"Latest Trends in

Multimedia Search Computing"

Version: 2.0

Last Update: December 25, 2012

Distribution Level: *PU = Public,*

Executive Summary

This report provides an overview of the existing technology in several areas related to search together with the major challenges and future trends. The areas of interest include Media search, Social and Mobile search, Enterprise search, Technology Transfer, and the Economics of Search. For each of these areas the document highlights the overlaps and the differences with the other areas as well as the recommendation for future activities.

Section 1 presents the most of the important aspects of Media search. This area is by definition very large and it touches practically all the other areas covered in this document. Nevertheless, the goal here was to cover the technological advances in multimedia indexing and retrieval, large scale search issues, distributed media, crowdsearching and search in the cloud.

Section 2 covers the important aspects of social and mobile search. While this area could be considered as an integral part of Media search, we have decided to present it separately in order to emphasize the important differences and issues pertinent to social and mobile aspects. This area is currently one of the most active areas and the document covers most of the hot topics ranging from the interfaces and services to convergences and divergences between mobile and social search and web search.

One of the key elements in increasing the competitiveness of the digital economy and therefore, a strategic market for the European Union is Enterprise search (Section 3). The term *enterprise search* (ES) is commonly used to describe the application of search technology enabling information retrieval within organizations. Therefore, as the practice of identifying and enabling specific content across the enterprise to be indexed, searched and displayed to authorized users, it should be differentiated from other type of search, such as web search.

Technology transfer covered in Section 4 is another important area that is inherently connected to search. The document covers the role of open benchmarking in fostering scientific excellence and technology transfer, the lessons learned from European collaborative R&D projects in technology transfer and the existing benchmarking initiatives.

Finally, the economics of search is discussed in Section 5. As can be inferred from previous sections, the search query and its parameters, in particular personal information used to 'enrich' search, have become an essential building block of Internet activities and its economics, as it were a new currency of the digital world. On the grounds on this currency, information goods and services are offered, and physical goods purchased and paid for. This section is covering all these important aspects which are usually neglected in most of the technologically oriented studies.

Table of contents

Rationale of this document

The objective of this document is to synthesize the global findings of the project in the different areas of search
- based on Chorus+ member expertise,
- incorporating the outcomes of the 5 previous Think-Tanks,
- with the validation and complements brought by experts attending the Think-Tank n°6.

This document is elaborated in two steps:
- a draft document has been prepared by Chorus+ members, and was distributed to the participants to the Think-Tank 6;
- the document was discussed during the TT6 and the 25 experts present expressed their opinions on the ideas presented in the document, and brought in corrections and alternative views they had on each of the subjects discussed;
- this version of the document was rewritten after the TT6, taking into account the outcomes of the discussion

The document is structured in a number of sections addressing the various search areas and challenges related to search. Each section aims at covering the following aspects for the area that it covers:
- background analysis, in terms of actors, business and societal aspects;
- current technical state of the art and R&D mainstreams;
- future trends, in terms of functionalities and technologies; grand challenges;
- relevant recommendations for future EC collaborative R&D in the area.

1. Media Search

1.1. Background and state of the art in Media Search

The explosion of multimedia content in databases, broadcasts, streaming media, etc. has generated new requirements for more effective access to these global information repositories. Content extraction, indexing, and retrieval of multimedia data continue to be one of the most challenging and fastest-growing research areas. A consequence of the growing consumer demand for multimedia information is that sophisticated technology is needed for representing, modeling, indexing, and retrieving multimedia data. In particular, we need robust techniques to index/retrieve and compress multimedia information, new scalable browsing algorithms allowing access to very large multimedia databases, and semantic visual interfaces integrating the above components into unified multimedia browsing and retrieval systems.

The aim of these systems is to handle general queries such as "find outdoor pictures or videos of an interview with James Cameron discussing the making of the Titanic film." Answering such queries requires intelligent exploitation of both speech and visual content. For multimedia retrieval, the combination of multiple integrated media types increases the performance of content-based retrieval. Available content analysis and retrieval techniques tailored to a specific media are therefore not adequate for queries as the one mentioned above.

Clearly, Multimedia Search is a very broad area covering both structural issues (e.g. framework, storage, networking, client-server models) and intelligent content analysis and retrieval. These all need to be integrated into a seamless whole which involves expertise from a wide variety of fields. Below we cover the state of the art in a variety of topics of interest.

1.1.1. Multimedia Content Indexing

A plurality of multimedia content indexing approaches has appeared in the literature over the past years, ranging from approaches based on low-level indexing features alone to others also using semantic information extracted from the multimedia content. Starting with the simpler ones, the methods are based on extracting low-level features that accurately and compactly represent properties such as color, texture, shape, etc. of the content. This is motivated by the assumption that similarity (or dissimilarity) calculated directly in a low-level feature space can be analogous to (or at least indicative of) the semantic similarity of multimedia content, an assumption that hardly ever holds. To bridge this semantic gap, more elaborate approaches have been developed, combining low-level features with higher-level semantic information that can be

extracted either from the textual modality of the multimedia content or by means of multimedia classification techniques. The latter rely on machine learning algorithms to associate each piece of multimedia information with one or more of a (usually limited) number of pre-defined semantic classes, which can range from simple ones (e.g. "outdoors") to considerably more meaningful ones representing real-life events (e.g. "person driving a car"). While this highlights the importance of event-level information for the indexing, search and retrieval of multimedia content, the drawback of current state-of-the-art (SoA) approaches for the extraction of such information is that they treat this problem as the problem of finding a direct correspondence between the low-level features and the event of interest (e.g. the separating hyperplane in a very high-dimensional low-level feature space). Thus, they do not consider how an event can be modeled, what are the elementary pieces of semantic information an event is made of and what are their interrelations, and how these could be reliably extracted from the multimedia content. Consequently, current SoA approaches are shown to perform acceptably only for a handful of mostly simple semantic classes (rather than complex events) and have very limited impact on the effectiveness of multimedia indexing.

1.1.2. Bridging the Local-Global Gap in Search

The idea of exploiting information from a user's activities and their local working context to improve search results has a long tradition in the area of personalized information retrieval. To take examples from widely deployed systems: the personalized search of Google makes use of a user's search history and behaviour and the pages that the user has clicked on to rerank search results. The interests of the user are not represented in terms of any sort of central classification system. Rather, they consist of a collection of more specific bits of information, such as the number and date of visits to particular websites, which are then used directly for the reranking of the search results. By contrast, a central organization scheme was used in the much earlier version of Google's personalized search that was tried out in its "laboratory" in 2005: each user was given an opportunity explicitly to indicate their interests in terms of a number of categories such as "Movies", but no use was made of information from the user's naturally occurring behaviour with the search engine. Some efforts have been made, however, to interpret a user's behaviour and work context in terms of a general classification scheme.

1.1.3. Large-scale Multimedia Search

Recent years have witnessed a consistent growth of content-aware and multi-modal search engines deployed on massive multimedia data. Popular multimedia search applications such as Google image, Youtube[1], Shazam[2], Tineye[3] or

---

[1] http://www.youtube.com
[2] http://www.shazam.com

MusicID[4] clearly demonstrated that the first generation of large-scale audio-visual search technologies is now mature enough to be deployed on real-world multimedia data. Google image enables content-based similarity search and near-duplicates retrieval on more than 50 billion images. Youtube's content-based video identification system analyses the equivalent of 100 years of videos each day. The database of the system contains 8 million reference video files provided by 3000 content owners. Tineye reverse image search is now running on more than 2 billion images. Shazam song's identification service works on about 10 million of songs whereas Music-Id exceeded the number of 38 million songs indexed in their alternative content-based indexing technology.

All these successful applications did greatly benefit from 15 years of research on *multimedia analysis* and efficient *content-based indexing* techniques. Figure 1 presents some of the most influential academic works of the last decades as a function of the number of items used in their experiments. It is interesting to notice that the only scientific publication reporting an experiment on more than 1 billion documents is a work of Google Research Lab (in 2007) which required the use of 2000 CPU's, far away from the hardware resources available for most other research players in the field. It shows that bridging the last orders of magnitude between algorithmic research and real-world applications is clearly a matter of large-scale infrastructures and distributed architectures. On the other side, fundamental research on breaking algorithm complexity has been a crucial pre-requisite. Whatever the efficiency of the implementation and the use of powerful hardware and distributed architectures, the ability of an algorithm to scale-up is strongly related to its time complexity and space complexity. High level tasks such as content-based image search, video copy detection or songs identification have indeed been scaled-up by reducing the complexity of low-level algorithms such as high-dimensional feature spaces quantization, approximate KNN search or embedding of feature sets.

On another side, *concept-based* and *object-based* indexing technologies are still not really integrated in real-world applications (except few items such as face, drawings or clip arts in Google Image or music genres in audio search engines). Effectiveness of automatic annotations methods on small data has indeed concentrated most efforts of the research community for a long period and only very recently did appear successful large-scale academic works. The most promising one is probably the automatic recognition of (semi)-rigid visual objects in large-scale image and video collections. Recent academic works did actually prove that buildings, trademark logos, paintings or manufactured objects can effectively and efficiently be retrieved in millions of images. Thanks to recent advances in scalable machine learning and the availability of huge

---

[3] http://www.tineye.com
[4] http://www.gracenote.com/products/musicid

training sets, large-scale concept-based indexing also starts to be a reality. The most noticeable achievements were initiated by Princeton and Stanford universities through the ImageNet[5] initiative. ImageNet is an image database organized according to the WordNet[6] hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images (resulting today in more than 14 million semantically tagged images). It created a beneficial stimulation of the computer vision community towards scalability and the best technologies evaluated in the 2011 ImageNet challenge got high classification rates (75%) on thousands of visual concepts. ImageNet being organized according to WordNet ontology, this also opens the door to effective *cross-media semantic indexing and search* that are so far not really successful.
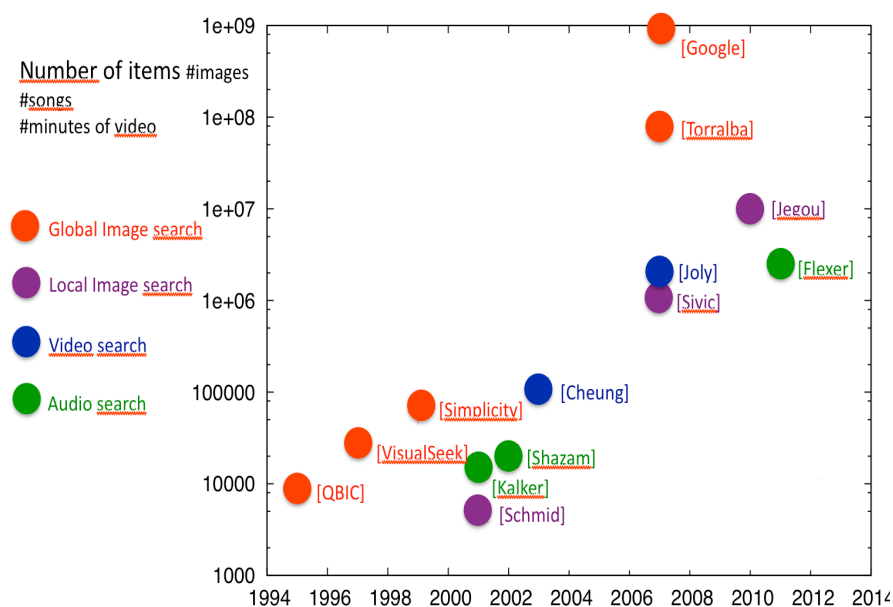
Figure 1: Some of the most influential works related to large-scale multimedia search as a function of the number of items used in the reported experiments. References of the works: QBIC [20], VisualSeek [21], Simplicity [22], Torralba [23], Google [24], Cheung [25], Joly [26], Kalker [27], Shazam [28], Flexer [29], Schmid [30], Sivic [31], Jegou [32]

Besides purely content-based techniques, another range of search technologies that start to be successfully deployed in real-world systems is spoken and written content indexing of audio-visual documents. Via *speech recognition* (audio channel) and *text recognition* (visual channel), most of the linguistic information encoded in audiovisual data can actually be accessed using efficient and scalable text-based indexing tools. This technology is still in the labs of search companies (e.g GAudi by Google, or Voxalead by Exalead) and at a relatively moderate scale. It is likely that they will be deployed very soon in

---

[5] http://www.image-net.org/
[6] http://wordnet.princeton.edu

large-scale productive applications since there is no technological gap towards scaling them in a similar way than text documents. This could enable a step forward for video search engines (Youtube, Dailymotion[7], etc.) that are so far desperately only text-based and with modest search performances.

Recent efforts of the Laboratory for the Recognition and Organization of Speech and Audio (LabROSA)[8] of Columbia University lead to the compilation of the Million Song Dataset (MSD) - a large collection consisting of music meta-data and audio features. This freely available dataset gives researchers the opportunity to test algorithms on a large-scale collection similar to real-world commercial settings. Besides audio features music metadata identifiers are provided that can be used to access complimentary metadata repositories such as Musicbrainz[9], 7Digital[10] and Rdio[11] as well connect to the actual audio files. Through the Rosetta Stone project by The Echonest[12] further data can be linked to the MSD including social media information provided by Last.fm[13]. Due to copyright issues the source audio files of the MSD cannot be distributed, which makes it impossible to develop new audio features. Recently further contemporary audio features have been extracted[14] from audio samples provided by 7Digital. This provides the possibility to facilitate comprehensive benchmarking on the MSD such as the recently finished Million Song Dataset Challenge – an evaluation campaign similar to the famous Netflix price to foster further enhancements in music recommendation systems. A further large scale music evaluation campaign is the MusiCLEF benchmarking initiative – currently hosting a multimodal music tagging task. These constitute the first large scale evaluation campaigns in Music Information Retrieval and complement established evaluation boards such as MIREX. Information about datasets used by evaluation campaigns such as VideoOlympics, TRECVid, ImageCLEF, MusiCLEF or MIREX as well as other established benchmark datasets are provided in the collaborative platform[15] of the CHORUS+ project.

1.1.4. Distributed Media

The machine learning literature has studied a variety of ensemble-based meta methods such as bagging, stacking, or boosting, and also combinations of heterogeneous learners. One current trend is exploring automatic data organization using machine learning methods in combination with ontology

---

[7] http://www.dailymotion.com

[8] http://labrosa.ee.columbia.edu

[9] http://musicbrainz.org

[10] http://www.7digital.com

[11] http://www.rdio.com

[12] http://the.echonest.com

[13] http://last.fm

[14] http://www.ifs.tuwien.ac.at/mir/msd/

[15] http://www.avmediasearch.eu/wiki/

alignment for large sets of users in distributed environments. Common recommender systems are usually used in one of two contexts: 1) to help users locate items of interest they have not previously encountered, and 2) to judge the degree of interest a user will have in items they have not yet rated. In both cases recommender systems serve to meet the user's information demand in a personalized way. In many aspects, recommendation algorithms for folksonomies may substantially differ from methods known from the Web IR scenario. Web retrieval primarily utilizes the content of hypertext documents and the link structure of cross-references between them. In contrast, the trend is to explore more enhanced web community structures and Web 2.0 systems that provide much richer, collaboratively edited, social metadata (comments, users and user groups, cross-annotations, etc.).

