of content with new forms of usage the multimedia field needs new business models which cross existing commercial boundaries. Content developers and owners, network operators and access providers, and device manufacturers do not today have common business goals.



.

Figure 4: **Google, Yahoo and Microsoft operate a number of services that render them key players as enables of content producers, as information providers and as communication facilitators**

3.2.1 Vision

The vision for future web search services is first and foremost the obvious extension from text to multimedia. This process is already underway – and involves attending to representation and identification of content to go from annotation-based search to content-based search and the potential for establishing transparent and interoperable mid-level descriptors for content for ad hoc search together with fine-tuned low-level descriptor for specific tasks and high-level descriptors for conceptual access services. Specialised services are already appearing at an increasing rate, put together from freely or simply available existing components and services. These services will not only be self-contained but are likely to leverage information from existing knowledge sources (cf. Section on Public Archives and Digital Asset Management below).

In a somewhat further perspective, the vision for web search involves technology for Disappearing search. Search is already present in many applications and web services, but often invisible to the user – information needs are inferred from concepts, actions, and events in interaction, and learnt from observation of both collective and personal behaviour, obviating need for specification of information need in situations where complex data entry would be cumbersome, leaving users to tasks such as verifying system assumptions or inferences. This trend will continue, as users will learn to expect that every service will provide relevant materials to complement the actions they are engaged in.

The vision also includes heightened awareness among users of the value of their interaction with services. Service providers provide free access to services through advertising are eager to collect personal interaction data as well as content and character of annotations made by users to provide more targeted recipient groups for their advertising customers. These data constitute a resource which holds economic value for the service providers: today, few users are aware of this, even when they provide valuable refinements to data collections they access by annotating and labelling information items. Providing for this vision involves creating the basis for business models which enable users to be more aware of the commercial value of data based on mining and extracting information about individual usage patterns and which allows the individual user as well as collectives of users to monetise information about them. This will indirectly address and allow questions about personal integrity to become more salient for the individual user.

3.3 PERSONALISED TV

The volume and diversity of the offering of audio-visual broadcast material for consumers has increased steadily from a small number of radio stations and TV channels to hundreds. For years the broadcasting[4] model has been

[4] Broadcasting: Initially analogous dissemination of radio signals along a one to many model..

the single available means for bringing the information to the consumers. In this model the content distributors compose a linear TV or radio broadcast designed to please a maximally large consumer base, funded by public funds, by sponsorship, or by advertisement segments interleaved with the broadcast program.

A number of telecommunication operators in the world have started deploying IPTV5 infrastructures during the past several years. Despite these efforts, in most countries, the market take-up6 remains rather low. There are several reasons for this resistance to adopt new technology and new distribution mechanisms:

- Compared to broadcast TV the current IPTV offering does not bring much new content to the consumer. With the exception of the distribution of a subset of mainly blockbusters via Video-On-Demand services (VOD), current IPTV offering is more or less the same as broadcast TV except that it uses an ADSL link rather then analogue or digital broadcasting signals.
- The new distribution mechanisms do not bring new functionality to the user. Integration with other digital services is non-existent: there is no possibility to share, annotate retain or communicate with other viewers.
- The business models for the new distribution channels are typically based on subscription, which in itself creates an adoption threshold. The expectation of users, based on their experiences from other internet services is that content can be had at no marginal cost and little or no marginal effort.

Meanwhile, the use of video in web services has dramatically increased. A number of new players have emerged, and have taken reached very significant user communities. As an example, the video sharing site YouTube ranks at second place for the number of search queries on the Web7. Traditional media companies try to keep momentum and audience by proposing Web oriented information services such as Catch-Up TV. The success factor in this case is largely dependent on the ease with which users can test and try out a new service without committing to it. Likely features consumers will expect is to be able to control their schedule e.g. through time-shift technologies, to be able to request entertainment and information in settings where they do not wish to take the initiative, and immediate and transparent connection to communication services which enable users to situate their usage in a social context.

3.3.1 Vision

Media consumers will have a wide offering of entertainment and information content, easily accessible, with little effort, low marginal cost, and with easy transitions from provider to provider. The attraction of broadcast material has been threefold:

- ease of use – with prototypical access queries of “What's on?” and “Entertain me!”;
- a simple business model from the consumer point of view; and
- a shared social context – incidental to the limited offerings of previous generations of broadcast TV services.

New services will have to address those advantages.

Professional and editorially produced media will retain their primacy as authoritative channels of entertainment and information in home settings, but the business models will be based less on blanket advertising than on directed information targeting identifiable groups of consumers, measured by household or consumer centric coverage rather than crude statistics over the entire population.

Media consumers will be able to share their viewing habits with peers, both friends, unknown but trusted parties, and the provider of the service they obtain their media from. They will also be able to stop sharing. When they elect to publish their viewing habits they will know when they are doing so, and they will be aware of the fact that they provide information of potential commercial value. Media providers will be able to use viewing habits to improve the quality of service to the degree that consumers will willingly share information about themselves and their viewing habits to obtain the improvements.

[5] IPTV: Offering of TV over managed networks using IP protocols.
[6] Around 9 millions users end 2007 according to a study by MRG.
[7] Comscore August 2008.

The entertainment component of the home will be integrated with communication and information technology; consumers will be able to obtain information about the offerings, and relevant side information with little extra effort. The media viewed on personal TV will be delinearized and annotated with relevant time-coded tags in order to allow users direct access to the content using metadata analyses available from the content provider, from annotations made by users themselves, as well as from third parties.

