Multimedia retrieval evaluation dimensions

Deliverable 3.3 – Report on evaluation dimensions

Distribution Level: PU

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
THOMSON	THOMSON	FR
JRC Institute for Prospective Technological Studies	JRC	EU

Document Identity

Title:	Deliverable 3.3
Subject:	Report on evaluation dimensions
Number:	3.3
File name:	D3.3.doc
Registration Date:	15.12.2011
Last Update:	15.12.2011

Authors

Alexis Joly (Editor)	**INRIA**
Henning Müller	HES-SO
Nicu Sebe	Trento university
Thomas Lidy	TU-WIEN
Henri Gouraud	INRIA
Pieter Van der Linden	Technicolor

Table of Content

1. Introduction

Evaluation is one of the key drivers of innovation for multimedia search components or integrated systems. As a result, many evaluation dimensions exists that cover technical (quantitative component evaluation), as well as industrial and end user side criteria. The impact of these evaluation dimensions on scientific & technological progress as well as on the commercial success of a given search technology is however not fully understood. This report is the result of a study conducted within CHORUS+ EU coordination action towards a better understanding of these relationships and with the goal of improving the current evaluation practices.
The report surveys and analyses existing evaluation dimensions by classifying them into 4 categories with regard to different contexts: (i) scientific literature, (ii) system-oriented benchmarks and evaluation campaigns, (iii) user-centered evaluations and user trials, (iv) industrial and business-oriented settings. As illustrated in Figure 1, these evaluation dimensions work at different levels in the innovation workflow, from fundamental research to software production and sales. As a consequence, each of them aims at evaluating different objects (from theoretical models to real-world systems) and usually concerns different categories of actors with different interests and evaluation criteria.

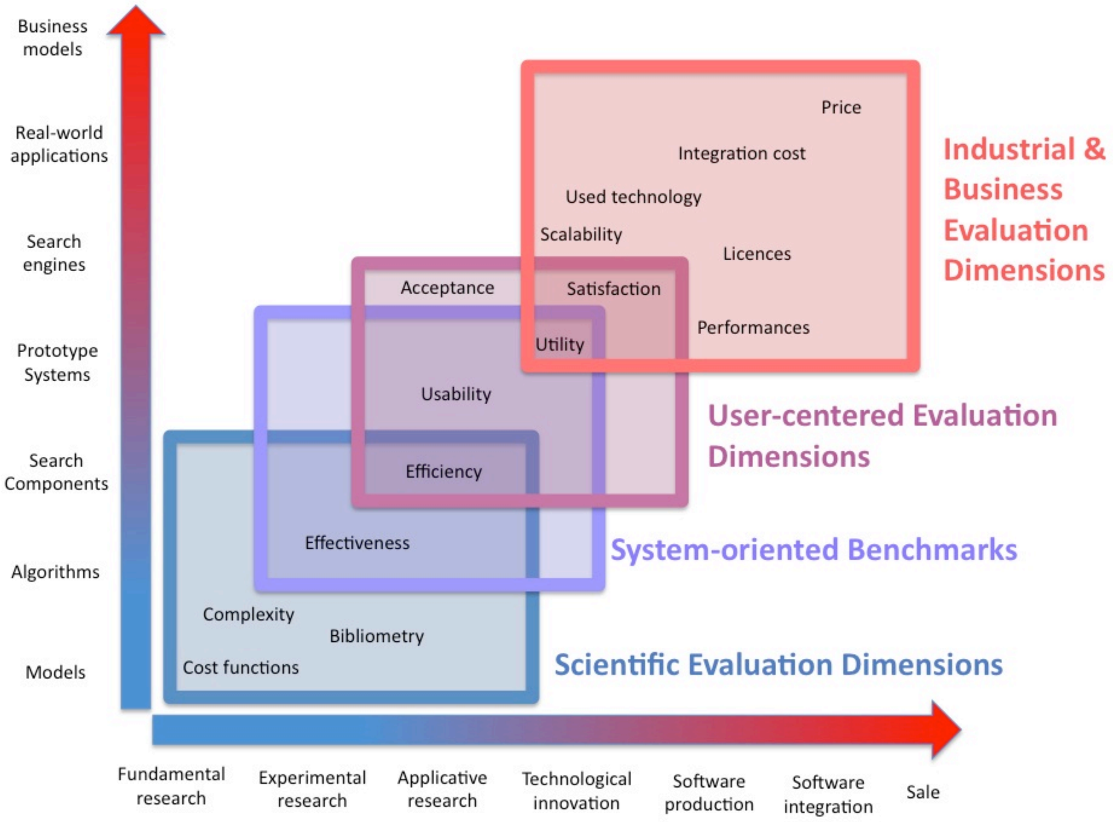

Figure 1 – Multimedia retrieval evaluation dimensions (positioned in the innovation workflow)

Of course and fortunately, some overlaps and interactions exist between these categories. Any actor in the evaluation workflow is usually concerned with several adjacent categories. This allows the evaluation results and conclusions at a given level to be naturally inferred to upper levels, usually in a more synthetic form. So that low level scientific evaluations may have some impact on the commercial success of a given technology, even if the actors at the end of the chain do not have any knowledge of these evaluations.

Understanding these interactions and the real impact of the different evaluation dimensions is however complex. We will partly address these issues in this report, thanks to existing studies and CHORUS+ network experience in the evaluation of multimedia search technologies. But to get complementary information to be analyzed, we also decided to set up a survey aimed at collecting feedback from the different actors of the innovation workflow (including both academics and industry). The produced questionnaire for this purpose is presented at the end of the report. It covers the 4 categories of evaluation dimensions we introduced in order to measure their relative impact at different levels of the innovation workflow. At the time of writing, the questionnaire has been delivered to several communities of actors including mailing lists of international conferences and evaluation campaigns as well as professional forums dedicated to search engines. The results and the analysis of this study will be presented in a further deliverable about evaluation needs (D3.4).

2. Overview of existing evaluation dimensions

2.1 Scientific evaluation dimensions

We refer to **scientific evaluation dimensions** as the ones used by the scientific community to publish, review, compare and analyze scientific results. We can distinguish between *explicit evaluation dimensions* that are reported in the publications themselves to quantify the performances of multimedia retrieval methods, and *implicit evaluation dimensions* that are used after the publication of research works to measure their quality and impact, without explicitly considering their content.

2.1.1 Datasets and quantitative evaluation metrics in the literature (explicit dimensions)

Experimental evaluation is the most widely used protocol for quantifying the performances of multimedia search methods and technologies. Most works in the literature evaluate and compare methods by reporting quantitative *evaluation metrics,* measured on specific experimental *datasets* and for specific *retrieval tasks*. The combination of these 3 *evaluation variables* lead overall to a strong **fragmentation** in the evaluation results reported in the literature. There are various reasons for this fragmentation, grounded in all of the 3 evaluation variables:

First of all **structural** reasons: There is no centralized information resource for evaluation metrics, datasets or a definition of tasks specifically for the multimedia search domain. A wide range of evaluation metrics exists and is being used in research and experimental evaluation. Also multimedia benchmarking and evaluation campaigns such as ImageCLEF and MIREX do not impose explicit usage of specific evaluation measures [Sanderson2010, MIREX]. Individual task organizers decide on the evaluation metrics to be used or define their own evaluation metrics with varying priorities and interests of what to actually measure [Sanderson2010]. Frequently, a set of metrics is reported for a single task (e.g. recall, mean average precision (MAP), mean reciprocal rank (MRR), etc.) to cater for the different interests of evaluation. In scientific literature the wide range of metrics reduces comparability of research output. This problem is increasing with the ongoing publication deluge. More and more research is being done and increasing competition leads to publication of even small progresses in experimental research. The problem of fragmentation is therefore rising. A risk is that it might even appear desirable for the experimental results not to be comparable (typically by using “exotic” metrics or datasets).

Resources are the second major reason for fragmentation. The use of “in-house” datasets for evaluation in literature is also caused by a lack of availability of standard datasets. For researchers in multimedia retrieval it is very difficult to select and use proper datasets for experimental evaluation. There is no consensus on existing experimental datasets because usually they are not representative of real-world data. Researchers in general do not have access to real-world data or, in the cases where they have, cannot process or manage them because of infrastructure issues. *Real-world* data is typically very large, associated with data transfer, data storage, and data processing cost that might not be affordable for a research institution. A particular problem to getting hold of real-world data in the multimedia domain is the one of copyrights: In order for data to be transferred and used for research it must be licensed through - there is no particular licensing rights in place in general for non-commercial usage of multimedia data in research. In order to apply newly developed algorithms to real-world scenarios, a researcher would need access to huge datasets - which are extremely difficult to license or to get hold of. Multimedia evaluation campaigns are facing the same problem, so that even if the size of the datasets has increased in previous years they are far from being real-world.