1.1.5. Crowdsourcing

Content sharing through the Internet has become a common practice for the vast majority of web users. Due to the rapidly growing of new communication technologies, a large number of people all over the planet can now work together in ways that were never before possible in the history of humanity. The collective intelligence that emerges from the collaboration, competition, and co-ordination among individuals in social networks has opened up new opportunities for knowledge extraction. Valuable knowledge is stored and often "hidden" in massive user contributions, challenging researchers to find methods for leveraging these contributions and unfold this knowledge. Maybe the most interesting part in this phenomenon is the potential to exploit the collective effort invested from web users for their own interest, in order to accomplish tasks that seem infeasible of extremely tedious. For instance, the Google's "Did you mean" tool corrects errors in search queries by memorizing billions of query-answer pairs and suggesting the one closest to the user query, similarly the Google Image Labeler (ESP Game) [34] exploits the effort invested by the users to play a game for obtaining accurate image annotations. This game-based approach is often called Game With A Purpose (GWAP[16]).

The idea of crowdsourcing has found applications in many different fields ranging from news (e.g. blogs) and fun content generation (e.g. YouTube), all the way to solving challenging scientific problems including media search based on speech recognition [42] and music analysis [43][16]. Indeed, driven by the abundant availability of content on the web researchers have shifted their interest away from devising more sophisticated algorithms to solve their problems, and turned into discovering ways on how to use the huge volume of available information in order to exploit the collective intelligence that emerges from the collaboration, competition, and co-ordination among individuals in

---

[16] http://www.gwap.com

social networks [17]. The vast majority of currently performed research efforts are now applied on social media, showing how some problems can be more easily addressed by resorting to this kind of intelligence. Irrespectively of the domain and despite the fact that the unstructured and noisy nature of social media raises serious obstacles, crowdsourcing is probably the first option to examine when faced with tedious tasks.

1.1.6. Search and the cloud

The problems around search and the cloud are closely related to both the enterprise search and big data areas, which are addressed in the previous two sections. Therefore a specific section is dedicated to this subject.

Search and could be analyzed from two distinct perspectives:
- from an architecture standpoint where the search engine is built using cloud technologies;
- from a data storage standpoint where the data being searched is stored and managed in a cloud.

The first perspective has been explored and amply demonstrated since the early days of Internet search, in particular with the solutions developed by Google. From this perspective, one could argue that the term parallel processing and distributed architecture applies as well as cloud based architecture. Such distributed/cloud based architectures were applied to various aspects of the search engine (crawling, indexing, matching, see Chorus for analysis of these functional terms) and this led to the development of tools and general purpose solutions applicable to storage, and to general purpose processing. Once it became apparent that such efficient distributed architectures could be built over TCP/IP layers, development of management and accounting tools were the last building block leading to the generalization of cloud based services and computing.

Actors in this "story" were the leaders of Internet search (Google, Yahoo!, Microsoft), joined by the large corporations developing specialized services or e-commerce over the Internet (Amazon, Facebook, Twitter, ...). In this list, Amazon is the most notable for having pushed the cloud logic all the way to the market. In this world race, mostly dominated by US enterprises, the presence of Europe based Exalead should be noted with a cloud based search solution which belongs to the same league (in terms of functionalities only, Exalead Internet search traffic being much lower).

The second perspective comes into play after the success of the first. With the development of cloud based storage and application processing, Enterprise Search needs to extend its reach into the cloud to crawl and index the data which

it holds. Note that access to enterprise data stored in a cloud is an issue for any enterprise search engine, independently of its own architecture.

In such a configuration, the issue is that of the existence of appropriate API allowing for third party applications (the search engine) to access searchable data while maintaining appropriate security. In this context, the provider of cloud storage is well positioned to offer a search function into the data it holds for its enterprise customers, and may develop its own search solution or purchase one from the leading developers (Exalead in Europe). In both cases, the cloud storage provider must address the problem of accessing and indexing data not stored into its cloud as it is unlikely that 100% of the searchable enterprise data is stored in the cloud.

The problem is similar, but not identical to that of unified search both on the personal desktop and into the corporate data store. No clear cut solution seems to have appeared to address this problem.

As cloud storage and computing develops, the competition for enterprise search solutions capable of searching into both spaces to maintain the "unified access point" paradigm is going to heat up and presents new opportunities. But the main challenge here is probably more of "software engineering" nature rather than "technology improvement and research".

The problem becomes more interesting when one takes into account the observed evolution of data volume growth and generalization of cloud based storage. With that perspective in mind, the tasks of "crawling" and "indexing" (as defined in the first Chorus project) are likely to be impacted as it will become nearly impractical to transfer the relevant data to the search engine site to perform these tasks (indexing in particular).
This problem is potentially augmented in the context of multimedia by the size of individual documents involved (images and video).
If the data volume prevents their transfer for processing, the alternative is to "transfer" the processing to the location where the data resides. This implies that the site, or cloud based storage service, be capable of accepting software components prepared by outside partners or customers to be executed locally to their service. This also implies that the API and relevant middleware proposed by the cloud service offer access and processing functions compatible with the distributed and expandable nature of the cloud service.

The problem at hand is not trivial as it is clear that the data volumes at hand, which prevented their transfer to the remote processing site also prevent a simple minded and brute force approach for the data access API offered by the cloud service. Recently, the observation that a significant class of data processing applications can be expressed in terms of a small set of primitives that are, in

many cases, easy to parallelize, has led to frameworks, such as MapReduce, which have been successfully applied in data processing, mining, and information retrieval domains. The case of multimedia data (but not only) however still poses significant challenges. Multimedia applications actually share some core operations performed on very high-dimensional features, including near-neighbor & similarity search, clustering, training & classification or graph analysis. When performed naively with classical cloud-based frameworks, these core operations are often very costly, because the number of objects and features that need to be shared between the nodes can be prohibitive. Next generation cloud-based multimedia processing and analysis systems should notably include awareness of the utilities of data and features to a particular task, notably for ranked processing operations such as k-NN search, k-NN joins, top-k skylines, top-k group, etc. [24]. This will more generally require building new scalable data processing middleware's adapted to the needs and requirements of multimedia applications.

To illustrate our point, one could describe the following scenario or use-case: Exalead holds within its cloud based architected search service a maintained crawl of the web tagged with a set of facets (domain of interest, country, language, etc.). A young start-up aiming at providing of a specialty and value added service wants to "re-crawl" and "re-index" a subset defined by one or more tags of this initial crawl. The result of this re-indexing phase would be a new targeted search service. The only practical approach is for the start-up to send its re-crawl and re-indexing software to Exalead's cloud which can be done only if Exalead offers a set of API allowing transfer and execution of such programs with the expected level of security and efficiency. Beyond the software engineering effort this implies, research into parallelizable and distributable algorithms is necessary.

The use-case above remains somewhat "centralized", and assumes for instance that duplicate detection has been resolved during the first larger crawl. One could imagine other scenarios where this is not the case, and where duplicate detection (by nature a "centralized" function) must be executed in a distributed fashion across the cloud.

It is also clear that the next evolutionary step is to consider not only that the indexing function is executed in the cloud, locally to the relevant data, but that the resulting meta-data are kept also locally. In this configuration, it is the "matching" function that becomes distributed across the cloud. Once again, we are faced with the problem of distributing across the cloud a function centralized by nature.

The COAST project funded by the European Commission explores avenues related to the issues of distributing crawl, indexing and meta-data storage across multiple sites.

As a conclusion of these comments on the relationship between search and the cloud, it appears that the major issue to be addressed is that of the opposition between algorithms and functions which are conceptually centralized and the objective of implementing and executing them in a parallelizable and distributed fashion. Efforts such as MapReduce have shown the power of solid conceptual and theoretical analysis of the proposed solutions. Further progress in this domain will most likely benefit from similar and ongoing theoretical studies of parallel and distributed algorithms.

1.2. Future trends and grand challenges in Media Search

1.2.1. Personalized access

Advanced models for personalization and conceptualization of information are required. Models should include those dedicated to the context of the original source media as well as those reflecting the situation of the user. User models need also to include structures to describe personal demographic profiles and preferences, local setting (e.g., time, place, and organization), task assignment and goal, and recursively, other domain specific ontological structures, to enable contextual reasoning. User context and profile should be dynamically updated and refined to reflect external input including the user retrieval and interaction history. The context models can be described in knowledge representation structures that adhere to the relevant international standards. For example, one of grand challenges for music media search is to provide the best music for each person by finding or even generating appropriate context-aware music [44].

1.2.2. Human Centered Methods

Whether we talk about the pervasive, ubiquitous, mobile, grid, or even the social computing revolution, we can be sure that computing is impacting the way we interact with each other, the way we design and build our homes and cities, the way we learn, the way we communicate, the way we play, the way we work. Simply put, computing technologies are increasingly affecting and trans-forming almost every aspect of our daily lives. Unfortunately, the changes are not always positive, and much of the technology we use is clunky, unfriendly, unnatural, culturally biased, and difficult to use. As a result, several aspects of daily life are becoming increasingly complex and demanding. We have access to huge amounts of information, much of which is irrelevant to our own local socio-cultural context and needs or is inaccessible because it is not available in our native language, we cannot fully utilize the existing tools to find it, or such tools are inadequate or nonexistent. Thanks to computing technologies, our options for communicating with others have increased, but that does not necessarily

mean that our communications have become more efficient. Furthermore, our interactions with computers remain far from ideal, and too often only literate, educated individuals who invest significant amounts of time in using computers can take direct advantage of what computing technologies have to offer.

1.2.3. Multimedia Collaboration

Discovering more effective means of human-human computer-mediated interaction is increasingly important as our world becomes more wired. In a multimodal collaboration environment many questions remain: How do people find one another? How does an individual discover meetings/collaborations? What are the most effective multimedia interfaces in these environments for different purposes, individuals, and groups? Multimodal processing has many potential roles ranging from transcribing and summarizing meetings to correlating voices, names, and faces, to tracking individual (or group) attention and intention across media. Careful and clever instrumentation and evaluation of collaboration environments will be the key to learning more about just how people collaborate.

1.2.4. Neuroscience and New Learning Models

Neuroscientists and cognitive psychologists are only beginning to discover and, in some cases, validate abstract functional architectures of the human mind. However, even the relatively abstract models available from today's measurement techniques (e.g., low fidelity measures of gross neuro-anatomy via indirect measurement of neural activity such as cortical blood flow) promise to provide us with new insight and inspire innovative processing architectures and machine learning strategies.

Machine learning of algorithms using multimedia promises portability across users, domains, and environments. There remain many research opportunities in machine learning applied to multimedia such as on-line learning from one medium to benefit processing in another (e.g., learning new words that appear in newswires to enhance spoken language models for transcription of radio broadcasts). A central challenge will be the rapid learning of explainable and robust systems from noisy, partial, and small amounts of learning material. Community defined evaluations will be essential for progress; the key to this progress will be a shared infrastructure of benchmark tasks with training and test sets to support cross-site performance comparisons. Different users exhibit different patterns when they interact with computer systems. Machine learning algorithms are also needed to recognize such regularities and integrate them into the system, to personalize the system's interactions with its user, and to build up user models. User models may seek to describe (1) the cognitive processes that underlie the user's actions; (2) the differences between the user's skills and expert skills; (3) the user's behavioral patterns or preferences; (4) the user's characteristics.

1.2.5. Big Multimedia Data: Towards more diverse and more complex data

The trend to *big data* and *linked data* in the database and the web communities are mostly due to the additional information derivable from the analysis of a single large set of related data, as compared to heterogeneous smaller sets with the same total amount of data, allowing correlations to be found to spot business trends, prevent diseases, prevent environmental changes, determine quality of research, combat crime, etc. Today's multimedia data are facing the same shift paradigm. First of all, they are produced massively through world-wild continuous and heterogeneous streams. The growing production of social data and UGC data will notably quickly make current technologies fail if no additional research efforts are provided (60 hours of video uploaded every second in Youtube, 3 billion photos uploaded every month in Facebook, etc.). But the *spread* and *diversity* of multimedia data to be integrated, analyzed, and searched has to go beyond the web for future challenging applications at a global scale. Enterprise multimedia data did notably received only few attentions from the research community whereas it involves even more complex challenges in terms of scalability and heterogeneity. Modern science such as medicine, agronomy, bio-informatics, physics or environmental science must also deal with overwhelming amounts of scientific multimedia data produced through medical exams, satellite's earth observation, naturalistic observations, empirical measurements, simulation, etc. Such data must be processed (cleaned, transformed, indexed, analyzed) in order search correlations, draw new conclusions, prove scientific theories and produce knowledge.

But volume and velocity are far from being the only difficulty related to big multimedia data. Heterogeneity and complexity of the data are actually as challenging regarding scalability. More and more modern applications will indeed rely on processing jointly a large number of decentralized and heterogeneous collections of multimedia objects rather than a single huge, homogeneous and well-structured collection of multimedia documents. And each collection comes with its own structure, format, metadata, tags, etc. This heterogeneity makes most current multimedia search technology fail since they require a well-defined data model to ingest raw audio-visual contents. Ad-hoc processing tools can be used to merge a given set of collections but they do not automate the integration of heterogeneous multimedia collections since they require specific additional software developments. The multimedia objects to be managed jointly will notably become increasingly diverse and complex. Primary indexed objects can actually vary a lot from a system to another (products, events, plants, etc.) and each of them can be composed of several images and audio-visual contents (with associated metadata), full text descriptions, geo-tags, author's information, date, ratings, RDF's semantic descriptions and even database records or numerical data in tabular. The additional integration cost to process such heterogeneous data slow down the experimentation and deployment of existing search technologies on large-scale applicative data

(medical multimedia data is a noticeable example). It is therefore time to recognize that heterogeneity and integration issues are part of the multimedia-indexing problem (and not something that will be solved by another community).

1.2.6. Specialized Search: Towards search systems optimized for specific needs

As powerful as a multimedia search engine such as Google Image is, it is useless for lots of practical usage, notably for professionals and scientists. Such players are actually searching for very specific and heterogeneous resources. It is for instance not possible to identify a human disease, a plant species, a media event or a design's trend by submitting an external query image to Google Image. The underlying visual search technology will probably improve in the upcoming years, but it rather shows the limit of the *one size fits all* approach. The impossibility of solving all needs with a unique tool is already confirmed by the proliferation of specialized text-based search engines in the last decade (news search engines, books search engines, games search engines, software search engines, cooking recipes search engines, lyrics search engines, etc.). And Google itself has already a range of different search applications dedicated to different entities. It is therefore likely that multimedia web applications will become specialized as well in order to cover the specific needs of a wide range of areas. Now, one of the big challenges is to develop generic and scalable multimedia search technologies as a foundation of a wide range of specialized applications involving complex and heterogeneous multimedia objects**.**

Following the *Web 3.0* and the *linked data* trends, a first issue towards processing big multimedia data will be to link heterogeneous and decentralized multimedia data in order to create very large knowledge databases. Back to machine learning and concept-based multimedia indexing, it is for instance noticeable that the main brake to the emergence of large-scale applications is today not the scalability of the algorithms but rather the inexistence of large training sets. As a concrete example, building a world-scale plant identification system on hundreds of thousand plant species is today impossible because the required training data does simply not exist. As for many other areas or application fields, a very high number of heterogeneous multimedia collections exist all around the world but there is no technological solution for automatically localizing these resources and extracting the right data from them. Beyond classical integration issues related to heterogeneous data models and formats, a more challenging issue for the multimedia community will be to automatically handle complementary and redundant contents. Real-world multimedia collections particularly suffer from the *long tail*: few objects or concepts are represented by a very large number of multimedia documents whereas the vast majority of objects of interests are represented by very sparse content. Automatically building balanced and large training sets will therefore require

scalable multimedia matching and clustering algorithms.

Advances towards linking large-scale and heterogeneous multimedia data will also be a key issue for knowledge discovery and recommendation applications.