Interaction technology will provide effortless and brief interaction with the system, obviating complex specification of information need, preserving the sense of entertainment and information being on offer rather than being on request, allowing users lean-back interaction rather than proactive search. This will to a large extent be based on information generalised from personal viewing habits and the habits of peer groups, using social filtering and recommendation mechanisms.

Convergence between content producers, content owners, media distributors, broadcasters, network owners, internet and telecommunication operators, and device manufacturers will create new business models and new business opportunities. Existing monolithic business interests may lose relevance to consumers.

Between the broadcasting business model based on mass advertising and the telecommunication operator business model based on individual subscribers, new business models will appear. Personalized TV operators will aim at increasing advertisement revenue. Knowledge of the users and households will allow enhancing the relevance of the advertisements to their audience, moving from a more qualitative evaluation of advertisement effect.

3.4 Enterprise Search

Enterprise search – search in information sources within some organisation for the purposes of that organisation itself – differs from other search market segments in that the size and editorial qualities of the content is monitored by organisational policies, the number of users is limited compared to public search services, the information sources can be fragmented to several types of repository within the organisation, some of which may be crucial to operations, yet technically behind any development curve. Most importantly, however, the quality requirements of enterprise search are much higher and more precise than those of public search services, and there may be external regulations, cameral and legal, which bind the information systems of the organisation in question and may have effects on the search service.

European industry is strong in locally targeted enterprise search solutions. Enterprise search needs to be sensitive to local requirements as regards legislation and regulation of various kinds, and to requirements with respect to local languages. This is a natural opening for locally based solutions, but provides no natural basis for growth beyond the market areas where local expertise is the competitive edge.

According to Gartner, revenue generated from enterprise search software increased by 15% between 2007 and 2008. Search has been growing rapidly since 2004, but is predicted to slow in coming years due to licensing and market consolidation. Nevertheless, growth is expected to remain healthy as organizations seek cost savings associated with improvements in business processes. (Gartner Press Release: February 2008)

Year	Millions of dollars	Percent increase over previous year
2006	717.2	
2007	860.6	19%
2008	989.7	15%
2009	1,108.5	12%
2010	1219.3	10%

The overall size of this Enterprise Search market is evaluated at 2 bn$ in 2009 by IDC, with a 20% to 25% growth rate in the last quarter of 2008. Enterprise search is growing in strategic importance this year due to the global recession and continuing trends towards cloud computing. It is expected that these two factors will drive further consolidation among vendors in 2009, giving larger players, such as Google and Yahoo, an opportunity to enter a market which has, until now, been dominated by smaller players.[8] Examples of enterprise search companies include Autonomy, Exalead, Endeca, Dieselpoint, and Teragram. It is worth noting that in its often quoted "Magic Quadrant", the Gartner Group lists, in addition to Autonomy (UK – leader) and FAST (No - leader), two other European companies of smaller size: Exalead (Fr - visionary) and Expert System (It – niche player).

Enterprise Search has not followed the same growth trend as has Internet search except in certain sectors. The growth of materials within organisations has not experienced the same explosive growth as the private usage and media markets have: the operation processes of information intensive organisations are already in place and have been influenced less by the information explosion; the organisations that are currently entering digital operations are able to utilise solutions already tested for fore-runner organisations. Multimedia documents do not yet play a significant role in the day-to-day operations of a business or public organisation.

Business solutions for the two technically similar markets of internet search and enterprise search are drastically different. While internet search is typically an advertisement based service, enterprise search is practically always based on licensing revenue.

3.4.1 Trends and vision

The main driver for Enterprise search is the more general trend of applying consumer oriented, Internet based technologies within the enterprise. This general trend has already been observed in many circumstances such as IP networking (early corporate networks were X25 – moved to IP to benefit price reductions offered by massive Internet growth); in GSM telephony (GSM phones developed first within consumer market. Professional users brought their personal phones to work, then requested corporate support for professional use); and in Web services (the whole intranet sector grew as an internal replication of services and tools found on the Internet).

Following this trend, professional users expect to find the same tools on their enterprise network that they are used to accessing on the Internet. This trend is not limited to search and can be observed for software development (open source), social networking, collective information creation and sharing, and usability of document processing applications.

Smaller businesses are also moving their information operations from single-user systems or even paper-based systems to networked multi-user systems. They are increasingly likely to turn to network based outsourcing services rather than installing large enterprise systems in-house.

3.4.2 Challenges

Within enterprises, professional users of information access tools expect the facility of installation and of usage that they observe on the Internet, but at the same time, demand from these tools accuracy and efficiency beyond what is often considered sufficient on for the general public on the Internet. Databases afford advanced professional users precise and reliable access to organisational information, and provide a high level of transactional security and editorial structure and oversight but do not offer the flexible and dynamic access to data, nor the scalability and structural independence that search engines offer. Transcending this gap between everyday and professional usage is a challenge for enterprise solutions.[9]

A second technical challenge encountered by Enterprise Search solution providers is related to the overall architecture and modularity of the solution. As search will more and more need to be closely integrated into other enterprise solutions, its "software engineering" qualities will become essential. Today, the lack of agreed upon standards and API for search components is one of the issues that needs to be addressed.

A third challenge has to do with customer protection. As less technology-savvy corporations outsource their information to external partners - how can they invest the appropriate effort to ensure their data to be protected from interference or leakage? How can they ensure that the counterpart has any permanence in the marketplace? What rules bind a purveyor of information services?