The third reason for fragmentation lies in the **intrinsic complexity** of evaluating multimedia retrieval, which is related to the richness of the scientific objectives in the context of growing and evolutionary use-cases and user needs. While for some of the tasks in multimedia retrieval standard evaluation metrics are applicable, some others are very specialized and therefore require specialized forms of evaluation. Specific applications require specific evaluation protocols that do not fit existing ones. For example, in the ImageCLEF photo annotation task recently Semantic R-Precision (SR-Precision) was introduced as a novel performance measure, which determines the semantic relatedness of misclassified concepts by considering the Flickr Tag Similarity measure. This is a reaction to novel forms of evaluation through large-scale annotations through users by tagging. In MIREX, and Music IR in general, there are metrics that were specifically introduced for tasks such as Chord detection (e.g. "Chord Average Overlap"). Overall, choosing an appropriate evaluation metric is a research area in itself and many articles have been addressing it. A recent study also stresses the need for standard and reliable IR experimentation methodologies [Blanco2011]. The study shows that random perturbation of scores can result in improvements in performance of a retrieval system and compares this with the typical approach in research of inventing a new feature or optimizing parameters alongside repeated experimentation. It points out that one must be extremely careful when evaluating new algorithms as random effects can yield similar improvements. This shall be ensured by preventing over-fitting and carrying out proper statistical significance tests. This means that not only the choice of proper evaluation metrics is a difficult one - in the study, mean average precision (MAP) proved to be a more robust measure against random improvements than mean reciprocal rank (MRR) and precision @ 10 (P@10) - but also proper significance testing. Typical statistical tests, such as the t-test, the Wilcoxon signed rank test and the sign test [Croft2009] have the problem that the assumptions about the data in general do not hold true in practice. Sanderson [Sanderson2010b] provided a counter-argument according to Hull's remark [Hull1993] that although data do not satisfy the assumptions of these tests, discrete measures might be well approximated by continuous distributions if there is **sufficient data**. But how large the data should be is usually unclear since no empirical examinations of these tests have been conducted for multimedia retrieval.

Fragmentation in the evaluation of multimedia retrieval research works has some consequences. Whereas the scholarly peer review process would ideally require impartial and clear indicators, many experts in the multimedia retrieval community agree on that fragmentation makes difficult the analysis of stand-alone experimental results. Improvements over existing methods are often suspects of being ad-hoc to a specific dataset, metric or task. The choice of baselines against which to compare can also be subject to discussion. On the other side, comparative studies conducted by third parties (i.e. not the one that designed the methods to be evaluated) are often perceived as more trustworthy and their impact (in terms of number of citations) is usually very high. Benchmarks and evaluation campaigns also help in reducing the fragmentation of metrics, datasets and tasks. This will be discussed in detail in the section 2.2 of this report. Even with new tasks (and therewith new datasets and metrics) continuously introduced, the tasks are well defined with metrics that are usually discussed and agreed among researchers. Datasets are frequently provided to researchers either before an evaluation campaign, or released afterwards, in order to enable researchers to work on improvements measured on the same dataset as used in the benchmark. While the use of these datasets and the agreed evaluation measures in the literature on the one hand reduces fragmentation and increases comparability it might also lead to overfitting, a well-known problem of optimizing algorithms on specific data. Thus, the call for more and larger datasets remains open. Recent answers to this are platforms providing resources for research and evaluation purposes, such as the PEIPA platform [PEIPA] or the CHORUS+ Collaborative Platform [CHORUSplat]. These are also intended to eliminate the barriers between the narrow individual research areas (image retrieval, music retrieval, video analysis, etc.) and to cross-fertilize the efforts in multimedia search systems by creating cross-multimedia platform for all researchers in the domain.

There are finally several recent and promising trends as an answer to the call for "real-world" and consensual datasets. Large data sets such as ImageNet (http://www.image-net.org) are evolving, building *knowledge data* rather than *experimental dataset*. Labeling large datasets has indeed become faster, cheaper, and easier with the advent of **crowdsourcing** web services. Not long ago, labeling large datasets could take weeks or months and required finding and training expert annotators on custom-built interfaces. Today, with crowdsourcing applications like Amazon Mechanical Turk, such laborious work can be outsourced to a large crowd of workers and get results back in a matter of hours. This opens the door to labeling huge datasets by a large number of end users and user communities. So, that the resulting data is perceived as more relevant by the community and might therefore reduce experimental data fragmentation. This trend of building large experimental resources is also supported by the LinkedData efforts (http://linkeddata.org/), connecting and linking various datasets on the web and enriching them semantically.

As a brave new idea towards reducing fragmentation, some scientific journals or conferences are investigating a shift from classical papers to executable or reproducible papers. The *Elsevier executable papers* initiative [ElsevierExe] is for instance a challenge aimed at answering questions like: how to make the components of the experiments available to the reader so that the experiment can be repeated and manipulated? How to make equations, tables and graphs interactive in such a way that reviewers and readers can check, manipulate and explore the result space? How to validate data and code without increasing to much the reviewer's workload? Another related initiative, specific to the database community, is the SIGMOD Conference Experimental Repeatability Requirements [SIGMODRep]. Since 2008, on a voluntary basis, authors of accepted SIGMOD papers could provide their code/binaries, experimental setups and data to be tested for repeatability of the experiments described in the accepted papers. In 2010, a college of experts has been in charge of reproducing the experiments provided by the authors (repeatability), and exploring changes to experiment parameters (workability). Their conclusion was that the received contributions are still far from the vision of executable papers (basically Linux packages accompanied by instructions on how to setup and run the experiments). Overall, these initiatives are still very prospective and it is difficult to evaluate the impact they could have if they were generalized.

2.1.2 Community evaluation dimensions (implicit dimensions)

In the context of increasing research competition for limited amounts of funding, publications deluge and fragmentation, researchers developed new strategies in identifying high-quality research and selecting most relevant works they should compare with. Exhaustive reviews of existing methods and their systematic experimental evaluation are clearly not the rules anymore. The selection of state-of-the-art methods and acceptance as such are more driven by community evaluation dimensions.

There are several ways of judging the quality of the research by looking either at the venue of the publication (journal, conference, etc.) which reflects the impact a particular publication is expected to have to the community or directly looking at the publication track record of a particular researcher with respect to the way his/her work is received by the scientific community. While the former is mostly measured using the **impact factor** (the case for the scientific journals and magazines) or the acceptance rate (in the case of conferences and workshops), the latter is mostly captured by a few measures (such as the **h-index),** which give a holistic measure of the scientific career of a particular researcher. In the following, we will summarize these two implicit evaluation dimensions and discuss their advantages and shortcomings. Beforehand, we notice here that both types of measure are based on the same axiom, i.e. that the **citation count** of articles is a good indicator of their impact and quality. So that the citation count can also be used directly to judge the quality of a given published work, without considering the impact factor of the journal or the h-index of the authors. Unfortunately, this can be done only several or many years after the publication date, when a sufficient amount of posterior works has been published. This limits its practical use for state-of-the-art research evaluation. But it might still be a very relevant criterion for actors interfering later in the innovation workflow (typically for technology identification and selection).

The **impact factor**, often abbreviated **IF**, is a measure reflecting the average number of citation to articles published in science and social science journals and magazines. It is frequently used as a proxy for the relative importance of a journal within its field, with journals with higher impact factors deemed to be more important than those with lower ones. The impact factor was devised by Eugene Garfield, the founder of the Institute for Scientific Information (ISI), now part of Thomson Reuters. Impact factors are calculated yearly for those journals that are indexed in the Thomson Reuters J*ournal Citation Reports*. In a given year, the impact factor of a journal is the average number of citations received per paper published in that journal during the two preceding years. 2011 impact factors are for instance published in 2012; they cannot be calculated until all of the 2010 publications have been processed by the indexing agency. Additionally, the new journals, which are indexed from their first published issue, will receive an impact factor after two years of indexing. The impact factor relates to a specific time period; it is possible to calculate it for any desired period, and the *Journal Citation Reports* (JCR) also includes a 5-year impact factor. The JCR shows rankings of journals by impact factor, if desired by discipline, such as Computer Science, Information Systems or Electrical and electronic engineering.

Numerous criticisms have been made of the use of an impact factor, including the more general debate on the usefulness of citation metrics. Criticisms mainly concern the **validity** of the impact factor and the **policies** that alter it. First, the impact factor is formally not a valid representation of the distribution predicted by theory (Bradford distribution) and therefore unfit for citation evaluation. Furthermore, it is highly discipline-dependent and affected by self-citations. Self-citation is common in journals dealing in specialized topics having high overlap in readership and authors, and is not necessarily a sign of low quality or manipulation. Overall, it has been shown that journal ranking lists constructed based on the impact factor only moderately correlate with journal ranking lists based on the results of an expert survey [Serenko2011]. The IF may also be incorrectly applied to evaluate the significance of an individual publication or to evaluate an individual researcher [Seglen1997]. This does not work well since a small number of publications are cited much more than the majority. The impact factor, however, averages over all articles and thus underestimates the citations of the most cited articles while exaggerating the number of citations of the majority of articles.

Editorial policies also affect the impact factor. Journals may publish a larger percentage of review articles, which generally are cited more than research reports. Therefore, review articles can raise the impact factor of the journal and review journals will therefore often have the highest impact factors in their respective fields. Conversely, journals may choose not to publish minor/short articles, which are unlikely to be cited and would reduce the average citation per article. Journals may also change the fraction of citable items compared to front-matter in the denominator of the IF equation. Which types of articles are considered citable is largely a matter of negotiation between the journals and Thomson Scientific. For instance, editorials in a journal are not considered to be citable items and therefore do not enter into the denominator of the impact factor. However, citations to such items will still enter into the numerator, thereby inflating the impact factor. In addition, if such items cite other articles (often even from the same journal), those citations will be counted and will increase the citation count for the cited journal. This effect is hard to evaluate, for the distinction between editorial comment and short original articles is not always obvious. "Letters to the editor" might refer to either class.