1.2.7. Meeting Reality: Towards large dataset and real-life settings

Current multimedia data mining technologies are mostly mono-modal (images or audio) and very few successful experiments were performed on more than 10,000 items. As a noticeable step forward, recent academic work enables the fully automatic discovery of the most frequent visual objects instantiated in large image collections up to 1M elements. But the recall and complexity of such methods still have to be improved for finding relatively smaller objects (e.g. visual logos or short sounds in large video collections). Dynamicity and multi-modality of the considered data are also still challenging (for instance discovering salient media events based on heterogeneous multimedia streams such as press agencies, blogs, social networks and multi-modal contents).

Yet, the real challenge lies in the ambition to satisfy much more challenging, complex information needs. An example might by a query for radical political art. This might include literature, visual arts, music, performances captured on video, etc. But even for each individual modality this can be further diversified considering that each of them again consist of multiple modalities. Music, for example, contains audio, text, image (e.g. album covers) and video information, which in turn might contain text information. The focus is usually on the primary domain (acoustic for music), combining lyrics information, album covers, social listening behavior and tags, artist biographies combined with geographical and cultural information may be necessary to comprehensively represent an information object. It is not clear up-front, which (combination of) information facets of the chosen target information object (music, in this case) need to be combined to satisfy the information need. Yet, all these information categories are acknowledged as separate research domains or at least as separate problem definitions. As mentioned before, only few research activities transcend categorical borders and combine multiple modalities to approach the semantic gap from different directions.

The predicate 'radical' was intentionally added to the example above. It demonstrates a divergent semantic concept that cannot be captured objectively and has different interpretations depending on the corresponding domain. "Radical music" can be defined by sound intensity, instrumentation and rhythm but also by lyrical content. Radical images and videos can be described by their visual content, as well as video shot sequences. Radicalism is often associated with aggression or highly polarizing and controversial opinions. Consequently, a query for radical political art cannot be efficiently defined on current approaches where each domain or modality requires a distinct problem solving environment

(e.g. Google Image Search).

Moreover, a general approach should be considered where multimodal information is provided as a network of information. Learning how to interlink this network to gain a wide coverage of semantic information is a challenging question. "Aggression", for example, is a subjective interpretation of a highly emotional expression. Almost every research domain in multimedia retrieval maintains an emotion recognition discipline. Thus, putting all this information into the network will provide a comprehensive description of ways of expressing emotions. Narrowing the query down to aggressive rock music or video art from the late 20th century the query definition has to address the proper selection of feature sets and machine learning algorithms that are specialized on detecting the genre/style as well as distinguished emotion recognition feature sets. Using keyword based decision rules is practically impossible due to the endless number of possible semantic combinations. Thus, the aim is to reveal this information intrinsically from the multimodal information provided.

More generally, the increasing complexity of the multimedia objects to be managed will be a major algorithmic challenge. Consider for instance a collection of medical records, each being composed of several images (MRI, PET, etc.), blood tests, full text practitioner's comments, personal information, etc. You may want to cluster these records to find some correlations. Clearly, feature extraction algorithms still need to be specialized for each modality. But on the other side, it will be crucial to converge to generic solutions for structuring, linking and searching the extracted features. Very recent progresses for building efficiently the k nearest neighbor graph of arbitrary objects with arbitrary similarity kernels for instance go in the right direction. More generally, generic and scalable methods for building content-based hyper-graphs might play an important role towards linking big multimedia data. Such an approach has indeed the advantage to model in a uniform way heterogeneous links between complex entities. It would therefore facilitate the search and the mining of the produced linked contents by applying scalable and generic hyper-graphs analysis methods.

Finally, the dynamicity of the developed structures and search algorithms will be a crucial aspect. Handling the high velocity of big multimedia data will actually require insertions and updates operations as a primary problem and not as a secondary one as in many current methods.

1.2.8. User Focus: Towards settings benefitting real users and industry

Industry and academia seem to be in agreement on their common interests, but new technologies often lack the expected acceptance of the intended audience. Due to the absence of appropriate real-world datasets and resources as well as lacking experience in new and evolving domains such as e.g. music search, access

and consumption or video distribution, artificial benchmarks were introduced to kick-start research. Such interpretations of real world problems as music genre, animal or face detection showed tremendous progress in their limited settings. These artificially chosen environments were necessary to gain experience in these complex domains. Yet, questions on the reliability of results achieved remain. Recent approaches where demands and research goals were initiated by the industry (e.g. Netflix price) showed remarkable results that found also wide acceptance by academia, industry and end users.

A need for new and more solidly defined benchmarks that are aligned with real world problems is also reported by recent evaluations. On the one hand a lack of significant progress is reported in retrieval campaigns due to inconsistent and arbitrary use of baselines in evaluations. On the other hand, even when significant improvements of algorithms are reported, this might not be recognized by the end user of such a system. Overfitting of algorithms and models on conventional benchmark datasets to win evaluation campaigns by a significant improvement of 0.01% does not contribute to user satisfaction. It was shown, that big improvements have to be accomplished to achieve an improvement that is recognized by the user and that are not simply artifacts of large data volumes and insufficiently applied concepts of statistical significance testing.

Targeting efforts of both sides towards the requirements, whishes or needs of customers, constitutes a challenging endeavor. The recently published EMI One Million Interview Dataset, which provides one million transcribed interviews EMI conducted in 25 countries, constitutes a step towards this direction. Such information promises high value to research communities' of all fields. It allows for detailed modeling of user profiles or archetypes and is expected to grow in importance in the near future.

1.3. Conclusions and recommendations regarding Media Search

Multimedia analysis is an emerging research area that has received growing attention in the research community over the past decade. Though modeling and indexing techniques for content-based image indexing and retrieval domain have reached reasonable maturity, content-based techniques for multimedia data, particularly those employing spatio-temporal concepts, are at the infancy stage. Content representation through low-level features has been addressed fairly, and there is a growing trend towards bridging the semantic gap. Monomodal approaches have proven successful to a certain level, and more efforts are being put for fusion of multiple media. As visual databases grow bigger with advancements in visual media creation, compaction, and sharing, there is a growing need for storage-efficient and scalable search systems.

Recommendations for the next future are listed below:

Establishing sustainable multimedia research infrastructures

Besides privacy and copyright issues, hardware resources and data management problems prevent many research groups from working on real-world and big data. We advocate for setting up a shared infrastructure at the European level adapted to research on information retrieval and data mining. Such infrastructure should allow hosting large-scale multimedia data as well as services developed by research projects (such as the services that could be evaluated in benchmarking campaigns). This could be done in collaboration with major content providers and owners of big infrastructures in Europe.

Establishing R&D initiatives focused on multimedia search architectures

In order to enable next generation multimedia services and meeting user & industry needs, transversal research with cross-disciplinary use-cases should be fostered. This will require targeting not specific techniques or modalities, but entire value chains from data collection/provision via techniques enabling new services, supporting business models, the consideration of legal aspects, and prototype deployments. Still, any such endeavour needs to be complemented by highly focused research on specific questions without the limitations of immediate commercial deployment, advancing science without limitations stemming from industrial involvement. Striking a balance between these two complementing settings will be necessary to drive R&D in this field.

Ensuring data openness in EU projects

Companies often refuse to share the large data they are using in their scientific publications, sometimes for competitive reasons and sometimes to protect customers' privacy. On the other side, as *big data* is becoming an important research area, this practice is criticized by many researchers for its secrecy and the risks of bad science, potential frauds, etc. A further aspect is the value of data. Industrial partners might be reluctant to join research projects where they are obliged to share valuable data. The problem occurs as well within EU funded projects. Our recommendation is therefore to condition EU funding to some guaranties on data openness, at least for the project's consortium, and possibly to the whole research community (typically through benchmarking campaigns).

2. Social and Mobile Search

### 2.1. Background and state of the art in Social and Mobile Search

2.1.1. Background analysis: actors, business and societal aspects

During the last decade the Information Society witnessed the rapid growth of social networks that emerged as the result of the users' willingness to communicate, socialize, collaborate and share content and the availability of technology and services. The outcome of this massive activity was the generation of a tremendous volume of user contributed resources that have been made available on the Web [1]. Moreover, the fact that users annotate and comment on the content in the form of tags, ratings, preferences, etc., and that these comments are provided on a daily basis, gives this data source an extremely dynamic nature that reflects the evolution of community focus. Combining the behavior preferences and ideas of a massive number of users that are imprinted in collaborative data can result into novel insights and knowledge [2], often called Collective Intelligence. By analyzing such data we manage to acquire a deep understanding of their inner structure, unfold the hidden knowledge and reveal new opportunities for the exploitation of collaborative data [3]. In this complex system the notion of "search" takes a radical new shape, since apart from the traditional dimensions for searching such as textual affinity and visual similarity, there is now a set of new dimensions available such as "friendshipness" (e.g., Facebook's open graph), timestamps, geo-location, tag co-occurrence, collaborative filtering, etc.

At the same time, high-end mobile devices (i.e. smartphones and tablets) have developed into capable computational devices equipped with high-quality color displays, high resolution digital cameras, and real-time hardware-accelerated 3D graphics. Smartphones are able to exchange information over broadband data connections, sense location using GPS, and sense the direction using accelerometers and an electronic compass. According to ITCandor [4], in 2011 the segment of smartphones in the market has reached 20,6% compared to 79,4% of basic phones, but the revenue is distributed as 47,8% for smartphones and 52,2% for basic phones (cf. Figure 2). By 2013, 54% of all phones globally will be smartphones [5]. In addition to smartphones, tablets also form a fast growing market that carries the advantages and challenges of mobile devices. According to Morgan Stanley Blue Paper on "Tablet Landscape Evolution"[17] the tablet is the fastest ramping mobile device in history. Cumulative tablet

---

[17] Tablet Landscape Evolution - Window(s) of Opportunity, Morgan Stanley Research Blue Paper, May 31, 2012

shipments in 2010 and 2011 were more than double the cumulative shipments of any other mobile device in its first two years (cf. Figure 3).
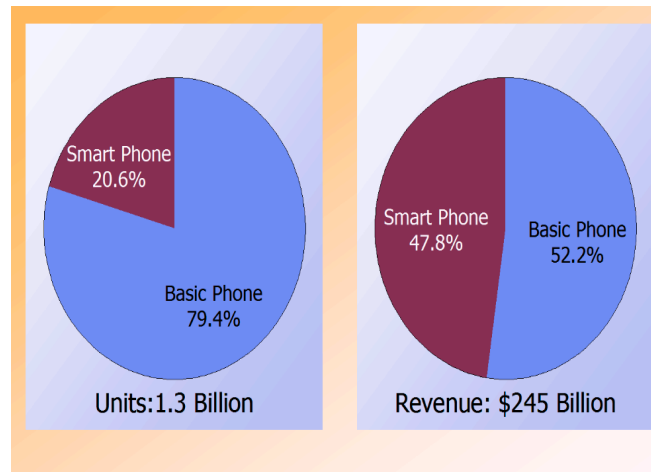
Source and Copyright: ITCandor, 2012

Figure 2- Mobile phone handset revenues ($US Billion) and shipments (Million) by type (April 1st 2011 – March 31st 2012)
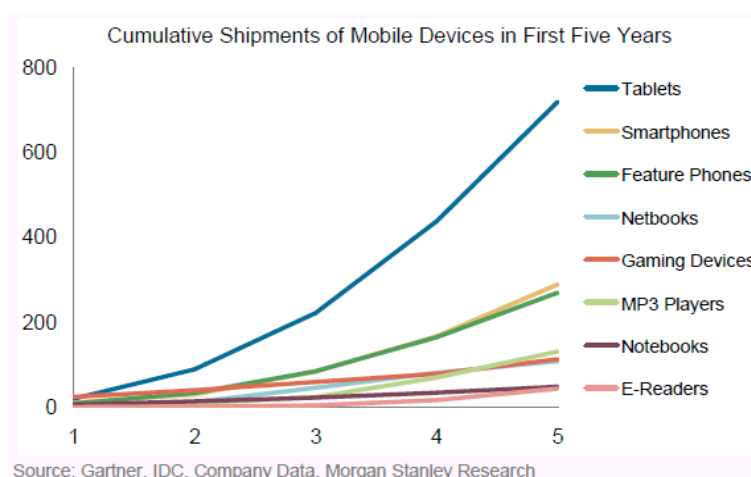
Source: Gartner, IDC, Company Data, Morgan Stanley Research

Figure 3 – Cumulative shipment of mobile devices in first five years of production. Tablets are the fastest ramping mobile device: Cumulative tablet shipments in 2010 and 2011 were more than double the cumulative shipments of any other mobile device in its first two years of production

Powered by user friendly devices users are gradually changing their habits in terms of search. Performics [6][7] predicted that mobile search would soon reach 10 percent of all the search impressions its clients were seeing. In the end of April 2011 the firm said that "mobile impressions accounted for 10.2% of all paid search impressions (desktop + mobile)". In addition, future trends in search interfaces will most likely reflect trends in the use of IT generally [41]. Today, there is a notable trend toward more "natural" user interfaces: pointing with fingers rather than using a mourse, speaking rather than typing, viewing videos rather than reading text, and writing full sentences rather than artificial

keywords. Not surprisingly, people are drawn to interfaces that allow them to think and move in a manner like what they use in their non-computing lives, but only recently has technology been able to support this. What is even more interesting is that innovations that originate from mobile search are getting adopted in traditional search. All the above clearly show signs that mobile search is moving mainstream and is gaining momentum. By considering this fact in conjunction with the constantly evolving landscape of social media we can claim that search technologies take radically different characteristics, both from the perspective of coping with the unstructured and noise nature of social media, as well as from the perspective of exploiting the new opportunities offered by mobile devices (smartphones and tablets) to deliver services of higher quality.

2.1.2. Current technical state of the art and R&D mainstreams

With respect to social media, Web 2.0 has effectively turned ordinary people into active members of the Web. Web users act as co-developers and their actions and collaborations with one another have added a new social dimension on Web data. User generated content can be viewed as a rich multi-modal source of information including attributes such as time, favorites and social connections. Current trends [8] in social media exploitation aim to facilitate a wide range of Information and Communication Technologies (ICT) by integrating research and development in: a) Media Intelligence: which refers to the intelligence originated from digital content items (images, video, audio, text), their contextual information and their merging. b) Mass intelligence: which analyzes user feedback on a large scale (i.e. information clustering and ranking) to enable potential events detection, trend analysis and prediction. c) Social Intelligence: which is the exploitation of information about the social relations between members of a community, so as to capture the social context and provide more advanced services.

On the other hand, current trends in mobile search are primarily related to innovative interfaces and location-based services [9]. For instance, by entering a word or phrase on their phone users are able to find a hotel in their vicinity after a tiring journey or call a taxi after a night out. In addition to the location based services several R&D efforts are trying to cope with the fact that mobile search involves a non-trivial processing/communication trade-off [10]. Even when the terminal is connected through a broadband mobile communication link, the possibility of transmitting huge amounts of media in upstream is usually not granted. On the other side, the on-board processing capabilities are limited by the nature of the processors, the memory, as well as by clock and power limitations. All these facts create strict limitations on the architectures that can be adopted and depend on the most acceptable compromise [10].

2.2. Future trends and grand challenges in Social and Mobile Search

2.2.1. Innovative interfaces for mobile search

The most recent trends in innovative interfaces typically originate from mobile search rather than the other way around, because the interface constraints for these types of devices are more important. The number of devices we use for mobile search has increased enormously. However, the properties of these devices are different and we need to understand better what devices are used in what context and how to optimize the interface to complete certain tasks. The mobile search domain moves rapidly and in order to keep up with the pace we need to identify the principal tendencies of mobile search, such as: mobile search as part of a larger group, and mobile search as something distinctive. It is important that we understand how these aspects develop in order to make forecasts about the growth of this tendency. Mobile search can be a component of an intelligent mobile system, such as car navigation systems, augmented reality applications or toys. Such systems make intensive use of search although it may not be directly apparent to the end user.

