Recommendations about benchmarking campaigns as a tool to foster multimedia search technology transfer at the European level

Deliverable 3.4 – Benchmarking impact and needs

Distribution Level: PU

"The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°249008"

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 3.4
Subject:	Benchmarking impact and needs
Number:	3.4
File name:	D3.4-v0.1.pdf
Registration Date:	
Last Update:	07.07.2012

Revision History

Version	Edition	Author(s)	Date
0	1	Alexis Joly	25/05/12
Comments:	First version based on think-tank5 and consortium discussions		
0	2	Alexis Joly	06/06/12
Comments:	Integration of first contributions and recommendations received by email (Henning Muller, Yannis Kompatsiaris)		
0	3	Alexis Joly	13/06/12
Comments:	Integration of new contributions, feedbacks and email discussions (Henning Muller, Andreas Rauber, Henri Gouraud, Yannis Kompatsiaris)		
0	4	Alexis Joly	20/06/12
Comments:	Internal Review version with last feedbacks integrated (Henning Muller)		
0	5	Alexis Joly	05/07/12
Comments:	Camera-ready version with reviewers feedbacks addressed (reviewers: Pieter Van der Linden, Joost Geurst)		
1	0	Alexis Joly	07/07/12
Comments:	After validation of the scientific coordinator (Yannis Kompatsiaris)		

Introduction

The main objective of this report is to provide recommendations to the European Commission concerning the funding and the steering of public evaluation campaigns as a tool to foster multimedia search technology transfer at the European level. These recommendations rely on the results of 3 complementary preliminary actions conducted within CHORUS+ coordination action:

1- **Report on Multimedia Retrieval evaluation dimensions**: Based on the literature and CHORUS+ consortium experience in the evaluation of multimedia search technologies, a report synthesizing the role of evaluation at different levels of the innovation workflow was published in November 2011 (deliverable D3.3). International benchmarks and evaluation campaigns were identified as a key component that might fill the gap between scientific criteria used by the researchers (to measure fundamental progress) and industrial & business-oriented criteria used by companies to identify relevant technologies and build innovative products.

2- **Survey on the evaluation of Multimedia Search technologies**: To better understand the suitability of benchmarking as a tool to foster exchange between academia and industry, a survey was set up and delivered to both communities. The content of this survey was already published within D3.3 deliverable (November 2011). It was subsequently distributed in December 2011 to several communities including ImageCLEF, TRECVID and MIREX benchmarking campaigns, ACM Multimedia 2011 conference, EU projects of the Media Search cluster and a LinkedIn group on enterprise search. The 20 participants to the 5th Think Tank on *Multimedia search technology transfer driven by benchmarking* were finally asked to fill the survey. The result's analysis of the survey was used as a support of discussion during the ThinkTank and is provided in the annex of this report.

3- **Thinktank on Multimedia Search technology transfer driven by benchmarking**: A *Thinktank* was organized by CHORUS+ on April 19th 2012 during the international conference WWW 2012 in Lyon. This event brought together experts and stakeholders of multimedia search related benchmarking efforts in order to exchange on lessons learned and to assess suitability of benchmarking to foster technology transfer. The results of the above mentioned survey were presented and discussed during the meeting. The question of whether EU should play a role in further strengthening and supporting these efforts was also addressed as an important issue. Around 20 people including leading industrials, expert SMEs, EC representatives and highly known researchers in the multimedia search technology field gathered in Lyon to discuss these subjects. The notes synthesising the whole discussions of this Thinktank are available as one CHORUS+ deliverable (D5.2.5, *Think-tank 5 Meeting Notes*).

The content of this report is organized in two main sections. The first part (section 1) synthesized the main conclusions and lessons learned from the three actions described above. The second part provides the recommendations of CHORUS+ consortium towards sustaining and/or improving European practices concerning public benchmarking.

1. Main conclusions and lessons learned about benchmarking campaigns

Benchmarking campaigns are suitable to foster exchange between academia and industry. Challenges measured in benchmarking campaigns are overall judged as relevant by both academia and industry. The motivations differ from an actor to another (e.g. between SME's, big companies and research institutes) but each of them finds an interest to participate in or to follow the results of benchmarking campaigns. Identified benefits for these actors include: (i) measuring and boosting global research progress (ii) increasing the visibility of good research (iii) facilitating access to evaluation data (iv) facilitating the emergence and the sustainability of research communities (v) fostering the convergence of evaluation methodologies (vi) fostering the emergence of private benchmarks modeled on public ones but using business-specific data

Benchmarking campaigns have a positive scientific, technical and economic impact. NIST (one of the largest evaluation campaign organized by the US National Institute of Standards and Technology) has measured significant technical, industrial and scientific impact of its campaign. In particular the TRECVID campaign has allowed to double performances of systems over 3 to 10 year span (depending on the topic). According to a study of RTI International return on investment reached a factor 3 to 5[1]. And finally TRECVID has generated more than 2000 publications. Similarly significant technical and scientific impact has also been described in publications about other campaigns.
The results of CHORUS+ survey as well as the discussions during the Thinktank confirmed that benchmarking campaigns became an important tool for companies to identify relevant research progress and select new technologies for their products. An increasing interest for participating in, and, organizing benchmarking campaigns in the future was measured in both academia and industry.

Benchmarking campaigns are criticized in some points. An important criticism is the implicit cost to participate in an evaluation campaign. Up to 10 additional man months over usual R&D costs are required to participate in an evaluation campaign for the first time. Even if this cost decreases for further participations, this expensive entry price has a negative impact on the participation of SME's as well as many research groups world wide. Another frequently mentioned shortcoming is related to the scale and scope of data used for benchmarking. Shipping real-world and big data is indeed logistically very difficult and limited by access rights. The consequence is that systems might converge to ad-hoc solutions and therefore generalize poorly when transferred to real-world content. A last criticism concerns the way technologies are evaluated in benchmarking campaigns, and notably the controversial question of user-centered vs. system-oriented evaluation. Some actors from both academia and industry complain that end-users of the technologies are not involved enough in the evaluation process. The large companies who participated in our survey particularly identified this point as critical. On the other side, user-centered evaluations strongly increase the evaluation cost and are suspect of being more subjective.