[8] IDC Press Release: 12 December 2008

[9] Cf. the workshop on "Using Search Engine Technology for Information Management" on August 24, 2009, Lyon, France

3.5 Public Archive And Digital Asset Management

Public collections of information have specific needs and provide specific services. The characteristics as compared to other types of information services are that the service is funded by the public, involves low or trivial cost for the user, and that the collection of information resources and cultural artefacts is organized and maintained by information specialists according to some established protocol. Examples are libraries, archives, and museums; whether the collection is owned by some private or public entity is less crucial than the perception of authority and permanence it engenders in its users; a special case is managing archives where the items – assets - themselves carry specific and explicit legal and commercial conditions for access.

Public collections are frequently characterised by the metadata associated with the information items, meaning data about the information items themselves, such as information about source, production factors, history and usage, and often includes annotations referring to the content of the item such as topics covered, entities mentioned, summaries, or suitability for some purpose. Defining the metadata protocol and annotating items in the collection with it is a demanding and intellectually non-trivial editorial task which requires manual intervention by skilled information professionals; metadata schemes are costly to maintain, extend, and transform. Metadata are used for organisation of a collection and are typically useful for search and retrieval of items from it. Metadata annotations can be the difference between an useful and useless collection of information items and are often valuable resources in their own right, apart from the value of the information items they describe.

Computer aided tools for media archive and asset management are readily available today, but the amount of aid provided by tools are not sufficient for fully automated asset management. Technology for automatically extracting content descriptions from audiovisual data is not yet reliably in place for large scale deployment – archive maintenance must rely on human annotation to a large extent. The resulting data structures are of high value and reliability.

3.5.1 Trends and vision

Public archives are moving towards a future process where manually created annotations are extended automatically and semi-automatically to new data, and where annotation schemes can be edited, managed, and enhanced using automatic tools. This will enable a collection to leverage existing schemes to rapidly extend coverage of a collection to new data and to incrementally improve the quality of the annotation scheme in face of a changing collection or evolving usage requirements.

Annotation schemes vary from one metadata setup to another. A serious effort is being put into achieving interoperability between annotation schemes to allow access between collections organised in different but compatible schemes or between different versions of the same annotation schemes.

Archives and authoritative collections will in the near future be available on line for public perusal, without adminstrative thresholds such as accounts or membership registration. This will afford the community of users a larger trust in publicly available information, and boost efforts such as Wikipedia and other user-generated information sources to base their claims on a legacy and heritage of previously accepted information sources, bridging the current divide between "new" information which is timely and "classic" which is well-established.

Similar accessibility will be true for corporate archives and privately owned collections, through publicly available publishing and collection management systems, which allow material to be published under controlled conditions, with public information clearly separable from proprietary.

Users will be able to use this information to e.g. curate their own exhibitions of museal and archival material, create their encyclopaedic information sources for other users and cater for special interests not foreseeable by the original creators and maintainers of the collection.

3.5.2 Challenges

Metadata is automatically generated by many recording devices today – location, data, orientation, technical data. With the advent of more competent recording devices, the material will contain ever more sophisticated analyses of the original data: face detection and recognition, text recognition are examples in commercially available devices today. These data are are most often stripped out and discarded in subsequent processing steps e.g. in postproduction of broadcast material or in transport and conversion of data from one system to another. This information is lost for archive management purposes.

Publicly available collections in archives, libraries, museums and similar institutions, whether private or public, struggle to find a place in the internet information landscape. While the charter of the organisations in question is to make their collections available to the public, subject to constraints motivated e.g. by preservation concerns, the open access policies inherent in the future internet risks lowering the visibility of the collection, the recognition of the editorial effort made by the specialists in creating the metadata, and the appreciation of the curation effort put into the collection. The brand strength of memory institutions such as archives, libraries, and museums must be leveraged to preserve their status even in an open access information environment.

3.6 Personal Archive Search

Personal information is being created at a rapid and growing pace through the widespread availability of competent near-professional quality recording devices, cameras, and storage capacity on personal information systems. The rate of creation is large enough at present for storage space to be envisioned as a coming bottleneck for availability of private and personal information. Personal information management solutions and personal archive search systems allow content to be indexed, searched, and displayed within a small collection of information resources and cultural artefacts organized and maintained by a non-professional primarily for personal purposes; content is normally restricted to a personal computer or network or a private account on the web.

Personal archives have historically been collected in unorganised shoeboxes, and only lately been moved to digital media on the personal computer. They are also becoming increasingly likely to be found on public networks such as the web. Examples of content within this search market include email, blogs, and photos. These pieces of personal information increasingly require search tools for better personal data management and has the potential to become an important commercial driver. Search utilities are usually tightly integrated with the service or product used for managing the personal information. Examples of organizations developing search in this market include all makers of operating systems – Apple, Linux, and Microsoft, makers of blog software such as Moveable Type, WordPress, and Textpattern, providers of archiving and sharing services such as Flickr, Picasa, and Photobucket, as well as third party providers such as PolarRose.

3.6.1 Trends and Vision

The vision of low-effort and reliable personal information access, management and search solutions is related to the previously related visions. The business model for personal archive search is today mostly based on free software, often bundled together in a purchase of some hardware device, and is likely to remain so at the entry level, as a lead-in for more capable professional editions. It is likely that network-based services will be available both via advertising based business models as well as for subscribers.

