Report on the Panels and Workshops of the Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation Frameworks Project

J. Stephen Downie
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
jdownie@uiuc.edu

Introduction:

Music Information Retrieval (MIR) and Music Digital Library (MDL) are two interrelated, multidisciplinary research areas with a growing community of involved parties. MIR/MDL research brings together computer scientists, audio engineers, librarians, musicologists, educators and business executives in a common effort to provide robust mechanisms for organizing, storing and accessing the world's ever-increasing volume of music. Newcomers to the world of MIR/MDL are invited to read Downie [8], Futrelle and Downie [10], and Byrd and Crawford [3] for overviews of issues currently being examined by MIR/MDL researchers.

The MIR/MDL community has recognized for several years now that it needs a more formal set of evaluation tools with which to scientifically compare and contrast the techniques its researchers have been developing. In the Summer of 2002, this author was able to secure funds from the Andrew W. Mellon Foundation to begin exploratory work on the development of evaluation standards for the MIR/MDL community. The formal name of the Mellon-funded project was "Establishing Music Information Retrieval and Music Digital Library Evaluation Frameworks: Preliminary Foundations and Infrastructures." The mandate of the "Evaluation Project" was to "…lay the foundation for the formation of meaningful and comprehensive MIR/MDL evaluation through the identification and or creation of standardized test collections, retrieval tasks and performance metrics" [7].

An integral part of the "Evaluation Project" are the two workshops and two panels from which recommendations and commentary have been (and will be) drawn. The first two workshops and one panel have been held. These were:

1. Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation. Joint ACM/IEEE Conference on Digital Libraries (JCDL '02), Portland, OR, 18 July, 2002.

2. Panel on Music Information Retrieval Evaluation Frameworks. 3rd International Conference on Music Information Retrieval (ISMIR 2002), IRCAM, Paris, 17 October, 2002.

3. Workshop on the Evaluation of Music Information Retrieval (MIR) Systems. 26th Annual International ACM SIGIR Conference (SIGIR '03), Toronto, 1 August, 2003.

A second panel session has yet to meet at the time of writing. The second panel session is to convene at the 4th International Conference on Music Information Retrieval (ISMIR 2003) in Baltimore on 30 October, 2003. This panel session is entitled, "Making Music Information Retrieval Evaluation Scenarios a Reality."

In this report we will highlight the central questions being explored at each of the aforementioned meetings. We will briefly highlight the emergent themes from those meetings that have already concluded. We will also highlight the goals and structure of the fourth, yet-to-be-held, ISMIR 2003 panel session.

This material is based upon work supported by the National Science Foundation under Grant Nos. NSF IIS-0340597 and NSF IIS-0327371.

Central Questions

Participation in the two workshops and the first panel occurred at two-levels. First, a series of formal solicitations went out asking for expert opinions in the form of white papers. These white papers, which are now available as "The MIR/MDL Evaluation Project White Paper Collection, 3rd Edition" (http://music-ir.org/evaluation/wp.html), were presented by their authors at the meetings. Second, audience members and fellow panelists were asked to respond the recommendations being put forward by the white paper presenters.

The writers of white paper submissions were asked to situate their commentaries within the framework of the central underlying questions of the project:

1. How do we determine, and then appropriately classify, the tasks that should make up the legitimate purviews of the MIR/MDL domains?
2. What do we mean by "success"? What do we mean by "failure"?
3. How will we decide that one MIR/MDL approach works better than another?
4. How do we best decide which MIR/MDL approach is best suited for a particular task?

To better contextualize the project's central questions, the following set of more specific sub-questions was also given to white paper authors:

1. How do we adequately comprehend the complex nature of music information so that we can properly construct our evaluation recommendations?
2. How do we adequately capture the complex nature of music queries so proposed experiments and protocols are well-grounded in reality?
3. How do we deal with the "relevance" problem in the MIR context (i.e., What does "relevance" really mean in the MIR context?)?
4. How do we continue to the expansion of a comprehensive collection of music materials to be used in evaluation experiments?
5. How do we manage the interplay between TREC-like and other potential evaluation paradigms?
6. How do we integrate the evaluation of MIR systems within the larger framework of IR evaluation (i.e., What aspects are held in common with traditional IR? What aspects are unique to MIR?)?

Meetings: Highlights and Themes

JCDL '02 Workshop (18 July, 2002)

Dr. Ellen Voorhees, Project Manager of the National Institute of Standards and Technology's, *Text REtrieval Conference* (TREC) (*http://trec.nist.gov*), presented the keynote address [31]. Dr. Voorhees' presentation outlined the potential applicability of the TREC evaluation paradigm to the needs of the MIR/MDL community. Fifteen other authors, presenting eleven white papers, also participated in the JCDL workshop. It was at this meeting that the term "TREC-like" was adopted, as attendees made it clear that MIR/MDL systems, because they deal with music, are not directly analogous to text retrieval systems. Other issues raised for more detailed examination included the successful integration of multiple formats (i.e., audio [22, 25], symbolic representations [1, 20], metadata and scores [17]), analysis of real-world queries (i.e., needs and uses) [4, 9], and the set of tasks to be examined [19] (e.g., recreational uses, educational uses, scholarly uses [15], etc.). The overall consensus was that work should proceed on developing TREC-like evaluations. However, such work would be contingent on the provisos that:

1. any TREC-like approach developed be centered on the unique nature of music information and not "artificially imposed" on MIR/MDL systems simply because of the perceived "convenience" of the approach;
2. the importance of integrating music metadata not be ignored; and,
3. the TREC-like approach not become the sole means of evaluating the performance of MIR/MDL systems.