[1] Please refer to: http://trec.nist.gov/pubs/2010.economic.impact.pdf

There is a lack of support to the organization of benchmarking campaigns in Europe. There is no dedicated funding in Europe to sustain the organization of public benchmarking campaigns at the international level. Large initiatives such as CLEF or MediaEval typically live through heterogeneous and opportunistic research funds including national and European projects, and volunteer resources from research institutes. In this context it appears particularly difficult to assess the impact of campaigns over longer periods (5-10 year area). On the other side, the American National Institute of Standards and Technology is in charge of organizing most benchmarking campaigns in US with significant permanent resources (complemented by contributing external researchers). There was a consensus during the Thinktank on that Europe should not simply leave the floor to NIST (for several reasons related to scientific, cultural and social diversity as well as economic strategy). As a result of its central role on stimulating research and innovation in Europe, the EU commission appears as a highly recognized candidate to efficiently set up and support a sustainable and efficient way to fund and synchronize benchmarking campaigns in Europe.

Steering of benchmarking campaigns is controversial. Selecting and synchronizing scientific challenges measured in public benchmarking campaigns is a complex process sensitive to impartiality and biases. In the US, NIST employs several mechanisms: sometimes the challenge is defined by an agency, and in other cases the challenge is defined collectively by the research community. Most European benchmarking campaigns such as CLEF and MediaEval are based on a bottom-up mechanism. New challenges are proposed by individual research groups or research projects, and the organizers of previous campaigns decide collectively whether this new task should be integrated. There does not exist a specific mechanism to synchronize the campaigns between each other's. Some participants to the Thinktank rather suggested a top-down approach where the challenges would be defined by public agencies. A EU based effort should concentrate on evaluating results that are funded by EU funds. Some other participants tempered this approach to avoid adding a layer of bureaucracy and to avoid fragmentation of research evaluation.

2. Recommendations

According to the conclusions and lessons learned about multimedia search technology benchmarking, CHORUS+ consortium believes that a more sustainable and efficient way to fund and synchronize benchmarking campaigns in Europe is required. Here is a list of recommendations that go in that direction:

Ensuring transparency, sustainability and efficiency of benchmarking campaigns funding. As long as EU benchmarking campaigns rely on opportunistic and unaccountable funds, efficiency won't be measurable (by both the funders and the organizers of these campaigns). EU funding for the organization of benchmarking campaigns should therefore be more centralized and conditioned to a clear budget and work plan (for instance through specific calls for projects or through a dedicated EIT service similarly to NIST). The additional cost will be compensated by the reduction of the current costs (that are split over several projects) and by an overall efficiency gain.

Ensuring that benchmarking campaigns steering is balanced. The definition of the challenges measured in benchmarking campaigns has to be done collectively by the research community, the industry and the authorities. It is in particular crucial to keep an important place for innovation diversity by ensuring that new task proposals come from both the research community and the industry. Acceptance mechanisms should also rely on a balanced pool of experts.

Ensuring that the costs to participate in benchmarking campaigns are eligible for funding in EU research projects. This would help covering the additional engineering costs required to participate in benchmarking campaigns and foster the participation of small organizations (PME's, small research groups, etc.).

Encouraging participation in benchmarking campaigns. Benchmarking campaigns are a tool to boost technological progress and foster exchanges between industry and academia. They should not be considered as a way to rate companies or research groups. The technical performances measured in these challenges are actually reflecting only partially the scientific excellence of the underlying works or the quality of the tested products. Conditioning funding to benchmarking results should in particular be avoided. The simple fact that an organization participates to a campaign will stimulate results and motivation to go ahead. Successful organizations are free to communicate on their results as an argument of scientific excellence or for advertising their products. Allowing anonymous submissions could be a rather good means to increase participation of companies. Companies have been reported to defer participation because they fear bad publicity in case of poor results. Opening up the number and reach of the participants will mechanically foster technology transfer (e.g. : Companies may take over ideas and algorithms from academic teams ranked better than them).

Besides these structural recommendations, CHORUS+ consortium also would like to highlight two key objectives towards improving current practices in benchmarking:

Moving to larger and real-world data. The consequence of too small or too narrow data is that technologies generalize poorly when transferred to real-world content. This gap between the performances measured in benchmarking campaigns and what can be expected at scale-one is weakening technology transfer. Integrating new technologies in large infrastructures without enough guaranties on performances is actually too risky for many industrials.

Allowing user-centric and external evaluations. System-oriented evaluation metrics used in current benchmarks are essential but not sufficient to cover a vast range of usage of the evaluated technologies. Furthermore, evaluation methodologies are often not scalable because of the huge human work required to build appropriate evaluation data. Complementary to current practices, a good evaluation framework should allow other research groups, companies or even end-users to evaluate a technology with their own criteria or in the context of their own workflow.

These two objectives are actually conditioned to more general concerns in the multimedia research community: data openness, availability of large-scale infrastructures and technology sustainability. CHORUS+ recommendations towards achieving these objectives therefore go beyond benchmarking issues but we believe making such recommendations can help converging to solutions:

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

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

Ensuring sustainability of technologies built within EU funded projects. When not exploited commercially, many relevant technologies built within EU projects are lost. New projects often re-develop the same piece of work and this results in a large waste of time and money. Ensuring the sustainability of the technical components developed in EU projects is therefore crucial. Our recommendation is that the developed components should be

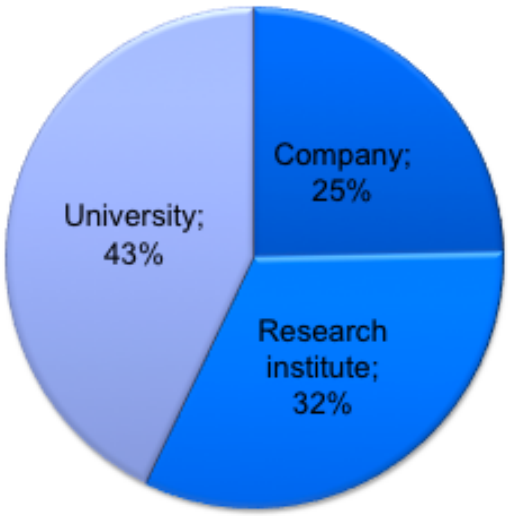
either commercially exploited or shared (with new EU projects and/or with the research community). A moratorium could be applied for making thinks easier, notably for industrial partners: any results may be locked up for one or two years, but then should be shared if no commercial exploitation occurred. An open infrastructure such as the one discussed previously could make such sharing easier.

