THE IMPACT (OR NON-IMPACT) OF USER STUDIES IN MUSIC INFORMATION RETRIEVAL

Jin Ha Lee
The Information School
University of Washington
jinhalee@uw.edu

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

ABSTRACT

Most Music Information Retrieval (MIR) researchers will agree that understanding users' needs and behaviors is critical for developing a good MIR system. The number of user studies in the MIR domain has been gradually increasing since the early 2000s reflecting the need for empirical studies of users. However, despite the growing number of user studies and the wide recognition of their importance, it is unclear how large their impact has been in the field; on how systems are developed, evaluation tasks are created, and how we understand critical concepts such as music similarity or music mood. In this paper, we present our analysis on the growth, publication and citation patterns, and design of 155 user studies. This is followed by a discussion of a number of issues/challenges in conducting MIR user studies and distributing the research results. We conclude by making recommendations to increase the visibility and impact of user studies in the field.

1. INTRODUCTION

Understanding users is a fundamental step in developing successful Music Information Retrieval (MIR) systems and services. Most MIR researchers will agree with this idea, and furthermore, it is not uncommon to hear various speakers at MIR related conferences specifically arguing for the importance of user studies, academically as well as commercially. Despite the growing number of user studies and the wide recognition of their importance in the MIR domain, it is unclear as to what impact these studies have really made. Have these studies in fact changed how MIR systems are developed or evaluation tasks are designed? Have they really changed how we understand critical concepts such as music similarity or music mood? For MIR researchers specializing in user studies to move forward in this domain, it is necessary to understand our past: what have we been doing and what kind of impact have we made or not? In order to lay the foundation for this discussion, we collected 155 user studies related to music, reviewed the content, and analyzed the publication and citation patterns, and research design of these studies.

2. STUDY DESIGN

2.1 Definition of "User Studies"

Our first challenge was to define and set the boundaries for "user studies." From our analysis of relevant literature, we found two major categories of user studies: "studies of users" (e.g., music information needs) and "studies involving users" (e.g., usability testing). Weigl and Guastavino [7], in their recent review article of user studies in MIR literature, defined user studies as "documents report(ing) on empirical investigations of user requirements or interactions with systems primarily aimed at providing access to musical information, including musical recordings, scores, lyrics, photography and artwork, and other associated metadata (p. 335)." In this study, we adopt a broader definition of "user studies" as studies reporting on 1) empirical investigation of needs, behaviors, perceptions, and opinions of humans, 2) experiments and usability testing involving humans, 3) analysis of user-generated data, or 4) review of the studies above. This is because a broader definition will allow for a comparison of these different types of user studies and enable us to see patterns of concentration with regards to particular types of user studies related to MIR.

2.2 Data Collection

We conducted an extensive literature search in multiple domains related to music (e.g., MIR, Library and Information Science (LIS), Human Computer Interaction (HCI), Computer Science (CS), Engineering, Psychology, Musicology) to identify these studies. We conducted searches in multiple databases including WorldCat, EBSCO, Web of Knowledge, IEEE Xplore, ACM DL, InfoPsych, and Google Scholar. We used the different combinations of the following search terms: music, user, human, people, need, use, behavior, testing, involvement, learning, interaction, design, accessibility, usability, user-centered, etc. After retrieving the relevant studies, we also followed the citations in order to broaden our search. In total, we found 155 studies related to music users.

3. PUBLICATION PATTERNS OF USER STUDIES

3.1 Growth of the Publications

First, we analyzed several aspects related to the publication patterns of the user studies. We examined the publications dates of the user studies in order to learn more about the growth pattern. Figure 1 shows the distribution

of the number of user studies published by year. We can observe the steady increase in the number of publications over the years. There were a small number of user studies pre-dating 2000, but the substantial growth started in early 2000s when the need for empirical user studies was pointed out in works such as [1], [2], and [3]. There was also a noticeable increase in 2009 and we expect that this growth pattern will continue for the coming years, at least in the near future. Although this growth pattern is encouraging, when compared with the number of studies focusing on the system aspect of MIR, the overall number of user studies is still relatively small [7].

