CHORUS Vision Document

(D3.4)

Deliverable Type *: PU
Nature of Deliverable **: R
Version: Released
Created: 24 Feb 2009
Contributing Workpackages: All
Editor: Jussi Karlgren, Markus Kauber
Contributors/Author(s): Jussi Karlgren, Nozha Boujemaa, Ramón Compañó, Christoph Dosch, Joost Geurts, Henri Gouraud, Paul King, Joachim Köhler, Pieter van der Linden, Robert Ortgies, Åsa Rudström, Nicu Sebe

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

Abstract:
The goal of the CHORUS Vision Document is to create a high level vision on audio-visual search engines in order to give guidance to the future R&D work in this area and to highlight trends and challenges in this domain. The vision of CHORUS is strongly connected to the CHORUS Roadmap Document (D2.3). A concise document integrating the outcomes of the two deliverables will be prepared for the end of the project (NEM Summit).

Keyword List: multimedia, search, research, vision, socio-economic aspects, trends, challenges

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/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

EXECUTIVE SUMMARY

Search and information access is a base technology for most human intellectual activities. In this report, 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 to address:

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

As a point of departure we discuss five identified defining trends in multimedia access:

- Increasing rate of data generation and storage
- Increasing innovation of usage involving active participation on the part of the user communities
- New content types with technological challenges to provide content analysis
- The issue of disappearing search – search being embedded in other applications and services
- Personalisation and context aware services

These in turn give rise to a number of challenges for future service providers:

- Issues related to privacy, integrity, data ownership, anonymity and identity
- Issues related to trust, reliability and outsourcing
- Issues related to establishing business models, best practice, and public policies
- A recommendation to initiate and uphold public debate on these issues

These challenges and 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 Transatlantic. 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.

1 INTRODUCTION

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 document 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.

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 digitalisation of heritage and previously inaccessible yet valuable material from archives, museums, private collections and corporate data warehouses. The dematerialisation 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 through 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 commodised 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 MARKET SEGMENTS

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 users of the search utility.*	Unrestricted, Restricted
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 characterised 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

2.1 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 monetise 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.

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.

2.2 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[1] model has been 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 IPTV[2] infrastructures during the past several years. Despite these efforts, in most countries, the market take-up[3] 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.

1 Broadcasting: Initially analogous dissemination of radio signals along a one to many model..
2 IPTV: Offering of TV over managed networks using IP protocols.
3 Around 9 millions users end 2007 according to a study by MRG.

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 Web4. 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.

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.

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

[4] Comscore August 2008.

relevance of the advertisements to their audience, moving from a more qualitative evaluation of advertisement effect.

2.3 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.5 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.

[5] IDC Press Release: 12 December 2008

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.

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.

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.6

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?

2.4 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

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

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.

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.

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.

2.5 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.

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.

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?

2.6 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.

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 heterogenous 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.

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).

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?.

3 CHALLENGES

The above market areas give rise to several general challenges for the future information landscape.

3.1 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.

3.2 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.

3.3 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.

3.4 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.

3.5 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.

3.6 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.

3.7 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.

3.8 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. Discussion

- End of document -