TEN YEARS OF MIREX: REFLECTIONS, CHALLENGES AND OPPORTUNITIES

J. Stephen Downie
University of Illinois
jdownie@illinois.edu

Xiao Hu
University of Hong Kong
xiaoxhu@hku.hk

Jin Ha Lee
University of Washington
jinhalee@uw.edu

Kahyun Choi
University of Illinois
ckahyu2@illinois.edu

Sally Jo Cunningham
University of Waikato
sallyjo@waikato.ac.nz

Yun Hao
University of Illinois
yunhao2@illinois.edu

ABSTRACT

The Music Information Retrieval Evaluation eXchange (MIREX) has been run annually since 2005, with the October 2014 plenary marking its tenth iteration. By 2013, MIREX has evaluated approximately 2000 individual music information retrieval (MIR) algorithms for a wide range of tasks over 37 different test collections. MIREX has involved researchers from over 29 different countries with a median of 109 individual participants per year. This paper summarizes the history of MIREX from its earliest planning meeting in 2001 to the present. It reflects upon the administrative, financial, and technological challenges MIREX has faced and describes how those challenges have been surmounted. We propose new funding models, a distributed evaluation framework, and more holistic user experience evaluation tasks—some evolutionary, some revolutionary—for the continued success of MIREX. We hope that this paper will inspire MIR community members to contribute their ideas so MIREX can have many more successful years to come.

1. INTRODUCTION

Music Information Retrieval Evaluation eXchange (MIREX) is an annual evaluation campaign managed by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the University of Illinois at Urbana-Champaign (UIUC). MIREX provides a framework and venue for music information retrieval (MIR) researchers to compare, contrast, and discuss the results of their MIR algorithms and systems, similar to what Text Retrieval Conference (TREC) has provided to the text information retrieval community [7]. MIREX has significantly contributed to the growth and maturity of the research community, and also affected the shaping of research priorities.

As this year marks the tenth running of MIREX, we take this opportunity to reflect upon its history, impact, and challenges over the past decade. We are at the critical point where a new funding model and distributed evaluation framework must be developed to ensure the sustainability of MIREX. We propose one of many possible solutions, and hope that this will spark the discussion in the MIR community. In addition, based on the feedback and criticisms provided by the members of MIR community, we make recommendations on future directions of MIREX, emphasizing more holistic user experience evaluation tasks.

2. HISTORY, STRUCTURE, AND IMPACT

The history of MIREX can be traced back to 1999 when the Exploratory Workshop on Music Information Retrieval was held as part of the ACM Special Interest Group Information Retrieval (SIGIR) Conference. Two years later, the attendees of ISMIR2001 passed the "Bloomington Manifesto" which called for a formal evaluation platform for MIR research. Afterwards, two additional workshops on MIR evaluation were held in ISMIR2002 and SIGIR2003, which led to funding for MIREX from both the Andrew W. Mellon Foundation and the National Science Foundation (NSF).

In 2004, the local committee of ISMIR, the Music Technology Group (MTG) of the University Pompeu Fabra organized an evaluation session, the Audio Description Contest (ADC) [1], which provided valuable insights for MIREX. After these preludes, MIREX officially began in 2005 and had its first plenary and poster sessions in ISMIR 2005, held at Queen Mary College, University of London. Building on successful runs of MIREX, three additional projects were funded to explore new technologies and models to uplift MIREX: Networked Environment for Music Analysis (NEMA), Structural Analysis of Large Amounts of Music Information (SALAMI) and MIREX: Next Generation (MIREX: NG). From 2000 to date, MIREX and related projects have received approximately $3,100,000 in funding from the NSF, The Andrew W. Mellon Foundation, University of Illinois, and the Korean Electronics Technology Institute (KETI). Table 1 summarizes these and other important events in the development of MIREX.

By 2013, MIREX has evaluated 1997 individual MIR algorithms over 37 different test collections, and has involved researchers from over 20 different countries with a median of 108 individual participants per year (Table 2). The tasks and subtasks evaluated in MIREX represent a wide spectrum of research interests among MIR researchers in the last decade (Table 3), including classical machine-learning train-test tasks (e.g., Audio Tag Classification), "low-level" tasks on which many MIR systems depend (e.g., Audio Beat Tracking), and tasks involving some types of user-issued music queries (e.g., Query-by-Singing/Humming). The evidence that MIREX has sig-

nificantly grown over the years is clear. The addition and retirement of the tasks reflect the shift in interests in the field.