Some of the quantitative aspects of personal archive search (manageable number of documents, limited number of users and of “person of interest”) is likely to allow for the appearance of some multimedia indexing and search techniques in this market faster than on the Internet or for Enterprise search. One example of such a move is the recent offering of Picasa, not only to detect faces in photographs 'which is available on the Internet), but also to “recognize” the same individual across a collection of photographs (provided the user has identifier him of her in a first few photographs). This kind of situation creates an opportunity for developers of focused multimedia search technology to showcase their capability, either directly of through association with larger service providers.

3.6.2 Challenges

Providing a solution that merges gracefully into Enterprise Search and Internet Search while preserving the private nature of the data held on a personal desktop (including in the enterprise context) is a challenge identified for PA search.

Search into family photo collections is a definite challenge, facilitated by the relative stability of the family context (small number of persons) and the potentially significant amount of tine available for indexing (1 minute per photo would take 2 or 3 hours to index a new batch of week-end photos!). The domain of “family photographs” is already well developed for its “storage and management aspects. It is therefore clear that search needs to be integrated into those existing tools, with a standards and API issue similar to the one discussed in the context of Enterprise Search.

Audio collection management could be considered as another potential domain for search from a personal archive point of view. The parallel with images is strong and the same arguments apply. This particular segment may nonetheless evolve in a very different fashion from photo storage and management as the consumption of audio content evolves over time. Instant availability of streamed audio content, new pricing models (and possibly the pressure against illegal download) may result in a significant reduction of personal audio archives in favour of on-line streaming services. Search into such services should not be compared with Personal Archives search, but rather as one application of search as an OEM component into a web based service.

Trust issues - similar to the Enterprise Search issues – related to outsourcing of personal data and annotations to them. If private or family information is stored in the cloud - who maintains it and what permanence has that entity, do rules binding it carry over to the next entity in that role?

3.7 Monitoring, Detection & Alert

This market area refers to the task of establishing divergence from an expected normal state. This has obvious application to security and surveillance tasks such as intrusion detection, control and command systems, public space monitoring, tracking potential security threats; to process monitoring such as maintenance scheduling, fault detection, and general process supervision; recent applications include epidemiological outbreak detection and monitoring natural processes such as flooding, volcanic and seismic activity, and wildlife tracking. However, monitoring can be equally applied to personal information management tasks, tracking the changes in a social network, the activities of family members, monitoring information sources of personal interest. Besides, content-based Image and video copy detection is a very promising technology helping for DRM (Digital rights management) Systems to arise with significant efficiency and opening a new niche market for indexing and search technologies.

These tasks involve the analysis of real-time signals from a multitude of different types of knowledge sources: multimedia information streams, physical sensors and signals. Some involve information on a high level of abstraction requiring analysis for task purposes; others are low-level signals which require aggregation. Anomaly detection is the primary technology for monitoring information streams for this type of applications. It is done typically done through establishing a model of normality with respect to a number of features of interest, detecting the anomalous situation with comparison to divergence from the normal state. Anomaly detection thus is the technology to separate a heterogeneous and vaguely delimited minority of observational data from a somewhat more regular majority of data. An alternative strategy for monitoring an information stream is not to base the analysis on normality, from which the system detects anomalies, but, if there is an expected set of situations of interest, the monitoring can proceed through targeted models of the expected situations, which then are detected when they occur.

These models can be statistical, giving a probabilistic interpretation of the situation at hand: if an observation has low probability given the model of normality it is tagged as an anomaly, or belonging to whatever category under consideration is most likely to have occasioned it. The models can also be knowledge-based (often termed "model based"), where previous observations have been clustered to form an explicit formulation of normality and anomaly respectively, with a decision procedure based on a similarity measure between the observation at hand and previously established model clusters. The distinction between a statistical and knowledge-based models is largely determined on whether the anomaly models can be assumed to conform to some previously understood statistical distribution or not.

Many monitoring application systems rely in practice heavily on human perception. The human visual competence is very high, and clustering visual data on a surface is a task human operators excel in. The system needs to provide the right features for the visualisation scheme.

3.7.1 Challenges

Defining standards for how such models can be established is a current research issue, especially with respect to situation awareness tasks, in e.g. command and control applications. This is a daunting task: the modelling of situational factors and features used to build the models is closely tailored to the task at hand and requires new levels of abstraction and generalisation to be useful. Current modelling languages do not have the generality required.

In all cases, feature selection and the appropriate knowledge representation is the major challenge for establishing truly useful methods, whether the task is to be performed automatically, semi-automatically or through manual monitoring by human operators.

Another challenge is to establish useful and effective quality criteria for monitoring: confidence (how certain is the system that the alert signal is real), adhering to real-time requirements (how soon after the observation is made can it be processed and analysis results presented), deception resistance (how robust is the system in face of adversarial behaviour, intentionally corrupted data or deliberate obfuscation).

3.7.2 Vision

Consolidating this diverse area is a longer-range vision than the other market segments treated in this report. The vision for future monitoring applications is establishing a common framework for portability and interoperability of monitoring tasks of various types. The technologies are not yet comparable to each other in seamless ways, and treating them using a common language is a challenge in itself: establishing shared resources and shared task for research projects is a first step towards achieving common frameworks across the application areas.

Future monitoring applications will be available in numerous domains: home surveillance, energy monitoring applications, information services, personal security, family health applications etc. In many of these cases users, whether private or public, will have low levels of technical competence. How can best practice across applications be transmitted from user to user?