The impact factor and all related values published by the same organization apply only to journals, not individual articles or individual scientists (unlike the h-index). The relative number of citations an individual article receives is better viewed as citation impact. It is, however, possible to measure the Impact factor of the journals in which a particular person has published articles. This use is widespread, but controversial. Garfield warns about the "misuse in evaluating individuals" because there is "a wide variation from article to article within a single journal". Given all these, impact factors have a large, but controversial, influence on the way published scientific research is perceived and evaluated.

Many measures have been proposed, beyond simple citation counts, to better quantify an individual scholar's citation impact. The best-known measures include the h-index [Hirsch2005] and the g-index [Egghe2006]. Each measure has advantages and disadvantages, spanning from bias to discipline-dependence and limitations of the citation data source [Couto2009]. An alternative approach to measure a scholar's impact relies on usage data, such as number of downloads from publishers.

The ***h*-index** is an index that attempts to measure both the productivity and impact of the published work of a scientist or scholar. The index is based on the set of the scientist's most cited papers and the number of citations that they have received in other publications. The index can also be applied to the productivity and impact of a group of scientists, such as a department or university or country. The index is based on the distribution of citations received by a given researcher's publications:

> *A scientist has index* h *if* h *of his/her* Np *papers have at least* h *citations each, and the other (*Np − h*) papers have no more than* h *citations each.*

In other words, a scholar with an index of h has published h papers each of which has been cited in other papers at least h times. Thus, the h-index reflects both the number of publications and the number of citations per publication. The index is designed to improve upon simpler measures such as the total number of citations or publications. The index works properly only for comparing scientists working in the same field; citation conventions differ widely among different fields.

The h-index works properly only for comparing scientists working in the same field; citation conventions differ widely among different fields. It serves as an alternative to more traditional journal impact factor metrics in the evaluation of the impact of the work of a particular researcher. Because only the most highly cited articles contribute to the *h*-index, its determination is a relatively simpler process. It has been also demonstrated that *h* has high predictive value for whether a scientist has won honors like National Academy membership or the Nobel Prize. The *h*-index grows as citations accumulate and thus it depends on the 'academic age' of a researcher.

The *h*-index is intended to address the main disadvantages of other bibliometric indicators, such as total number of papers or total number of citations. Total number of papers does not account for the quality of scientific publications, while total number of citations can be disproportionately affected by participation in a single publication of major influence (for instance, methodological papers proposing successful new techniques, methods or approximations, which can generate a large number of citations), or having many publications with few citations each. The *h*-index is intended to measure simultaneously the quality and quantity of scientific output.

However, there are a number of situations, most of these though not exclusive to the h-index, in which *h* may provide misleading information about a scientist's output [Wendl2007]:

The *h*-index does not account for the number of authors of a paper. In the original paper, Hirsch suggested partitioning citations among co-authors. Even in the absence of explicit gaming, the *h*-index and similar indexes tend to favor fields with larger groups, e.g. experimental over theoretical.

The *h*-index does not account for the typical number of citations in different fields. Different fields, or journals, traditionally use different numbers of citations.

The *h*-index discards the information contained in author placement in the authors' list, which in some scientific fields is significant.

The *h*-index is bounded by the total number of publications. This means that scientists with a short career are at an inherent disadvantage, regardless of the importance of their discoveries. This is also a problem for any measure that relies on the number of publications. However, as Hirsch indicated in the original paper, the index is intended as a tool to evaluate researchers in the same stage of their careers. It is not meant as a tool for historical comparisons.

The *h*-index does not consider the context of citations. For example, citations in a paper are often made simply to flesh out an introduction, otherwise having no other significance to the work. *h* also does not resolve other contextual instances: citations made in a negative context and citations made to fraudulent or retracted work. This is also a problem for regular citation counts.

The *h*-index gives books the same count as articles making it difficult to compare scholars in fields that are more book-oriented such as the humanities.

The *h*-index does not account for confounding factors such as *gratuitous authorship*, the so-called Matthew effect, and the favorable citation bias associated with review articles. Again, this is a problem for all other metrics using publications or citations.

The *h*-index can be manipulated through self-citations, and if based on Google Scholar output, then even computer-generated documents can be used for that purpose, e.g. using SCIgen.

Alternatively to the h-index, the ***g*-index** is an index for quantifying scientific productivity based on publication record. The index is calculated based on the distribution of citations received by a given researcher's publications. In simple terms, an author that produces **n** articles is expected to have, on average, **n** citations for each of them, to have a **g-index** of n. In this way, it is similar to the **h-index**, with the difference that the number of citations per article is not explicit. Practically, the g-index is highly correlated with the h-index [Serenko2010]. However, these indices are conceptually distinct, and the g-index attempts to address shortcomings of the h-index.

Given all these, automated citation indexing has changed the nature of citation analysis research, allowing millions of citations to be analyzed for large-scale patterns and knowledge discovery. The first example of automated citation indexing was CiteSeer, later to be followed by Google Scholar. More recently, advanced models for a dynamic analysis of citation aging have been proposed [Yu2010, Bouabid2011]. The latter model is even used as a predictive tool for determining the citations that might be obtained at any time of the lifetime of a corpus of publications. An important recent development in research on citation impact is the discovery of *universality*, or citation impact patterns that hold across different disciplines in the sciences, social sciences, and humanities. For example it has been shown that the number of citations received by a publication, once properly rescaled by its average across articles published in the same discipline and in the same year, follows a universal log-normal distribution that is the same in every discipline [Radicchi2008]. This finding has suggested a *universal citation impact measure* that extends the h-index by properly rescaling citation counts and resorting publications, however the computation of such a universal measure requires the collection of extensive citation data and statistics for every discipline and year. Social crowdsourcing tools such as Scholarometer have been recently proposed to address this need but are still not mature enough to derive conclusions about their usage.

2.2 System-oriented benchmarks & evaluation campaigns

Large-scale worldwide experimental evaluations provide fundamental contributions to the advancement of state-of-the-art techniques through common evaluation procedures, regular and systematic evaluation cycles, comparison and benchmarking of the adopted approaches, and spreading of knowledge. In multimedia retrieval, several benchmarks have had an important impact on the field by making databases available and providing a unified framework for comparing system performance. Benchmarks have made evaluation procedures more standardized and common, so many researchers are able to follow similar procedures, not creating their own methodologies (that risk to be non-comparable to the state of the art).

Large amounts of test data have been made available facilitating the validation of tools and techniques by academics and industrial researchers. Most benchmarks follow a yearly cycle of making data and tasks available, submitting results, comparing the performance and then discussing the results at a workshop. Such cycles allow researchers to adapt their timing in terms of research work at each period of the year. Starting with TrecVid (http://trecvid.nist.gov/), a video retrieval benchmark started at TREC (Text Retrieval Conference, http://trec.nist.gov/) to INEX (INformation retrieval Evaluation in XML data) and going towards ImageCLEF (http://www.imageclef.org/), the image retrieval part of CLEF (Cross Language Evaluation Forum, http://www.clef-campaign.org/), and MediaEval (http://www.multimediaeval.org/) the most recent initiative in multimedia evaluation. All benchmarks have increased the amounts of data massively gearing towards a realistic size in terms of the amount available at least for most applications. Usually, the participation in the benchmarks has also risen strongly over the past years, showing the significant impact and importance for research. TRECvid as well as ImageCLEF have had well over 100 registrations in the past years. Researchers from all continents and all countries participate equally in the benchmarks with a focus on North America, Europe and Asia, with poorer countries sometimes lacking the infrastructures to effectively participate in the benchmarks.

2.2.1 Overview of existing multimedia benchmarks

Table 1 gives an overall picture of existing benchmarks (also called evaluation campaigns) related to multimedia retrieval. We do not mention here benchmarks dedicated to pure text retrieval systems (notably TREC and CLEF), but only the ones dealing with multimedia content (image, video, audio, 3D models). It does not mean that these benchmarks ignore the textual modality since most of them make use of textual features extracted from the meta-data of the targeted media. The provided list is not claimed to be exhaustive but it covers most popular initiatives by showing their complementarity as well as the diversity of the evaluation tasks. For a more complete and more detailed state-of-the-art of benchmarking initiatives, the reader can refer to [CHORUSsoa].

Most of the existing benchmarks are **system-oriented**, in that they usually do not evaluate the interaction of real users with information retrieval systems (they actually use the same evaluation metrics as the one discussed in 2.2.1). TRECvid as well as ImageCLEF have had user-centered evaluation tasks but often the participants are only participating in these user-oriented evaluations in small numbers limiting the impact. Researchers are much more used to tuning the tools towards a specific task and many systems do not have a user interface but are rather scripts that combine existing tools that are sometimes also too slow for interactive work. User-centered evaluation contests will be discussed later in this report, in the section dedicated to user-centered evaluation dimensions (more precisely in 2.3.2).