1999	Music retrieval workshop at SIGIR proposed a range of evaluation scenarios
2000	First ISMIR held at Plymouth with participants holding brainstorming sessions
2001	ISMIR at Indiana University; "Bloomington Manifesto" on evaluation published
2002	Planning grant from the Andrew W. Mellon Foundation awarded
2002	ISMIR at Paris hosted special evaluation workshop
2003	SIGIR at Toronto held Workshop on the Evaluation of Music Information Retrieval Systems
2003	Andrew W. Mellon Foundation and NSF funding awarded
2004	Audio Description Contest run at ISMIR Barcelona
2005	First MIREX plenary session held at ISMIR London
2008	NEMA project funded by the Andrew W. Mellon Foundation
2009	SALAMI funded by the NSF, SSHRC and JISC
2012	MIREX:NG project funded by the Andrew W. Mellon Foundation

Table 1. Important Events in MIREX History

	Datasets	Individuals	Countries	Runs
2005	10	82	19	86
2006	13	50	14	92
2007	12	73	15	122
2008	18	84	19	169
2009	26	138	15	289
2010	31	152	21	331
2011	32	156	16	296
2012	35	109	20	302
2013	37	116	29	310

Table 2. Descriptive Statistics for MIREX 2005-2013

Due to restrictive intellectual property issues surrounding music materials, the test data used in MIREX cannot be distributed to participants. This distinguishes the structure of MIREX from those of other major evaluation frameworks such as TREC. MIREX has been operated under an "algorithm-to-data" or "non-consumptive computation" model: researchers submit their MIR algorithms to IMIRSEL which are then evaluated by IMIRSEL personnel and volunteers against the ground truth data hosted in IMIRSEL.

Beyond the technical infrastructure, the communications infrastructure is also critical for MIREX as it is a community-driven endeavor. The MIREX wikis were set up for the community to collaboratively define the evaluation tasks, metrics, and general rules in every spring, and to publish and archive results data for each task and associated algorithms in every autumn. Besides being used by participants for preparing their mandatory presentations in the annual MIREX poster session in ISMIR, the MIREX results data also provide unique and valuable materials for publications in the field. In addition, the MIREX "EvalFest" mailing list is used for discussions about evaluation issues. To date, 531 people have subscribed to EvalFest. IMIRSEL also creates task-specific mailing lists where researchers can have detailed discussions about metrics, collections, and input/output formats.

From its inception, MIREX has had a clear (and growing) impact on MIR research. Updating an earlier analysis of MIREX-related publications in [3], as of April 2014, 314 MIREX extended abstracts and 1,070 publications based on MIREX trials and results can be found through Google Scholar (Table 4). These publications have received a total of 18,239 citations (Table 5). We limited the analysis period to the seven years ending in 2011, as there is a considerable lag between the publication of a document and its appearance in Google Scholar (and then a similar lag before the paper can be cited). The growing number of Master's and PhD dissertations building on MIREX results—and in many cases, participating in MIREX trials—is particularly significant; MIREX has clearly become a fundamental aspect of MIR research infrastructure. In addition to this impact on academic research, 13 patents have explicitly referenced MIREX extended abstracts [4].

	'05	'06	'07	'08	'09	'10	'11
Tech. report	0	4	4	3	10	5	11
Book chapter	0	2	1	2	8	9	20
Dissertation	1	17	13	25	22	35	48
Conference	12	46	68	88	127	144	137
Journal article	1	15	27	21	29	50	65
Total	14	84	113	139	196	243	281

Table 4. Publication Types for MIREX-derived Papers

To elicit further, less easily measured contributions of MIREX to the research community, interviews of 18 influential MIR researchers were conducted in the MIREX Next Generation project [10]. From these, four key contributions were identified: 1) ***Benchmarking and evaluation:*** MIREX was born from the recognition that the field could not progress unless MIR researchers could benchmark their work against each other's; 2) ***Training and induction into MIR*****:** Emerging researchers and graduate students gain hands-on experience with MIR research and development, and build a reputation with potential employers within both the music industry and academia; 3) ***Dissemination of new research*****:** The annual MIREX trials and subsequent MIREX session at ISMIR provide a natural focus for the research community, and allow researchers to showcase their work to the MIR community at large; 4) ***Dissemination of data*****:** MIREX has been an important venue for the community to access previous high-quality evaluation datasets created by MIREX team or donated by researchers.