Data streams of various types can be expected to be publicly available, both by regulation and by design: individual users can install their own sensors, monitoring mechanisms and observation logs for perusal by the general public; corporations and public bodies may be required to publish their logs or action sequences to the public. Our vision is that these data streams can (in the words of Barack Obama) be used by the general public to "derive value and [enable us to] take action in [our] own communities"? How can monitoring tools and kits be commodised? How can we facilitate bi-directional data streams for monitoring and data aggregation applications between citizens and organizations?

4. CROSS-DISCIPLINARY CHALLENGES AND RECOMMENDATIONS

After conducting gap analysis studies, CHROUS identifies several directions that deserve European effort toward more efficient search engines in order to reveal the implicit knowledge and makes it reachable in fair and attractive ways to the user. From cross-disciplinary viewpoints, CHORUS recommend to:

4.1 TECHNICAL

Gap analysis studies[10] conducted by CHORUS allow identifying directions that deserve more European research and development effort. Hence, CHORUS recommendations toward more efficient search engines that will help to makes the implicit knowledge reachable in attractive ways to the user address mainly seven technical directions and are listed below.

4.1.1 Content enrichment

Content Enrichment is defined as all techniques that allow creation of new generation of metadata.

General definition of **Metadata**: all information besides the raw content that make the content searchable. It could be derived from different sources and present various types as follows.

- Manual annotations: "traditional" metadata provided by archivists or more generally by content owners,

[10] See D2.2 deliverable on "Gap Analysis" for more details

- Automatically generated low-level signatures: digital signatures based on automatic content analysis and description mostly represented by vectorial expression,
- Automatically generated semantic labels: after content analysis and digital signature generation, it is possible to learn concepts related to multimedia "entities" (visual/audio/video) based on learning corpora. Through statistical learning methods, it is possible to predict, with a level of confidence, the presence of a given entity's (visual/audio/video) concept in a multimedia document. It is related to what we call "concept detection". It hence provides new conceptual type of annotations automatically generated.
- Device generated information on media context: geographic/time, sensors parameters…
- …

We can state that the new definition of metadata includes: the traditional one (manual annotation) and computer (or device) generated metadata. Hence it is composed of two types: the human and machine generated metadata.

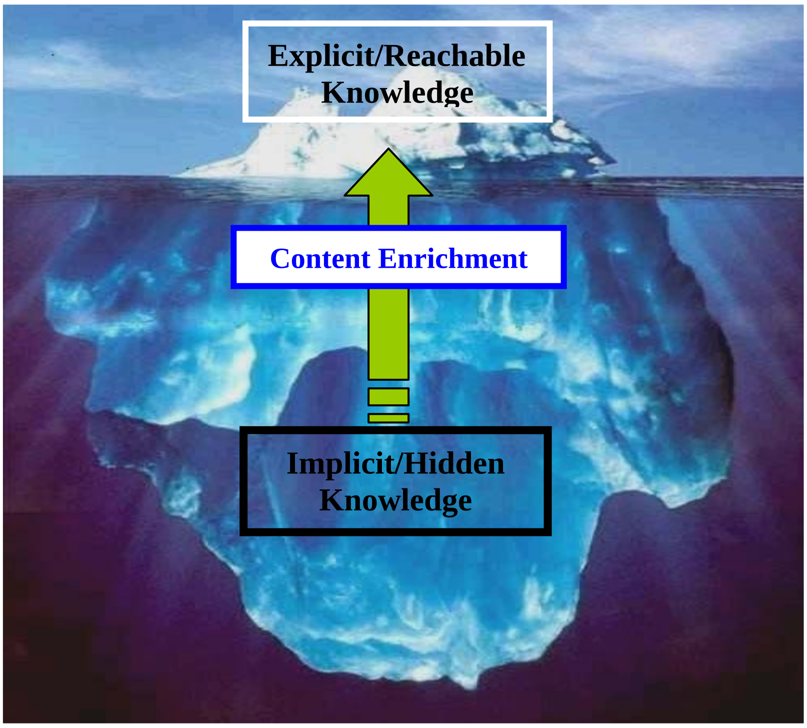


Figure 5 – Importance of "Content Enrichment" technologies to reveal hidden knowledge and it make reachable by new generation of multimedia search engines

In this section, we are concerned by the automatically generated metadata and the related methods with their progress. The scientific community needs to continue progress in this crucial direction to allow facing:

- the numerical gap: fidelity of the numerical content signature description to content ,
- the semantic gap: differences between the search engine results and the user intention through achieving more efficient indexing techniques for multimedia content enrichment and automatic meta-data creation through robust and efficient learning methods.

Socially-enriched automated indexing together with **folksonomies** will empower the robustness of the description and indexing techniques.

4.1.2 Events structuring and mining

New multimedia search paradigms based on content/context/event have to be developed to meet with user's semantic search result expectation and to go beyond current retrieval systems that are merely keyword-based or query-by-example-based. **Event structuring** is expected to be the main driver for media contextualization and to support queries by concept. Such structuring involves advanced content mining methods allowing similarity learning and powerful entities grouping methods as we are most of the time managing large data collections. Different issues need to be tackled:

- The problem of criteria selection for event definition and construction (cross-media event description, dynamic concepts with the temporal dimension…)
- The event mining from multiple and heterogeneous multimedia sources.

4.1.3 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.

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 organization.

Besides the amount of input data and generated features, complexity need to be managed for other growing quantities such as number of users, number of information sources, and number of data attributes / features dimension. Several directions have to be investigated: breaking algorithm complexity, development of technology aware algorithms, generalizing multidimensional indexes and similarity search structures, etc.