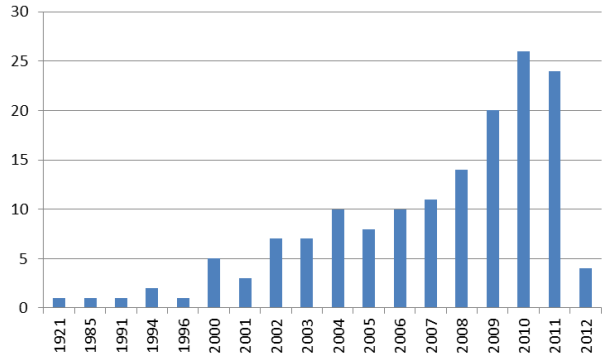

Figure 1. Distribution of the number of user studies by the year of publication

3.2 Types of Publications

We also examined the publication venues of these studies. Of the 155 studies, there were 91 conference publications, 56 journal articles, 6 workshop papers, 1 book chapter, and 1 white paper. There were a total of 83 different venues where music user studies appeared. The primary source of user studies was the ISMIR conference proceedings with 41 user studies, and all the other journals and conference proceedings included 5 or fewer user studies. 65 of the 155 user studies (42%) were the only music user study published in the particular venue. This pattern of concentration in a small number of core publications can be explained by Bradford's law which characterizes the pattern of diminishing returns in searching for references in scholarly publications [1]. The concentration of MIR user studies in the ISMIR proceedings is perhaps stronger than Bradford's predicted $1:n:n^2$ ratio of journals (where each proportion contains approximately the same number of articles). The top sources in the order of number of relevant papers are: ISMIR (41); ACM Conference on Human Factors in Computing Systems (5); ACM International Conference on Multimedia (4), ACM/IEEE-CS Joint Conference on Digital Libraries (4), International Conference on Information Visualization (4); International Conference on Mobile and Ubiquitous Multimedia (4), Journal of New Music Research (3), Music Perception (3), Psychology of Music (3), IEEE International Conference on Multimedia and Expo (3), etc.

The skewed distribution of publications poses a challenge for researchers of user studies as well as readers who are interested in finding these studies. We confirmed that it is in fact impossible to find all these studies using a single database or search engine. Also many researchers tend to conduct their literature search in their own domain, which will exclude many relevant works published in other domains (e.g., psychology scholars not citing MIR literature in CS domain). Although the ISMIR proceedings are freely available on the Web, a large number of other publications are fee-based. Unless the researchers' or readers' institutions have subscriptions to these different publications, it will be difficult and expensive to access these works. This also raises a question about distributing our knowledge to the general public who are simply interested in music and also people who are in music industry. Much of the MIR research aims to not only contribute to improving the general knowledge of music and how people interact with music, but also to create better systems and services related to music. If there is a barrier for general public and people outside of academia to access these works, then without a doubt, the impact we can make in the field will also be diminished.

3.3 Co-authorship Analysis

We performed a co-authorship analysis to further understand the patterns of publication. Figure 2 shows the co-authorship graph generated by using NodeXL, a tool for visualization and exploration of networks [6]. The graph's vertices were grouped based on the Clauset-Newman-Moore cluster algorithm and the graph was laid out using the Harel-Koren Fast Multiscale layout algorithm. The nodes represent the authors and the line connecting the nodes represents the co-authorship between the two authors. The size of the node is scaled based on the number of publications by a particular author, and the width of the line connecting two nodes is scaled based on the number of times the pair of authors have co-authored a user study.

A few strong networks emerged. The most notable network is grouped around Sally Jo Cunningham, J. Stephen Downie, Jin Ha Lee, David Bainbridge and 20 other scholars. The two networks formed around Jukka Holm and Arto Lehtiniemi, and Charlie Inskip, Andrew MacFarlane, and Pauline Rafferty are also very prominent. These strong networks seem to be forming based on the particular lab/university and regions: University of Illinois at Urbana-Champaign and University of Waikato for the first group, Finland for the second group, and UK for the third group. Another notable network formed around Adrian C. North and David Hargreaves in UK represents many user studies published in psychology. Another aspect to note is that the network is very disconnected, with a large number of small components, each consisting of a small number of authors. Part of the reason for this pattern could be because MIR is still a relatively new field, and there have not been many opportunities for cross-institutional ties to be formed. Or, it may reflect the widespread appeal of music as a subject for research (which is corroborated by the number and diversity of publication venues surveyed for this study). Further analysis will be necessary to determine the reasons for seeing this kind of co-authorship patterns.