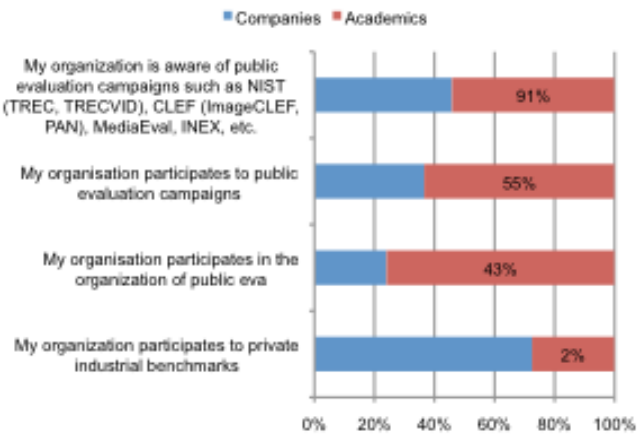
Annex 1 – Results of CHORUS+ survey on the evaluation of multimedia retrieval technologies

CHORUS+ survey on the evaluation of multimedia retrieval technologies, described in D3.3 deliverable (November 2011), was published as an online questionnaire[2] in December 2011. It was then advertised across several communities including ImageCLEF, TRECVID and MIREX benchmarking campaigns, ACM Multimedia 2011 conference, EU projects of the Media Search cluster and a LinkedIn group on enterprise search. The 20 participants to the 5th Think Tank on *Multimedia search technology transfer driven by benchmarking* were finally asked to fill the survey as well. At the time of compiling the results (mid-April 2011), 80 respondents had filled the online questionnaire.

Respondent's profiles

Organizations

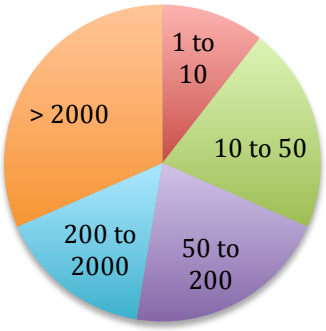




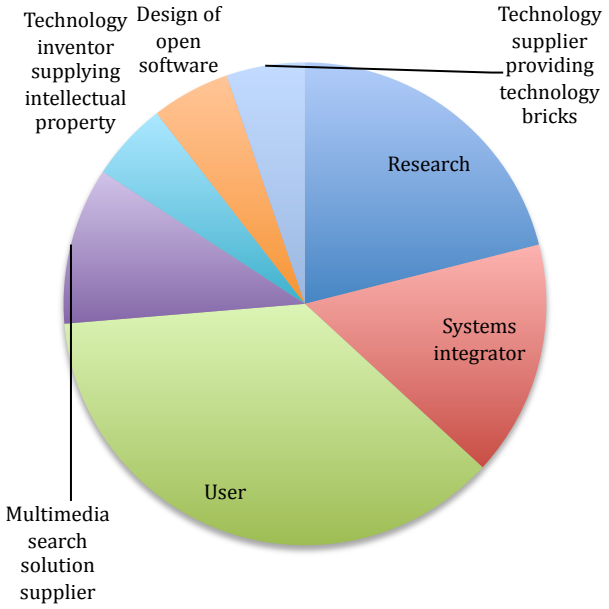
A large fraction of the respondents come from academia (75%). In the detailed result's analysis we refer to this group as **academics**. The remaining 25% coming from industry still represent a tolerable number of respondents (20 companies). A great majority of respondents from academia are *benchmarking aware.* Less than half of the respondents from industry are aware about public benchmarking campaigns.

[2]https://docs.google.com/spreadsheet/viewform?formkey=dEJodDJPRG1HcW94NHNfOUgxcXBBYWc6MQ#gid=0

Companie's profiles



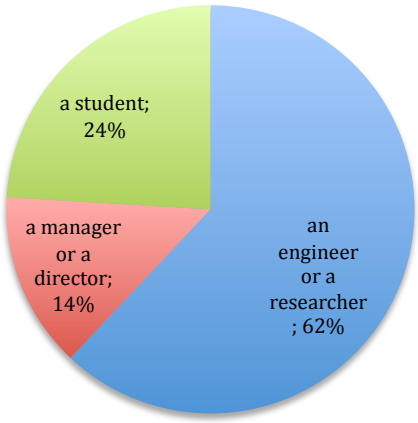
Companies Size



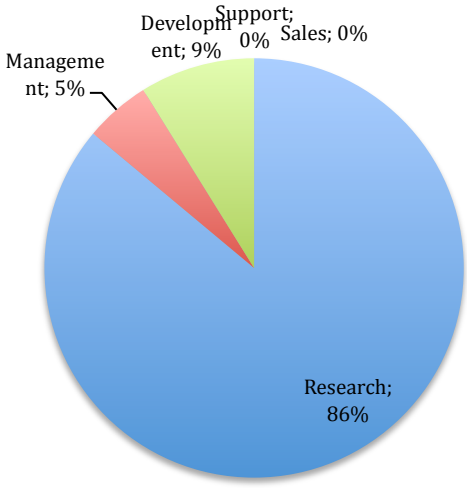
Companies activity

Small, intermediate and big companies are well represented. In the detailed result's analysis we refer to **small companies** as the one having up to 50 salaries and to **big companies** as the one having more than 200 salaries.

People's profiles



Are you?



Your main activity?

