THE IMPACT OF MIREX ON SCHOLARLY RESEARCH (2005 – 2010)

Sally Jo Cunningham
University of Waikato
Hamilton, New Zealand
sallyjo@cs.waikato.ac.nz

David Bainbridge
University of Waikato
Hamilton, New Zealand
davidb@cs.waikato.ac.nz

J. Stephen Downie
University of Illinois
Urbana-Champaign, USA
jdownie@illinois.edu

ABSTRACT

This paper explores the impact of the MIREX (Music Information Retrieval Evaluation eXchange) evaluation initiative on scholarly research. Impact is assessed through a bibliometric evaluation of both the MIREX extended abstracts and the papers citing the MIREX results, the trial framework and methodology, or MIREX datasets. Impact is examined through number of publications and citation analysis. We further explore the primary publication venues for MIREX results, the geographic distribution of both MIREX contributors and researchers citing MIREX results, and the spread of MIREX-based research beyond the MIREX contributor teams. This analysis indicates that research in this area is highly collaborative, has achieved an international dissemination, and has grown to have a significant profile in the research literature.

1. INTRODUCTION

In this paper we report on the results of a study investigating the scholarly impact of the Music Information Retrieval Evaluation eXchange (MIREX), an annual formal evaluation of MIR systems and algorithms. A detailed examination of the structure of the MIREX trials and the results of the initial three years of the MIREX program is presented in [2]. In this present work, we look back on the MIREX publication literature to develop a rich picture of patterns of publication, collaboration, and dissemination of MIREX research (Section 3). Our analysis is based on a set of MIREX-related publications gathered via Google Scholar (Section 2). Issues encountered in building our MIREX document set indicate the existence of barriers to the dissemination of MIREX results. These issues are further explored in Section 4, where we also describe proposals to reduce these barriers—specifically, by providing a digital library of MIREX extended abstracts (thereby pulling the scattered abstracts together into a single repository that supports searching and browsing), and by recommending the development of referencing conventions for MIREX-related documents, datasets, and evaluation frameworks.

2. BIBLIOGRAPHIC DATA GATHERING

In this present paper, the impact of the MIREX trials is measured through both the number of MIREX-related papers published and the number of times that these papers have been cited. The MIREX publications include both the brief descriptions of the MIREX algorithms submitted to a given trial (referred to in the MIREX trials as 'extended abstracts') and the papers derived from the MIREX extended abstracts and MIREX results. As relying solely on sheer quantity of papers has obvious drawbacks, additional analysis focuses on the citation counts to round out the picture by indicating the degree to which each publication "has made a difference" [8] [9].

Three document sources have been commonly used in previous bibliometric studies: the ISI Web of Science (Thomson Reuters), Scopus (Elsevier), and Google Scholar (Google). The three have very different collection policies. The differences most significantly impacting this present study are that ISI restricts its computer science conference proceedings coverage more heavily than the other two; Scopus provides a more comprehensive coverage of both publishers and what they term 'quality web sources' than ISI; and Google Scholar includes the majority of the ISI and Scopus offerings as well as books, technical reports, and white papers.

In choosing Google Scholar as the source for this present study, we were influenced by issues of *coverage* and *user preference*. MIREX-flavored research is based strongly in computer science and engineering, two fields that place a greater emphasis on conference publications and technical reports than other sciences—and both ISI and Scopus do not include these publications types to the extent of Google Scholar [3] [4]. As the MIREX extended abstracts are not formally published, they are not included in the ISI and Scopus databases, and so their impact could not be measured through those resources. Further, we are specifically interested in exploring the documents most readily visible from the viewpoint of researchers interested in MIREX (rather than obtaining comprehensive coverage by hunting down MIREX related publications through all possible sources). For a given topic, Scopus, ISI, and Google Scholar are each likely to cover some content unavailable to the other two. Google and Google Scholar are the resources of preference for researchers in computer science and other science fields

[5]—and so by basing this study on documents drawn from Google Scholar, we build up a picture of the world of MIREX research that more closely resembles the viewpoint of MIREX researchers.