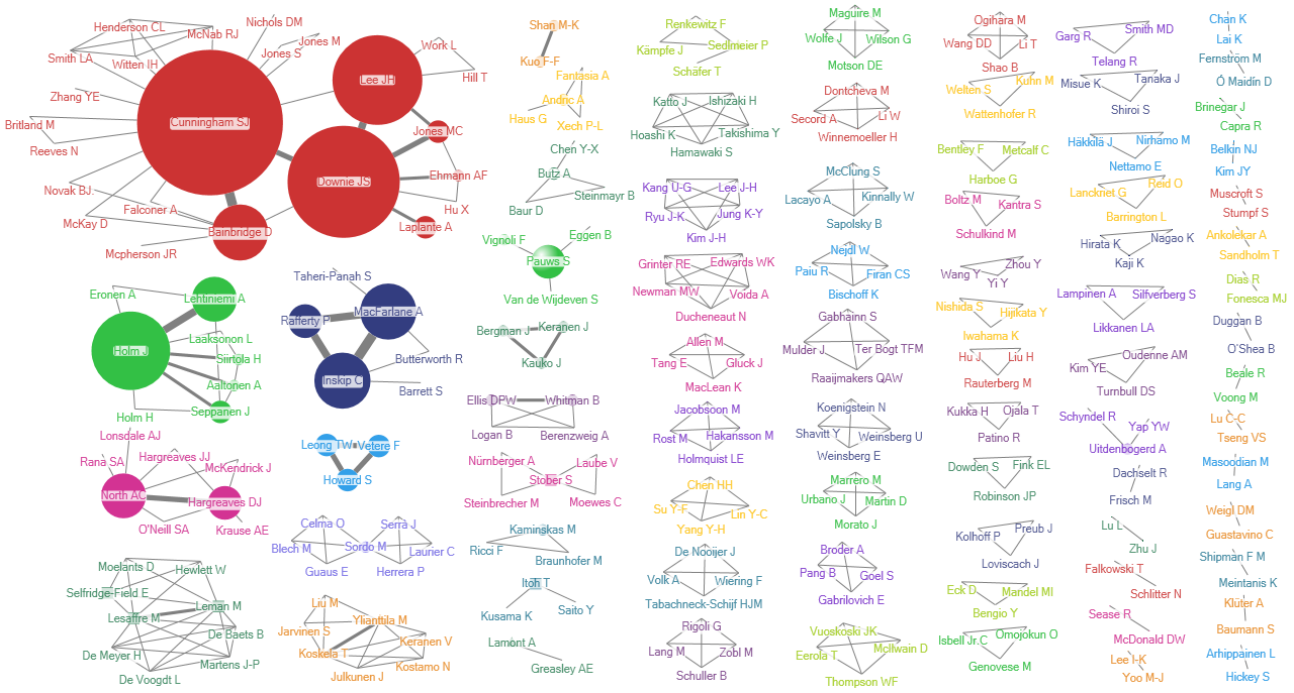

Figure 2. Co-authorship network among the authors

4. CITATION PATTERNS OF USER STUDIES

As part of the effort in understanding the impact of these studies, we investigated how often they were cited as of April 24, 2012 using the citation data from Google Scholar (GS). The reason for using GS is because the major publications in the field such as ISMIR conference proceedings are not indexed in other major databases such as EBSCO, Web of Science, etc. Also since we are interested in the scholarly as well as commercial impact of the user studies, being able to search for patents in addition to scholarly work on GS was deemed useful. We found a total of 3097 citations of 154 user studies in research publications and patents (one study did not show up). Figure 3 shows the distribution of the citation counts of the user studies in other materials. The X-axis represents the number of citations and the Y-axis represents the number of user studies that had the specified range of citation counts. The average number was 20.1 with the standard deviation of 44.5, the median of 5.5, and the maximum of 348.