	MIREX extended abstracts				MIREX-derived publications			
Year	No.	citations	mean	med.	No.	citations	mean	med.
2005	55	418	7.60	5	14	879	62.79	17.5
2006	35	217	6.20	2	51	2656	31.62	13
2007	32	403	12.60	4	113	1449	21.27	8
2008	39	136	2.61	3	139	3560	26.61	8
2009	48	144	3.00	0	196	2790	14.23	5
2010	61	135	2.21	0	243	3093	12.73	6
2011	44	63	1.43	1	281	2296	8.17	2

Table 5. Overview of MIREX Citation Data, 2005-2011

TASK NAME	2005	2006	2007	2008	2009	2010	2011	2012	2013
Audio Artist Identification	7		7	11					
Audio Beat Tracking		5		15(2)	22(2)	26(2)	24(2)	60(3)	54(3)
Audio Chord Detection				11	18(2)	15	18	22(2)	36(3)
Audio Classical Composer ID			7	8	30	27	16	15	14
Audio Cover Song Identification		8	8		6(2)	6(2)	4(2)		2
Audio Drum Detection	8								
Audio Genre Classification	15		7	26(2)	65(2)	48(2)	31(2)	31(2)	26(2)
Audio Key Detection	7					5	8	6	3
Audio Melody Extraction	10	10(2)		21(3)	12(6)	30(6)	60(6)	24(6)	24(6)
Audio Mood Classification			9	13	33	36	17	20	23
Audio Music Similarity		6	12		15	8	18	10	8
Audio Onset Detection	9	13	17		12	18	8	10	11
Audio Tag Classification				11	24(3)	26(2)	30(2)	18(2)	8(2)
Audio Tempo Extraction	13	7				7	6	4	11
Discovery of Repeated Themes & Sections									16
Multiple Fundamental Frequency Estimation & Tracking			27(2)	28(2)	26(3)	23(3)	16(2)	16(2)	6(2)
Query-by-Singing/Humming		23(2)	20(2)	16(2)	12(4)	20(4)	12(4)	24(4)	28(5)
Query-by-Tapping				5	9(3)	6(3)	3(3)	6(3)	6(3)
Real-time Audio to Score Alignment (a.k.a Score Following)		2		4		5	2	3	2
Structural Segmentation					5	12(2)	12(2)	27(3)	26(3)
Symbolic Genre Classification	5								
Symbolic Key Finding	5								
Symbolic Melodic Similarity	7	18(3)	8			13	11	6	6
Total Number of Runs per Year	**86**	**92**	**122**	**169**	**289**	**331**	**296**	**302**	**310**
Total Number of Runs (2005-2013)	**1997**								

Notes: 1) Superscript numbers represent the number of subtasks included. 2) Since 2009, the Audio Classical Composer ID task, Audio Genre Classification task, and Audio Mood Classification task have become subtasks of Train-Test Task.

Table 3. MIREX Tasks and the Number of Runs

3. CHALLENGES

3.1 Sustainability of Current Administration Model

The current model for administrating the evaluations is costly and unsustainable. Since its inception, all MIREX tasks have required manual execution of submitted algorithms. As algorithms are written in different languages and require a range of executing environments, running one algorithm takes about 5 hours of focused attention on average, including but not limited to the time spent on communicating with participants, debugging algorithms, reconfiguring input/output interfaces and execution environment, etc. More often than not, algorithms may have to be updated by participants and tested by IMIRSEL for multiple rounds before they can be executed correctly. Besides the algorithms, some tasks require ground truth data in every iteration of MIREX (e.g., similarity tasks, further discussed in Section 3.4), which takes a significant amount of time to build. To meet all these demands, IMIRSEL has been relying on a small number of graduate students fully devoted to running MIREX in each fall. Nonetheless, participants sometimes still have to wait for a long time to receive evaluation results.