As citations must build up over time, we restricted the scope of this study to the years 2005 – 2010 rather than coming up to date (with the expectation that the 2009 and 2010 will show a 'Groos droop' [7]—the noticeable 'droop' in the right hand tail of the distribution—as citations are still accumulating for these later years). Each year was individually searched by using the date restriction facility in Google Scholar Advanced Search, and the search criterion used was "MIREX AND music" ('Mirex' is also a widely used insecticide, and so a further restriction to the music domain was necessary to filter out agricultural research).

Each paper in these initial results sets was then examined to gauge its relevance to this study. To be included in the study, the paper had to use / reference the results of a MIREX trial, a MIREX technique, MIREX data, or MIREX software. MIREX extended abstracts present in the Google Scholar results were also retained (extended abstracts not available through Google Scholar were not included in this study). Papers only tangentially related to MIREX were eliminated (for example, papers mentioning the MIREX trials as one example among many of retrieval evaluation exercises). Further documents were culled because they were not formal research papers (for example, undergraduate student assignments). Documents that were not publicly available were, when possible, downloaded for examination through the researchers' university library facilities (for example, the ACM publications). Some papers were not readily accessible, and for these the abstract and search snippet were examined; if these did not indicate a significant relationship to MIREX then they were also eliminated. Finally, duplicates were identified and merged (citation counts for copies were added together). Table 1 shows the document counts for both the raw and cleaned datasets.

	2005	2006	2007	2008	2009	2010
Raw	74	154	186	246	281	330
Cleaned	64	87	131	139	134	196

Table 1. Number of documents in the initial search results (raw) and final datasets (cleaned).

For each document retained, we recorded: author names, authors' institutional affiliations, title, abstract, publication type (journal article, book chapter, conference paper, thesis, technical report), abstract, source (eg, conference name), and citation count. As Google Scholar provides only the raw citation count, we were not able to filter for self-citations. Not all of this metadata was available for every document; specifically, a small number of institutional affiliations were absent and so the analyses of author geographic distribution and collaboration (Section 3.6) may be slight underestimates.

3. ANALYSIS OF MIREX PUBLICATONS

This section examines the impact of MIREX through publication and citation counts, the extent of collaboration within the MIREX research community, and the geographic distribution of MIREX research efforts.

3.1 MIREX Publication Set

For 2005 - 2010 we identified a total of 752 publications: 236 MIREX extended abstracts, and 516 more formal publications based on the MIREX trials and results (Table 2). Theses and dissertations are treated separately in Section 3.2. Note that this dataset does not provide exhaustive coverage of either category, and coverage of the MIREX abstracts in particular is patchy when viewed through the lens of Google Scholar. We return to this point in Section 4 with an explanation of this phenomenon and a partial solution to the relative invisibility of some MIREX documents.

Table 2 shows an overall increase in the number of MIREX-derived publications—a ten-fold increase in the first three years of the trials, and another large increase in 2010. The MIREX trials are clearly seen by the research community to have value, as expressed through the growth of literature that builds on MIREX.

However, MIREX extended abstracts can be seen to receive relatively fewer citations than the publications deriving from the MIREX trials (and even at that, the citation average for MIREX extended abstracts is in most years heavily skewed by one or two abstracts that received large numbers of citations). In contrast, a comparable analysis of the TRECVid (video retrieval) [8] [9] and ImageClef (image retrieval) [10] evaluations show the papers for those evaluation trials to have a similar citation profile to their respective derived literature. Again, in Section 4 we explore possible reasons for the lower citation counts and offer a tactic to counter this effect.