4.1.4 Personalization and recommendation

Recent breakthroughs in Web 2.0 technology have led to unprecedented growth of social networks for interacting with multimedia content. Rich and massive media data are constantly being posted to the Web, including social networking websites, photo and video sharing websites, and photo forums. Multimedia information Retrieval techniques will benefit in robustness and reliability while integrating the individual and collaborative users' actions and social indexing concepts to optimize access to multimedia content in realistic use cases.

Hence, modeling efficiently both the implicit and explicit feedback (HCI) to improve personalization and recommendation abilities of a search engine including collaborative tagging, filtering, user preference detection (as well as the relationships between the users, content and tags) present vry high potential for MSE technologies progress.

4.1.5 Informative user interfaces

A major challenge, in this regard, is the development of more informative user interfaces (UI) for future applications (too little overlap between networked media technology providers and UI designers today): toward smart visualization of media delivery and **enhanced user quality of experience** (QoE).

Potential large amount of 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 relevant one (for instance ranking the results by popularity), or helping them refine their query with additional information or new terms that will result in fewer, hopefully more significant results (clustering by type, by theme, by additional features, etc.). This is imbedded in the results structuring phase prior to results display to the user. The GUI responsible for the actual display and interaction with the user should be carefully addressed in a unified fashion since the goal of the "result presentation" step is to facilitate the task of users in their query preparation, refinement and content consumption. Beyond, what ever the level of technology efficiency, this GUI part is crucial as it is be the only visible part of a search engine to the user. Usability and commercial success is greatly dependant to this dimension..

4.1.6 Interoperable standards:

Interoperability arise as one of the most challenging and crucial issue for the next coming years. It is central concern for the development and the deployment of advanced search technologies. Interoperability has two folds:

- Human-generated metadata: develop **interoperable standards for meta-data** usually associated to proprietary content by professional content-owner through diverse format over time (intra-content owner) and across different professional models (inter-content owners). The issue is to handle and open ended content metadata format which is associated to objects/event allowing the information preservation through its life,

- **Automatically generated meta-data**: in this regard, several initiatives attempt to elaborate standards toward description and representation of such newly content enrichment methods. Most of these initiatives do not encounter the expected success in the community (technologists and users). The search engine technologies domain is evolving very rapidly and most often outstrips the establishment of such standards which become more constraining factor than enabling factor. Nevertheless, another promising standardization effort will target **interoperability of search services/technologies** allowing operating on **distributed and heterogeneous information sources** with diverse community of users.

4.1.7 Open multimedia corpora and evaluation framework

The availability of open multimedia corpora for the community of researchers (academia and industry) represent a key enabler for multimedia search engine technology at both levels: scientific and commercial success.

While multimedia search techniques witnessed significant advances, paradoxically, our capacity to simply identify the most performing techniques among numerous solutions is hindered by sheer number of proposed alternatives and by the multiplicity of evaluation campaigns. Nowadays, there are no established methodologies fostering technology uptake as there are no common performance landmarks. We need developing methodologies and open evaluation framework allowing identifying the most promising technologies facing the new challenges of multimedia search engines at a **large scale.**

We need to elaborate a hierarchical structuring typology of all evaluation dimensions considering the diversity of success criteria depending on the component-based technological assessment or usage-based assessment where the user criteria should be considered (as usability in addition to efficiency). This hierarchy will give an objective picture of the current evaluation landscape and allows identification of non-covered dimensions to be developed as **new paradigms**. The **open evaluation framework** have to be **interoperable** with existing benchmarking initiatives to ensure comparable evaluations results.

In this regard, developing and gathering of a consolidated multimedia content repository that collects, at a single access point, large-scale standardized collections of multimedia data with associated annotation and ground truth together with terms of use will foster the **comparability of evaluation results**. Besides the known benchmarking campaign data, this has the advantage to make more visible less known collections used by individual EU projects or individual communities without sharing it with the community at large (especially when terms of use are loosely constraining). This will avoid fragmentation efforts in content collections gathering.

4.2 SOCIO-ECONOMIC AND LEGAL ISSUES

4.2.1 Integrity, privacy, data ownership

For individual users, it is important that information about them is used appropriately. This is true both in cases where the usage itself involves potentially private details about the user, but also in cases where the usage data may seem innocuous but still have commercial value for profiling, for directed advertising or for customer relations purposes. How to control the distribution, storage and retention of usage data may be solved variously, but affording users a sense of control and an appropriate level of awareness of what information their actions give rise to needs to be addressed.

A related long-term issue is that of data permanence. What permanence will data about a user have for instance in the specific but entirely predictable situation where users decease and their estate wishes to modify, continue or discontinue a service based on those data?

The risk of misuse of data is exacerbated by the trends given in the introduction, with more tailored, knowledge-based interaction, the trend towards greater user participation, and the trend towards search being part of other services, where the user may be less aware of the component technologies.

4.2.2 Identity and anonymity

Given the trend towards greater user participation in information services, there is an attendant demand to afford users the warranted right to remain anonymous at will and access to mechanisms to ensure that the anonymity holds. For certain types of actions, whether socially embarrassing, politically controversial, or commercially timely, the right to be anonymous may be the deciding factor whether one wishes to engage or not.

The converse issue, important for commercial, public, and private services alike, is that of identification: how one can ensure that one's communicative counterpart is indeed the person or organisation it claims to be, how one can be able to certify one's identity at will, and how one can ensure that one's identity is safe from encroachment from others.