The main activity of 86% of the respondents is research but at different levels of responsibility (62% are researchers or engineers). In the detailed result's analysis we refer to the 9% of respondents whose main activity is development as **developers** and to the 14% of

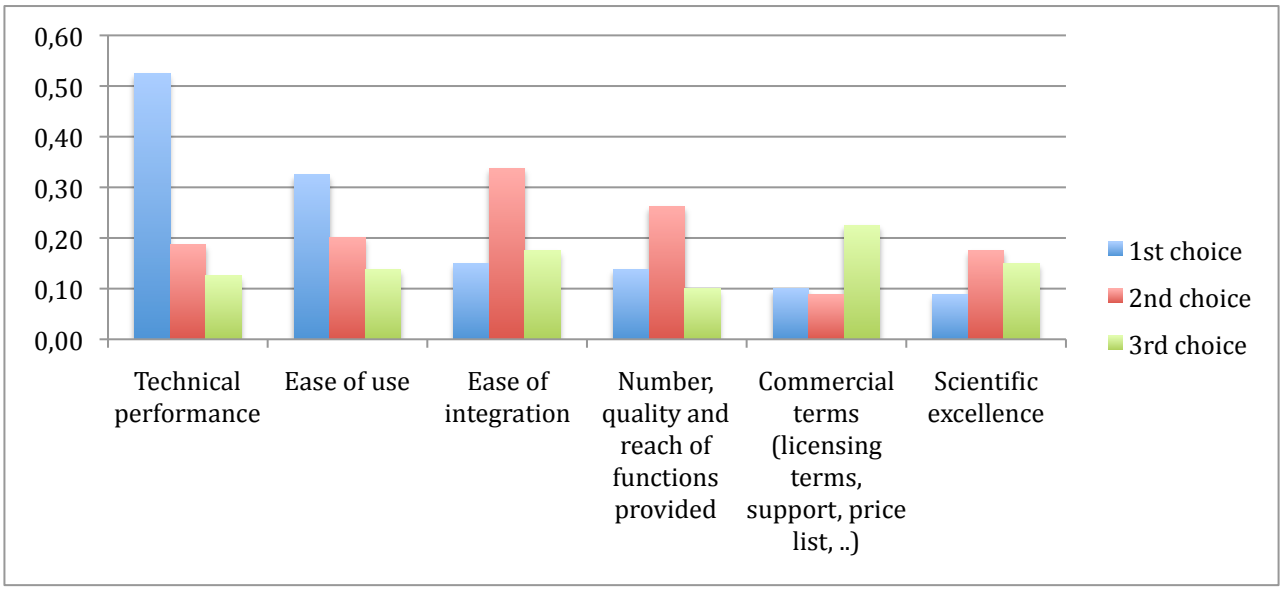
respondents who are either manager or director as **managers**. The group referred as **student** in the detailed results is simply composed by the 24% of the first diagram.

PART1 of the questionnaire

Using benchmarking for technology transfer

Q1.1 Which aspects are likely to contribute to the commercial success of a technical component ?

Overall results



Detailed results

	Big comp	**Small comp**	**Develop**	**Accadem**	**Manager**	**Student**
Top-1	Technical performance	Technical performance	Technical performance	Technical performance	Technical performance	Technical performance
Top-2	Nb & Quality of functions	**Ease of integration**	**Ease of integration**	Ease of use	Ease of use	Ease of use
Worst	Commercial terms	Scientific excellence	Scientific excellence	Scientific excellence	Commercial terms	Scientific excellence

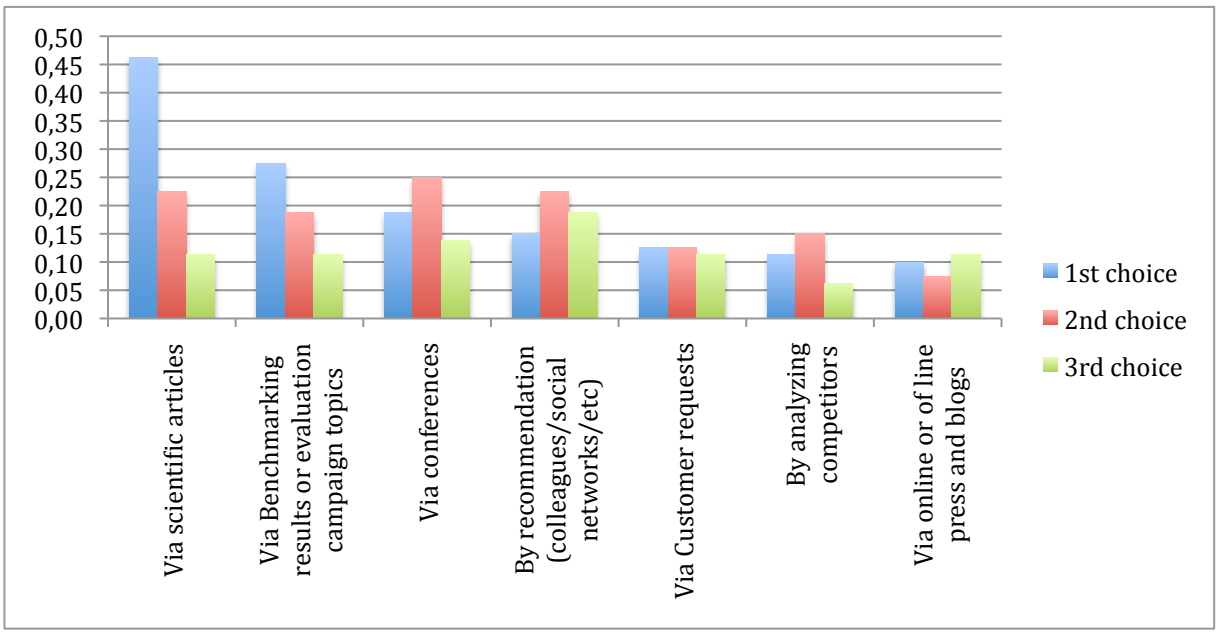
The results show that according to all groups of respondents, **technical performances** is the most important aspect that contributes to the commercial success of a technical component. **Ease of integration** is an important aspect for small companies and developers in general.

Interestingly *commercial terms* appear to be the less contributive aspect for big companies whereas reciprocally, *scientific excellence* is considered as the less contributive aspect by academics.

Scientific excellence is overall the less contributive factor to commercial success.

Q1.2 How do you identify new technical components that you would like to experiment and/or benchmark?