Acronym	**Running period**	**Number of participants**	**Media**	**Tasks**
TRECVID (NIST)	2001-2011	**73** (2011)	video	Shot boundary detection (2001-2007) Known-item search (2001-2011) Semantic indexing (2002-2011) Story Segmentation (2003-2004) Rushes (2005-2008) Video-surveillance (2008-2011) Copy detection (2008-2011) Instance search (2010-2011) Multimedia event detection (2010-2011)
ImageCLEF	2003-2011	**43** (2011)	image	Photo retrieval (2003-2009) Medical image retrieval (2004-2011) Medical image annotation (2005-2009) Photo annotation (2006-2011) Wikipedia image retrieval (2008-2011) Robot Vision (2009-2010) Plant identification (2011)
MIREX	2005-2011	~ **40** (2011)	audio	Artist Identification (2005-2011 Genre Classification (2005-2011) Melody Extraction (2005-2011) Onset Detection (2005-2011) Tempo Extraction (2005-2011) Music Similarity (2005-2011) Key detection (2005;2010-2011) Beat tracking (2006-2011) Cover Song Identification (2006-2011) Query-by-Singing (2006-2011) Mood Classification (2006-2011) Melodic similarity (2006-2011) Chord detection (2008-2011) Audio tag classification (2008-2011) Query-by-tapping (2008-2011) Structure segmentation (2009-2011)
MediaEval	2010-2011	**39** (2011)	multi-media	Geo-tagging (2010-2011) Video annotation (2010-2011) Video affect detection (2010-2011) Rich speech retrieval (2011) Spoken web search (2011) Social Event detection (2011)
Pascal VOC	2005-2011	~ **25** (2011)	image	Visual objects classification (2005-2011) Visual objects detection (2007-2011) Visual objects segmentation (2009-2011) Person layout detection (2009-2011) Action classification (2010-2011) Large-scale classification (2010-2011)
SHREC	2006-2011	~ **15** (2010)	3D-shape models	Watertight models (2006-2008) Correspondence/matching (2007-2011)

				Protein models (2007-2010) CAD models (2007-2008) Face models (2007-2011) Generic 3D models (2008-2011) Architectural models (2009-2010) Large scale retrieval (2010) Range scans (2011) Non-rigid shapes (2011)
INEX Multimedia	2005-2007	**4** (2007)	multi-media	Multimedia XML retrieval (2005-2007) Wikipedia image retrieval (2006-2007)
MusiCLEF	2011	**2** (2011)	audio	Music categorization (2011) Cover Identification (2011)

Table 1 - overview of system-oriented benchmarks related to multimedia retrieval

2.2.2 Impact analysis of existing benchmarks

TREC and TRECvid have received important funding from the American government and despite an importance in terms of participation and recognition of researchers, it was regarded as important to put the **impact into a measurable form**. Impact can be measured in several dimensions.

Economic impact can be measured by the amount of effort that is reduced by creating common resources or by the performance improvement of retrieval systems that results in a more productive workforce and/or new products that again generate additional revenue. Most such analyses need to take into account assumptions and it is not always easy to estimate these figures in a realistic way. TREC started in 2010 with **the first economic impact analysis** [Rowe2010] showing the enormous economic impact that TREC has had over the years. Every dollar spent on TREC has generated 3-5 dollars in revenue for the international industry and for academia.

Scientific or scholarly impact can be measured by the number of publications resulting from the benchmark or based on the produced data sets. The number of scientific articles directly produced by a benchmark and its organizers is often easier to obtain than all publications that are using the data sets as the sources are not always cited in a unified fashion and sometimes not at all. Most benchmarks ask participants to send new publications but this is also not always done. Measuring citations is also hard, as no agreed-upon standard exists. Publication and citation search tools such as Google scholar (http://scholar.google.com), Microsoft Academic Search (http://academic.research.microsoft.com/), Scopus (http://www.scopus.com/), and ISI web of science (http://apps.webofknowledge.com/) all have their advantages and inconveniences. ISI web of science is very incomplete and does not index conferences, which are of high importance in computer science. Working notes of CLEF or TREC are notably not indexed in this repository, for example, severely underestimating thus the impact. Scopus has a similar problem, being very incomplete and not indexing most of the conferences such as the working notes of benchmarks. Microsoft Academic Search is still in the process of its creation and the coverage is currently not higher than Scopus. It does include several conferences but on the other hand the results quality is not always perfect with users being able to add their data. Google scholar currently has by far the largest coverage and it includes all major computer science conferences and also working

notes of CLEF, TREC and TRECvid. On the other hand, the results quality of Google scholar is not always high and even non-peer-reviewed publications are added, sometimes leading to overestimated citation numbers. There are also other mistakes such as citations being separated due to spelling mistakes and sometimes similar sounding citations are combined into a single one. No tool is thus perfect and allows for a very good analysis.
In 2011, TRECvid published the **first scholarly impact analysis** [Thornley2011], well describing the problems of comparing citation counts in Google scholar and Scopus. The important impact of the benchmark was however clearly showed through the large number of generated publications and citations, particularly of the task overview papers. ImageCLEF then performed a similar analysis to TRECvid and published the results a few months later [Tsikrika2011]. In a first step only the papers published in the CLEF post-workshop proceedings and the papers published by the organizers were taken into account, still showing a similar impact in terms of citations counts per paper as TRECvid but on a smaller number of publications. It is clear that scientific impact analyses are taking into account papers describing usually mature techniques as research groups will publish novel techniques and approaches usually in more important journals and conferences to reach a higher impact. Papers in the working notes often also use very similar techniques to previous years, just adapting the tools to the novel data and tasks. This also limits the overall number of citations that papers obtain.

Finally, the **main shortcoming** in **all impact analyses** is the baseline that needs to be taken into account. It is not clear what would have happened if the resource had not been available. Would the research groups have published in any case but maybe on other data? Would repetitive papers have been avoided? Or would major breakthroughs have been missed due to poor evaluation and comparison between techniques? This is all not 100% clear but would be important to analyze in a proper system analysis to be complete for scholarly as well as economic impact. Performance analyses for CLEF and INEX are currently under way and should help estimating the real impact that standardized evaluation has had on information retrieval.

2.2.3 Evaluation infrastructures

With the strong impact of benchmarks and with the number of citations using standard resources increasing, it is important to think more about ways to reduce manual effort in evaluation and to automate part of the process to make evaluation using standards easy and to allow comparing techniques to strong baselines [Armstrong2009]. Propositions for **component-based evaluation** have been made [Hanbury2010] as this has the potential to actually better understand which factors have an influence on the final performance of a system. Component-based evaluation is on the other hand hard to set up and enforce as it can limit the flexibility of researchers.
Example systems for component-based information retrieval evaluation are the DIRECT system [Ferro2009], and for data mining the e-lico system (http://www.e-lico.eu/). DIRECT indexes a collection, topics the submitted runs and then allows executing the relevance judgments in the system, and performing evaluations automatically using the *trec_eval* tool (http://trec.nist.gov/trec_eval/). The goal of DIRECT is also to make results available in a sustainable way and allowing to compare performance also in the future on old data sets, showing long term performance improvements. For industry, the possibility to continuously evaluate performance and thus check system changes and their influence is important. The evaluateIR system [evaluateIR] allows just this, making past TREC collections available and allowing comparing performance continuously to strong baselines, meaning that each system

changes and the difference in performance can be measured and tracked over time. Still, the system did not receive enormous attention and it is thus not maintained anymore.

Other systems for the components evaluation of multimedia retrieval systems were presented in the Vitalas project [Vitalas]. Such a framework would be required if component-based evaluation is to become reality. This means more impact on the researchers and potentially a heavier architecture and slower response times, which are both not optimal for system developers. It needs to be shown whether the problems outweigh the advantages.

2.2.4 Shortcomings of existing benchmarks

Despite the shown impact and the enormous benefits of standardized evaluation there are also problems and shortcomings of existing campaigns. We list here the main identified shortcomings:

- One criticism is that pure focus on performance can **favor small changes** to existing components **rather than radically new approaches**. Components obtaining good performances are re-used and slightly improved by many participants the following years, with the risk to converge to consensual but non-optimal solutions. Benchmarks should thus not only evaluate performance but give more place to interesting techniques and approaches, for instance in the related workshops.
- Another frequently mentioned shortcoming is related to the **scale** and **scope** of the data used for benchmarking. For some domains, such as enterprise search, the existing databases are still too small compared to real repositories that can contain several terabytes. Shipping thus large databases is indeed logistically very difficult and limited by access rights that are hard to obtain with many of the data being often confidential. The consequence of a limited scope of the data is that systems might converge to ad-hoc solutions. After some years of benchmarking activities, several algorithms are able to perform particularly well for certain types of content, but generalize poorly when transferred to real-world content.
- The fact that almost no system-oriented benchmarks (beside maybe the Videolympics, http://www.videolympics.org/) include the **users in the loop** also means that only technology is tested and not working systems. User requirements might be much more important than pure system performance.
- For **companies to participate in benchmarks** is also not always easy as companies could be penalized if performance is not as high as of other competitors. This may require anonymous participation or the possibility to withdraw runs, which both are on the other hand not optimal for a benchmark.

2.2.5 Coordination of existing benchmarks

It is obvious that no single initiative could be by itself satisfactory by offering the context to test all the tasks addressed by the multimedia retrieval community. Also, due to the richness of the scientific objectives of multimedia search engines, corresponding to growing and evolutionary use-cases and user needs, benchmark initiatives have to follow the field dynamic.