4.2.3 Reliability of third parties

A critical step towards the creation of new services and leveraging existing ones is the trust invested in basing one's offerings on those of others. In a digital economy with a great deal of technological and commercial churn, a service which is relied on at one time may have dissolved at some other time. How can a company base its services on e.g. a metadata annotation service if it has unacceptable levels of downtime? If it is purchased by a competitor? If its servers may crash and leave its users without their data? The possibility to build business opportunities based on creation of common goods is based on reliance on other services and on open standards. These only emerge over time, but a public audit and certification of reliability and quality metrics based on customer experiences would be of use.

4.2.4 Public policy – best practice

The role of public bodies in ensuring development and provision of future information access services is first and foremost to be an informed customer and to observe best practice by providing information services designed according to best principles. What, in each given situation, can be identified as best practice may not be within the competence of each specific public body; there must be provision for consultation with competence to assess these issues.

4.2.5 New business models - prevent customer lock-in

The general and underlying question in many of the above issues is that of lack of new business models. Previous content, service, network and device providers must join in providing standards for new businesses to emerge; if not, they risk being sidestepped in an economy where the infrastructure investments for providing innovative services are relatively low. As a special case of this, subscription based services such as cable television cannot expect to be able to control network access to steer customers to their offerings; they will be bypassed by services on open networks instead.

4.2.6 Cooperative efforts

Encourage the share of data pools and make use of available know how to promote innovation through experimentation; focus research funding on more strictly delimited use cases, to provide comparability between research efforts – simultaneously assessing potential for market take up of technology solutions. Promote formation of cooperative research clusters. Integrate the craft of interface design with technology development and content ownership.

4.2.7 Public consultation

By initiating public discussion on information policy, including certification instruments, integrity, accountability, transparency and reliability issues, socio-economic debate can be addressed by relevant experts and the general public made aware of topics under debate, and simultaneously factors it out of technology development projects.

4.2.8 An observatory of search – European certification procedures

The suggestion from CHORUS think tank on socio-economic issues is to introduce a transparent and revisable procedure for privacy and integrity certification supervised by independent authorities. For instance, the European Privacy Seal, along the lines already introduced in Germany, is awarded for IT-products and IT-based services that have proven privacy compliance in a two-step certification procedure. The user benefits from a certified quality product or service, the manufacturer profits from market advantages, and added reliability of service, the privacy protection authorities benefit from a relief in control tasks.

A further suggestion is to establish a permanent entity that includes 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 European 'observatory of search', whereby a number of relevant issues are monitored and regularly updated, including providing a guidance for establishing best practice in various fields of public and commercial activity. 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. An observatory of search would channel discussion on information policy across the union and would provide an arena for stakeholders of various types to congregate and establish practices for future services.

5. SEARCH ENGINES IN THE PERSPECTIVE OF THE FUTURE INTERNET

Media will be plentiful all over the Future Internet (FI), often in distributed form. Making all media **searchable and accessible** (metadata generation and structuring) is the major challenge for the FI. Users will be accustomed to retrieving what they need delivered just-in-time. Some users will choose to deal with data, applications, storage from the "cloud" as a service infrastructure.

Search is at the **application** level: the network should be neutral (agnostic) with respect to applications. One exception can occur for the media delivery optimization stage (network layer) which impacts the user QoE. In this regard, content-based routing as well as QoS control (through API requirements) can have significant impact on the quality of the content delivery and presentation to the content consumer.

Search engines will be omnipresent, often embedded. Hence, they will not have significant differentiating impact on FI: either **unique** or **multiple Internets**: Internet of Things, Internet of Services, and Internet of Content.