The h-index is a measure that attempts to encapsulate both the quantity and visibility of a set of publications [1]. It is calculated as the number *h* that is the largest number of papers in the set that have each received at least *h* citations. In Table 2 we see further evidence that the MIREX-derived publications have a far higher profile than the MIREX extended abstracts; in a given year the h-index for the derived publications is roughly three to four times higher than that of the extended abstracts.

3.2 MIREX-derived Publications: Publication Types

The derived papers are published formally as chapters in edited books, as conference papers, and in journals, and are less formally made available as technical reports. The publication venues follow the profile typical of computer science and engineering: there is a greater emphasis on conference than on journal publications, with a smaller

Year	MIREX extended abstracts				MIREX derived publications			
	No.	Citations	Mean citations	h-index	No.	Citations	Mean citations	h-index
2005	54	302	5.59	10	10	358	35.80	–
2006	36	226	6.28	6	51	1308	25.65	20
2007	33	242	7.33	9	98	1453	14.83	21
2008	38	99	2.61	6	101	1754	17.37	22
2009	33	34	1.03	3	101	802	7.94	14
2010	42	35	0.83	3	155	914	5.90	14

Table 2. Overview of citation data, 2005 – 2010**.**

(but not completely negligible) number of book chapters and technical reports (Table 3).

The MIREX annual results are reported through a special session in the ISMIR conference, and ISMIR is the focal conference for music retrieval research—so it is to be expected that ISMIR would be a significant publication venue for the MIREX-derived research. As Table 4 illustrates, once past the inaugural year over three quarters of the MIREX-derived papers are published outside of ISMIR, and that spread to other conferences and journals increased in the final year of this present study.

	2005	2006	2007	2008	2009	2010
Technical report	0	1	3	3	2	1
Book chapter	0	2	1	2	2	7
Conference	10	37	67	79	83	106
Journal article	0	11	17	17	14	41

Table 3. Publication type for MIREX-derived papers.

2005	2006	2007	2008	2009	2010
5 (50%)	14 (27%)	29 (30%)	26 (26%)	25 (25%)	28 (18%)

Table 4. Number and percentage of MIREX-derived papers that are published in ISMIR conferences.

3.3 MIREX Theses and Dissertations

Table 5 shows the number of research theses and dissertations that are based to some extent on the MIREX trials—typically by referencing MIREX annual results, by testing a novel algorithm against published MIREX datasets, or reporting more fully on the researcher's own MIREX entry. The uptake of MIREX as a degree focus bodes well for the future of research in this area, as Masters and PhD students move into research positions.

The theses and dissertations are cited less than the other MIREX-derived publications (Table 5), but that is to be expected—in the science fields, theses/dissertations are commonly re-worked into journal or conference publications, which are both more visible to other researchers and more visibly peer-reviewed (and hence more likely to be noticed and cited).

Year	Degrees	No.	Total Citations	Mean citations
2005	Masters: 1 PhD: 1	2	7	3.5
2006	Masters: 8 PhD: 5	13	90	6.92
2007	Masters: 10 PhD: 3	13	114	8.77
2008	Masters: 14 PhD: 9	23	90	6.92
2009	Masters: 4 PhD: 10	14	19	1.36
2010	Ugrad: 1 Masters: 9 PhD: 11	21	46	2.19

Table 5. MIREX-related theses and dissertations.

3.4 Collaboration in MIREX Research

The mean number of authors per paper is presented in Table 6 and the distribution of author numbers per paper is presented in Figure 2. The research teams submitting to the original MIREX trials were small—the vast majority comprised one or two researchers—but over the years the number of participants in a MIREX submission has grown. The number of co-authors for papers based on MIREX has shown steady growth to 2010. Both trends likely reflect the maturing of this area of research, as stable research groups develop from the interests of one or two key researchers.

The size of the collaborative teams for both categories of paper are larger than might be expected; typically the mean number of co-authors for a computer science or engineering paper hovers around two [6].