Consequently, an important question is the **coordination** of the domain, with **fragmentation** or long-term **monopolies** being the main risks. The legitimacy and persistence of a given evaluation task is indeed highly conditioned by the number of participants, i.e. the number of international research actors interested in it. For benchmark organizers, it is difficult to find a balance between a sufficiently large spectrum of tasks and a sufficiently large number of participants in each of the tasks. And this is even more difficult when considering global interactions between the different existing benchmarks: research actors often switch from a benchmarking task to another, depending on their evolving competences, interests or chances to get good results. So that the global picture of existing benchmarks and evaluation tasks is determined by *supply and demand* in the worldwide research community.

The risk of fragmentation mainly comes from similar reasons to those mentioned for the evaluation dimensions used in the scientific literature (see section 2.1.1): increasing research competition, performance obligation, high diversity of multimedia retrieval applications, richness of the scientific objectives of multimedia search engines and intrinsic complexity of evaluating them. Too much fragmentation can limit the actual benefits of world-wide benchmarks and evaluation campaigns (common evaluation procedures, readable results, etc.) and in the end reduce their impact.

On the other side, fragmentation should not be reduced too much to preserve application diversity and enable radically new tasks to have their place in the global picture. If one specific evaluation task or benchmark became too much predominant, then its orientation choices would become problematic in terms of legitimacy and representativeness (researchers would not have choice anymore). Avoiding such monopoly requires that several large-scale and piloted benchmarks co-exist with funding being an essential variable.

2.3 User-centered evaluation dimensions & user trials

As previously discussed, multimedia retrieval evaluation in international benchmarks has been dominated by *system-oriented* evaluation dimensions which typically measure the relevance of the media retrieved by the different search components to be compared. But search-engine users usually expect more than just relevance. Are the results fresh and timely? Are they from authoritative sources? Are they comprehensive? Are they free of spam? Are their titles and snippets descriptive enough? Do they include additional elements a user might find helpful for the query (maps, images, query suggestions, etc.)? And even evaluating the relevance of the results is a tricky task in real-user's context. Understanding what a user really wants when typing or submitting a query (the user's intent) can be indeed very difficult. The scientific domain trying to tackle these questions is known as *user-centered evaluation*. This paradigm usually requires setting up some *user trial*, i.e. to create an environment that enables the interaction between the search engine and some users to be systematically examined and measured. It is therefore closely related to research works on *interactive information retrieval* [Ingwersen2005].

2.3.1 User-centered evaluation as a research topic

Evaluation of interactive information retrieval in the text domain has been an active research area in the last decades [Ingwersen2005]. More and more researchers have been arguing that *relevance* is not sufficient in evaluating search systems and that the *Cranfield paradigm* has undergone. The Cranfield paradigm was designed in the early 1960s (before the web) when information access was realized through Boolean queries against manually indexed documents. Implementation of this paradigm has undergone extensive modifications over the years (notably in TREC evaluation campaigns) and other evaluation forums as the data and tasks have gotten more complex. For straightforward searching tasks where clicks and dwell times can be used to predict relevance it is still often used for both academic as well as commercial evaluations. However the world of information access has exploded in recent years to encompass online shopping, social networking, personal desktop organization, etc. Developing scientifically valid evaluation paradigms for these novel domains is an important challenge.

Overall, the successful adoption of a system may depend on many heterogeneous criteria, including, but not limited to, the functional quality of the system, its ergonomic quality or even the costs involved to integrate the system in an existing workflow. Researchers ([Nielsen1993], [Shackel1991], [Dillon1996], [Tricot2003], [Rogers1995]) agree on 3 typical criteria, as discussed in [Vitalas2007]:

- **Acceptability**, which concerns the system in its broadest context. It therefore integrates most other criteria as sub-criteria including *utility* and *usability,* as well as others related to user motivation, cultural background or even socio-economic aspects. Shackel's approach [Shackel91] to acceptability has been the most popular and several modifications have been proposed afterwards. The approach proposed by Nielsen [Nielsen93] notably made some consistent refinements by adding the influence of *social acceptability*. Besides these two there are the models of [Dillon96] or [Davis89], who take into account the perceived *usefulness* in their criteria. [Rogers95] includes additional criteria in the definition of acceptability, which compares the relative advantage of the new system according to already existing ones.
- **Utility**, which concerns the functionality of the system. Senach [Senach90] defines it as: "Utility determines if the user may achieve his/her task with the system. Utility covers the

functional capacities, the performances, and the quality assistance of the system". In other words, does the software contain all the announced functionalities?
- **Usability**, which concerns the user friendliness of the system. Arguably usability is the most delicate criterion to evaluate because of the inherent subjectivity of the test subjects. On the other hand, it may also be considered one of the most important criteria as it is critical for adoption of the system.

From a practical point of view, we can broadly group user evaluations of IR systems by those that imply real users (user-centered approaches), and those that do not interact directly with users (user-data driven approaches) [Vallet2006]. Inside these two approaches there is a broad spectrum of evaluation techniques. User-centered approaches include user questionnaires [Dumais2003, Martin2004, White2005], side-by-side comparisons [Thomas2006], explicit relevance assessments [Finkelstein2001]. User-data driven approaches normally exploit query log analysis [Dou2007, Finkelstein2001] or test collections [Shen2005].

Beyond traditional information retrieval systems, several papers advocate a similar paradigm shift from system-centered evaluation to user-centered evaluation for *multimedia information retrieval* systems. In their pioneer work, Mad Donald et al. [McDonald2001] discussed a user-centered approach for content-based image retrieval. The results of their studies indicate that users were able to make use of a range of visual search tools, and that different tools are used at different points in the search process. It also did show that the provision of a structured navigation and browsing tool can support image retrieval, particularly in situations in which the user does not have a target image in mind. As another example, in [Leelanupab2009], the authors discuss a user-centered evaluation of a recommendation based image browsing system. As a last example regarding audio content, in [Hu2010], the authors discuss a general schema for user-centered evaluation of music information retrieval. Due to the entertaining nature of music, they suggest new user-oriented evaluation criteria such as *entertainability* or *social life support*.

2.3.2 User-centered evaluation in benchmarks and research projects

Although user-centered evaluation of multimedia retrieval systems is an attractive paradigm, very few concrete user-centered evaluations are actually reported in the literature. The main reason is that the feasibility of user-centered evaluations is much more complicated than classical experiments purely based on *relevance* and *performances*. The cost of recruiting a sufficiently large number of relevant users is indeed prohibitive for most research actors. The analysis of user trials results is also much more challenging than the analysis of classical quantitative evaluation metrics. As discussed in previous sections, an ideal user-centered evaluation requires compiling heterogeneous and high level feedbacks, such as interviews, open questions, user recommendations, etc. Another reason is that the acceptance of such evaluations in the peer-review process of scientific articles is still difficult. Many reviewers in the community are actually not used to judge whether a user-evaluation study was carried out scientifically correct. For both reasons, user-centered evaluations are far from being generalized to every day research in labs and universities. They are rather conducted in the context of large research consortiums (including international benchmarks, EU research projects, etc.).

And even within such large international initiatives, user-centered evaluations remain difficult. As mentioned before, TRECVID as well as ImageCLEF, have had user-centered evaluation tasks but often the participants were only participating in small numbers limiting the impact. Furthermore, it was up to the participants to organize their own user trials without a clear evaluation protocol. Results were consequently heterogeneous and difficult to analyze (different types of users across systems, different duration of the trials, …). In recent years, few live contests have been organized to solve these issues with some successful participation. Specifically VideOlympics [VideOlymp] was initiated in 2007 and was co-organized with CIVR conference until 2009 (now it is expected to appear again in ICMR conference). VideOlympics was a live contest of interactive video search engines, in which the participants compete at real time dealing with TRECVID-like topics. In order to involve regular users in the loop, VideOlympics 2009 included a run with novice users, while in the previous years the users of the systems were the researchers and the developers themselves. The idea of VideOlympics has given rise to similar contests as *PatOlympics* [PatOlymp] and *Video Browser Showdown* [VideoBrows]. PatOlympics started in 2010 with two *PatSports* (*ChemAthlon* and *CrossLingual Retrieval*) consisting of interactive prototype evaluation sessions. Video Browser Showdown will be held for the first time during MMM 2012 conference as a live video browsing competition where international researchers, working in the field of interactive video search, evaluate and demonstrate the efficiency of their tools.

Some user-centered evaluations of multimedia search engines were also conducted in research projects involving international consortiums. But still, only few of them are actually visible or reported in the literature. Some might actually remain unpublished for confidentiality reasons. One of the authors of this report was involved in the VITALAS EU project [Vitalas], which conducted an in-depth user study of a complete multimedia search engine designed to provide advanced solutions for indexing, searching and accessing large scale digital audio-visual content. Several evaluation cycles were conducted all along the project and finally resulted in more than 10 evaluation sessions involving about 30 users from a wide range of profiles (press agency journalists, professional archivists, TV broadcasters, general public). Each evaluation session was related to concrete and professional user-defined scenarios and users were asked to solve a large number of search tasks by using VITALAS search engine. The number of search functionalities that the users could use and combine to solve the tasks was quite impressive (text search, multimodal concepts search, image and video similarity, visual objects search, speech-to-text retrieval of videos, similarity maps, etc.). Conclusions of these evaluations were very positive and helped understanding the added value of multimedia retrieval components compared to classical text-based search. Some of the findings of VITALAS user trials were published in [Rode2010]. Before VITALAS, over 40 user interviews were also conducted and analyzed within AceMedia project [AceMedia]. Collected comments were mainly assessments, concerns and requirements, from which by careful analysis a condensed set of 160 essential user requirements has been identified. In the sequel, such user requirements have been translated into system requirements, and design solutions have been devised. More specifically about *cross-language information retrieval*, different user-centered evaluations were used during the life cycle of the Clarity Project [Clarity]. By aggregating the results of all the evaluations (in total 43 people were involved) it was possible to build a macro-view of how cross-language retrieval would impact on users and their tasks.