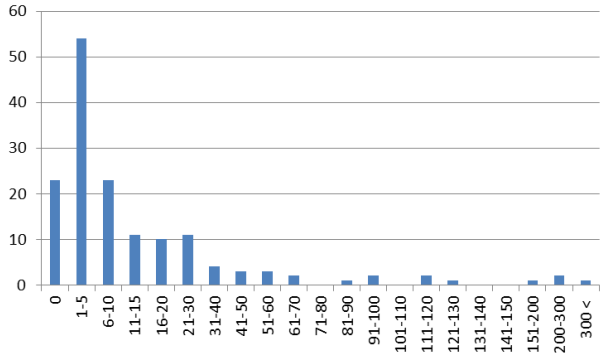

Figure 3. Distribution of the number of references of the user studies in other articles and patents

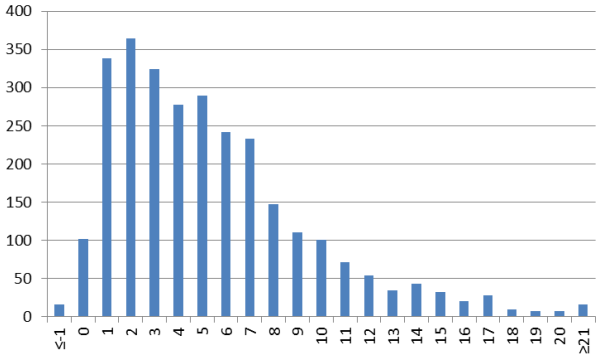

Figure 4. Distribution of the number of years for user studies to get cited

We were also interested in how long it takes for user studies to get cited. Figure 4 shows the distribution of the number of years it took for the user studies to get cited. This is based on the publication dates of the 2864 out of the 3097 citing articles and patents we were able to retrieve on GS. The X-axis represents the number of years passed after the publication of user studies and the Y-axis represents the number of citing articles/patents. The negative numbers (-1,-2) represent the cases where the author was self-citing a study that was yet to be published, or citing a study that was made available online before the print publication. The mean number of years was 5.48 with the standard deviation of 0.10, median of 5, and maximum of 90. About 40% (1144 out of 2864) were cited in 3 years or less after the user study was published and about 60% (1710) in 5 years or less. The citation pattern gradually decreases, and only about 15% (422) were cited after 10 years or more, and about 4% (119) after 15 years or more. The citation pattern suggests that the "perceived" relevance of the results quickly diminishes over

time. Since the majority of the user studies were published after 2000, for a more complete picture, this analysis will have to be replicated in 10 or 20 years.

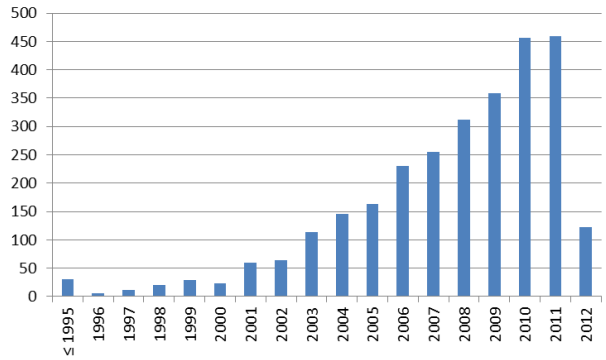

Figure 5. Distribution of the citing articles/patents by the publication year of citing articles/patents

Figure 5 shows the distribution of the citing articles/patents by their publication dates. Overall the numbers of citing articles/patents are showing a pattern of steady increase. Figure 3, 4, and 5, altogether seem to suggest that the user studies are in fact making growing impact to the field, although the impact of the studies tend to quickly diminish over time based on citation patterns.