	2005	2006	2007	2008	2009	2010
Extended abstracts	1.75	1.75	2.39	2.47	2.85	2.79
Derived papers	2.3	2.31	2.62	2.96	2.95	3

Table 6. Mean number of authors per paper.

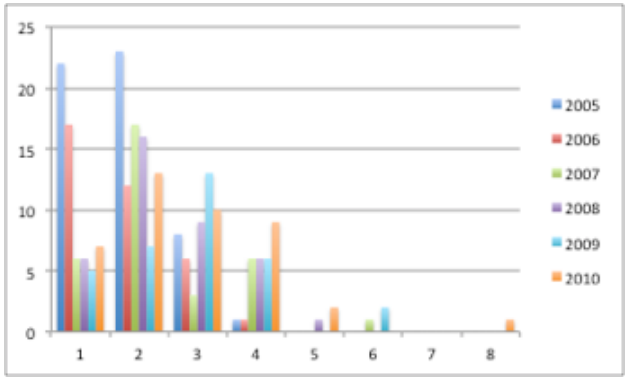

Figure 2a. Number of authors per paper for MIREX extended abstracts.

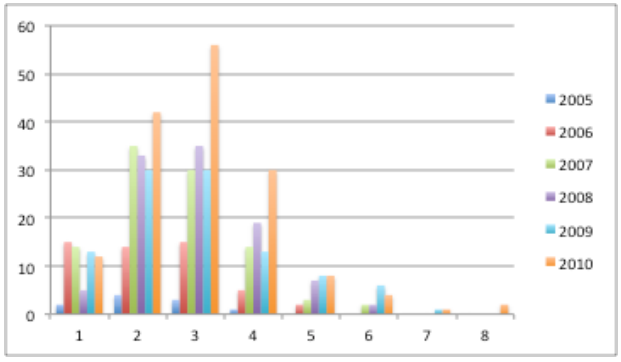

Figure 2b. Number of authors per paper for MIREX derived publications (excluding theses and dissertations).

3.5 Geographic Distribution of MIREX Researchers

Thirty-six countries have contributed at least one publication in the 2005 – 2010 MIREX document set (Table 7 presents the league table of the top contributors, and Figure 3 presents a map-based visualization of this geographic distribution). Participation in the MIREX evaluations is clearly not restricted to a small inner circle, and the MIREX results are seeing similarly widespread application.

Examining more closely the national affiliations for authors of the papers under study, we see that the research is surprisingly collaborative across national boundaries and between institutions within a single country (Table 8). The percentage of papers involving co-authors from two or more countries seems to have stabilized at 12% from 2007 – 2009, and then to have increased sharply in 2010 to 18%. The increases in these cross-boundary collaborations may reflect the increasing maturity of the field, as researchers move to new positions while maintaining research ties in their former institutions, or perhaps the personal connections made through ISMIR / MIREX conferences are encouraging greater collaboration outside the researcher's home institution. A further drill-down into the publications dataset (and likely follow-up survey of MIREX researchers) is necessary to clarify the factors contributing to this effect.

Country	MIREX abstracts	Derived papers	Theses	Total
USA	33	130	19	182
France	29	61	5	95
Spain	27	48	11	86
UK	22	50	11	83
Canada	14	32	7	53
Austria	16	23	6	45
Finland	16	16	2	34
Germany	15	19		34
China	14	19	1	34
Japan	8	22		30

Table 7. Number of publications by country for the top ten contributors, 2005-2010.

	Avg no. of countries per paper	% of multi-national collaborations	Avg no. of institutions per paper
2005	1.2	20.0%	1.13
2006	1.04	3.9%	1.08
2007	1.16	13.3%	1.22
2008	1.15	12.9%	1.32
2009	1.14	12.0%	1.43
2010	1.18	18.3%	1.46

Table 7. Summary of international and cross-institutional collaborations.