To mitigate the problem, the Networked Environment for Music Analysis (NEMA) project was established to "construct a web-service framework that would make MIREX evaluation tasks, test collections, and automated evaluation scripts available to the community on a yearly basis" (p.113, the so-called "Do-It-Yourself" model) [6]. However, due to the large variety of execution environments of algorithms, the built framework has not been widely adopted in the MIR community, except for the automated evaluation package in the NEMA framework which has been used in recent iterations of MIREX to automate the evaluation of tasks such as Train-test and Audio Tag Classification. This has greatly improved the efficiency of MIREX and productivity of IMIRSEL personnel, but such procedures still require manual input of raw results produced by the algorithms. The sustainability of MIREX calls for new technology and structures that can streamline the entire process of data/algorithm ingest, evaluation code generation/modification, and results posting, so that the evaluations can not only be effective, but also efficient, robust, and scalable.

3.2 Financial Sustainability Challenges

The fact that MIREX has been providing significant value to the MIR community is clearly evident. However, IMIRSEL has effectively offered MIREX as a free service to the community. This model is unsustainable; in January 2015, the current Mellon funding concludes, leaving MIREX with no financial support for the first time in its history. A back-of-the envelope calculation using the amount of grant funding ($3,100,000) divided by number of runs (1997) gives an estimate of the cost per run of $1,552. Cost estimates per participant (960 total) come in at $3,229. These rough numbers illustrate the general magnitude of the funding challenge MIREX is facing.

3.3 Knowledge Management and Transfer

Over the past decade, the leading task organizers of MIREX have left IMIRSEL, including Dr. Andreas Ehmann (now at Pandora.com) and Dr. Mert Bay—both in-

strumental in creating MIREX processes and techniques. Considerable time and energy are being expended in reconstructing past practices to help new IMIRSEL members and new task organizers complete their assigned duties. MIREX needs more effective mechanisms to manage corporate memory so as to successfully transfer knowledge to new lab members and external volunteers. Notwithstanding recent efforts to more thoroughly document MIREX technologies and procedures, more work needs to be done to support hands-on training sessions for all who manage and run MIREX tasks.

3.4 Ground Truth Data Shortage

The lack of ground truth data is one of the primary obstacles facing the field of MIR. There is a strong demand for large, high-quality ground truth datasets for various evaluation tasks. However, generating any kind of user data is expensive. Crowdsourcing has been suggested as a possible solution by a number of MIR researchers (e.g., [12][16]). Although previous studies have shown that the user evaluation results collected by crowdsourcing and from music experts in the conventional MIREX framework are comparable, the issues of representativeness and noise in data still exist.

In order to generate the ground truth data, human evaluators must listen to sample music pieces and manually input their responses. The task must be carried out by individuals who have had a baseline level of training, making the data even more expensive to collect. Currently, most ground truth data is generated within academic institutions through the use of graduate and undergraduate student labor. Funding opportunities for generating ground truth data are limited, and the fact that audio data is often not transferrable between multiple researchers or labs due to copyright restrictions further complicates dataset creation.

There are a variety of sources for ground truth data, some released by MIREX, and also by other researchers in an ad hoc fashion. However, academic scholars as well as researchers in industry have difficulty identifying and obtaining relevant datasets. Currently, there is no organization or lab that is taking the role of creating, maintaining, and sharing ground truth data. In other IR domains, there are central organizations that fulfill at least part of this responsibility to support evaluations [10]. For example, ground truth data in TREC is created and/or managed by National Institute of Standards and Technology (NIST) and is released after each evaluation [7]. In the field of speech recognition, the Linguistic Data Consortium (LDC) creates ground truth datasets that can be purchased for use by individual labs [10]. This cycle of refreshed data allows the research community to conduct high-quality evaluation. As this has not been the case for MIREX, the same ground truth data must sometimes be used for multiple years.