Overall, the richness of results that were acquired through user-centered evaluations in international benchmarks and research projects did stimulate researchers into considering user-centered evaluations as a flexible, adaptable and comprehensive technique for investigating non-traditional information access systems. On the other side, the scholarly and economic impacts of such studies still remain unclear in most cases. The obtained results are actually highly dependent on user context, data context and technological context. So, that they can generally not be considered as definitive conclusions about a given search components or paradigms. They are rather proofs of concept and drivers for designing improved systems. Furthermore, the scale of these user studies remains a matter of discussions, particularly for large public applications. Involving several tens of users faced to hundreds of query tasks is clearly a consistent effort. But it is still far from many potential real-world applications. And, as discussed in next section, this can often bias the conclusions.

2.3.3 User-centered evaluation of real-world systems

Evaluating real-world search applications is difficult for several reasons. Clearly, one can never measure anything close to all the queries a world-wide search engine would get. Every day, for instance, Google gets many millions of queries that were never seen before, and that will never be seen again. Therefore, measures have to be done statistically, over representative samples of the query-stream. Google is quite secretive regarding the details of their ranking algorithms and the way they are evaluated (mainly for competition and abuse reasons). It is however clear from their communication that *search quality evaluation* is mainly driven by user-centered considerations. As claimed by one engineer at Google in charge of Search Quality [google1], the evaluation of daily new improvements to Google is "done in many different ways, but the goal is always the same: improve the user experience. This is not the main goal, it is the only goal.". A specific team of the group is notably responsible for search evaluation, i.e. the process of measuring the quality of Google search results as well as the users' experience with using Google search engine [google2]. According to the director of this team, they employ both *human evaluators* and *live traffic experiments*, similarly to the *user-centered* and *data driven* approaches to evaluation reported in the scientific literature (as discussed in previous section).

More generally speaking, the heterogeneity, the complexity and the huge number of queries to be considered in real-world search applications highlight the difficulties of any user-centered evaluation. Since evaluation has to be done statistically, *bias* effects between the considered queries and the real ones are indeed very challenging. Any change of a search component might actually improve the performances on a subset of the query-stream but it will at the same time degrade the performances of many other searches. So, that a 100% improvement is essentially impossible. This makes in practice very difficult to validate the positive contribution of a new search component: how to sample the query stream without introducing too much bias? Should all queries be considered as equivalent? Should some of them be considered as more essential? Is the frequency of a given query a good indicator of its importance? Or is the number of users issuing this query a better one? Should some users be considered as more relevant? These simple questions reveal the social dimensions underlying the evaluation of real-world search engines and some potential risks. This also highlights the difficulty of comparing two different systems. A given search application might indeed target a specific community of users with some specific needs. It would therefore be unfair to compare it the same way than a generalist search engine.

Somehow, the above-mentioned scale issues and bias effects are to the detriment of most existing evaluation practices discussed in this report (including the one reported in the literature or used in benchmarks). Most of them are indeed based on very limited sets of queries that are usually built from the expertise of few users (or even randomly from the media themselves). The supremacy of a method on a given evaluation task might therefore be strongly biased compared to what it would get within a real usage scenario. But to the defense of current practices, this remark has to be mitigated by several concerns. First, new functionalities will change the usage itself. For instance, a visual object's search functionality will not be used to solve the same tasks than a text-based search. So that trying to define unbiased query samples is not possible until the functionality itself is integrated in a real world search engine. Secondly, many multimedia search tasks are less sensitive to query bias. Or at least, their current performances are not mature enough to consider such high-level concerns. If a system returns only pictures of dogs when it is asked to retrieve pictures of cars, anyone agrees it is not working well, whatever the query distribution. Same remark could apply to a speech-to-text search engine that would match too much false positives. In other words, the *relevance* evaluation criteria used in the literature and in evaluation campaigns are a primary goal before considering high-level and user-centered criteria such as *usefulness*.

2.4 Industrial & business evaluation dimensions

This section covers the industry point of view relative to benchmarking and evaluation criteria of search technologies. In this context, it is important to make explicit the difference of point of view and of goals of this community as compared to the scientific community covered elsewhere in this report. The purpose of evaluation and benchmarking in the scientific context has the main purpose of measuring progress in a given field by ranking technologies relative to one another. Of course, this ranking and progress measurement goal is somewhat irrelevant in the case of fully innovative technologies, which are first in their domain, but this is in fact seldom the case. By contrast, the purpose of evaluation and benchmarking from the industrial point of view is to achieve technology transfer from science to technology. For that reason, the criteria that are likely to be used by industry, when evaluating a specific technology, include those used in the scientific evaluation, but include also criteria specific to this technology transfer goal.

Industry is very familiar with evaluation and benchmarking, in particular in the context of "Competitive Bids" and "Request for Tender" which very often follow rules imposed by market regulators. The goal of these evaluations is to achieve the purchase of a product or a service from another industrial vendor. In this context, evaluation is performed against a "Requirement Document" which lists all the evaluation criteria for this specific bid. But this evaluation situation is most of the time in relation with finished products or services, which differs from the situation discussed here.

The discussion proposed here focuses on the case of technological components participating to the construction of a larger search system. Such components are likely to be benchmarked both by the scientific community that built them and by the industrial community that is likely to use them. During the Chorus project, of which Chorus+ is a continuation, a functional analysis [CHORUSfunctional] of search engine was performed, identifying the various technological components whose assembly resulted into a full-fledged search system.

In this more specific context, the industry actors of interest are either:
- **end-users** capable of acquiring and integrating technological components into larger applications that they develop in house.
- s**ystem integrators** capable of integrating a technology component into multiple systems that they are developing for their customers
- **software package developers** which wish to integrate into their larger solution a specific technological component.

We have omitted here one specific type of industrial actor, which nonetheless demonstrates the ultimate evaluation and technology transfer situation: that of the "start-up" created by the scientific team responsible for the development of the technology. It would take a full report to analyze the various criteria that come into play in the evaluation and decision process resulting in the creation of a start-up by a scientific team.

Each of these three type of actors is faced with similar problems and is likely to use similar evaluation criteria, but with a slightly different balance. We will point out specific situations where the organizations above may have a diverging point of view relative to evaluation.

The remainder of this section will focus on the territory where scientists and industry meet, with the following questions in mind: Are current evaluation methods and campaigns, conducted by scientists, relevant to the decision process of industrial actors? Which additional evaluation criteria, not covered by current practices, are used by industrial actors in their evaluation process?

In order to progress towards this goal, Chorus+ has been preparing a questionnaire that is proposed both to scientists and to industrial actors, asking them to rank by importance various evaluation criteria, which we believe to be relevant. Of course the questionnaire offers the possibility for the respondent to propose additional criteria. The final deliverable of this report will analyze the results of this questionnaire and propose a ranked list of evaluation criteria both from the point of view of scientific and industrial actors.

As the oldest of the well-established evaluation and benchmarking efforts, TREC has conducted a study on the "business impact" of its activity [Rowe2010] (as already discussed in section 2.2.2). This study concludes at significant direct and indirect business impact (3 to 5 $ for every invested $). This study is based on questionnaires filled in by the various actors of the Information Retrieval field, estimating their perceived impact of the TREC campaigns on their activity. It does not analyze the various evaluation criteria that might be taken into account. When discussing evaluation criteria, we will take the position of an industrial organization, which wants to incorporate a technological component into a system under construction. In this process, the organization under consideration can either be an end-user performing in-house implementation of a product organization incorporating a technology into its product.

Most of the time, such an evaluation is likely to be conducted in the presence of competing technological alternatives, which would be the sign of a reasonably mature domain with a stable architectural and component breakdown structure. We believe this to be the case for the multimedia search domain, based on the functional description elaborated during the Chorus project.

This does not exclude an innovative situation where a new technology is introduced as a novel component into a larger system. It should be noted that in this particular case, much less frequent than the previous, scientific evaluation is somewhat at a loss, having no data to compare against. It maintains significant value in setting base values for future evaluation of this technology. On the other hand, business oriented evaluation remains fully useful in the context of innovative technology in its aspects that do not imply comparative evaluation, but try to measure the impact of such a technology on a broader system or market.

The evaluation dimensions taken into account can be regrouped into five categories:

1/ **technical criteria identical to those used by the scientific community**
(See e.g. sections 2.2.1 and 2.2)

2/ **additional technical criteria specific to industrial organizations**
Beyond the classical evaluation criteria, businesses may want to take into account criteria which are technical by nature, but are more connected to implementation rather that function:
- quality of the implementation: can the technology be reused "as is", or does it need to be re-implemented by the importing organization - modularity: is the technology packaged into modules, with appropriate API facilitating its integration into a larger system - scalability: in relationship with point 5 below, it is important to evaluate whether the technology can scale up to the size of its effective target market.