ISMIR 2002 Panel (17 October, 2002)

Dr. Edie Rasmussen, Professor at the University of Pittsburgh's, School of Information Sciences, delivered the keynote white paper [24]. Dr. Rasmussen further developed the TREC-like evaluation through her personal insights on the merits of the TREC paradigm. Twelve authors also contributed eight ISMIR 2002 white papers. Almost every paper addressed issues surrounding the requisite components of the large-scale test collections needed for TREC-like evaluations (e.g., [13, 28, 29]). One paper extended the large-scale test collection notion to encompass multiple test collections housed in multiple locations and interconnected via a Music GRID [6]. The importance of delineating the nature of music-specific retrieval tasks — and their related queries — to be used in evaluation testing was another significant theme (e.g., [18, 26, 30]). The idea that the TREC-like evaluation scenario not be the sole evaluation approach used was iterated in [26]. Notwithstanding the caveats expressed in [26], the central theme of the panel can be summarized as "How do we make TREC-like MIR/MDL evaluations a reality?"

SIGIR 2003 Workshop (1 August, 2003)

Given the strong momentum of community support for the TREC-like evaluation paradigm, participants in the SIGIR 2003 workshop were prompted with some more specific topics for consideration. Participants were asked to provide their thoughts on:

1. How best to ground evaluation methods in real-world requirements.
2. How to facilitate the creation of data-rich query records that are both grounded in real-world requirements and neutral with respect to retrieval technique(s) being examined.
3. How the possible adoption, and subsequent validation, of a "reasonable person" approach to "relevance" assessment might address the MIR "relevance" problem.
4. How to develop new models and theories of "relevance" in the MIR context.
5. How to evaluate the utility, within the MIR context, of already-established evaluation metrics (e.g., precision and recall, etc.).
6. How to support the ongoing acquisition of music information (audio, symbolic and metadata) to enhance the development of a secure, yet accessible, research environment that allows researchers to remotely participate in the use of the proposed testbed.

Dr. Beth Logan of HP Labs presented the workshop keynote white paper, which she co-authored with Daniel Ellis and Andrew Berenzwieg of Columbia University [16]. Logan's multifaceted presentation touched upon a wide range of evaluation projects that she and colleagues are pursuing that deal with issues of testbed establishment, the use of extracted features (as opposed to raw music) to facilitate community sharing of materials (to overcome the copyright issues that are stifling MIR/MDL research) and the grounding of automatically-generated similarity comparisons in real-world, human-generated, similarity data. Seven other white papers, written by fourteen contributing authors were presented. Four papers specifically considered issues surrounding the establishment of TREC-like evaluation scenarios [5, 11, 23, 27]. Issues addressed included the need for specific TREC-like tracks[23], the lessons to be learned from previous benchmarking and evaluation work[5, 11, 27], and the role that the proposed Music GRID could play in evaluation testing [23]. One paper provided strong evidence for requiring that the TREC-like evaluation testbed include real-world music (i.e., think "popular" and/or "well-known") as opposed to "artificially" created music [14]. (Some research teams had previously proposed that they could overcome copyright issues by artificially generating or custom composing music without compromising the evaluation results that would be based upon such "artificial" music [12].) Two papers were premised on the significance of human factors in MIR/MDL research and evaluation. The first human-factors paper stressed the importance of modeling the human errors involved in the use of Query-by-Humming (QBH) systems [21]. The second human-factors paper convincingly argued that the human perception of music information is richly multifaceted and as such it must be evaluated from many disciplinary perspectives if the results are to be valid[2].

ISMIR 2003 Panel (30 October, 2003)

Based upon the preliminary work conducted under the auspices of the three meetings discussed above, this author and his colleagues at the University of Illinois at Urbana-Champaign (UIUC) have recently been

awarded research funding to begin construction of a large-scale, internationally-accessible, MIR/MDL testbed system. The testbed is being established with substantial short-term funding from the National Science Foundation (approximately $100,000 over one year) and longer-term funding from the Andrew W. Mellon Foundation ($390,000 over four years).

To begin populating the testbed with real-world music materials, agreements-in-principle from two significant music content rights-holder have been obtained. HNH Hong Kong International, Ltd. (owner of the Naxos and Marco Polo recording labels (http://www.naxos.com)), has agreed to let the MIR community have access to its entire catalogue of Classical, Jazz, and Asian digital recordings via the secure, yet accessible, testing and development system we are developing at UIUC's National Center for Supercomputing Applications (NCSA). This generous gesture represents approximately 3,000 CDs or about 3 Terabytes of digital music information. All Media Guide (http://www.allmusic.com), has also agreed to let us add to the collection its vast database of music metadata including descriptive catalogue records, discographies, and recording classifications.