Author/Year	Title	Ref
McNab et al./96	Towards the digital music library: tune retrieval from acoustic input	348
Berenzweig et al./04	A large-scale evaluation of acoustic and subjective music-similarity measures	230
North et al./00	The importance of music to adolescents	224
Levitin, D. J./94	Absolute memory for musical pitch: evidence from the production of learned melodies	184
Voida et al./05	Listening in: practices surrounding iTunes music sharing	121
Ellis & Whitman/02	The quest for ground truth in musical artist similarity	111
North et al./04	Uses of music in everyday life	111
Boltz et al./ 91	Effects of background music on the remembering of filmed event	100
Pauws & Eggen/03	Realization and user evaluation of an automatic playlist generator	100
Lee & Downie/04	Survey of music information needs, uses, and seeking behaviours: preliminary findings	82

Table 1. The top 10 most cited user studies

Table 1 presents the top 10 most cited user studies in the field. There is a mix of user experiments, evaluation of particular systems, studies of information behaviors and user-generated data, etc. The most heavily cited user study was by McNab et al. In this study, 10 users were asked to sing 10 songs from memory which were taped for analysis of key, pitch, contour, etc. The article was published in Proceedings of the First ACM International Conference on Digital Libraries and was cited widely in various papers on content-based music retrieval systems and measures. We believe that the heavy citation of this paper and also Levitin was at least partly due to the fact that they were early papers which dealt with content-based MIR, a topic which has dominated MIR research for the past decade. Studies by Berenzweig et al. and Ellis & Whitman explore measures for generating ground truth based on user data which is strongly relevant to the evaluation of algorithms, another big accomplishment of the past decade (i.e., MIREX). Studies of more general user needs and behaviors (North et al., Lee & Downie) may have had a broader impact to multiple areas related to music. The popularity of particular music application (Voida), association of music and other multimedia (Boltz et al.), and particular organizational measures (Pauws & Eggen) also seem to affect the heavy citation patterns.

5. RESEARCH DESIGN OF USER STUDIES

Lastly, we examined the studies more deeply in order to learn more about the research design of these user studies. We analyzed the content of the studies to discover the types and frequency of the various methods used (Fig. 6).

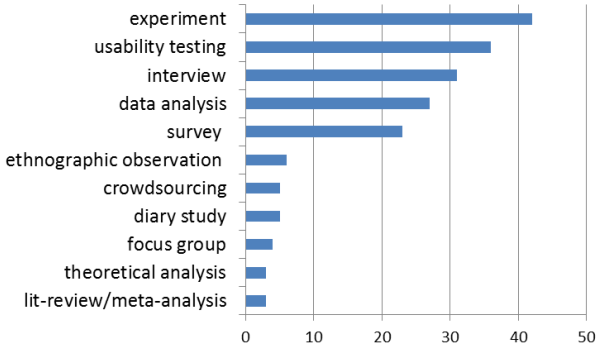

Figure 6. Research methods used in user studies

Experiment and usability testing were most commonly used (42%). The predominance of these methods may suggest that we are heavily focusing on evaluating what is out there rather than focusing on deeper problems or questions, a similar issue noted in other areas such as HCI [5]. These studies are primarily evaluating performance (e.g., error rate/time to perform task with a new system); identifying usability issues (i.e., interface design problems); or investigating acceptability of new system/interface. The full user-centered design process should include stages supporting coming to an understanding of the users, development of system prototype(s), and evaluation of the prototypes with users. However, relatively few papers presenting a new system include both an initial user requirements elicitation study and a follow-up performance/usability/acceptability study.

We also investigated the scale of these user studies by tabulating how many human subjects were involved in these studies. 124 user studies involved human subjects, and 26 analyzed human-generated data such as queries, tags, etc. 7 studies did not directly involve human subjects or human-generated data, as they were papers based on literature review, meta-analysis, or theoretical reasoning. Figure 7 shows the distribution of the number of human subjects included in the studies of real users. Many studies are of fairly small scale: 57 of the 124 studies (46%) involve 20 or fewer human subjects, and 102 stud-

ies (82%) involve 100 or fewer subjects.

Note that the active involvement of participants is limited for lab experiments and usability tests, which typically run at most a couple of hours. Ethnographic observations are constrained by the time available to the researcher to conduct observations. Interviews, surveys, and focus groups are attractive in that they may invite introspection and comment on music-related behavior over the long term, but at the cost of relying primarily on retrospection rather than direct, measurable experience. Only data analysis and the diary study naturally offer the opportunity to examine authentic music information behavior over the long term, though 'long term' studies are mainly of one to four weeks. In evaluations of specific systems, the common finding is that the users like the new system and find the new interface entertaining or novel—but it is generally unclear how or whether participant behavior may change after the novelty effect wears off.