3.5 Intellectual Property Issues

Another major problem facing the MIREX community is the lack of usable music data upon which to build realistic test collections, due to intellectual property issues surrounding music materials. The datasets used in MIREX are very limited in terms of size, variety, recency, and novelty. Moreover, the fact that datasets cannot be distributed after being used in MIREX effectively prevents researchers from replicating the evaluation and benchmarking their newly developed algorithms on their own. To tackle this issue that has plagued MIR research since day one, the MIR community needs to work together to explore possible solutions such as negotiating with copyright holders collectively, using creative audio and/or music in the public domain, and running algorithms against multiple datasets hosted in different labs. The latter approach has been attempted by projects such as NEMA. However, none of the possibilities is straightforward and this battle is likely to exist for many years to come.

3.6 System vs. User-centered Evaluations

MIREX has followed the conventional, Cranfield IR system-centered evaluation paradigm [2]. Recently, this evaluation approach has been criticized by multiple researchers for excluding users from the evaluation process. To name a few, Hu and Liu [9], Hu and Kando [8], Lee [11], Schedl and Flexer [15], and Lee and Cunningham [13] all argued that the goal of MIR systems is to help users meet their music information needs, and thus MIR evaluation must take users into account. For instance, a number of MIR researchers have questioned the validity of system-centered evaluation on tasks that involve human judgments such as the similarity tasks [12], [15], [16]. Music similarity may be interpreted differently for different people, yet the variance across users is simply ignored in the current evaluation protocol. As noted by Lee and Cunningham [13], a result of system-centered evaluation "may not be effectively translated to something meaningful or practical for real users (p. 517)." They suggested introducing tasks that "seems closer to what would be useful for real users" such as playlist generation, known-item search, or personal music collection management.

Notwithstanding the importance of traditional system-centered tasks, some suggestions have been made to MIREX to bridge the gap between system-centered and user-centered evaluation (e.g., incorporating user context in test queries, use terms familiar to users, combine multiple tasks in [11][9]), although they are yet to be reflected in the MIREX tasks. As the field matures, in order to move forward, it is vital to explore user-centered and realistic evaluation tasks.

4. FUTURE DIRECTIONS

4.1 Developing a User Experience Task

In keeping with our desire to expand MIREX beyond its current system-centered paradigm, we are conducting the first user-centered grand challenge evaluation task. The "Grand Challenge '14 User Experience" (GC14UX)[1] task is unlike any previous MIREX task. The GC14UX is directly inspired by the grand challenge idea proposed in Downie, Crawford and Byrd [5], which noted the persis-

[1] http://www.music-ir.org/mirex/wiki/2014:GC14UX

tent absence of complete MIR systems presented at ISMIR that could be released to the public for music searching and discovery. Thus, the GC14UX has two underpinning goals: 1) to inspire the development of complete MIR systems to be shared at ISMIR; and 2) to promote the notion of user experience as a first-class research objective in the MIR community.

The choice of "Grand Challenge" to describe our first UX task was made, in part, to signify that MIREX will be entering into uncharted evaluation territory. By finally undertaking a user-centered evaluation task, the GC14UX will require the MIREX team (and the MIR community) to come up with new evaluation methods and criteria that will be made manifest in ways significantly different from our now standard MIREX operation procedures. We argue that the current state of the art in conventional MIREX tasks is sufficient to support an acceptable degree of efficiency and effectiveness for most of the now classic MIREX system-centered tasks. It is now time to look towards the more holistic user experience: subjective explorations of hedonic aspects of use such as satisfaction, enjoyment, and stimulation. To that end, the MIREX team is proposing several radical departures from MIREX tradition that promise to better support the focus on the user experience. The most radical changes include: 1) no submission of algorithms to IMIRSEL; and 2) distribution of audio data to participants.

To ensure that the GC14UX does not become a system-centered evaluation in disguise, the process is designed to remain as agnostic as possible concerning the technological means by which participating systems create and deliver their experiences to the users. This deliberate indifference suggests that the GC14UX has no need to run or evaluate the underlying system code that delivers the content to the users. Since the GC14UX will not be evaluating the system-code per se, it makes sense that the GC14UX does not follow MIREX's usual practice of running code on behalf of the submitters. There are obvious benefits to this non-submission approach, including greatly reduced system requirements and significantly reduced MIREX staff time requirements for debugging and administration.