Personal and contextual information is another very important element of innovative search interfaces in a mobile setting. Indeed, context (e.g. location) is particularly important for media search. This is because in order for mobile visual search to be efficient enough, vertical search engines (i.e. engines that search a specific industry topic, type of content such as travel, movies, images, blogs, live events, painting, piece of data, geographical location, and so on) will have to be employed. In this case, automatic context (and therefore engine) selection could greatly improve the obtained results by narrowing the search space. Thus, in the extreme case, the user doesn't really have to search; the results are pushed to him according to the constantly monitored context, e.g. Google Now[18].

When the internet arrived in 1995, this completely changed the way we thought about search. With the arrival of cloud, mobile internet and the availability of personal data, we are probably at a similar point now. Traditional methods of doing search may no longer be appropriate and we have to completely rethink how we perform search. Innovative interfaces for mobile devices are expected to stay at the forefront of this endeavor.

2.2.2. Smart-phones and social media as sensors for reality mining

Reality Mining defines the collection of machine-sensed environmental data pertaining to human social behavior [11]. This new paradigm of data mining makes possible the modeling of conversation context, proximity sensing, and spatio-temporal location throughout large communities of individuals. Thus, the

---

[18] http://www.google.com/landing/now/

goal of this research trend is to collect and aggregate the data streams generated from smartphone installed sensors (e.g. physiological measures, GPS tracks, proximity, location, and activity information) to gain insight into the dynamics of both individual and group behavior [12]. Such dynamics can help us predict what a single user will do next, model the behavior of large organizations, detect trends, spot emerging phenomena and events, mine public opinion, etc. What is particularly interesting in this case is the fact that mobile users can act as the living sensors of their city, its governmental and social aspects, its cultural events or even its mobility status by doing no more than their everyday activities. Following the principles of un-obtrusive and ubiquitous computing the reality mining trend promises a plentiful and particularly rich content allowing to grasp the actual footprint of a living society almost in real-time.

In achieving the ambitious objective of reality mining there are two main obstacles that need to be addressed. The first has to do with capturing and processing the large volume of environmental data that are expected to grow as the number of living sensors increases. The number of smartphone users is growing very rapidly and the reliable and unobtrusive access to the generated volume of data poses strict requirements on the storage and processing infrastructure. The other important obstacle relates to ethics, since data privacy and protection should be seriously taken into consideration when capturing the user's environmental and activity context. It is very important that all necessary measures should be put into force in order to protect users' rights and prevent data misuse by the collector.

2.2.3. Mobile search for linking between the physical and digital world

During the last years the mobile content has changed its media direction from keyword-based to visual-based search [13]. In this case the user is offered the functionality to capture a photo in his/her cell phone and find relevant information on the Web. Visual-based search works like traditional search engines but without having to type any text or go through complicated menus. Instead, users simply turn their phone camera towards the item of interest and the mobile search engine returns relevant content based on its interpretation of the user's visual query. Thus, visual-based search becomes the natural way to link between the physical and digital worlds by making the visual surroundings of the user "searchable" and objects in visual proximity "clickable". If we consider this in combination with the phone's sensors supporting location-based services and with the fact that mobile phones are always on and always with us, we understand that the range of services that can be built upon these functionalities could completely restructure the way users consume multimedia content on the web.

Towards this goal, considerable research effort is being invested not only on improving the results of visual recognition but also on the architecture that will

facilitate this functionality in real-time. Ranging from server-based approaches where the service is completely provided by a remote server, to client-based approaches where almost the entire processing is performed on-board and hybrid approaches that try to combine the advantages of both, the architecture for mobile image search has become the subject of very intensive research.

2.2.4. Building innovative services exploiting the added-value of social media

Ranging from politics and marketing to economics, tourism and news the potential to collect, aggregate and process large volumes of user-generated content has prompted many developers to design and implement innovative solutions. When considered in combination with smartphones that offer the means to generate and share content on the fly, we understand the great deal of opportunities arising in this new market. However, current approaches are mostly ad-hoc and fail to describe a complete ecosystem that starts from how to generate a sufficient user-base all the way to capitalizing the user's loyalty for the benefit of the enterprise or the organization. In bridging this gap substantial effort is being invested both towards the development of the necessary infrastructure for capturing the related information and performing data mining to obtain the desired knowledge, as well as towards the direction of building innovative business models that could turn this knowledge into direct profit [14].

There are many application fields where the added value of social media can be exploited to offer a competitive advantage of direct profit. In personalization, social-based indexing and user profiles can incorporate information about the users' social network directly into the personalization, search and ranking process. In tourism and culture, uploaded media can reveal "off-the-beaten-path" points of interest and events, otherwise difficult to discover through usual Web sources [15]. In politics, web data may contain the most up-to-date information that can reflect the popularity of elections candidates and their support among populations. In economics, marketing and brand monitoring, the number of related media uploaded online can reflect the number and locations of products sold in the market. It is evident from the above that companies and SMEs have now at their disposal a new, rich source of valuable source of data that can be used to evaluate their strategic decisions, design new plans and establish novel communication channels with their customers.

2.2.5. Enterprises adjusting to the 'new normal' of mobile

Based on the results from the European biannual sample of IDC's 2012 EMEA Enterprise Mobility CIO Survey, fielded in April/May of 2012[19] that was posted

---

[19]http://www.alacrastore.com/storecontent/IDC-Western_Europe_Enterprise_Mobility_Biannual_Survey_1H12-2107-6875

by Nicholas McQuire, Research Director, Enterprise Mobility EMEA, IDC, we can draw the following conclusions about enterprise mobility:

- European enterprises expect mobile working to accelerate in the next 12 months, spurred on by greater flexibility, customer service and productivity benefits.
- 35 percent of those surveyed believe their mobile workforce will grow over the next year, with 15 percent believing it will grow more than 10 percent, according to the results.
- 25 percent of those surveyed said they will spend more on mobile over the next 12 months as a percentage of their IT budgets than they did in 2011. At the same time, close to half of those surveyed (48 percent) do not yet have a mobility strategy in place for their business.
- While 10 percent plan to create one in the next 12 months, the results indicate that despite an increase in mobile working and spending, most European firms including SMBs, German, Italian and Spanish enterprises in particular, largely view mobility tactically within their organizations.

Cost is a growing barrier to many first generation mobile solution deployments

When asked "what are the biggest mobile deployments issues faced in the past 12 months?" "Cost Overrun" was listed as the top problem faced by 34 percent of respondents (and almost half of all French CIOs). This beats security and compliance problems (29 percent), complexity of management and support (27 percent) and challenges with database integration (26 percent).

Most firms are currently immature but they are on the process of evaluating mobile apps

Interest in security and management is contrasted by immaturity regarding mobile applications. According to the results, while close to 40 percent of the survey are currently evaluating their mobile application strategies, just 16 percent have launched their first mobile app and 21 have no plans to develop or deploy applications at all in the next 12 months. When asked those evaluating or deploying mobile applications "how many mobile OS platforms do they plan to write to?", 32 percent didn't know, the highest selection. Still, 60 percent do plan to invest in developer resource to deploy apps in the next 12-18 months, 40 percent with their own IT staff according to the results, suggesting that wider scale deployment of mobile applications is imminent in 2013. The types of apps listed as highest priority by respondents included office apps, file management and collaboration, field service and unified communications.

The market activity during the past months is indicative of the "new normal" in mobility which many European enterprises are struggling to adapt to

There is a lot of work taking place around converting enterprise search applications to be of value to users of mobile devices when accessing enterprise content, with some indicative examples including ISYS-Search[20] and IDOLme[21]. This requirement is also driving the development of cloud-based search solutions[22] like CloudMagic[23]. However, the results show us that while mobility is accelerating, so are its costs (and risks) and that many European firms, operating in a tough macro-economic climate, haven't yet made the investments in critical mobile technologies such as mobility management, security and mobile applications. The survey also reveals, perhaps most importantly, that most firms don't yet fully understand how mobility can strategically drive their business goals and how a mobility strategy is necessary to achieve their aims.

2.2.6. Convergences and divergences between mobile search and web search. Implications of a potential divergence and possible recommendations

What search for mobile search?

Applications on mobile devices relying upon search technology can be grouped into a) those that adapt or emulate existing web search services to the mobile environment and b) those that also exploit the unique features of mobile devices or the environment in which they operate. In the first 'search-as-usual' case a), mobile search applications evolve by migrating established PC-based Internet search tools to a mobile environment; possibly complementing it with some "mobility" enhancement function such as refining the results taking into account the user's location. Such migration can pose severe technical challenges, mainly on enhancing the efficiency to retrieve relevant content in all digital formats, particularly in an audiovisual context, and on responding as accurately as possible to natural language queries. In fact, web search engines enjoy a unique position as an entry point for end-users to retrieve, subscribe and use content and applications. Provided they succeed in producing the necessary adaptations to the mobile environment, steering mobile platforms or connecting them to their existing portfolio of services in the wired Internet is their ultimate goal.

Web/mobile search divergence #1: intensive use of contextual information

However, in authors' view, mobile search will increasingly turn more sensitive to a logic –case b)- driven fundamentally by personalized and context-based

---

[20]http://www.isys-search.com/products/enterprise-server
[21]http://www.autonomy.com/content/Products/idol-modules-idol-me/index.en.html
[22]http://wwwen.zte.com.cn/endata/magazine/ztecommunications/2011Year/no1/articles/201103/t20110318_224551.html
[23]https://cloudmagic.com/

services offering a search experience well beyond of today´s wired Internet world. Mobile search will make effective use of contextual information (relevant data embedded in the mobile device, information in the surrounding environment, users' profiles or behavioral patterns) to improve the relevance of search results and/or to provide a more valuable and entertaining user experience as the "engine within" new types of applications. Context-relevant information is typically derived from sensors (both users' bio-parameters and their physical environment) and from cognitive technologies. It is further expected that the use of context opens undiscovered needs and interactions. For instance, as mobile devices have rich sensing capabilities, they allow augmenting the real world commons with the Internet. Mobiles will gradually turn into the natural device to bridge the physical world surrounding us with the wealth of information on the net, whereby search engines operate in the background to provide the link between both worlds.

Web/mobile search divergence #2: search as a building block of mobile applications

Following the promising mobile market prospects, many actors are seeking to develop the next 'big' search application. Nevertheless, how to gain advantage from existing (and future) mobile technologies and where to position themselves (in a naturally intermediary application) to enjoy sufficient revenues are far from obvious problems. A major reason is that search applications are neither simple nor autonomous building blocks. Rather, search functionalities are tightly embedded into the value chain of wider mobile services, which themselves can be numerous and of diverse natures. In techno-economic terms, we could portray the search functionality as a key constituent in an "ecosystem" where industrial players compete and/or collaborate to generate successful and scalable mobile applications.

Web/mobile search divergence #3: advertising based on keyword bidding not the only business model

The real challenge in mobile search is how to monetize it. Advertising seems the intuitive response when comparing with current Internet search, particularly for class a) mobile search applications. For class b) applications two factors will influence the business model: first, the feasibility to monetize the added value provided by mobile search within a given applications, and, second, the economic value of the search functionality with regard to the totality of the value chain of the service. It is the combination of these two elements that determines the sustainability of new types of mobile search ventures.

Following the success in the wired Internet, advertising is the most obvious candidate for a business model also in mobile search. Search results are provided free-of-charge to final users and revenues are generated from third-party advertisers. Advertising models come in various facets and business tactics. Traditionally, banner ads on search results (sponsored links) where included in

the results, usually including a direct response method as well (a link to a micro-site, a click-to-call link, or a short code). Other options are off-portal campaigns for specific services, such as travel, restaurants, automotive, or consumer electronics, or click-to-call text links connected to search results as a simple way to leverage the voice capabilities of mobile devices. Off-portal keyword bidding, especially for marketers offering digital content is another example. Advertising is considered by many as the cash cow in the deployment of advanced mobile applications. But evidence on how exactly the revenue stream will be generated is still unclear, rather it seems that many firms bet on advertising only because alternative concepts to render mobile applications profitable are even more vague. In spite of the diffuse picture advertisers perceive that "business-as-usual" ads are not going to deliver the expected impact and sustainability. Therefore, the focus is shifting towards more sophisticated approaches, like targeted product placement or dire personalization which also obeys to offer a better approach to an increasingly fragmented audience across media. Another relevant trend is advertising within the mobile application itself, rather than guiding users via the browser to sponsored links. This approach would allow including search capabilities into a number of mobile applications while keeping the advertising business model.

However it is also possible to generate revenues from final users simply based upon the premise that the value offered is attractive enough for users to pay for the search service. This includes pay-as-you-go (pay as-you-need) when there is a high need or urgency for particular search results (examples include necessary search services in unknown environments). Also premium services could provide more value than their free basic instances (this added value could be in the interface, the range of inputs and/or results, or in additional information provided in the results to the query, for instance a telephone number ready to be called to).

In addition, value-added services offer an improved version requiring an additional contract for particular services, like niche applications or professional services (for instance, search of audiovisual content from an image taken from a mobile device in a health service). In the subscription model, the search service is supplied for a specific period of time (a month, a year) or in specific circumstances (a travel, a city, etc).

Due to the success of application stores, business models which could step on their particular features are also on the vogue. Beyond the traditional models above, value-added applications, i.e. applications downloaded from an application store and from which -during its use- new functionalities can be accessed, may incorporate time-based billing for subscriptions to services, event-based billing (getting some results from a search service) or even item-based billing (e.g. payment as a function of the usefulness of the results of a query).

This type of business model is expected to become a relevant part of the overall revenues of the mobile content and applications market.

There is also the possibility of packaging mobile search with some other products or services. This business model appears when search applications are considered just an additional commodity for other products and services; its usefulness of search being perceived as low added-value. In this case, the revenue for search-related parties would be basically limited to developing and licensing 'white label' search applications or a specific search application to be embedded in other product or service. Typical scenarios are packaging the search application with the mobile device for suppliers or telecom operator services (for on-portal or on-device search, for instance) or with a non-mobile product or service (as part of a holiday package). As search would merely constitute an add-on to existing products or services, users would unlikely be willing to pay separately for search functionalities, rather they would be charged for the complete package. Autonomous revenues obtained through search (through any of the business models previously considered) would typically go to the outsourcing player rather than the outsourced search developer.

In addition, complementary business models to the above could be considered that provide indirect revenue source with regard to the search process; the most important being those exploiting user profile derived from the queries and results for marketing purposes. Their implementation is delicate given legal restrictions and growing privacy concerns by policy makers and users. Anonymising data to prevent tracing back to individuals and groups is technically viable nevertheless the value of information diminishes as it gets more generic and less individualized. Striking the right balance between privacy, personalization services and sustainable business model remains a foremost challenge.

Web/mobile search divergence #4: innovations from the mobile apps economy

In both cases a) and b), techno-economic factors pushing mobile Internet are obvious direct enablers. These include the (still) growing mobile penetration, the increasing mobile broadband availability and affordability, the improvement in usability and affordability of smartphones, and the availability of useful mobile content and applications. Obviously, the more the mobile Internet grows, the more mobile search intensifies. The combination of a rapidly growing market with fierce competition among applications will bring many innovations.

However, the app economy has a serious contender in the –mobile- web browser. If mobile applications go back to be provided through a browser, as typically happens in the wired web, it is probable that an extension of the existing approaches in the conventional Internet will be the winner.