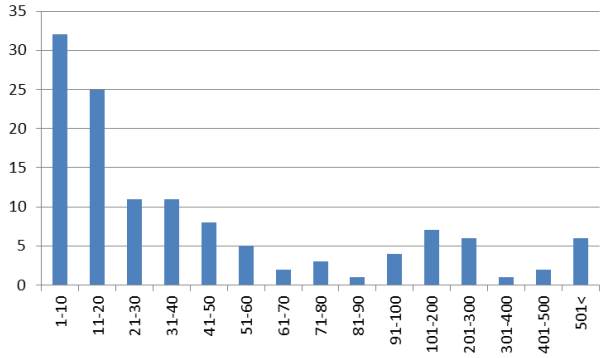

Figure 7. Number of subjects in user studies

6. DISCUSSION

Based on the results of our analyses as well as our own experiences in conducting music user studies, we provide a list of challenges/issues facing researchers who conduct music user studies which require further discussion. We believe that these issues are stemming from the uniqueness of the subject and the research domain.

6.1 Fast-changing Field

We believe that the speed with which the MIR field has evolved has had a strong affect on both the scale of user studies as well as the longevity of the research findings of these studies. The rapid development of tools and technologies for music storage, distribution, and experience in the past few decades has been remarkable. Some of the most popular music related services today such as Spotify or YouTube launched less than 6 years ago. This implies that how our users envision and expect from music services are most likely changing rapidly as well. Most of the young adults today probably never had to deal with physical media and grew up with various music streaming services. The results from studies that investigated how people find and purchase music on such physical media will have limited applicability today.

We conjecture that the fast-changing field is at least one of the reasons for the prevalence of small-scale studies. Large-scale studies take longer, in terms of recruiting human subjects, as well as collecting and processing data, in particular if researchers want to incorporate a qualitative component. Longitudinal studies are by definition time-consuming. Due to the rapidly changing environment, researchers are constantly under pressure to conduct and publish studies swiftly. This can be especially true for those who are trying to test a particular system or methods for providing access to music, as there is a good chance that by the time the research gets published in a journal, the results are already outdated. This may also explain a large proportion of user studies being published in conference or workshop proceedings.

6.2 Issue of Generalization

A large proportion of MIR user studies are small to moderate scale studies investigating a limited number of users. How does the scale of the study affect the generalizability of its results? Can we in fact make any reasonable inferences from studies of this scale that are generalizable to a larger user population? In addition, at least in certain parts of the world, it is not possible to obtain a comprehensive list of email addresses for the purpose of survey due to privacy concerns. This means that we often have to resort to convenience sampling, and study participants are in fact most frequently drawn from students or co-workers of the researchers which again can negatively affect the generalizability of our findings.

A point worth noting here is that researchers of music users are trying to grapple with this nebulous idea of users. Who really are our users? Where do we draw the boundaries? Music is so pervasive in our lives that it is difficult to know who is and is not affected by music. Moreover music is often enjoyed and sought out across different regions and cultures. Many of the MIR systems and services are now being used by global user base. Thus researchers of music users, in some sense, are expected to derive findings that can potentially have global implications on a wide range of users across space and time. Then how do we define and randomly sample this population in a practical sense? Even if we draw an artificial boundary and try to sample a smaller population, the subjects who participate in our studies will most likely be people who are interested in music to some degree. In this sense, the results are *always* likely to be biased.

Due to these issues, we believe that rather than aiming for generalizing the research findings, it might be helpful to take an alternative approach to understanding the purpose of these studies that each of these studies is discovering some piece of information about the users that is correct, but not comprehensive. When multiple pieces are put together, common themes emerge which we can generalize over multiple groups of users, as well as unique themes that can only apply to a particular user group.

6.3 Lack of Systematic Synthesis of Research Results

Although a large proportion (26%) of user studies were published in the proceedings of ISMIR conference, other studies were published in journals and conferences in multiple domains including LIS, HCI, Musicology, Psy-

chology, etc. We had to repeat our search in multiple databases in order to retrieve all these studies scattered in multiple domains. Despite of our best efforts, we would not be surprised if there were studies we were not able to find. We suspect that this is probably one of the reasons hindering the synergic impact of these studies. Without being able to easily find all the previous user studies that have dealt with similar research questions and user populations, we will essentially reinvent the wheel every time. In order to resolve this issue, there is a need for additional review articles such as [7] and also an archive of all the citation information of user studies related to music. As the first step, we made our list of user studies with full citation available on the web[1]. However, a static webpage is far from an ideal way for collecting and sharing this type of information. We believe a more sustainable solution is needed, managed by multiple stakeholders.