Overall results



Detailed results

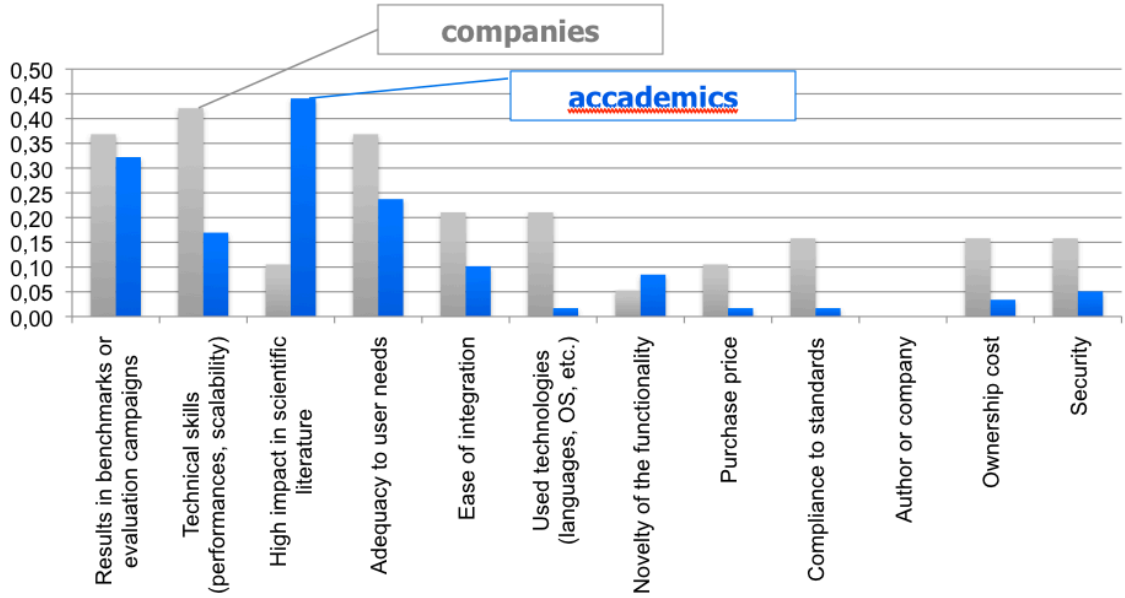
	Big comp	**Small comp**	**Develop**	**Accadem**	**Manager**	**Student**
Top-1	Scientific articles	Scientific articles	Scientific articles	Scientific articles	Scientific articles	Scientific articles
Top-2	Benchmarking, Competitors	==	==	Benchmarking	Benchmarking	**Recommend, conferences**
Worst	Press & blogs	Competitors	Competitors	Press & blogs	Competitors	Competitors

The results show that according to all groups of respondents, **scientific articles** are the best way to **identify** new technologies, followed by **benchmarking campaigns**.

Interestingly, *watching competitors* is a good source of information according to big companies whereas small companies and academics ranked it the worst criterion.

Q1.3 What criteria do you use for selecting technical components (for experimentation, proof of concept or integration in products) ?

Overall results



Detailed results

	Big comp	**Small comp**	**Develop**	**Accadem**	**Manager**	**Student**
Top-1	Adequacy user needs	Technical skills	Technical skills	Scientific impact	Adequacy user needs	Scientific impact
Top-2	Technical skills, **Benchmarking**	**Benchmarking results**	Adequacy user needs	**Benchmarking results**	**Benchmarking results**	**Benchmarking results**
Worst	**Novelty**	Security	Purchase price	Security	Security	Security

Overall results show that **benchmarking campaigns** are on average the best criteria to **select new technical components** for integration or deeper testing (whereas in previous question scientific articles were judged as the best way to identify/discover new components). But it is important to notice that benchmarking campaigns are ranked as the second best criteria by almost all groups when looking at the detailed result's table. **Top-1 criteria for academics** is actually **scientific impact** whereas the **top-1 criteria for companies** is **technical skills** (e.g. scalability, response times, portability, etc.). In between, **benchmarking appears as the best compromise between research (technology suppliers) and exploitation (technology integrators).**

PART 1 Synthesis

- Scientific literature is the best way to prospect and **discover new technologies**

- **Technical performances are the best key of commercial success** whereas scientific excellence is judged as the worst one

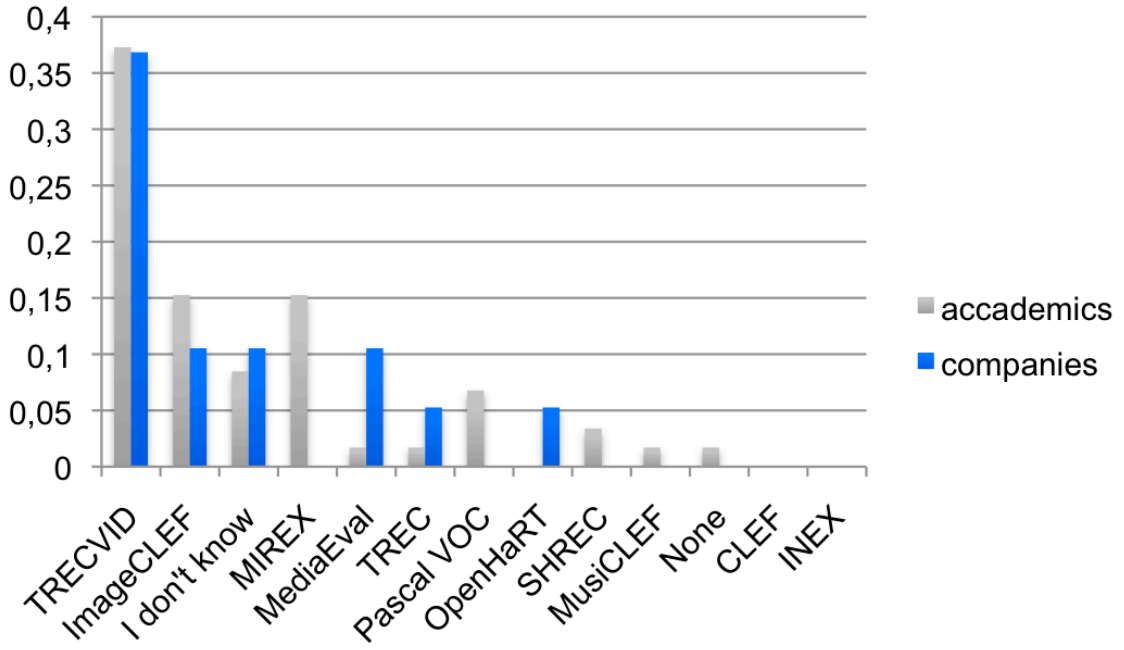
- **Academics & Companies differ on how they select technologies in practice** (for integration or testing): **scholarly impact vs. technical skills**