Dropping the usual algorithm-to-data procedures does, obviously, beg the question about data sources for the systems to use. All the usual copyright reasons why music distribution is problematic for MIREX still apply and therefore we need data sources that are amenable to distribution. For the first running of GC14UX, the test collection will be drawn from Creative Commons music. We believe that a set in the magnitude range of 10,000 songs would strike a nice balance between being non-trivial in size and breadth while not posing too great of a data management burden for participants. A common dataset helps mitigate against the possible user experience bias induced by the differential presence (or absence) of popular or known music within the participating systems.

The GC14UX task is all about how users perceive their experiences with the systems. We intend to capture the user perceptions in a minimally intrusive manner under as-realistic-as-possible use scenarios. To this end, all participating systems are required to be constructed as websites accessible to users through normal web browsers. For user evaluation, we also do not want to burden the users/evaluators with too many questions or required data inputs. Our main goal is to determine whether each system was able to provide a satisfying user experience ([14], [17]). Thus, a question asking about the level of overall satisfaction is posed to each user for each system. An option for open-ended responses is provided so as to capture the expressions of the users in their own words.

There are many potential challenges that could prevent GC14UX from being the progenitor of future MIREX UX evaluations. For example, the utility and possible side-effects of using Creative Commons music as the common dataset have yet to be ascertained. Also, the effectiveness of the current GC14UX user inputs will most likely spark lively debate among MIR researchers after our first round of data is collected. Notwithstanding these known problems, as well as the challenges currently unknown, we are eager to see GC14UX proceed and inspire new evaluations. It is well past time that MIREX act to create a real user-centered evaluation stream. If we allow perfection to be the enemy of the good, MIREX might never be able to launch a vibrant UX evaluation thread.

4.2 Funding Models

In order to continue providing benefits to the MIR community, MIREX must explore a range of funding options. In order to reduce the dependencies and burdens placed upon any one funding source, it is necessary to seek multiple sources of income. Some of the current possibilities include:

- **Lab Memberships:** MIREX is exploring the possibility of setting up a lab membership system for labs that are active in MIR. Member labs would be represented on MIREX's governing committee, and would have access to the new datasets that MIREX creates.
- **Sponsorship:** MIREX would also like set up a sponsorship program for leaders in industry. A sponsorship program would give companies a chance to support and/or discover interesting new MIR work by emerging researchers. Identification of recruiting opportunities is a valuable benefit that industry currently derives from MIREX (Section 2).
- **Institutional Support:** The University of Illinois has provided significant in-kind support for MIREX in the past. MIREX seeks to extend this partnership into the future. However, budget shortfalls at the State level are diminishing the prospects of ongoing University support.
- **Data Creation and Curation:** The MIREX team completed a collaborative project developing ground truth genre and mood data for, and funded by, Korea Electronics Technology Institute (KETI) in 2013. The data created is being folded into the MIREX task pool. The success of the KETI project, combined with the precedent set by the LDC (Section 3.4), inspires future data creation actions. In a similar line, we are exploring the possibility of providing fee-based data

curation and management services to those who have data sets that require long-term preservation.

While it will need to seek more support from its participants, MIREX recognizes the need to balance this with openness and accessibility. MIREX aims to remain open to any researcher who wants to participate, with a healthy funding mix making this goal more likely to be achieved.

4.3 Distributed Management Model: Task Captains

MIREX is pursuing a more decentralized model in order to reduce the strain on IMIRSEL and to more actively involve the entire MIR community in task creation, organization and delivery. Under this model, multiple labs can run particular tasks while IMIRSEL functions as a central organizer and algorithm submission point. This model was piloted in 2012 with Query-by-Singing/Humming (QBSH) and Audio Melody Extraction (AME) run by KETI. In MIREX 2013, Audio Beat Tracking (ABT), Audio Chord Estimation (ACE), Audio Key Detection (AKD), Audio Onset Detection (AOD), Audio Tempo Estimation (ATE), and Discovery of Repeated Themes & Sections (DRTS) were led by non-IMIRSEL volunteer "Task Captains" who managed the tasks from start to finish. While shortcomings in MIREX documentation were evident, the Task Captain initiative was successful and will be developed further.