4. BUILDING A GREATER PROFILE FOR MIREX EXTENDED ABSTRACTS

Early in the data gathering process it became apparent that a substantial proportion of the MIREX extended abstracts were not being harvested by our Google Scholar searches—for example, a manual count of the 2008 extended abstracts on the MIREX wiki (http://www.music-ir.org/mirex/wiki/) yielded 51 submission abstracts, where our Google Scholar search identified only 38. Further, several extended abstracts appeared as multiple, but not identical, versions of the same intellectual content (obviously revised versions of a single submission). We later discovered that yet other extended abstracts were indeed present in the Google Scholar collection, but as they did not include MIREX in the document text or extracted metadata, they were not returned in our searches.

Perhaps more troublingly, Google Scholar was unable to extract meaningful bibliographic metadata for a number of the extended abstracts that did appear in the MIREX searches. For these latter extended abstracts, the researchers verified that they were indeed part of the MIREX trials by traversing backwards through the file hierarchy in which the document was stored, until we could determine that it was indeed a legitimate contribution to a MIREX evaluation cycle. For an extended abstract lacking metadata, a researcher unfamiliar with MIREX, but interested in the intellectual content of the paper, would not know the extent to which the results presented in the paper could be trusted—was this paper

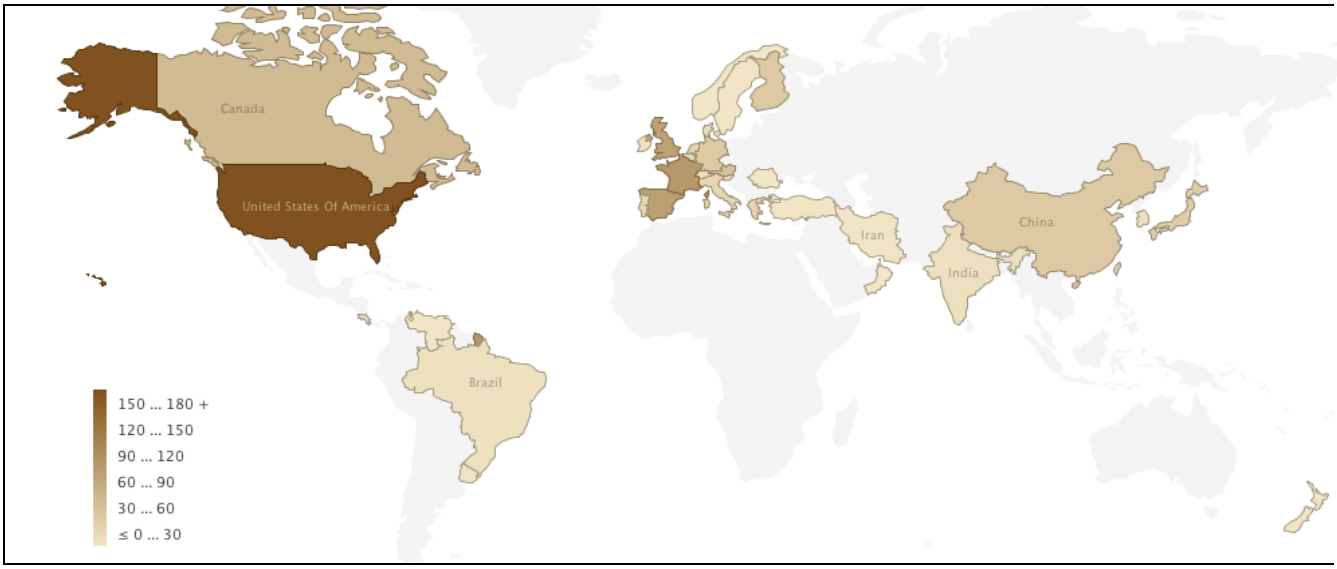

Figure 3. Geographic distribution of MIREX researchers.

peer reviewed? Was it a technical report, less formally 'published' but still endorsed by the authors' institutions? Or was it a student assignment accidentally harvested by Google Scholar?