2.2.7. Grand challenges

Both mobile and social search are two rapidly evolving areas, facing a series of challenges before mainstream adoption. In the following we discuss some of these challenges and indicate the level of consensus achieved among the certified experts.

2.2.7.1. Mobile Search

With respect to *mobile search* the emphasis is primarily placed on technological future trends and directions. Some of the key questions that typically arise are: a) How is mobile information needs changing? b) How are mobile search usage patterns changing? c) What are the main technological and economic challenges ahead? d) Which are the major bottlenecks? e) How is the mobile search market likely to evolve? and f) How is Europe placed with regard to the rest of the world? Based on these questions some of the grand challenges that we have been identified for Mobile Search are:

Mobile Search: what is it about?

There is some consensus concerning the fact that the mobile device (at least in the developed world) is felt by its user as the most personal and confidential piece of technology. Beyond being the most personal device, the mobile is also always on and always with the user. Nevertheless, the distinctiveness of mobile use is hard to model, therefore, a generic definition of what mobile specific is cannot be given, without the risk of being reductionist.

Technology: is it a bottleneck?

According to [18] the main technological bricks enabling mobile search are considered already available. However, this is debatable by many experts in the field, who believe that the main challenge for the full deployment of mobile search services is still of technological nature. The absence of a Europe-wide ubiquitous Wi-Fi infrastructure is also a barrier to the deployment of mobile services.

Privacy: does it constitute a barrier for the development of mobile search services?

Given that relevance is individual-specific, personal data gathering is crucial to provide filtered results that are relevant to the individual. However, there are many privacy issues that are raised. The availability of personal information for service provision and the related privacy concerns are considered as one of the critical issues for the deployment of mobile search. Nevertheless, there is still no clear consensus for possible mitigation actions, since "there is not much you can do against people disclosing their personal information"[24].

---

[24]Round-table discussion during the Chorus+ Workshop on "Exploring the Future of Mobile Search"

Search Service: how can value be monetized?

According to [18] the majority of experts identify the major challenges for Mobile Search to be in the economics arena. It is generally accepted that one of the main bottlenecks for the development of EU search market is represented by the absence of data flat rates and data roaming flat rates (or at least of predictable rates). Today's technology enables a paradigm shift in product/service promotion leading to a model where customer engagement is the main key to success. However solutions have to be tailored to the type of service as well as to the target audience. It is unlikely that a single winning revenue model will emerge fitting all business models.

2.2.7.2.   Social Search

With respect to *social search* the scientific interest usually evolve around the following subjects: a) Which search tools are needed for the social networks? b) Do we envision new tools and services to emerge anytime soon? c) Do we envision new applications and services to emerge from the combination of automatic information retrieval and social tagging and comments from the social networks? d) What is the future for real time trend and opinion analysis? Motivated by those questions some of the challenges that we can identify concern:

- The importance of the "Social Search" theme being debated, indicating that although social media is seen as the main enabler of advanced information services, several threads and challenges exists, including spam and lack of objectivity.
- Social media should facilitate access to relevant information.
- Specialization towards a reduced set of specific key areas such as for instance Healthcare and Education is foreseen.
- The application of social context and social media in professional environments is often cited.
- In addition to the statistical and linguistic methods, several innovative experiences are also reported such as analysis of social relationships in photo albums, use of real time sensors, etc.

With only some years of existence social networking is still in its infancy and is expected to bring the biggest evolution in search technologies since 1990, simply because dramatic changes are taking place in the Web topology.

2.3. Conclusions and recommendations regarding Social and Mobile Search

Mobile and social search exhibit a very high level of penetration in IT users both for the purpose of business and fun. What is common in both cases is that the high amount of potentials promised by these two rapidly evolving search paradigms is coupled with a great number of challenges that differentiate them from the typical cases of desktop and web search. Towards addressing these

challenges and overcoming the existing obstacles we have the following recommendations:

- **Establish a Europe-wide policy for handling private information**: Given the nature of mobile and social search the vast majority of potential services involve the storage and processing of user's private information. The fragmentation of existing policies on handling private data and the absence of a clear Europe-wide directive on this issue makes difficult the exploitation of the data that are generated during mobile and social search. As a consequence users are reluctant to hand-over their personal information while companies and research organizations face many difficulties in collecting them at the appropriate scale.

- **Consulting services on how the value of social and mobile search can be monetized by the existing companies**: As became apparent in the description of the aforementioned topics, it is not always obvious for the companies to develop innovative solutions that exploit the added-value of social media, as well as for the enterprises to adopt their working model in the new norm of mobility. Support should be provided both in terms of financial resources, as well as targeted expertise that will help existing companies to understand and exploit the real benefit of user contributed content, as well as to leverage the great potentials offered by mobile devices.

- **Encourage research on large-scale social media and data analytics**: The fact that users currently annotate and comment in Social Networks through tags, ratings, social connections and preferences, and that these activities are performed on a regular basis, gives social media data an extremely dynamic nature that reflects topics of interests, events, and the evolution of community opinions. However, current technologies have largely focused on enabling the production of large volumes of media and piecemeal consumption of tweets, images, or songs. In contrast, the utilization of aggregated collections of such media and the combination of their multimedia, spatio-temporal, and social context provide the ingredients for deep understanding of events, patterns, and situations emerging from data. The need for aggregation of different sources and for the combination of their rich context requires scalable and multi-modal approaches that are able to handle the massive amount of available data (big data) and transform Social Networks to inference engines of topics, events, and ultimately facilitate planning, prediction and action.

- **Bring user experience to the forefront of research**: Since the evolution of touch displays into the main handling instrument of mobile devices, user experience has gained a leading role in the success of every product. This fact is also true for the web based applications that deal with user generated content. This important dimension has been, to some extent, overlooked by the EC. However, in order to develop successful services

for mobile and social search, the user interface should be as intriguing as possible offering a unique experience to the user.

- **Ensure the openness of data generated within Social Networks**: Social context is most probably the major differentiating factor between social and web search. Additional dimensions like "friendshipness" (e.g., Facebook's open graph), timestamps, geo-location, tag co-occurrence, etc., are now offering new opportunities for smarter and more personalized search. However, despite the existence of Application Programming Interfaces (APIs) that allows you to get access to some fraction of the collected data, most of the major players in the field (e.g., Facebook, Google, Yahoo) pose some limitations on the use of the content collected from their users. An indicative example of this attitude are the changes that were recently introduced in the new Twitter API that severely restrict how much and how often third-party Twitter clients and other services can access Twitter's information [19]. Indeed, these companies often refuse to make such information public, sometimes for competitive reasons and sometimes to protect customers' privacy. But to many scientists, this practice is an invitation to bad science, secrecy or even fraud[25]. Unless there is significant pressure on these companies to make their data public, the potentials for research and innovation will be limited only to the research labs of these players, excluding all other players from industry and academia.

- **Ensure that Intellectual Property Rights are not hindering innovation**: Large industrial players gather huge patent portfolio in this area. Although patents are meant to protect innovation, in practice they are often used as a legal tool to limit competition in this market. A European entrepreneur that combines several technologies in an innovative way does not stand a chance because he will be attacked on patent infringement.

3. Enterprise Search

3.1. Background and state of the art in Enterprise Search

3.1.1. The Enterprise Search business: current state of the art

The term *enterprise search* (ES) is commonly used to describe the application of search technology enabling information retrieval within organizations. Therefore, as the practice of identifying and enabling specific content across the enterprise to be indexed, searched and displayed to authorized users, it should be differentiated from other type of search, such as web search.

It includes the search of the organization's intranets and other electronic text held by the organization, such as emails, database record and documents on file sharing, often referred to as 'unstructured information' and it's external web sites.

In particular, ES encompasses two sub-domains:
- Search within the enterprise private network; For this sub-domain, the "customers" are the enterprise employees. The main issues in this regard are: high expectation (precision/recall), diversity of sources (formats/repositories, etc), security (must be built in beforehand).
- Search as a component of a larger application developed by the enterprise; In this context, the client can be either the enterprise employee (internal applications), or the general public (public facing services such as e-commerce, yellow pages, etc.). The main issues in this regard are: architecture, flexibility, development skills

Being a technology that retrieves information within organizations, Enterprise search solutions render business processes more efficient. Enterprise search is therefore a key element in increasing the competitiveness of the digital economy and as such constitutes a strategic market for the European Union.

Motivators for the development of an enterprise search market, as emerged from the workshop "Exploring the future of Enterprise Search", organised by the IPTS in the framework of the CHORUS+ project and held in Seville in October 2011[26] are:
- There is increasing information everywhere; more than 200 billion of emails per day; 80% of the enterprise information is unstructured.

---

[26]http://is.jrc.ec.europa.eu/pages/TFS/search/workshop2011.html

- The digital data growth is enormous, expected to be in the order of 35 zettabytes in 10 years. In particular, 94% of organizations are collecting and managing more business data than few years ago and there is an increase of 86% in business information collected/managed in the last few years [20]. In addition, business data nowadays comes in a variety of types, media, and formats. The cost of poor data management: organizations are seemingly losing revenue each year (on average, 14%) as a result of not being able to fully leverage the information they collect. That translates to circa $130 million in lost opportunity each year for a $1 billion organization[27].
- Legal compliance of the enterprise: obligation to store and find all enterprise documents, business communications for legal reasons Enterprise data is all over the place. ES has to federate all the information existing in both structured data (databases) and unstructured data (text, reports, mail).

In other words, if one reason for adopting ES is the growth in digital data generation, more worrying than the huge amount of information is its structure or rather the lack of it - it is estimated that about 80% of the information stored is either unstructured (a figure widely deployed by Autonomy) or has no adequate metadata for the needs of employees. A recent survey (2012) carried out by MarkLogic[28] suggests that the figure of 'unstructured' information used by enterprises is lower than 80% but still significant. In another survey by AIIM[29] in 2012 over 60% of respondents answered that they would find it very useful if they could correlate text-based data with transaction data, but only 2% are able to do so at present.

World-wide there are probably no more than 200 companies providing enterprise search software. A forthcoming IPTS report [45] provides probably the most comprehensive and up-to-date listing of ES providers and analyzes the most significant ones. The report lists some 60 enterprise search vendors which represent about 90% of enterprise search software sales. Amongst all providers examined in the report, three different groups of companies can be differentiated:

- The top five vendors, all of them multinational IT companies (*Autonomy/HP*, *Google*, *Microsoft*, *IBM* and *Oracle*), play a major role in the market strategy and thus have a major impact on the development of search technology. These can be called *Type 1* vendors, the few ones that

---

[27]http://is.jrc.ec.europa.eu/pages/TFS/search/workshop2011.html

[28]'Big Data is Real and It is Here', MarkLogic 2012, http://info.marklogic.com/post-relational-reality-dbta-survey-2012.html.

[29]Big Data – Extracting value from your digital landfill – AIIM, 2012, http://www.aiim.org/Research-and-Publications/Research/Industry-Watch/Big-Data-2012

sell enterprise search as part of an overall enterprise application suite (rather as standalone products).

- The next set of companies (*Type 2*) comprises highly dynamic mid-sized (in terms of software) companies that are growing and trying to find their place in the very competitive ES market. These are companies exploiting unique proprietary software and have (often) received funds by venture capital and private equity placements to improve their software solution(s) and to promote their products. Examples of type 2 companies headquartered in Europe include *Exalead/Dassault* (France), *Fabasoft* (Austria) and *Sinequa* (France).
- The last group of companies (*Type 3*) includes small enterprises that specialize in specific niches. Some of these companies use Apache/Solr or some other open source application as the basis of their enterprise search offering, with *Intrafind* (Germany) being an example. Other companies that provide add-ons for ES applications, like the Helsinki-based *Documill* are also part of this group.

3.1.2. The European market for Enterprise Search: a SWOT analysis

Although European enterprises are beginning to realize the importance of ES for improving the internal work process and the efficiency of the company in general, this does not result yet into making a straightforward business case for ES. This paradox arises from the observation that – although the ES's potential is of value to most, if not all, employees – no single enterprise department wishes to take responsibility for making a business case. The growth potential for ES is good but the market remains largely unexplored.

Albeit the general consensus on the strong potential of ES amongst experts, the low market size seems to indicate significant barriers for enterprise search to move mainstream.

Unfortunately there are neither reliable data nor revenue analyses for the EU enterprise search market; therefore, extrapolating the market potential for the EU enterprise search is not easy due to the lack of solid data and a rigorous statistical basis. However, we can assume, by correlating different sources, that the market potential for ES solutions is still considerable, particularly for mid-sized companies that could benefit from using ES solutions [45]. Such companies would increase their ability to manage (i.e. to discover, reuse, modify, extract and combine) all kinds of information assets rendering them more efficient and thus more competitive.

The fact that enterprise search solutions are poorly deployed in European enterprises is not good news. Several studies show the economic value of rapid and reliable access to internal and external company information. This value is both in terms of opportunity loss (e.g. time spend in searching rather than doing

alternative more productive work) and of added value (e.g. identifying useful relationships – for example to make a better offer to clients, detecting inconsistencies in internal data, etc.). A higher use of search applications and other state-of-the-art business processes tools is likely to result into a higher productivity of individual companies and to increase the competitiveness of whole sectors of the European economy.

3.2. Major challenges and future trends in Enterprise Search

3.2.1. Major challenges

A major challenge for enterprise search solutions is, first of all, providing an integrated and value-adding retrieval of both structured and unstructured information and merging them into a sort of 'Hybrid Structured data'. The vision is thus towards a Unified Information Access, providing end-user with a user-friendly interface retrieving heterogeneous data sources and providing added value making use of semantic modeling.

A second challenge is organizational and related to the market dynamics. Compared to web search, the value chain of the enterprise search business is more complex. In addition, established commercial products are facing the pressure from open-source software which is catching up quickly in terms of functionality and coping well with 'big data'.

From the perspective of a client organization, we can distinguish several ways in which an enterprise search application can be procured and implemented. In the realm of commercial products, ES can be provided by a) a system integrator, b) a specialist integrator, c) a commercial software company (which may also provides services support for the implementation) and d) an application developer. In the case of open source products such ES can be provided either by a) system or b) specialist integrator. Future assumptions on the development of the enterprise search business should take into account that the business models for the above cases differ.

Another major challenge and a reason for the slow adoption of ES is the skill gap and the low investment in ES training programs. Experts consider a lack of trained people in Europe in the field of ES as a major problem and they point out the very low number of specific academic programs on ES offered by European universities. Though there is a general agreement amongst stakeholders on the skill gap in ES, opinions diverge on what kind of teaching and skills should be provided by academia to suit the industry's needs. While some experts consider training "information retrieval specialists" in fully fledged master courses on ES to be the best option, others consider training IT experts on the particularities of search related applications in dedicated courses to be sufficient. In the USA and

other parts of the world the training of 'data scientists' has become a priority for many educational and other institutions.

Another bottleneck for the technological evolution of ES is the knowledge transfer from academic research to commercial products. While relevant European research projects in ES are not missing, there is limited support and efforts to fill in the transfer gap between research and industry. A closer relationship between the two would also favor a better strategic approach in research in the area. Experts observe that – if further research is needed on enterprise search solutions – it is not always clear to the industry and professionals in which directions it should be directed (i.e. novel approaches to tackle the semantic gap, interfaces between modeling blocks, data-sets and collections for benchmarking, linguistic filters, etc.).

Another main challenge the ES market has to face is: how to define a stable and open architecture over which a European eco-system could develop and mature? How a component technology testing could be performed (see benchmark section), and which would facilitate technology transfer to industry (see technology transfer section)?