- But **Academics & Companies agree on** that **Benchmarking** is a good way to select technologies in practice. So that **benchmarking** appears as the **best compromise** between research and exploitation**. This central position** makes it a powerful tool for **boosting technology transfer.**

PART2 of the questionnaire

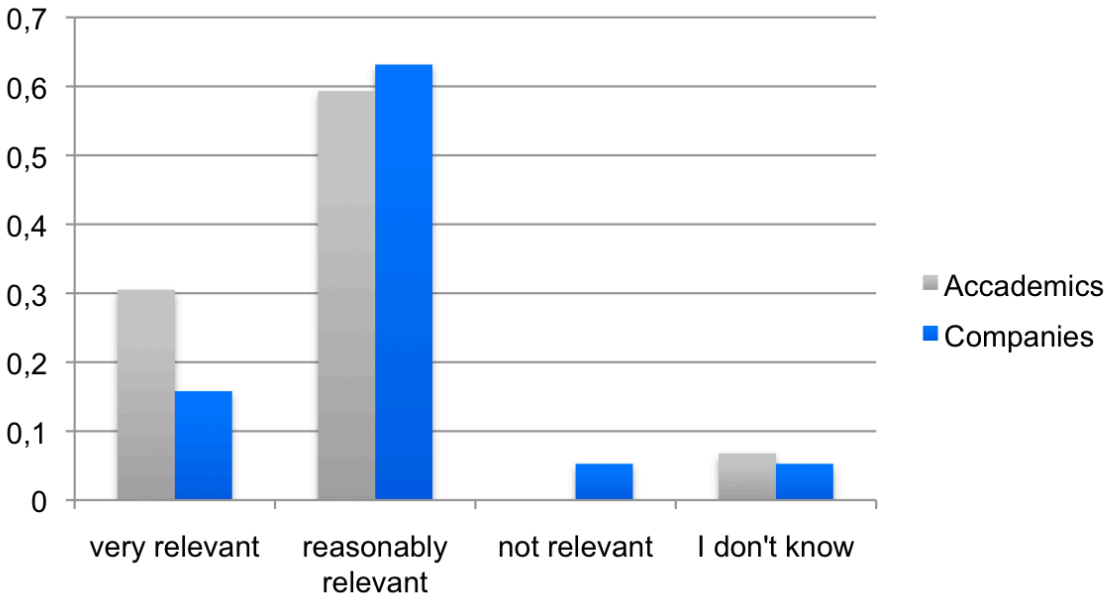
Public Evaluation campaigns

Q2.1 Which evaluation campaign is the most suitable for your business or research activity ?

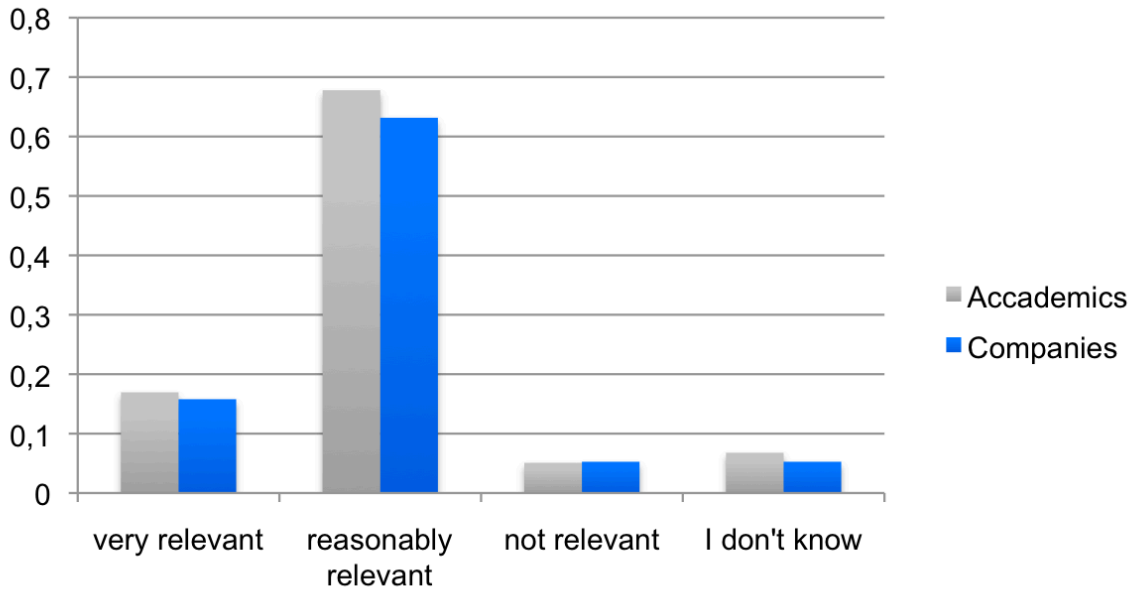


For both academics and companies, TRECVID is far away the most suited evaluation campaigns for their business or research activity, followed by ImageCLEF, MIREX and MediaEval. Notice that these results might be biased by the proportion of respondent's coming from the TRECVID community. But still, according to other statistics on the number of participants to these different campaigns (provided in D3.3) it is highly believable that TRECVID is the most popular one. In 2011 for instance, the number of participants was 73 at TRECVID, 43 at ImageCLEF, 40 at MIREX, 39 at MediaEval, 25 at PASCAL VOC, 15 at SHREC.