These issues with identifying both the existence and provenance of MIREX extended abstracts in Google Scholar are likely explanations for the relatively low citation counts for the extended abstracts identified in this present study (Table 2). To mitigate these issues and, we hope, provide a mechanism for the MIREX evaluation documents to gain a higher profile, we have developed a digital library of the extended abstracts using the open source digital library software Greenstone [11]. Figure 4 shows a snapshot taken from this resource. The figure shows the result of searching for "F0" using the full-text index of the abstract texts. Each matching document displays the title, year of publication, and the authors, along with a link to the PDF document. Also provided for each document is a "Locate @ Google Scholar" link. Clicking on this takes the title of the paper and initiates a search for this on Google Scholar. While not guaranteed to find a match, we found it worked reliably well in practice, and a convenient way to locate citation information about the extended abstract. Features also include browsing by title, author and date, as well as search by these metadata fields. The resource can be accessed through http://music-ir.org/mirex-dl/library).

While this digital library provides improved access facilities to the extended abstracts, it is worth noting that the some of the metadata for each abstract may be provided through the digital library interface and is not apparent on the document itself. Searchers may stumble across an extended abstract via any number of mechanisms—a Google Scholar search, a general search engine query, a link from another website—and there is no guarantee that the specific path a particular user takes in locating a given document will provide any cues as to the document's provenance *beyond those included in the text of the document itself*. For this reason, we recommend that each extended abstract should include a header providing the citation for that abstract.

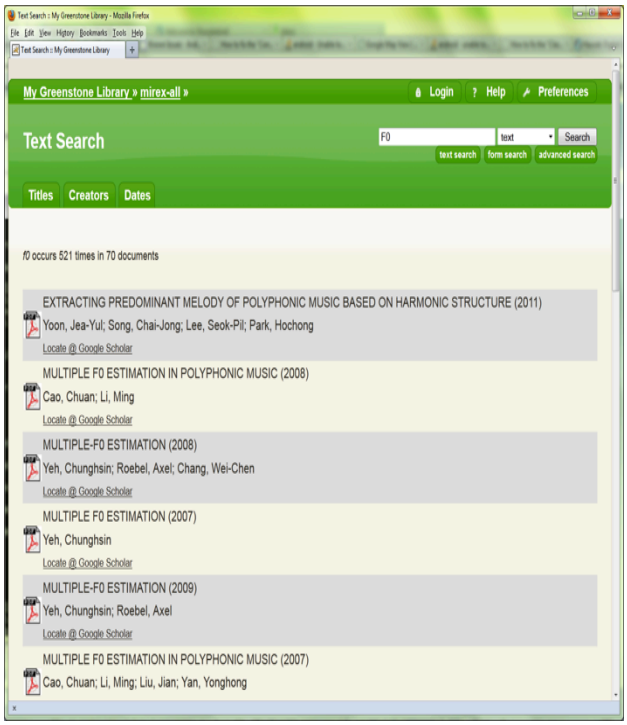

Figure 4. Sample search results display in our prototype digital library of MIREX extended abstracts.

Close examination of the MIREX-derived literature also uncovered difficulties that some authors had obviously experienced in knowing how to cite the results of the MIREX trials (for example, the relative performance of specific algorithms). While an overview of the year's MIREX evaluations generally appears in the proceedings of the annual ISMIR conference, this document does not provide comprehensive results from all tasks. Exhaustive

summaries of results are available on the MIREX wiki[1], but these are not provided in a form that is recognized as being suitable for indexing by Google Scholar—and no guidelines are given on the wiki as to how to cite these results. A straightforward solution would be to issue these results summaries as technical reports and store them in repositories indexed by Google Scholar and other scholarly indexing systems.