6.4 The Disconnect Between System/Evaluation Task Designers and User Studies Researchers

In MIREX, the evaluation task is typically proposed by researcher(s) who are involved in developing algorithms related to the task. In the MIR domain, however, researchers who conduct user studies are not always algorithm developers themselves; this is especially true for researchers engaged in studies of music users focusing on information needs or behaviors. This disconnect may be one of the reasons why we have not seen a significant change in the way evaluation tasks have been run over the past seven years since MIREX started in 2005. Some of the suggestions made in the user studies might be logistically impossible to implement, or the evaluators might not even agree with those suggestions. Without a more thorough investigation asking the system developers and the organizers of evaluation tasks, it will be premature to determine what the exact reasons are.

7. CONCLUSION AND FUTURE WORK

In this paper, we reflected on how music user studies have been conducted and published, and what impact these studies have had on the field. Findings from our analysis suggest that there may be multiple layers of barriers for the user studies to make a strong impact: lack of findability due to the scattered patterns of publication, weak connections among scholars, dominance of small scaled studies that are difficult to generalize, etc. The purpose of this work is to provide an opportunity for starting a discussion at the ISMIR where many stakeholders involved in MIR research can together explore potential solutions to the issues raised in this paper. Thus, we want to conclude our paper with a set of questions that need further discussion:

For researchers conducting user studies:

- How can we provide systematic and intelligent access to the work we produce? Is there a sustainable method? Maybe a collaboratively managed resource?
- Is it necessary to change our research questions, methods, study populations, or venues in increase impact and affect change in the field?

For system and evaluation task designers/developers:

- How do you find out about new research on users and keep yourself updated? Are there particular kinds of publications do you seek often?
- What kinds of user studies do you find most and least useful? What do you see as the grand challenge in the area of MIR user studies?

In our future studies, we plan to survey and interview designers/developers of music related services and systems as well as organizers of MIREX evaluation tasks in order to more deeply understand the impact of these user studies. Specifically, we are interested in how the information on users are disseminated and diffused in the MIR domain, and how that knowledge may or may not affect the ways music services/systems are designed and modified. A deeper understanding on what kind of user information is actually sought by system designers/developers will be significant for researchers of MIR user studies.

8. ACKNOWLEDGMENTS

We thank Gary Gao and Tiffany Huang at University of Washington for their valuable contributions to this project.

9. REFERENCES

[1] S. Bradford: “Sources of information on specific subjects,” *Journal of Information Science,* Vol. 10, No. 4, pp. 173-180, 1985.

[2] S. J. Cunningham: “User studies: a first step in designing an MIR testbed,” In *The MIR/MDL Evaluation Project White Paper Collection* (3rd ed.). Champaign, Illinois: GSLIS, pp. 19-21, 2003.

[3] J. S. Downie: “Music information retrieval,” *Annual Review of Information Science and Technology*, Vol. 37, ed. B. Cronin, Medford, NJ: Information Today, pp. 295-340, 2003.

[4] J. Futrelle and J. S. Downie: “Interdisciplinary communities and research issues in music information retrieval,” *Proceedings of the 3rd ISMIR Conference*, pp. 215-221, 2002.

[5] S. Greenberg and B. Buxton: “Usability evaluation considered harmful (some of the time),” *Proceedings of the CHI’08*, pp. 111-120, 2008.

[6] D. L. Hansen, B. Schneiderman, and M. A. Smith: *Analyzing social media networks with NodeXL: Insights from a connected world*, Burlington, MA: Morgan Kaufmann, 2011.

[7] D. M. Weigl and C. Guastavino: “User studies in the music information retrieval literature,” *Proceedings of the 12th ISMIR Conference*, pp. 335-340, 2011.

[1] http://www.jinhalee. com/miruserstudies