Search operates on **metadata** of **searchable entities.**

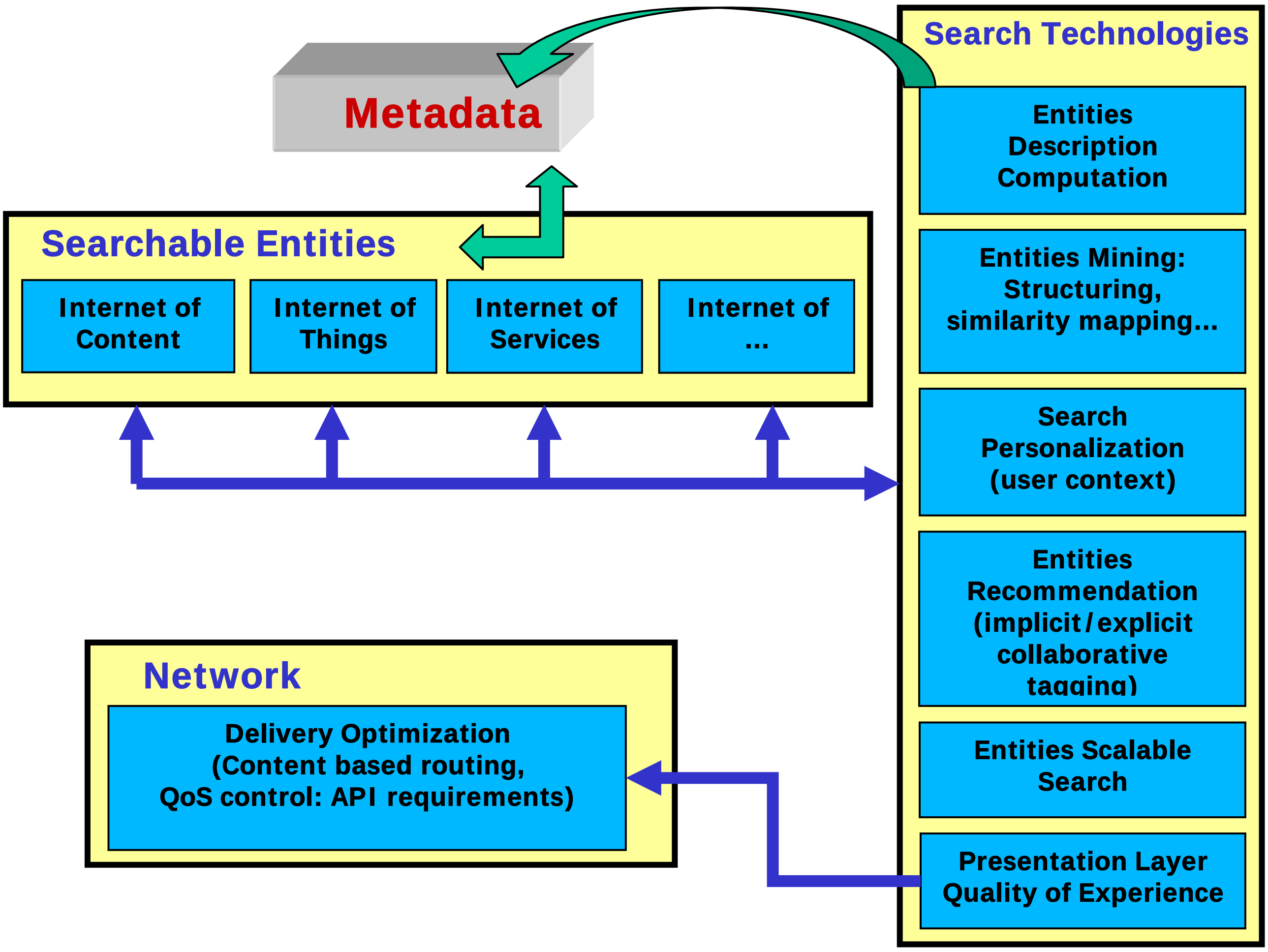

Metadata
Searchable Entities
Internet of Content
Internet of Things
Internet of Services
Internet of ...
Network
Delivery Optimization (Content based routing, QoS control: API requirements)
Search Technologies
Entities Description Computation
Entities Mining: Structuring, similarity mapping...
Search Personalization (user context)
Entities Recommendation (implicit/explicit collaborative tagging)
Entities Scalable Search
Presentation Layer Quality of Experience

Figure 6 – **Search Services in the perspective of Future Internet Infrastructure**

Two major issues need to be highlighted in this perspective:

- The derived knowledge from the **growing power of social networking** needs to be highlighted. It may represent an alternative to information retrieval through trusted recommendation mechanism.
- **Interoperability,** in this context, appears as a major concern. It will need to be carefully addressed in the coming years beyond the Internet of Media reaching all networked entities including "Services" and "Things".

Finally, in the perspective of Future Internet development, we have identified five major challenges which are listed below:

1. Heterogeneity and diversity issues, concerning resources, information sources, infrastructure, user communities, etc;

2. Human factors, concerning "the user in the loop", feedback, personalization, recommendations, interaction with the information, emotional characteristics, etc;

3. Real time issues, concerning content/media indexing and retrieval, visibility, aggregation, re-composition of services, etc;

4. Security and privacy issues;

5. Evaluation and benchmarking issues.

6. CONCLUSION

Coordination effort through CHORUS allows positioning various EU efforts among the technological landscape dimensions. The mapping reveals that **Europe** is **rich with very sharp expertise** in many vertical topics in the field of search engines. It was even pointed out that innovative commercial services are feasible today using exiting pieces of research results for some niche markets (mainly for business search market – not necessarily web information).

While European search technology research is of internationally established quality and some corporations deliver services based both on repackaging existing technology and developing in-house technology, the large consumer services are mostly Transatlantic. The challenges and trends outlined in the document provide an opportunity for service providers in Europe: the field is open for new entrants to provide services for specific cases, for new media, for new service contexts. Scalability and service reliability based on continuous progress of technology effectiveness, requiring annotated reference content availability, are currently the most crucial bottlenecks for establishing take-up. From a European perspective, the industrial base for search engines lacks actors.

Regarding consumer search market, Europe is lacking today an integrated program that gathers all needed expertise for building competitive real life search engines. CHORUS recommends to:

1. **Empower aggregation and orchestration of such expertise into an operational end-to-end search organizational structure.** Europe needs a **strategic initiative** such as creation of an "Institute" or a "Centre" for Information processing. This recommendation deserve interest, investigation and focus from governmental, industrial and academia players for the emergence of a new generation of multimedia search engines. The existing European funded projects framework allow fragmented technical results for a given period of time with no follow-up or aggregating procedure. Hence the **existing expertises represent separate pieces of a puzzle with no "Big Picture" to guide toward a significant achievement**.

2. **Foster user-centric design** (market-pull) **requirements for EU funded projects against the technology-driven design** (techno-push). The latest is one of the major draw-back we have identified in the so-far conducted efforts. The market success is much dependant from user **acceptance** and the **usability** of whatever advanced/revolutionary technology presented to the end-user who is finally the consumer. It is crucial to integrate these constraints and requirements from the very early specification and design stages.