Similar difficulties were apparently experienced in providing formal acknowledgment of the MIREX trials, experimental setup, or datasets (the MIREX wiki does not provide a canonical reference form for these). While these papers did use the term "MIREX" in describing the results and datasets in the paper body (and so our Google Searches did return these papers), these mentions were not tied to entries in the papers' reference sections—and consequently no MIREX entity receives citation credit. Contrast this situation with that of the TRECVid evaluation series, which suggests standard references for many aspects of this programme (http://www-nlpir.nist.gov/projects/t01v/trecvid.citation.html). We encourage the MIREX organizers to develop similar referencing guidelines, and will include them in the home page of our extended abstracts digital library.

5. CONCLUSIONS

Our examination of the MIREX literature (the extended abstracts and papers referring to / referencing MIREX results, datasets, and evaluation trials) portrays a thriving international research community, characterized by collaboration. We have identified barriers to the accessibility of the MIREX extended abstracts, and present a prototype digital library for these documents that we believe can improve the MIREX profile in the larger research community. We also provide recommendations for modifications to the format of extended abstracts and the information presented in the MIREX wiki, to increase the visibility of MIREX to search engines and to make it easier for researchers to locate citation information for MIREX documents.

We believe that these small changes have the potential for a large payoff: MIREX can follow in the steps of the successful TRECVid and ImageCLEF series by providing the MIREX extended abstracts and citation information in formats that are readily indexed by Google Scholar and other resources, easily located by interested researchers, and easily cited in relevant publications.

6. REFERENCES

[1] L. Bornmann and H-D Daniel: "What do we know about the h index?", Journal of the American Society for Information Science and Technology, Vol. 58, No. 9, pp. 1381-1395, 2007.

[2] J. S. Downie: "The music information retrieval evaluation exchange (2005 – 2007): A window into music information retrieval research", Acoustical Science and Technology, Vol. 29, No. 4, pp. 247-255, 2008.

[3] J. Freyne, L. Coyle, B. Smyth, and P. Cunningham: "Relative status of journal and conference publications in computer science", Communications of the ACM, Vol. 53, No. 11, pp. 124-132, 2010.

[4] A.W. Harzing: "Citation analysis across the disciplines: the impact of different data sources and citation metrics", http://www.harzing.com/data_metrics_comparison.htm

[5] B.M. Hemminger, D. Lu, K.T.L. Vaughan, and S.J. Adams: "Information seeking behavior of academic scientists", Journal of the American Society for Information Science and Technology, Vol. 58, No. 14, pp. 2205-2225, 2007.

[6] M.E.J. Newman: "The structure of scientific collaboration networks", Proceedings of the National Academy of Sciences, Vol. 98, No. 2, pp. 404-409, 2001.

[7] R. Rousseau: "A bibliometric study of Nieuwenhuysen's bibliography of microcomputer software for online information and documentation work", Journal of Information Science, Vol. 16, pp. 45 – 50, 1990.

[8] C.V. Thornley, A.C. Johnson, A.F. Smeaton, and H. Lee: "The scholarly impact of TRECVid (2003 – 2009)", Journal of the Ame4rican Society for Information Science and Technology, Vol. 62, No. 4, pp. 613-627, 2011.

[9] C.V. Thornley, S.J. McLoughlin, A.C. Johnson, and A.F. Smeaton: "A bibliometric study of video retrieval evaluation benchmarking (TRECVid): A methodological analysis," Journal of Information Science, Vol. 37, No. 6, pp. 577-593, 2011.

[10] T, Tsikrika, A.G. Seco de Herrera, and H. Muller: "Assessing the scholarly impact of ImageCLEF", Multilingual and Multimodal Information Access Evaluation, Lecture Notes in Computer Science, Vol. 6941, pp. 95-106, 2011.

[11] I.H. Witten, D. Bainbridge, D.M. Nichols: How to build a digital library (2nd edition), Morgan Kaufmann, 2010.

[1] eg, "MIREX 2008 Overall Results Poster , http://www.music-ir.org/mirex/results/2008/MIREX2008_overview_A0.pdf