There are only a few venues and forums where ES professionals can meet and discuss major challenges, progress in the field, and present their newest products. These include the Enterprise Search Summit in New York and the Enterprise Search Europe conference in London, as well as several academic conferences on information retrieval. Chorus+ activities like the Think Tanks or the specific thematic workshops, are amongst the few events[30] trying to fill in this gap and provide such a forum. Increasing the awareness about ES and the use and usefulness of ES solutions will certainly help its wider adoption.

3.2.2. Multimedia and Enterprise Search

Within the Enterprise Search domain, multimedia does not play yet a major role. The main reasons for this are: a) multimedia data is not yet widely used within the enterprise; and b) current technologies for MM Search are not yet performing well enough to trigger the development of widely accepted applications and services.

However, the situation seems to be changing, with more and more enterprises collecting and using multimedia data and considering it as enterprise data. Some niche markets where MM Search is beginning to be used in the enterprise context include photo repositories, medical data repositories, etc.

---

[30]http://avmediasearch.eu/

Improved technologies for automatic creation of metadata (tagging, geo-location, etc) for audio, images and video owned by enterprises on a large scale will soon add new even larger volumes of data to the ES world.

3.2.3. Future trends in Enterprise Search

According to a Delphi-type study conducted in 2011 by Intranet Focus Ltd. in collaboration with JRC-IPTS [45], the technologies that are regarded by experts as most important to the demand and adoption of ES are Search Based Applications (SBA) and integrated search platform (unified access platform). At the Enterprise Search Europe 2012 conference there was considerable unanimity that some of the main factors influencing recent developments in ES were Big Data, mobile search and cloud services[31].

Despite the need to improve technological building blocks, it is not the absence of effective technologies that prevents ES from wider deployment and use. Stakeholders and experts do not consider a lack of specific ES technologies to be the limiting factor for the take-up of enterprise search.

This does not mean that there are no technological challenges. On the contrary, improvements in re-use of components or better interoperability are examples of missing elements that would be beneficial for the development in the short to medium term. In the long term, the growth of semantic web applications, the 'Internet of things' and cloud computing will all impact on ES, both in terms of new applications and in terms of innovative business models to be developed.

One notable challenge could derive from the fast increasing use of mobile devices and in particular tablets, as the forecasts for shipments are of over 100 million in 2012. By 2016 there are forecast to be close to 1 billion tablet devices in use. A substantial proportion of this use will be for enterprise applications[32] and this will have an impact on enterprise search application development. Finally, visual enterprise search, though still quite limited and applied for specific scopes, is increasingly demanded, as enterprises, as already mentioned, have more and more multimedia corporate assets.

A common future trend for ES we identified is represented by "domain specific semantics", i.e., the development of the intelligent query interpretation (intelligent analysis of query to extract relevant terms). Moreover, a possible future evolution of ES could be towards the application of cross-sessions search models on which there is an increasing interest, regarding, so far, the general

---

[31]M. White, Enterprise search – trends and developments, Intranet Focus Ltd Research Note 06/12, June 2012, http://www.intranetfocus.com/wp-content/uploads/Enterprise-search-trends-and-developments-RN06012.pdf
[32]IDC press release http://www.idc.com/getdoc.jsp?containerId=prUS23833612

web search. The assumption underlying these models is that while some simple needs can be satisfied with a single query, others require a series of queries over the long term, depending on the task users are trying to accomplish. In other words, users may be interested in resuming and grouping search results in the future, when they have to do with cross-session search tasks (e.g. a vacation planning task). Searchers behaviors extending over multiple search sessions are analyzed so that knowledge of previous queries on the same long-term task enables the search tool to support the user in return to his task [22].

The evolution we foresee in ES as 'search within the enterprise private network' is towards the single information access portal, Search Based Applications and mobile ES, while the evolution of 'search as a component of a larger application developed by the enterprise' is more projected towards: noSQL, SBA, cloud-based and multimedia search.

A general issue to be addressed in the future is the fact that boundaries between enterprise search, text and data mining, information discovery, business intelligence and content analytics are becoming very blurred [23].

3.2.4. Open questions regarding the topics addressed in this section

Some important open questions that need to be addressed in the near future include:
- As there are no academic and specialized courses in enterprise search in the EU, how organizations are going to deal with the lack of technical skills in search implementation?
- How to define a stable and open architecture over which a European ES eco-system could develop and mature, taking into account the fact that boundaries between enterprise search, text and data mining, business intelligence and content analytics are becoming increasingly blurred?
- How are ES solutions going to develop in the mobile, multimedia, cloud computing world of tomorrow?

3.3. Conclusions and recommendations regarding Enterprise Search

The *main reason* for the adoption and development of ES solutions is the difficulties companies have to manage the vast and rapidly increasing volumes of data they collect, store and process, and the need they have to retrieve relevant information from this data in an efficient way.

From an industrial policy point of view, the prime interest is to propel the adoption of efficiency improving technologies in all European enterprises in order to increase Europe's economic competitiveness as a whole. ES is certainly such an ¨enhancing technology and the question arises whether and what policy measures need to be taken.

However, here we have to differentiate two levels of potential activity. One is to assess the importance of ES as an industrial sector *per-se*; the other is the impact of ES on other industrial sectors.

With regard to ES as an industrial sector *per-se*, we have observed two major waves of market consolidation and it seems that the mergers and acquisitions will probably continue. Extrapolating the trends of the past several years, the disappearance of dynamic smaller players, many of them based in Europe, seems likely, either by acquisition by larger companies (Type 1) or not being able to resist the market pressure. The question arises whether it is better to let the market consolidate on its own or take active steps and specific actions to influence the consolidation. The second case would be justified if ES would be considered highly strategic for Europe's economy. One option would be to promote specific research programs for smaller ES companies (specializing in some niche ES markets or developing specialized ES tools) to develop further their technologies and research capacity in order to be able to compete with larger players, or being able to collaborate with these on more favorable terms. This might reduce the probability being acquired by non-European 'big fishes', and eventually to operate in collaboration with bigger European companies (or integrated within them).

Policy challenges emerging from our techno-economic analysis of enterprise search in Europe [45] relate mainly to the need to stimulate the awareness of the usefulness of ES and its potential to increase the efficiency and competitiveness of enterprises, and to further increase European presence in the global context. A transversal issue regards whether and to what extent the EU might propose and intervene with ad–hoc policy options, considering that the opportunities and challenges for the ES market are often seen differently by different players and experts in the field.

4. Technology transfer

4.1. Future trends and challenges regarding Search and Technology Transfer

4.1.1. The role of open benchmarking in fostering scientific excellence and technology transfer

Benchmarking campaigns stimulate excellence in scientific production

- By providing comparative objective assessment of progress and performance level for all research actors in a domain.
- By stimulating technical exchanges between the different teams.
- By contributing to set up a 'competition' scheme between all research actors.
- By giving access to (training and) evaluation data to research teams which contributes to a better scientific production of these teams.

Benchmarking campaigns foster exchange between academia and industry

Benchmarking campaigns contribute to improve the visibility on good research, and exhibit performance of alternative scientific approaches on a set of common and relevant evaluation criteria. Furthermore benchmarking campaigns give industrial actors the opportunity to define technical requirements and performance objectives that fit to their target applications. By sharing that with academia, industrial actors can on one side influence scientific developments so that the technologies produced better fit to potential industrial applications and on the other side improve their knowledge and objective assessment of how good the technologies available are for their industrial needs. This positive influence on transfer can be further reinforced if industrial partners also contribute directly to the production of the training and evaluation corpora that are provided in the frame of the benchmarking campaigns.

The positive scientific, technical and economic impact is attested by international studies performed (significant impact on more than 2000 publications claimed by TRECVID, similarly over 1000 publications and around 8 citations of each publication by ImageCLEF and a factor of 3 to 5 return on investment according to a study by RTI International on TREC).

There are however some obstacles to the development of benchmarking campaigns. The cost of participation in an evaluation campaign, in particular for a first participation can mean up to 10 extra man months and can be a show

stopper, in particular for SMEs, and also for many research groups. Preparing representative corpora of sufficient size for benchmarking campaigns is challenging for various reasons (cost of annotating the data, infrastructure for collecting and maintaining big data corpora, limitations related to access rights, etc.). The risk of having sub-optimal corpora is to bias the evaluation, causing systems to converge to ad-hoc solutions that can generalize poorly when transferred to real-world content. This challenge related to corpora will increase in the future, due to the continuous and massive increase of the corpora size that will be requested to accompany the technical challenges related to big data. Another open challenge related to benchmarking campaigns is about including in the evaluations some user-centered criteria, to address problematics related to end users, without increasing substantially the cost of evaluations and reducing the objective quality of the evaluation outcomes.

4.1.2. Technology transfer: lessons learned from European collaborative R&D projects in the search area

Beyond progress on the research front, one of the goals of EU funded projects is to foster transfer of research results to effective industrial activity. To that effect, projects include a specific work-package covering "dissemination, exploitation and technology transfer". Although the comments below do not constitute a full-fledged study on technology transfer in EU funded projects, they come from analysis of this activity in several projects associated with Chorus and Chorus+. They also do not cover all possible aspects of the topic, but concentrate on those appearing most salient.

Overall, it can be said that technology transfer is very hard to achieve, and that success is rarely achieved. As said by Jim Mitchell, head of Sun Microsystem Research Laboratories, "technology transfer is a contact sport". This implies both personnel and activities.

On the "personnel" front, the EU project organization is expected to achieve technology transfer by requiring that "industrial actors" participate to the projects. To be efficient one should ensure that this participation is endorsed by the operational management of the industrial actor as opposed to its internal research organization. Presence and argumentation of marketing and sales managers of the industrial organization at the submission defense of the project is potentially a good measure of this success factor. Beyond the submission and defense stage, participation of the operational actors of the industrial partner to the project management activity is a second requirement.

During the course of the project, a sufficient "activity" should ensure that the technology transfer has an opportunity to happen. One should note here that technology transfer should not happen "after" but "during" the project. Frequent and effective contacts between the researchers and the potential operational

industrial partners should both create opportunities for the former to educate the latter, and for the latter to influence the former. It is often said that demonstrable research results must be available at hand to be effective in the "selling" effort of research results. This is indeed desirable or preferable, but positions this activity too late in the project. Researchers must describe potential results ahead of time, and verify that such results are of real interest to the industrial partner. On his side, if contacted sufficiently early, the industrial partner may integrate possible research results in his overall strategy.

Lastly, the form in which research results are made available plays an important role and must be discussed very early in the project. For instance, the system versus components approach may have a significant impact on the facilitation of technology transfer, and must be discussed very early in the project with the potential industrial partner(s). Patents and intellectual property rights associated with the results must be clarified at the earliest possible stage, in spite of the fact that the results are of hypothetical form. This unfortunately leads very often to IP terms which are too broad and too restrictive (we keep all rights) for an efficient technology transfer, essentially leaving actual discussion to "later". At this "later" stage, it is rarely the case that there is enough motivation and resources for a new round of discussion. The Open Source approach can alleviate some problems here. Automatic release of the results in case of absence of exploitation by one of the partners is an alternative, but as its definition shows, it assumes failure of initially planned technology transfer!

A good case illustrates some of the issues discussed above: the Vitalas project for a multimedia search engine. Among the partners, two were qualified as industrials: one "customer" for the proposed system, one "system integrator". The second partner expressed from the start his disinterest in commercializing the future system while the first one expressed interest in acquiring such a system, but not for supporting or commercializing it. In the middle of the project, the hypothesis for a "start-up" capable of carrying the system towards the markets lost credibility and a component approach was tested with an industrial partner active in the search domain. In spite of the demonstrated interest of this new partner, the lack of clear IP rights statements and licensing terms prevented from achieving results. The situation remains the same two years after the end of the project!

4.1.3. The benchmarking gap in Europe

There is no dedicated funding in Europe to sustain the organization of public benchmarking campaigns at the international level. Large initiatives such as CLEF or MediaEval typically live through heterogeneous and opportunistic research funds including national and European projects, permanent resources from research institutes and even volunteer resources. On the other side, the American National Institute of Standards and Technology (NIST) is in charge of

organizing most benchmarking campaigns in the US with significant permanent resources (and still with the contribution of external researchers). Academic and industrial actors think that Europe should not simply leave the floor to NIST (for several reasons related to scientific, cultural and social diversity as well as economic strategy). The role and the force EU has in funding research at the European level makes it the best candidate to set up a sustainable and efficient way to fund and synchronize benchmarking campaigns in Europe.

4.2. Conclusions and recommendations regarding Search and Technology Transfer

Triggering the participation in benchmarking campaigns for research activities benefiting from EU funding

Benchmarking campaigns are a tool to boost technological progress and foster exchanges between industry and academia. Collaborative R&D projects benefiting from EU funding should be incited, by contractual terms, to take part in international benchmarking campaigns in the domains that they are addressing. In order to support this incitation and make it broadly acceptable by the academic and industrial research actors, the following conditions should apply:

- funding should not be conditioned to results obtained in benchmarking campaigns;
- the extra costs incurred to participate to benchmarking campaigns should be funded at 100%, up to a ceiling amount to be defined with the research community and benchmarking structures;
- anonymous participation should be allowed for participants requesting to participate under this condition (anonymous participation meaning that the results and ranking are communicated without the name of the participant, the identity of the participant being known only to a few trusted organizers of the campaign bound to confidentiality).

Contributing to a sustainable benchmarking framework in Europe

The EC should play a leading role in setting up and coordinating a sustainable benchmarking framework at European level. It includes in particular the following aspects:

- Coordinating the various initiatives within Europe, to avoid duplications, ensure actions are taken to fill identified gaps, and guarantee optimal use of resources;
- Ensuring that benchmarking campaigns steering is balanced. The definition of the challenges measured in benchmarking campaigns has to be done collectively by the research community, the industry and the authorities. It is in particular crucial to keep an important place for innovation diversity by ensuring that new task proposals come from both the research community and the industry. Acceptance mechanisms should also rely on a balanced pool of experts;

- Bringing financial support to set up benchmarking actions to fill identified gaps, or to ensure sustainability of existing benchmarking actions that cannot rely on committed resources;
- Ensuring data openness in EU projects. Companies often refuse to share the large data they are using in their scientific publications, sometimes for competitive reasons and sometimes to protect customers' privacy. On the other side, as big data is becoming an important research area, this practice is criticized by many researchers for its secrecy and the risks of bad science, potential frauds, etc. The problem occurs as well within EU funded projects. Our recommendation is therefore to condition EU funding to some guaranties on data openness, at least for the project's consortium, and possibly to the whole research community, therefore ensuring that the investments made in the production of corpora in individual projects benefit to some extent to the building of a benchmarking framework at European level.
- Funding large-scale infrastructures. Besides privacy and copyright issues, hardware resources and data management problems prevent many research groups from working on real-world and big data. We advocate for setting up a shared infrastructure at the European level adapted to research on information retrieval and data mining. Such infrastructure should allow hosting large-scale multimedia data as well as services developed by research projects (such as the services that could be evaluated in benchmarking campaigns). This could be done in collaboration with major content providers and owners of big infrastructures in Europe;

Contributing to improve the efficiency of benchmarking campaigns

We believe that some support should be brought at EU level in order to improving current practices in benchmarking along the following two key objectives:

- Moving to larger and real-world data. The consequence of too small or too narrow data is that technologies generalize poorly when transferred to real-world content. This gap between the performances measured in benchmarking campaigns and what can be expected at scale-one is weakening technology transfer. Integrating new technologies in large infrastructures without enough guaranties on performances is actually too risky for many industrials.
- Allowing user-centered and external evaluation. System-oriented evaluation metrics used in current benchmarks are essential but not sufficient to cover a vast range of usage of the evaluated technologies. Furthermore, evaluation methodologies are often not scalable because of the huge human work required to build appropriate evaluation data. An open evaluation framework would rather allow any other research group or company to evaluate a technology with its own criteria or in the

context of its own workflow. On top of such open evaluation framework, user-centered evaluations would allow evaluating technologies according to end-users acceptance.