5. CONCLUSIONS

In this paper, we reflect on ten years of experience of MIREX. As the major community-based evaluation framework, MIREX has made unprecedented contributions to the MIR research field. However, MIREX also faces a number of significant challenges including financial sustainability, restrictions on data and intellectual property, and governance. Future directions of MIREX are proposed to meet these challenges. By moving towards the evaluation of entire systems and emphasizing holistic user experience, MIREX will allow us to compare and evaluate startups and experimental systems, as well as commercial MIR systems. We hope this paper will serve as a catalyst for the community to come together and seek answers to the question: what is the future of MIREX? More importantly, we hope this paper will inspire MIR community members to actively engage in and contribute to the continuation of MIREX. MIREX has always been a community-driven endeavor; without the active leadership and involvement of MIR researchers, MIREX simply cannot exist.

6. REFERENCES

[1] P. Cano, E. Gomez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, N. Wack: "ISMIR 2004 audio description contest," MTG Technical Report, MTG-TR-2006-02 (Music Technology Group, Barcelona, Spain), 2004.

[2] C. W. Cleverdon and E. M. Keen: "Factors determining the performance of indexing systems. Vol. 1: Design, Vol. 2: Results," Cranfield, UK: Aslib Cranfield Research Project, 1966.

[3] S. J. Cunningham, D. Bainbridge, and J. S. Downie: "The impact of MIREX on scholarly research," *Proceedings of the ISMIR*, pp. 259-264, 2012.

[4] S. J. Cunningham and J. H. Lee: "Influences of ISMIR and MIREX Research on Technology Patents," *Proceedings of the ISMIR*, pp.137-142, 2013.

[5] J. S. Downie, D. Byrd, and T. Crawford: "Ten Years of ISMIR: Reflections on Challenges and Opportunities." *Proceedings of the ISMIR*, pp. 13-18. 2009.

[6] J. S. Downie, A. F. Ehmann, M. Bay, and M. C. Jones: "The music information retrieval evaluation exchange: Some observations and insights." In *Advances in music information retrieval*, pp. 93-115, Springer Berlin Heidelberg, 2010.

[7] D. Harman: "Overview of the Second Text Retrieval Conference (TREC-2)." *Information Processing & Management*, 31(3), 271–289, 1995.

[8] X. Hu and N. Kando: "User-centered Measures vs. System Effectiveness in Finding Similar Songs," *Proceedings of the ISMIR,* pp.331-336, 2012.

[9] X. Hu and J. Liu: "Evaluation of Music Information Retrieval: Towards a User-Centered Approach". *Proceedings of the 4th Workshop on Human-Computer Interaction and Information Retrieval,* 2010.

[10] Ithaka S+R: "MIREX Consulting Report and Proposed Business Plan," 2013.

[11] J. H. Lee: "Analysis of user needs and information features in natural language queries for music information retrieval," *Journal of the American Society for Information Science & Technology,* 61(5), pp. 1025-1045, 2010.

[12] J. H. Lee: "Crowdsourcing Music Similarity Judgments using Mechanical Turk," *Proceedings of the ISMIR*, pp. 183 - 188, 2010.

[13] J. H. Lee and S. J. Cunningham: "Toward an understanding of the history and impact of user studies in music information retrieval", *Journal of Intelligent Information Systems,* 41, pp. 499-521, 2013.

[14] H. Petrie and N. Bevan: "The evaluation of accessibility, usability and user experience," *The Universal Access Handbook,* pp. 10-20, 2009.

[15] M. Schedl and A. Flexer: "Putting the User in the Center of Music Information Retrieval." *Proceedings of the ISMIR*, pp. 385-390, 2012.

[16] J. Urbano: "Information Retrieval Meta-Evaluation: Challenges and opportunities in the Music Domain," *Proceedings of the ISMIR,* pp. 609 - 611, 2011.

[17] A. Vermeeren, E. L-C Law, V. Roto, M. Obrist, J. Hoonhout, and K. Väänänen-Vainio-Mattila: "User experience evaluation methods: current state and development needs." In *Proceedings of the 6th Nordic Conference on HCI*, pp. 521-530, 2010.