Q2.2 To your opinion, the challenges measured in public evaluation campaigns are

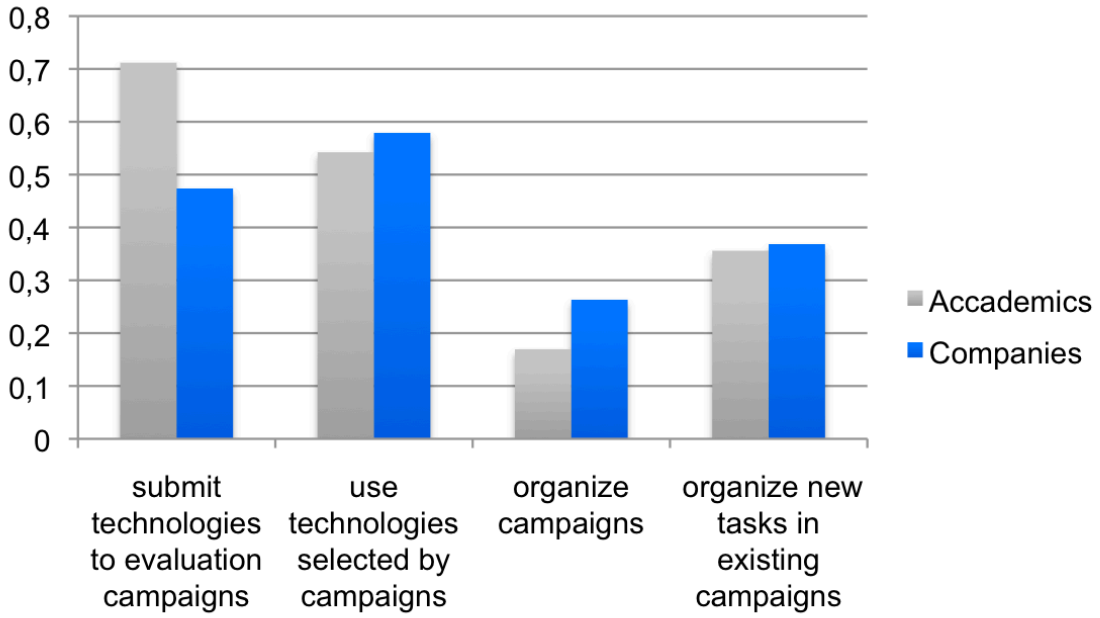


Q2.3 To your opinion, the evaluation criteria used in public evaluation campaigns are



Both academics and companies consider that challenges measured in public evaluation campaign as well the used evaluation criteria are reasonably relevant and very relevant for about 20% of them.

Q2.3 In the future do you plan to



An important conclusion of this graphic is that almost **60% of the companies** who responded to the questionnaire **plan to use technologies selected** as the best ones **within benchmarking campaigns**.

Future intentions of respondents about their participation in benchmarking campaigns show a stable interest compared to actual participation (45% of respondent's companies and 70% of respondent's academics did participate in a campaign in a the past).

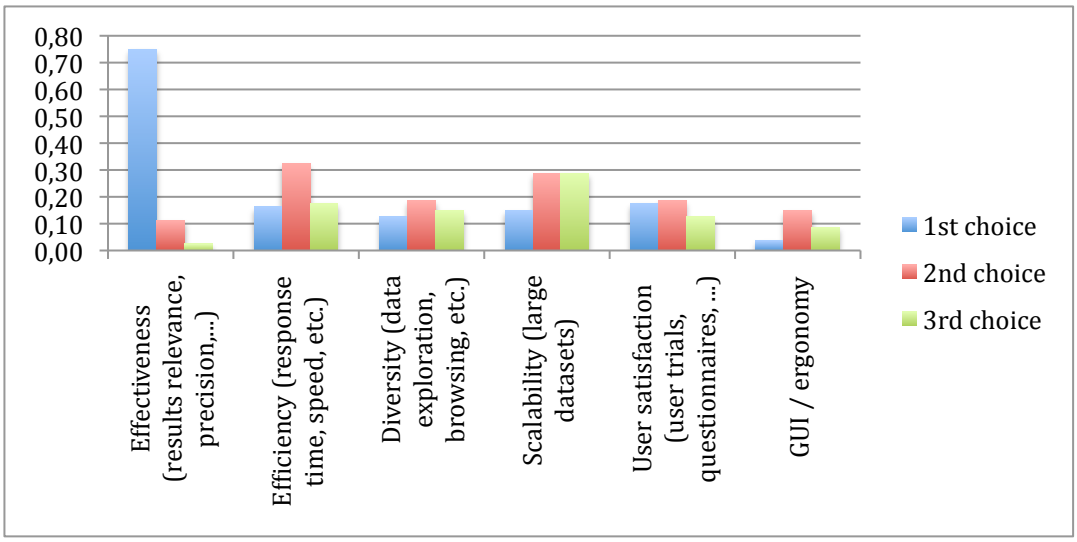
PART 2 Synthesis

- There is an **agreement on the relevance of existing benchmarks**
- **60% of companies plan to use technologies selected by benchmarks**
- **Attractiveness** to participate in and organize public benchmarks **is still there**

PART3 of the questionnaire
Scientific Evaluation Criteria

Q3.1 What are the best criteria that you think should be taken into account when benchmarking multimedia IR components ?

Overall results



Detailed results

	Big comp	**Small comp**	**Develop**	**Accadem**	**Manager**	**Student**
Top-1	Effectiveness	Effectiveness	Effectiveness	Effectiveness	Effectiveness	Effectiveness
Top-2	Scalability	Efficiency	Scalability	Scalability	Scalability	Efficiency
Worst	**GUI/ergonomy**	User satisfaction	Diversity, exploration	**GUI/ergonomy**	**GUI/ergonomy**	**GUI/ergonomy**