Improving the efficiency of technology transfer and the return on investments in collaborative R&D

In order to address some of the issues discussed previously, the following recommendations could be taken into consideration:

- ensure that operational members of the industrial partners are involved in the preparation, defense and management of the projects
- give a higher priority to the exploitation and technology transfer section of the project (both at the initial defense stage and during periodic evaluation)
- provide aggressive directives regarding IP rights of the work performed during the project

When not exploited commercially, many relevant technologies built within EU projects are lost. New projects often re-develop the same piece of work resulting in a large waste of time and money. Ensuring the sustainability of the technical components developed in EU projects is therefore crucial. Our recommendation is that the developed components should be either commercially exploited or shared (with new EU projects and/or with the research community). A moratorium could be applied for making thinks easier, notably for industrial partners: any results may be locked up for one or two years, but then should be shared if no commercial exploitation occurred.

5. The economics of search

5.1. What is personal in search: the economic value of personal data and user empowerment

As can be inferred from previous chapters, the search query and its parameters, in particular personal information used to 'enrich' search, have become an essential building block of Internet activities and its economics, as it were a new currency of the digital world. On the grounds on this currency, information goods and services are offered, and physical goods purchased and paid for. It is not an exaggeration to say that search underpins the Internet of services: it would be hard to identify an alternative to search and to indexing based on the search query as an entry point to these specific markets.

It is also known that online and offline companies today collect, store and use increasingly more user personal information in order to offer online targeted products and services. But only major companies have in-house resources and incentives to run their own customer analytics based on personal information. Most companies rely on other, third parties to reach consumers based on their preferences and behaviors. In this respect, search as an information good generates value over and above its gate-keeping, indexing role.

Along the click stream, users disclose increasing amounts of data for different purposes and in different settings, in particular, via online social networking. As emerges from [28], more than a third of European citizens access to SNS and more than one half of those also use websites to share pictures, videos, movies, etc. As the main use of SNS is to enable online socializing, this necessarily implies a disclosure of social (personal) information.

The important point is that these settings are increasingly integrated and interdependent, such as the case of search services bundled with video services, with email services and with advanced clouds services.

Integration is progressing both cross-platform and cross-device, opening new opportunities for monetization of personal information profiles. Search functionalities on the Web, in mobile communications, in audio-video and in online social networking stress the relevance of personal information flows; meanwhile, they have raised significant privacy and data protection issues.

As search engines became an unavoidable gate to digital goods and services well beyond information, their success relies on the ability to provide searches that closely match users' preferences beyond search terms, which in turn relies on personal information. The business model of most successful search services is

based on targeted adverting, which requires profiling users and vast use of personal information. Personal information has economic value for search companies and for their advertising customers, though the precise estimation of this value is still to be defined.

On the one hand, bottom line costs/benefit analysis and system development lead companies to disregard users' privacy, as they are routinely accused of doing by a large section of the press. Companies claim that innovation in this domain rests on the capacity to take alternative, fresh looks at whatever information users are happy to disclose. On the other hand, the rigorous privacy and data protection framework in Europe is proving to be inadequate to safeguard these fundamental rights and to allow, meanwhile, a better experience/service for the users, 'user surplus' in economic terms. The lack in the US of a general, comprehensive data protection framework coincides with a more utilitarian approach.

The European Data Protection framework does not seem to provide individuals with a satisfactory level of control, and data protection requirements are difficult to implement [25] or create a *search* environment excessively burdensome for the user herself [26]. The framework could raise a set of legal barriers to the development and innovation of search computing, data minimization being the rule. Some of the requirements for lawful data processing (consent, legitimate interest of the providers, limited scope and duration) are at odds with current Internet practices and users' behaviour, and risk to restrain users from obtaining economic and social advantages of data use (especially from search).

The consent requirement, for instance, shows its limits with regard to user profiling by search providers and related behavioural advertising, because of difficult implementation and because most users seek relevant search results and many are happy with targeted advertising [46]. Legal instruments based on the consent requirement, such as the 'cookies laws' or the codification of the right to be forgotten[33] have already attracted criticisms and risk of being difficult to be implemented, because grounded mainly on the "data minimization" principle. Moreover, given the amount of user-generated content and data disclosure, the definition of a controller (and its obligations) meets even more problems when applied to search services [27].

Is it possible in Europe to favouring techno-economic innovation by adopting a more flexible data protection system? Lately, there has been increasing pre-regulatory pressure on companies to comply with minimum standards and safeguards in the market for online services, in the shape of prior privacy impact

---

[33] See the Proposal for a General DP Regulation of the 25/01/2012, http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf

assessment of new business processes and technologies, binding corporate rules and other 'soft' measures.[34] If data protection right means for the individuals to keep a high level of control over their data, this should imply the *power* to choose about the data disclosure and graduate/accommodate their use (what we can call user' *empowerment*).

Even though, how users benefit from the disclosure and processing of personal data by search engines is less known and less interesting in Europe, more focused on considering privacy challenges.

5.2. Search engines as key players in two-sided markets: issues and implications

According to a traditional definition, two-sided markets are markets in which one or several platforms enable interactions between end-users and try to get the two sides "on board". Platform businesses exploit positive, indirect and network or cross-group effects between various customer groups. They add value by facilitating interactions between customers and/or firms who are attracted to the platform at least in part by the number of agents on the other side of the market. These interactions are important in many key industries but particularly in digital goods markets. Since its birth, the digital contents market has been considered a canonical example of a two-sided market. It has been also the case for industries whose business model is based on advertising. Newspapers and media firms have been usually cited on the literature on two-sided markets based on advertising. Advertising is also the main business model for many online services and virtually the only one for search providers, as search markets are becoming a canonical example of a two-sided market. Some particular features of search can introduce, however, interesting variations on the general theory of two-sided markets, limiting considerably their usefulness to explain the dynamics of search markets or even voiding it completely.

On the bottom line, the utility of advertisers depends on the size of demand in the media market: an advertiser will be willing to pay more for a slot the larger the readership of a newspaper (as a proxy for impact). But even if the attitude of media consumers towards advertising cannot be unambiguously ascertained, it is widely recognized that a large part of the audience is generally negative to the quantity of advertising contained in the media (undermining impact). Thus, the double direction of cross-group externalities does not appear as clear. Back to search, little is known about the annoyance caused by advertisements to searchers. It may be that advertisements do not interfere with 'organic' search results, except perhaps for mobile search, given the reduced screen space. Of course, some searchers are interested in advertisements as they look to purchase

---

[34] Such as Commissioner Almunia's letter to Google stating the European Commission concerns about Google's perceived discriminatory behaviour with respect to 'organic' search results.

specific goods for which there is a thinner advertising market, especially ones that are harder to find. In this case, we would go back to general two-sided markets theory, with positive externalities growing on both sides as the number of participants increases. However, this only holds for transactional queries – that account for about 10% of the total queries. For informational queries, the perceptions of searchers on advertisements range from neutral to negative – also because they remind consumers about their personal information being used.

Secondly, there is the issue of externalities from consumers to advertisers and the use of a pay-per-click mechanism. Within pay-per-impression costing, advertisers value the number of searchers as there is a higher likelihood that some will find a suitable match that will lead to a beneficial interaction. On the contrary, if advertisers pay a platform only in the event of a successful interaction (pay-per-click), they should in principle be indifferent to using a platform with few or many searchers so long as the value exceeds the cost of each click the advertiser gets, and therefore externalities are lessened with per-transaction charging.

In addition, the use of personal information by search engines leads to rather different outcomes from those of usual two-sided market strategies. First, there is no way today to balance strategies. Most two-sided markets subsidize participation on one side and recover the loss on the other side. As targeted advertising is more profitable for search providers, they should try to have more users willing to disclose their personal information. But they cannot offer a discount on the price to be paid as this price (personal information) is at the same time a (the) production factor for the improved version of the search result. Neither can search providers offer a different version of their content differing in advertising quantity. Most probably, users would not accept a change in the results page they are used to on top of having disclosed their data. This is why search providers that bundle search with other incentives for the users (such as free email, for instance) have some more leeway that pure search players in this market.

The crucial question follows: is two-sided market theory a fitting framework for targeted advertising search mechanisms or are there players with large market power practicing first-degree discrimination of its market by exploiting information asymmetries? Two rejoinders with policy implications.

The first is whether the information obtained from consumers should be limited, as search providers have an incentive to extract all possible information from users and monetize it in all possible ways. This is something unusual in conventional two-sided markets where one side (consumers) is typically subsidized to extract value from the other side. In agreement with this view, the

EU seems to have taken this regulatory 'road', and compensate the loss of a possible surplus with the avoidance of a probable privacy loss.

The second consequence refers to market dimension, market power and level of competition tests. As the real competition is for consumers' personal information to be later monetized, the contenders are all the providers able to profile users and sell targeted advertising. If this is the right definition of the market, search engines and social networks, to name the foremost types of providers, are to be found in the same market competing both for advertising money. No need here for two-sided markets theory, just for usual checks on market power and potential abuse of dominant positions in the value chain.

5.3. Barriers/opportunities of theoretical models based on personal information as an asset

The shape and size of welfare effects deriving from limits imposed on the unfettered exchange of personal information in markets for goods and services is to date unclear. Consumer and market welfare depend on complex equilibrium of personal data disclosure, requests and withholding. For instance, personal information at the fringe – social data, activities, pictures - are not very valuable to individuals[28], but they may become so when unintended consequences of their loss or exploitation are factored in [47]. In particular, there is no proof that consumption linked to online advertising generates positive social externalities. When looking into the parties involved in the transaction, [48] concludes that 'ultimately, the economic consequences of information sharing for all parties involved (the data subject and the actual or potential data holders) can be welfare enhancing or diminishing'. Indeed, different theoretical models can be used to gauge the impact of the granularity of profiling on welfare. For neoclassical economists, efficiency is achieved through a free market of personal information where administrative interference introduces distortions and inefficiencies [49]. Privacy costs emerge for firms and individuals when insufficient information is shared with third parties. Therefore, enforcement of privacy is welfare diminishing. The futility of personal information protection is also stated by [50]: the allocation of rights on personal privacy depends on the relative valuation of parties involved. If the user values privacy more than the on-line firm, the data will remain protected even in the absence of privacy regulation[35].

The main critics to neoclassical approaches note that the assumption of rational choice does not capture the complexity of consumers' decisions on personal information and privacy and that, in particular, consumers decide with incomplete information, have bounded cognitive ability to process the available

---

[35]Although, game theory may arrive at very different conclusion from the very same assumptions, depending on who is seen as best placed to protect one's personal information.

information, and have also behavioral biases that depart from rational decision making. They also stress that the secondary usage of personal information gives users little control upon how the firm will later use that data, including potential hidden costs in terms of privacy loses when a third party reuses the data. It has also been argued that the exploitation of personal information for unsolicited marketing can constitute a negative consumer externality [51]. As a consequence, 'the market equilibrium will tend not to afford privacy protection to individuals, and therefore privacy regulation may be needed to improve consumer and aggregate welfare' [49]. The absence of a framework for contextual integrity of personal information [29] adds to the necessity of balancing the market equilibrium to protect the weaker side in the transaction, while keeping the opportunities in the provision of innovative services.

5.4. Regulators in the hot seat

A number of solutions are advanced in the literature to mitigate some of the problems exposed here: possible market asymmetries and lack of effective protection of people's personal information. Some scholars acknowledge the economic value of personal data and recommend governmental regulatory action in order provide users' with higher control on their data. Others sustain the development of ad hoc privacy enhancing technologies; others suggest introducing in the current European legal framework some elements of the theory of the economics of privacy.

A recent appraisal of the economic value of personal data and consequent implications in terms of regulatory options argue for the need for tailored consumer protection, the user's right to rescind enrollment in a service, i.e., the ability to delete and export their information. The claim refers to the SNS context, but the economic considerations of privacy could be applied to Search services proven the increasing interplays between them[36]. These options would be a partial solution for user-generated content, but would encounter more problems as regards data collected and aggregated by third parties, especially in the search context, where profiles are superimposed to users rather than being generated by them. Portability is far more difficult in an opaque "data maximization" environment, where data disclosure by the users is seamless.

Other solutions may be emerging from the market. A market of personal data management tools is already emerging. These start-up companies, that promise to work as data lockers are Azigo, Mydex, the Data Banker, Personel.com, Connect.me. They work as cyber-lockers, safe-deposit box, that would allow users to store own personal data and meanwhile as personal digital assistant: a sort of cloud-based centre to store and manage all users' digital stuff, all their

---

[36] See the recent Google case of shifting policies, integrating all its products/services and the new feature of its SNS, Google+, that would include content shared privately in search results.

online lives (financial information, medical records, music etc.). What has been less explored so far is what (and how) users can do to effectively control their search data, preferences and behaviour (i.e., to decide who can access to their personal information) and to even profit from them. According to the Data Protection Ecosystem Consortium[37], there are at least three business models linked to this new concept of 'data control': 1) users data stored in lockers or accessed via real time browsing can be directly sold by 'owners'; 2) or can be provided to third parties who will sell the data to advertisers; 3) or users can put together their data with others' into a larger repository that is sold or shared with companies or public entities.

Industry self-regulation might go in the same direction: reducing the burden for users, allow a significant degree of disclosure, ensure transparency and build in actionable remedies for users whose privacy may be compromised. The recent debate in the US about the 'Do not Track' policy (DNT) gives the idea of the issues at stake. DNT is a proposed HTTP header field that would request a web application to disable either their tracking or their cross-site tracking of the users. The header can be enabled or disabled in the browser, and needs website agreement to function. It thus aims at granting consumers greater control on their personal data online and at limiting the tracking power of online companies. The first hurdle for its generalized adoption to become a reality is the lack of an agreement on what 'Do not Track' should mean, as privacy advocates and Internet firms have divergent opinions on that.[38] According to privacy groups, the system should enable stopping data collection as such, so that consumers are not continuously spied in their online activities by Internet companies for their economic gain; according to web companies and advertisers, the DNT tool means not targeting advertisements to consumers on the basis of their online activities, but still keeping data collection for other purposes.[39]

---

[37] This Consortium is an international community of more than 30 companies, aiming at "connecting entrepreneurs building new businesses around user-centric personal data and advocating for individuals having the tools and rights to access and manage their own data", available at: http://pde.cc.

[38] The US administration assigned the W3C the production of a DNT tool after talks with Internet firms, online advertisers and privacy advocates. See J Melvin, Do not Track Internet spat risks legislative crackdown, Business News, 24/07/2012.

[39] Privacy advocates' proposal of a DNT tool would bar third parties (i.e. web companies, advertisers that track users across the visited websites) from collecting data from a user he opt out to be tracked (limited exception would be security and fraud prevention, so there would be a sort of selection, limitation based on the subjects authorized to the collection-and here we would have again the thorny issue of blurring private and public contexts); advertisers point to a self-regulatory model that places an icon on ads that consumers see thanks to behaviour advertising (they will be linked to a website explaining the reasons of that ads and giving them the option to opt-out from receiving targeted ads): in other words, it would only give the users control on the type of adverting (targeting) not on the data collection or on the fact of receiving generalized ads.