3/ **criteria relative to effective access to the technology**

Provided the technology satisfies the criteria above, is it accessible by an industrial organization? Is there a contractual or licensing environment allowing such an organization to effectively import the technology into a commercial system. Prior to actual licensing, trials and experimentation may be useful for an "in-situ" evaluation of the technology. Does a "trial license" mechanism exist? Beyond the legal aspect of licensing, does a price structure exist?

It is often the case that those evaluation aspects are absent at the time of first contact between a research organization and its potential industrial partner, often creating a "chicken and egg" situation. In terms of practical organization, the existence of a structure capable of actively contributing to the initial licensing and pricing discussions, and later, helping on the matter of technical support of the transferred technology constitutes an additional evaluation criteria that industrial organizations will take into account.

4/ **usefulness of the technology, at the user level, at the system level**

When evaluating a technology with the goal of integrating it into a larger system, industrial organizations will want to take into account the potential impact of such technology on their system. This impact plays at two levels: -User-level, measured in terms of either new functionality, or new user interface. -System-level measured in terms of overall system performance improvement. In both cases, assessing the improvement in user perception or system performance is important as it contributes to the justification the overall cost of acquisition and transfer of the technology.

5/ **marketability of the technology**

The last evaluation domain pertinent for technology evaluation with the aim of transferring it into an industrial organization is its market value. Several aspects come into play: - are the users of the technology well identified, both in terms of end-users and intermediate integrators? Is the target for the technology the general public, or specialist/professionals? The quantitative - what is the value proposition to these users? - as the technology is unlikely to operate in a stand-alone and isolated fashion, - are the data elements on which the technology operates in sufficient volume, and effectively available

3. Conclusion and perspectives

This report was dedicated to the overview and analysis of existing evaluation dimensions with regard to different contexts: scientific literature, international evaluation campaigns, user-centered evaluations and industrial side criteria. It did show that these evaluation dimensions work at different levels of the innovation workflow, all of them with distinct interests for different actors and possible ways to be improved. The relationship between these different dimensions and their relative impact on innovation is however much more difficult to analyze and to understand from existing studies and data. We therefore set up a survey whose aim is to collect feedbacks on these evaluation dimensions and their impact, from both academia and industrial communities.

Practically, a web questionnaire has been designed and distributed to different communities through mailing lists, conferences and social networks. At the time of writing, the targeted audience included ImageCLEF conference participants, ACM multimedia conference participants and a Linkedin group on enterprise search. It will be further disseminated in the future, probably at ICMR conference and among the industrial participants to the next CHORUS+ think-tank. The full questionnaire is reproduced in the annex of this report. Besides preliminary questions about profiles and organizations, the questionnaire follows a top-down approach, starting from business and industrial evaluation dimensions and ending with evaluation criteria used in the scientific literature. The results of the survey and their analysis will be presented in a further report about *evaluation needs* (CHORUS+ deliverable D3.4).

4. References

[AceMedia] http://www.acemedia.org/aceMedia/project/work_breakdown/wp2.html

[Armstrong2009] Timothy G. Armstrong, Alistair Moffat, William Webber, Justin Zobel, Improvements that don't add up: ad-hoc retrieval results since 1998, Proceeding of the 18th ACM conference on Information and knowledge management, 2009.

[Blanco2011] Roi Blanco, Hugo Zaragoza. Beware of relatively large but meaningless improvements. Technical Report YL-2011-011, Yahoo! Research, Barcelona, Spain.

[Bouabid2011] Bouabid, H. "Revisiting citation aging: A model for citation distribution and life-cycle prediction". Scientometrics 88 (1): 199–211, 2011.

[Clarity] Daniela Petrelli, "On the role of User-Centred Evaluation in the Advancement of Interactive Information Retrieval", in Information Processing and Management, 2007

[CHORUSfunctional]http://chorusgapanalysis.wetpaint.com/page/Functional+description+of+a+generic+multimedia+search+engine

[CHORUSplat] http://avmediasearch.eu/wiki

[CHORUSsoa]http://jcpconsult.wetpaint.com/page/SOA+of+existing+benchmarking+initiatives+%2B+who+is+participating+in+what+(EU%26NI)

[Couto2009] Couto, F., Pesquita, C., Grego, T. and Veríssimo, P., "Handling self-citations using Google Scholar", Cybermetrics 13 (1), 2009.

[Croft2009] Bruce Croft, Donald Metzler, and Trevor Strohman. Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company, USA, 2009.

[Dillon and all, 1996] Dillon, A & Morris, M (1996) User acceptance of information technology: theories and models, Annual review of information science and technology, 3-32

[Dou2007] Dou, Z., Song, R. and Wen, J., A Large-scale Evaluation and Analysis of Personalized Search Strategies. in Proceedings of the 16th international World Wide Web conference (WWW2007), (Banff, Alberta, Canada, 2007), 572-581.

[Dumais2003] Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. 2003. Stuff I've seen: a system for personal information retrieval and re-use. In Proceedings of ACM SIGIR 2003. ACM, New York, NY, USA, 72-79.

[Egghe2006] Egghe, L., "Theory and practise of the g-index", Scientometrics 69 (1): 131–152, 2006.

[ElsevierExe] http://www.executablepapers.com/

[evaluateIR] http://wice.csse.unimelb.edu.au:15000/evalweb/ireval/

[Ferro2009] Ferro, N., Harman, D.: CLEF 2009: Grid@CLEF pilot track overview. In: Working Notes of CLEF 2009. (2009)

[Finkelstein2001] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E., Placing search in context: the concept revisited. in World Wide Web, (2001), 406-414.

[google1] http://googleblog.blogspot.com/2008/05/introduction-to-google-search-quality.html

[google2] http://googleblog.blogspot.com/2008/09/search-evaluation-at-google.html

[Hanbury2010] Allan Hanbury, Henning Müller, Automated Component-Level Evaluation: Present and Future, CLEF 2010, Springer Lecture Notes in Computer Science, Padova, Italy, pages 124-135, 2010.

[Hirsch2005] Hirsch, J. E., "An index to quantify an individual's scientific research output", PNAS 102 (46): 16569–16572, 2005.

[Hu2010] Xiao Hu, Jingjing Liu. Evaluation of Music Information Retrieval: Towards a User-Centered Approach, Microsoft technical report, 2010.

[Hull1993] D. Hull, “Using statistical testing in the evaluation of retrieval experiments,” in Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 329–338, New York, NY, USA: ACM, 1993.

[Ingwersen2005] Peter Ingwersen and Kalervo Järvelin. The Turn: Integration of Information Seeking and Retrieval in Context, The Information Retrieval Series. Springer, 2005.

[Leelanupab2009] Leelanupab, T., Hopfgartner, F. and Jose, J.M. (2009) User centred evaluation of a recommendation based image browsing system. In the 4th Indian International Conference on Artificial Intelligence , 2009.

[Martin2004] Martin, I. and Jose, J., Fetch: A personalised information retrieval tool. In Proceedings of RIAO 2004, 2004, 405-419.

[McDonald2001] Sharon McDonald, Ting-Sheng Lai, and John Tait. 2001. Evaluating a content based image retrieval system. In Proceedings of ACM SIGIR 2001. ACM, New York, NY, USA, 232-240.

[MIREX] http://www.music-ir.org/mirex/

[Nielsen1993] Nielsen J. Usability Engineering. Academic Press, Boston, 1993.

[PatOlymp]http://www.ir-facility.org/events/irf-symposium/irf-symposium-2011/patolympics

[PEIPA] http://peipa.essex.ac.uk/

[Radicchi2008] Radicchi, F.; Fortunato, S. and Castellano, C., "Universality of citation distributions: Toward an objective measure of scientific impact", PNAS 105 (45): 17268–17272, 2008.

[Rode2010] Henning Rode, Theodora Tsikrika, and Arjen P. de Vries. Differences in Video Search Behaviour between Novices and Archivists. In Proceedings of the 8th International Workshop on Adaptive Multimedia Retrieval (AMR 2010), August 17-18, Linz, Austria, 2010.

[Rogers, 95] ROGERS, E. (1995). Diffusion of Innovations. (New York: Free Press).

[Rowe2010] B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni. Economic impact assessment of NIST's Text REtrieval Conference (TREC) Program. Technical Report Project Number 0211875, RTI International, 2010.

[Sanderson2010] Mark Sanderson. Performance measures used in image information retrieval, 2010, ImageCLEF book.

[Sanderson2010b] Mark Sanderson. Test Collection Based Evaluation of Information Retrieval Systems, Foundations and Trends in Information Retrieval, Vol. 4, No. 4, pp. 247-375 (2010).

[Seglen11997] Seglen PO. "Why the impact factor of journals should not be used for evaluating research". BMJ 314 (7079): 498–502, 1997.

[Senach1990] B. Senach, “Evaluation ergonomique des interfaces home-machine: une revue de la literature”, Rapport de recherche INRIA n1180, Programme 8, Communication home-machine, mars 1990.

[Serenko2011] Serenko A, Dohan M., "Comparing the expert survey and citation impact journal ranking methods: Example from the field of Artificial Intelligence". Journal of Informetrics 5(4): 629–648, 2011.

[Shackel1991] Shackel, B. (1991). Usability – context, framework, design and evaluation. In Shackel, B. and Richardson, S. (eds.). Human Factors for Informatics Usability. Cambridge University Press, Cambridge, 21-38

[Shen2005] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Context-sensitive information retrieval using implicit feedback. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05).