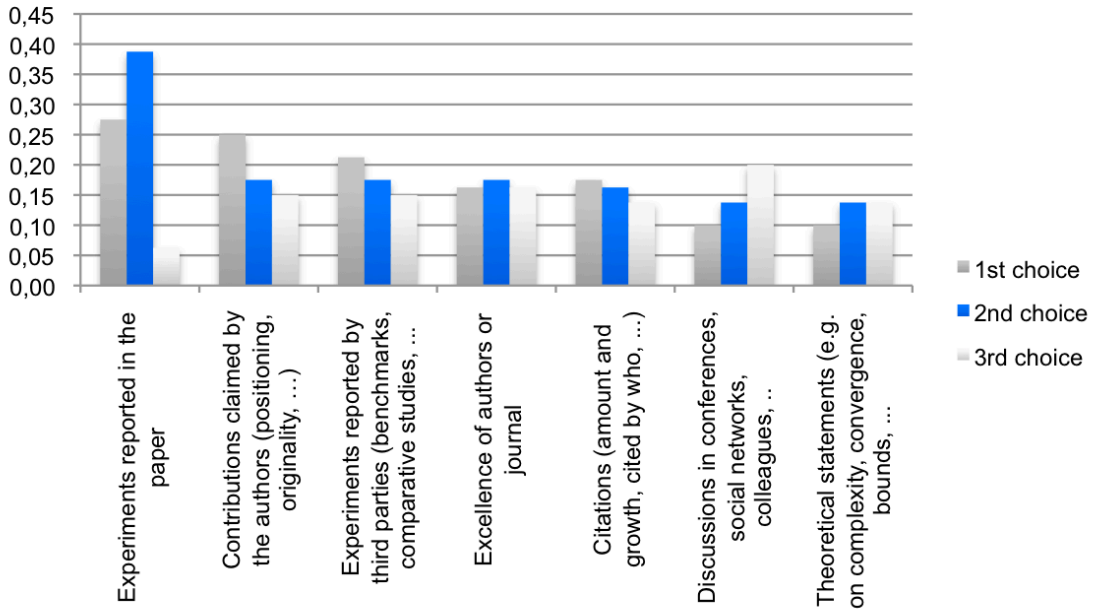
User satisfaction criteria alone

In top-2	**0,56**	**0,20**	0,43	0,36	0,33	0,37

The results show that according to all groups of respondents **effectiveness** is the best criteria to evaluate multimedia IR components, followed by **scalability** and **efficiency**. *Ergonomy* of the Graphical User Interface is not considered an important criterion in such evaluations. Looking at the user satisfaction criterion alone, we see that a majority of **big companies** would like to see **some user trials** in benchmarking campaigns.

Q3.2 What criteria do you use to judge that a scientific article is an important contribution ?

Overall results



Detailed results

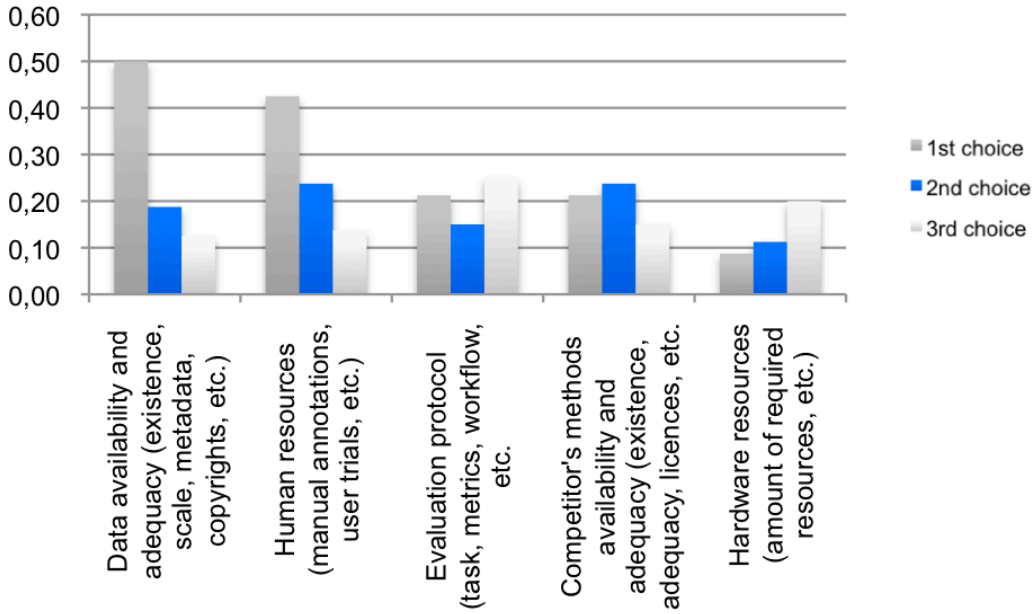
	Big comp	Small comp	Develop	Accadem	Manager	Student
Top-1	**Scientific excellence, citations**	**Experiments**	**Experiments**	**Experiments**	**Scientific excellence, citations**	**Claims**
Top-2	**Third parties expes**	**Claims**	**Claims**	**Claims**	**Experiments**	**Experiments**
Worst	**Claims**	**Theoretical statements**	**Discussions in cinferences**	**Theoretical statements**	**Theoretical statements**	**Discussions in cinferences**

Scientific excellence and biblio-metrics (number of citations, H-index, etc.) are ranked first by managers and big companies. On the other side, experimental results are ranked first by small companies, developers and academics. Finally, claims of the authors of a paper are ranked first by students whereas they are among the worst criterion for big companies and managers. Overall, we can remark that:

- Confidence in claims decreases with financial impact of the respondents
- Confidence in research community increases with with financial impact of the respondents
- Relevance of experimental results is quite stable over the different groups and on the average the best criterion

Q3.3 What are the greatest difficulties in the scientific evaluation of multimedia retrieval ?

Overall results



Detailed results

	Big comp	Small comp	Develop	Accadem	Manager	Student
Top-1	**Data**	**Data**	**Data**	**Data**	**Data**	**Data**
Top-2	**Human resources**	**Evaluation protocol**	**Human resources**	**Human resources**	**Human resources**	**Human resources**
Worst	**Harware resources**	**Harware resources**	**Harware resources**	**Harware resources**	**Harware resources**	**Harware resources**

The results clearly show that according to all groups of respondents **data availability is most critical issue** in evaluating multimedia retrieval technologies. Human resources appear as the second main limitation whereas hardware resources do not appear as a problem. This last point has to be mitigated by the fact that the scale of currently available data is relatively small compared to real-world data. So that if the main limitation (data availability) was solved, it is probable that hardware limitations would become more critical to process (to process very large amount of data).

PART 3 Synthesis

- **Effectiveness** is considered as **the top-1 evaluation criterion** and this validates the approach of current benchmarking campaigns. It is followed by **scalability** and **Efficiency** concerns. Only big companies are convinced by **human-centered evaluation** as a complementary criterion to be used in benchmarking campaigns.
- Criteria used to evaluate scientific publications are diverse and evolve with the financial impact of the underlying decisions to be taken. **Experimental results** are the **most consensual criteria**.
- **Data availability is most critical issue** in evaluating multimedia retrieval technologies