Regarding technological remedies, most of the many privacy-preserving technologies that have been deployed in the last decades [52], aimed at on limiting data disclosure and data collection by the user. But these technologies also place a burden on the user and remove some of the direct benefits of controlled disclosure, thus reducing user welfare. Technical solutions that rely on the advantages of personal data disclosure may be most effective in terms of optimal balance between privacy preserving and quality of the search performance.[40]

Increased obligation of transparency may also assist the market without diminishing user surplus and companies' revenue. Consumers may agree to provide their information without knowing that a search provider would sell this info to others providers and they may act differently if they would have known. Legal intervention that is more relaxed on data collection but that posed stricter transparency obligations may be appropriate in these cases.

Finally, some scholars debate the possibility to adopt proprietary rights over personal data. If certain conditions are met, it is possible to introduce elements of the theory of property rights also in the European personal data framework [30]. This does not mean that users can 'sell' their personal data, but rather that, under certain contractual conditions, it may be welfare enhancing to allow the commercial exchange and, to some degree, free flow of personal information. The data lockers idea, discussed above, falls into this category. But the compatibility of this approach with the European data protection tradition, more oriented towards a system of unalienable rights and liberties, is still under debate. The norms foreseen by the Proposal for a General Regulation on Data Protection (2012) retain major safeguards and traditional data protection mechanisms - data minimization, consent, clear distinction between controller and processor.

Open questions
A lot of legal-economic and technical issues still need to be addressed as regards to the topics of this section.
Is a stricter governmental regulatory action needed in order to guarantee a free flow of information and, meanwhile, consumers' privacy protection[41]? Or, a self-regulation system, created by the same online businesses, would bring to simpler, immediate and more effective data control tools?

---

[40] A. Erola and J. Castellá-Roca [53] propose a protocol for distorted profiles relying on the user' SNS connections: the search queries sent by a user are obfuscated by the queries of his peers, so that the search system is able to recognize the interests of the user but not to identify him.
[41] See the recent Proposal for a General Regulation on Personal Data of the 25 January 2012, in which stricter rules aiming at strengthening the controllers liability and users' rights (including right to be forgotten) have been introduced and that has already raised several concerns among online companies.

Would the identification and awareness of the value of personal data facilitate the definition of adequate policies and/or regulatory actions?[42] Would this awareness bring to enhanced legal-technical solutions for a higher user control (including economic gain) over the content she posted online?

Would it be possible to introduce some elements of the theory of proprietary rights on personal data in the European traditional regulation, in order to ensure a development of the European regulation more in line with the development of technologies and the societal changes, keeping, nevertheless safe the values underlying privacy and data protection regime?

A specific market of user data management tools is evolving (e.g., data lockers). What should the European policy be on this regard?

Conclusions

This topic of the economic value of personal data is relevant for the future of Search Computing and requires further research and a deeper policy analysis. A pilot project run at the European Commission JRC-IPTS (code names 'Economics of Personal Information'), posits that higher disclosure of personal data offers, in certain circumstances, economic and social benefits for the user of web service (particularly search services); and that more flexibility and focused data protection, embedding the concept of economic value of personal information into the European legal context may be beneficial. However, establishing a trust relationship maybe more natural and important to users than a monetary value.

Of course, there is evidence that consumer protection needs to be rethought and reinstated vis-à-vis the specificities and challenges of cloud computing and 'big data' analytics [31].

Though overall, a more nuanced approach to data protection, less prohibition-based and more projected to account the economic and social value of personal data, could be beneficial for both businesses and users in order to strike a balance between the aspiration to benefit from the economic value of personal data (for businesses and users) and the need to safeguard core data protection values.

---

[42] The Universidad Politécnica de Madrid, in collaboration with the JRC-IPTS, developed a method to measure the economics value of personal data, which has been presented in October 2012, at the Amsterdam Privacy Conference (http://www.apc2012.org/). The method (a discrete choice experiment) allows putting some figures on information 'assets'. The results of the related pilot projects ('Economics of personal data') will be published as JRC-IPTS Report (forthcoming 2013).

References

[1] T. O'Reilly, *What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software.* O'Reilly Media Inc., 2005.

[2] T. Segaran, *Programming Collective Intelligence.* O'Reilly Media Inc., 2007.

[3] S. Nikolopoulos, E. Chatzilari, E. Giannakidou, S. Papadopoulos, Y. Kompatsiaris, and A. Vakali, "Leveraging Massive User Contributions for Knowledge Extraction." N. Bessis and F. Xhafa, [ed.]. *Next Generation Data Technologies for Collective Computational Intelligence.* s.l. : Springer, 2011, Vol. 352.

[4] ITCandor., [Online] http://itcandor.net/2012/07/19/phone-shares-q112.

[5] IHS iSuppli. [Online] http://gigaom.com/mobile/by-2013-it-will-be-a-smartphone-majority-world/.

[6] Performics., [Online] http://searchengineland.com/performics-mobile-impressions-cross-10-percent-threshold-74984.

[7] —. [Online] http://blog.performics.com/search/2011/04/mobile-paid-search-impression-share-crosses-10-threshold.html.

[8] S. Diplaris, A. Sonnenbichler, T. Kaczanowski, P. and Mylonas, "Emerging, Collective Intelligence for personal, organisational, and social use." N. Bessis and F. Xhafa [ed.]. *Next Generation Data Technologies for Collective Computational Intelligence.*: Springer, 2011, Vol. 352, pp. 527-573.

[9] G. Takacs, V. Chandrasekhar, N., Y. Xiong, W-C. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, B. Girod "Outdoors augmented reality on mobile phone using loxel-based visual feature organization", ACM International Conference on Multimedia Information Retrieval (MIR'08). pp. 427-434, 2008.

[10] B. Girod, V. Chandrasekhar, D.M. Chen, N.M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S.S. Tsai, and R. Vedantham, "Mobile Visual Search", IEEE Signal Processing Magazine, 28(4):61-76, 2011.

[11] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems." Journal of Personal and Ubiquitous Computing, 2006, Issue 4, Vol. 10, pp. 255-268.

[12] A. Thiagarajan, J. Biagioni, T. Gerlich, J. Eriksson, "Cooperative Transit Tracking Using Smart-phones", International Conference on Embedded Networked Sensor Systems, SenSys. pp. 85-98, 2010.

[13] Nikolopoulos, S., Nikolov, S.G. and Kompatsiaris, I., "Study on Mobile Image Search." NEM Summit: Implementing Future Media Internet, 2011.

[14] H. Zhao, *Emerging buisness models of the mobile internet market, M.S. thesis.* Helsinki University of Technology, 2008.

[15] S. Papadopoulos, S., C. Zigkolis, Y. Kompatsiaris, A. Vakali, "Cluster-based Landmark and Event Detection on Tagged Photo Collections." IEEE Multimedia Magazine, 2011, Issue 1, Vol. 18, pp. 52-63.

[16] A. Foncubierta-Rodríguez and H. Müller, "Ground Truth Generation in Medical Imaging: A crowdsourcing-based iterative approach.", International ACM Workshop on Crowdsourcing for Multimedia, 2012.

[17] E. Chatzilari, S. Nikolopoulos, I. Patras, Y. Kompatsiaris "Leveraging social media for scalable object detection." Pattern Recognition, 2012, Issue 8, Vol. 45, pp. 2962-2979.

[18] JRC-IPTS., IPTS Mobile Search Survey. [Online] 2010. http://www.ist-chorus.org/public/files/CTTE workshop files 09062010/14 - IPTS-Mobile-Search-Survey-2010-Results-WEB.pdf.

[19] Arstechnica., [Online] http://arstechnica.com/buisness/2012/08/new-api-severely-restricts-third-party-twitter-applications/.

[20] Oracle Survey., *From Overload to Impact: an industry scorecard on big data business challenges.* 2012.

[21] G. Grefenstette and L. Wilber, *Search-Based Applications.* s.l. : Morgan & Claypool, 2011.

[22] A. Kotov, P. Bennett, R. White, S. Dumais, J. Teevan, "Modelling and Analysis of Cross-Session Search Tasks." ACM SIGIR 2011.

[23] PWC., "Technology forecast 2012." [Online] 2012. http://www.pwc.com/us/en/technology-forecast/2012/issue1/index.jhtml.

[24] K.S. Candan, "RanKloud: scalable multimedia and social media retrieval and analysis in the cloud." Workshop on Large-scale and distributed informational retrieval (LSDS-IR '11), 2011.

[25] K. Jones and D.Tahri, "An overview of EU data protection rules on use of data collected online." Computer Law and Security Review, Vol. 27, 2011.

[26] N. Andrade and S. Monteleone, "Digital Natives and the Metamorphosis of the European Information Society. The emerging behavioral trends regarding provacy and their legal implications." *European Data Protection: coming of age?,* Springer, 2012.

[27] R. Wong, "Data protection: The future of privacy." Computer Law& security Review, 2011, Vol. 27.

[28] "Eurobarometer 359/2011." [Online] http://ec.europa.eu/public_opinion/archives/ebs/ebs_359_en.pdf.

[29] H. Nissenbaum, "Privacy as contextual integrity." Washington Law Review, 2004, Vol. 79.

[30] N. Purtova, *Property Rights in Personal Data. A European perspective.* s.l. : Kluwer Law International, 2012.

[31] W. Hon, C. Millard, I Walden,"Personal Data in cloud computing. What information is regulated?", Queen Mary University of London, School of Legal Studies, Research Papers n. 75/2011.

[32] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker., "Query by Image and Video Content: The QBIC System." Computer, 1995, Issue 9, Vol. 28.

[33] J.L. Gómez-Barroso, R. Compañó, C. Feijóo, M. Bacigalupo, O. Westlund, S. Ramos, A. Jaokar, F. Álvarez, R. De Waele, G. Mateos-Barrado and M.C. García-

Jiménez *Prospects of mobile search.* Institute for Prospective Technological Studies. European Comission, 2010.

[34] L. von Ahn and L. Dabbish, "Labelling images with a computer game." SIGCHI conference on Human factors in Computing Systems (CHI'04). pp. 319-326, 2004.

[35] B. McFee, B., T. Bertin-Mahieux, D. Ellis, G. Lanckriet, "The million song dataset challenge." 2012. International Conference Companion on World Wide Web. pp. 909-916.

[36] A. Schindler, R. Mayer, and A. Rauber, "Facilitating comprehensive benchmarking experiments on the million song dataset.", ISMIR 2012.

[37] T. Bertin-Mahieux, et al.,"The million song dataset." 2011. International Society on Music Information Retrieval Conference (ISMIR 2011).

[38] H. Müller, et al., *ImageCLEF - Experimental evaluation of visual information retrieval.* Springer, 2010.

[39] T. Tsikrika, A. García Seco de Herrera, and H. Müller, "Assessing the Scholarly Impact of ImageCLEF." Lecture Notes in Computer Science (LNCS 6941). pp. 95-106, 2011.

[40] J. Melvin, "Do not Track Internet spat risks legislative crackdown." *Business News.* 07 24, 2012.

[41] M. Hearst, 'Natural' search user interfaces. *Commun. of the ACM* 54(11):60-67, 2011.

[42] M. Goto and J. Ogata, "PodCastle: Recent Advances of a Spoken Document Retrieval Service Improved by Anonymous User Contributions", Interspeech 2011, pp.3073-3076, 2011.

[43] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano, "Songle: A Web Service for Active Music Listening Improved by User Contributions", ISMIR, pp. 311-316, 2011.

[44] M. Goto, "Grand Challenges in Music Information Research", In M. Muller, M. Goto, and M. Schedl, editors, Dagstuhl Follow-Ups: Multimodal Music Processing, pp.217-225, 2012.

[45] M. White and S. Nikolov, JRC-IPTS Report, "Enterprise Search in Europe. A techno-economic analysis", Eds: R Compano, S Monteleone, I Maghiros, forthcoming 2013.

[46] W. Lusoli et al. "Pan-European Survey of practices, attitudes & policy preferences regarding personal identity data management." Sevilla, EC JRC Institute for Prospective Technological Studies, EUR- Scientific and Technical Research series, Luxemburg: Luxemburg Publications Office, 2012.

[47] S. Berthold and R. Böhme, Valuating Privacy with Option Pricing Theory, in: Economics of Information Security and Privacy, 2010.

[48] A. Acquisti, The economics of personal data and the economics of privacy, OECD OECD Conference 30 Years after the OECD Privacy Guidelines, 2010 http://www.oecd.org/sti/interneteconomy/46968784.pdf

[49] G. Stigler, An Introduction to Privacy in Economics and Politics, Journal of Legal Studies, 9, 1980.

[50] E Noam, Privacy and Self-Regulation: Markets for Electronic Privacy. In U.S. Dept. of Commerce, Privacy and Self-Regulation in the Information Age, 1997.

[51] K.L. Hui and I. Png, Economics of privacy, in T. Hendershott, Ed., Handbooks in Information Systems, Vol. 1, 2006.

[52] V. Senicar, B Jerman-Blazic, T Klobucar, Privacy Enhancing Technologies-approaches and development, Computer Standard and Interface, 25, 2, 2003.

[53] A. Erola, J. Castellá-Roca, A. Viejo and J. Mateo-Sanz, Exploiting Social Networks to provide privacy in personalized web search, The Journal of System and Software, 84, 1734, 2011.

Contributors

Section	Name	Organization
Media Search	Vincenzo Croce	Engineering
	Antonio Focubierta Rodriguez	HES-SO
	Piero Fraternali	Politecnico di Milano
	Henri Gouraud	INRIA
	Alexis Joly	INRIA
	Henning Mueller	HES-SO
	Andreas Rauber	TU Wien
	Alexander Schindler	TU Wien
	Nicu Sebe (coordinator)	Univ. of Trento
Social and Mobile Search	Claudio Feijoo	Universitad Politecnica de Madrid
	Yiannis Kompatsiaris (coordinator)	CERTH
	Shara Monteleone	JRC
	Spiros Nikolopuolos	CERTH
Enterprise Search	Henri Gouraud	INRIA
	Shara Monteleone (coordinator)	JRC
	Henning Mueller	HES-SO
	Stavri Nikolov	Imagga
Technology Transfer	Henri Gouraud	INRIA
	Alexis Joly	INRIA
	Pieter van der Linden	THOMSON
	Serge Travert (coordinator)	THOMSON
The Economics of Search	Jose-Luis Gomez-Barroso	Universitad Politecnica de Madrid
	Claudio Feijoo	Universitad Politecnica de Madrid
	Shara Monteleone (coordinator)	JRC
	Stavri Nikolov	Imagga

The Chorus+ Project Consortium groups the following Organizations:

Partner Name	Short name	Country
JCP-Consult	JCP	FR
The French National Institute for Research in Computer Science and Control	INRIA	FR
Centre for Research and Technology Hellas - Informatics and Telematics Institute	CERTH-ITI	GR
University of Trento	UNITN	IT
Vienna University of Technology	TUWIEN	AT
University of Applied Sciences Western Switzerland	HES-SO	CH
Engineering Ingegneria Informatica SPA	ENG	IT
Technicolor	THOMSON	FR
JRC Institute for Prospective Technological Studies	JRC	EU

Revision History

Version	Edition	Author(s)	Date
0	1	Serge Travert	14/09/2012
Comments:	First version integrating contributions from Andreas Rauber, Alexander Schindler, Nicu Sebe, Henning Müller, Shara Monteleone, Yiannis Kompatsiaris, Spiros Nikolopoulos, Alexis Joly, and Henri Gouraud.		
0	2	Antonio Foncubierta Rodríguez	06/12/2012
	Harmonized format version, added references.		
1	2	Nicu Sebe	15/12/2012
	Version incorporating contributions from Alexis Joly, Henning Mueller, Shara Monteleone		
Final		Nicu Sebe	25/12/2012
	Final version		