[SIGMODRep]
http://www.sigmod2010.org/calls_papers_sigmod_research_repeatability.shtml

[Thomas2006] Thomas, P. and Hawking, D., Evaluation by comparing result sets in context. in CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management, 2006, ACM, 94-101.

[Thornley2011] C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee. The scholarly impact of TRECVid (2003-2009). JASIST, 62(4):613-627, 2011.

[Tol2008] Tol, R.S.J. A rational, successive g-index applied to economics departments in Ireland, Journal of Informetrics, vol. 2, pp. 149–155, 2008

[Tricot2003] Tricot A., Plegat-Soutjis F., Camps J.-F., Amiel A., Lutz G., Morcillo A. (2003). Utilité, utilisabilité, acceptabilité : interpréter les relations entre trois dimensions de l'évaluation des EIAH, In C. Desmoulins, P. Marquet et D. Bouhineau (dir.). Environnements informatiques pour l'apprentissage humain, 391-402, Paris : ATIEF – INRP

[Tsikrika2011] Theodora Tsikrika, Alba Garcia Seco de Herrera, Henning Müller, Assessing the Scholarly Impact of ImageCLEF, Springer Lecture Notes in Computer Science (LNCS), CLEF 2011, Amsterdam, The Netherland, 2011.

[Vallet2006] David Vallet, Miriam Fernández, Pablo Castells, Phivos Mylonas, Yannis Avrithis. Personalized Information Retrieval in Context. AceMedia EU project report. 2006.

[VideoBrows] http://mmm2012.org/vbshowdown/

[VideOlymp] http://www.videolympics.org/

[Vitalas] http://vitalas.ercim.eu/

[Vitalas2007] A. Saulnier, M.L. Viaud, P. Altendorf, T. Tsikrika, C. Martinez, N. Boujemaa, A. Joly, J. Geurts, Report on success criteria definition, VITALAS EU project deliverable D1.3.1.

[Wendl2007] Wendl, Michael, "H-index: however ranked, citations need context". Nature 449 (7161): 403, 2007

[White2005] White, R., Ruthven, I. and Jose, J., A study of factors affecting the utility of implicit relevance feedback. In proc. of SIGIR 2005, ACM, 35-42.

[Woeginger2008] Woeginger, G.J. An axiomatic analysis of Egghe's g-index, Journal of Informetrics, vol. 2, pp. 364–368, 2008

[Yu2010] Yu, G.; Li, Y.-J. "Identification of referencing and citation processes of scientific journals based on the citation distribution model". Scientometrics 82 (2): 249–261, 2010.

Annex 1 – CHORUS+ questionnaire on the evaluation of multimedia retrieval technologies

CHORUS+ questionnaire on the evaluation of multimedia search technologies

Your organization and profile

Are you?

○ a student

○ an engineer or a researcher

○ a manager or a director

What is your main activity within your organization?

○ Research

○ Development

○ Management

○ Support

○ Sales

How many years experience do you have in multimedia search?

○ 1 to 4

○ 5 to 10

○ 11 to 20

○ > 20

Your organization is

○ University

○ Research institute

○ Company

What is the size of your organization?

○ 1 to 10

○ 10 to 50

○ 50 to 200

○ 200 to 2000

○ > 2000

What is the main activity of your organization relative to multimedia search?

○ Research

○ Technology inventor supplying intellectual property

○ Design of open software

○ Technology supplier providing technology bricks

○ Multimedia search solution supplier

○ Systems integrator

○ User

Within you organization, how many people are concerned with multimedia search activities?

○ 1 to 10

○ 10 to 50

○ 50 to 200

○ 200 to 2000

○ > 2000

What is the level of awareness of Benchmarking in your organization ?

Please check one or more of the following proposals

□ My organization is aware of public evaluation campaigns such as NIST (TREC, TRECVID), CLEF (ImageCLEF, PAN), MediaEval, INEX, etc.

□ My organisation participates to public evaluation campaigns

□ My organisation participates in the organization of public evaluation campaigns

□ My organization participates to private industrial benchmarks

□ My organization organizes private industrial benchmarks

Within you organization, how many people are directly involved in benchmarking of multimedia search technologies (benchmarking of software, organizing a campaign or other)?

○ 1 to 3

○ 4 to 10

○ 10 to 30

○ 30 to 100

○ > 100

« Back   Continue »

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

CHORUS+ questionnaire on the evaluation of multimedia search technologies

Using benchmarking for technology transfer

Which aspects are likely to contribute to the commercial success of a technical component ?

Check your top 3 criteria among the following, let the other lines empty

	1st choice	2nd choice	3rd choice	ignore/reset
Technical performance	○	○	○	○
Number, quality and reach of functions provided	○	○	○	○
Ease of integration	○	○	○	○
Commercial terms (licensing terms, support, price list, ..)	○	○	○	○
Ease of use	○	○	○	○
Scientific excellence	○	○	○	○

How do you identify new technical components that you would like to experiment and/or benchmark?

Check your top 3 criteria among the following, let the other lines empty

	1st choice	2nd choice	3rd choice	ignore/reset
Via Customer requests	○	○	○	○
By analyzing competitors	○	○	○	○
Via Benchmarking results or evaluation campaign topics	○	○	○	○
Via online or of line press and blogs	○	○	○	○
Via scientific articles	○	○	○	○
By recommendation (colleagues/social networks/etc).	○	○	○	○
Via conferences	○	○	○	○

What criteria do you use for selecting technical components (for experimentation, proof of concept or integration in products)?

Check your top 3 criteria among the following, let the other lines empty

	1st choice	2nd choice	3rd choice	ignore/reset
High impact in scientific literature	○	○	○	○
Results in benchmarks or evaluation campaigns	○	○	○	○

Technical skills (performances, scalability)	○	○	○	○
Author or company	○	○	○	○
Used technologies (languages, OS, etc.)	○	○	○	○
Ease of integration	○	○	○	○
Purchase price	○	○	○	○
Ownership cost	○	○	○	○
Compliance to standards	○	○	○	○
Security	○	○	○	○
Novelty of the functionality	○	○	○	○
Adequacy to user needs	○	○	○	○

« Back Continue »

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

CHORUS+ questionnaire on the evaluation of multimedia search technologies

Public evaluation campaigns

In which public evaluation campaign(s) have you been participating in ?
Please check none or several of the following proposals

□ TREC
□ TRECVID
□ CLEF (PAN, LogCLEF)
□ ImageCLEF
□ MusiCLEF
□ INEX
□ MIREX
□ MediaEval
□ Pascal VOC
□ SHREC
□ Other:

Which evaluation campaign is the most suitable for your business or research activity ?

○ TREC
○ TRECVID
○ CLEF (PAN, LogCLEF)
○ ImageCLEF
○ MusiCLEF
○ INEX
○ MIREX
○ MediaEval
○ SHREC
○ Pascal VOC
○ I don't know
○ None
○ Other:

To your opinion, the challenges measured in public evaluation campaigns are

○ very relevant
○ reasonably relevant
○ not relevant

○ I don't know

To your opinion, the evaluation criteria used in public evaluation campaigns are

○ very relevant

○ reasonably relevant

○ not relevant

○ I don't know

In the future do you plan to

□ submit technologies to evaluation campaigns

□ use technologies selected by campaigns

□ organize campaigns

□ organize new tasks in existing campaigns

« Back Continue »

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms

CHORUS+ questionnaire on the evaluation of multimedia search technologies

Scientific evaluation criteria

What are the best criteria that you think should be taken into account when benchmarking multimedia IR components ?

Check your top 3 criteria among the following, let the other lines empty

	1st choice	2nd choice	3rd choice	ignore/reset
Effectiveness (results relevance, precision,...)	○	○	○	○
Efficiency (response time, speed, etc.)	○	○	○	○
Diversity (data exploration, browsing, etc.)	○	○	○	○
Scalability (large datasets)	○	○	○	○
User satisfaction (user trials, questionnaires, ...)	○	○	○	○
GUI / ergonomy	○	○	○	○

What criteria do you use to judge that a scientific article is an important contribution ?

Check your top 3 criteria among the following, let the other lines empty

	1st choice	2nd choice	3rd choice	ignore/reset
Excellence of authors or journal	○	○	○	○
Citations (amount and growth, cited by who, ...)	○	○	○	○
Contributions claimed by the authors (positioning, originality, ...)	○	○	○	○
Theoretical statements (e.g. on complexity, convergence, bounds, ...)	○	○	○	○
Experiments reported in the paper	○	○	○	○
Experiments reported by third parties (benchmarks, comparative studies, ...)	○	○	○	○
Discussions in conferences, social networks, colleagues, ...	○	○	○	○

What are the greatest difficulties in the scientific evaluation of multimedia retrieval ?

Check your top 3 criteria among the following, let the other lines empty

	1st choice	2nd choice	3rd choice	ignore/reset
Data availability and adequacy (existence,				

scale, metadata, copyrights, etc.)	○	○	○	○
Competitor's methods availability and adequacy (existence, adequacy, licences, etc.)	○	○	○	○
Evaluation protocol (task, metrics, workflow, etc.)	○	○	○	○
Hardware resources (amount of required resources, etc.)	○	○	○	○
Human resources (manual annotations, user trials, etc.)	○	○	○	○

« Back   Submit

Powered by Google Docs

Report Abuse - Terms of Service - Additional Terms