Because of the unique opportunity that these rights-holders have afforded us, it is important that MIR/MDL testbed be constructed with three central features in mind:

1. security for the property of the rights-holders;
2. accessibility for both UIUC and external researchers; and,
3. sufficient computing and storage infrastructure to support the computationally- and data-intensive techniques being investigated by the various research teams.

The "Virtual Research Labs" (VRL) being developed at NCSA to make the testbed secure, yet accessible, will be instrumental in supporting the multidisciplinary nature of MIR/MDL research and development. We also hope that the VRLs will also help retain and encourage the interest and participation of such non-computer experts as librarians, musicologists, educators and business executives. These tools, based upon NCSA's, "D2K Framework," should prove powerful enough to support the advanced research of technology graduate students and their supervisors, yet flexible enough to permit non-technology graduate and undergraduate students to successfully interact with the music collection in accordance with their particular needs. A brief outline of the project's goals and structures can be found at http://music-ir.org/evaluation/wp3/wp3_appendixC.pdf.

It is the purpose of the ISMIR 2003 panel to bring together experts in each of the testbed's central aspects: the infrastructural supercomputing resources at NCSA, the music in its symbolic notation formats, the music in its audio formats, and the metadata associated with the music. Each domain expert has been invited to contribute thoughts, suggestions and guidelines so as to facilitate the successful construction of the testbed for use by members of the MIR/MDL research community. Audience members will also be encouraged to engage with the expert panelists and offer their own thoughts, suggestions and guidelines. All expert and audience input will be considered in future testbed design decisions. The expert panelists include:

David Tcheng, NCSA—computational infrastructure issues, requirements for the VRLs
Perry Roland, University of Virginia—symbolic music issues
Jon W. Dunn, Indiana University—music metadata issues
Brad Eden, University of Nevada at Las Vegas—music metadata issues
Mark Sandler, Queen Mary College, University of London—audio issues
Josh Reiss, Queen Mary College, University of London—audio issues

Concluding Comments and Acknowledgements

Our thanks to everyone who has helped to make the MR/MDL Evaluation Frameworks Project a success. The ongoing MIR/MDL community support and involvement that has brought the establishment of scientific MIR/MDL evaluation scenarios ever closer to reality has been nothing short of remarkable. The consistent thoughtfulness of the formal white papers and the commentary they have generated from community members is simply astounding. Support from the JCDL and SIGIR programme committees has

also played a significant role by allowing a wider range of researchers to participate in the project. The generous and farsighted involvement of HNH and All Media Guide in providing valuable music materials for the testbed must be applauded. Finally, we must give special thanks to the National Science Foundation and the Andrew W. Mellon Foundation for their important financial support.

Reference List[1]

[1] Bainbridge, D. (2002). Towards a Workbench for Symbolic Music Information Retreival. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 14-16). Champaign, IL: GSLIS.

[2] Batlle, E., Guaus, E., & Masip, J. (2003). Open Position: Multilingual Orchestra Conductor. Lifetime Opportunity. The MIR/MDL Evaluation Project White Paper Collection (3rd ed., pp. 86-89). Champaign, IL: GSLIS.

[3] Byrd, D., & Crawford, T. C. (2002). Problems of Music Information Retrieval in the Real World. Information Processing and Management, 38, 249-272.

[4] Cunningham, S. J. (2002). User Studies: A First Step in Designing an MIR Testbed. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 17-19). Champaign, IL: GSLIS.

[5] Doraisamy, S., & Rüger, S. M. (2003). Emphasizing the Need for TREC-like Collaboration Towards MIR Evaluation. The MIR/MDL Evaluation Project White Paper Collection (3rd ed., pp. 90-96). Champaign, IL: GSLIS.

[6] Dovey, M. J. (2002). Music GRID: A Collaborative Virtual Organization for Music Information Retrieval Collaboration and Evaluation. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 50-52). Champaign, IL: GSLIS.

[7] Downie, J. S. Establishing Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation Frameworks: Preliminary Foundations and Infrastructures. In The MIR/MDL Evaluation Project White Paper Collection (pp. 3-6). Champaign, IL: GSLIS.

[8] Downie, J. S. (2003). Music Information Retrieval. Annual Review of Information Science and Technology, 37, 295-340.

[9] Futrelle, J. (2002). Three Criteria for the Evaluation of Music Information Retrieval Techniques against Collections of Musical Material. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 20-22). Champaign, IL: GSLIS.

[10] Futrelle, J., & Downie, J. S. (2002). Interdisciplinary Communities and Research Issues in Music Information Retrieval. In Third International Conference on Music Information Retrieval (pp. 215-221). IRCAM.

[11] Goodrum, A. A. (2003). If It Sounds As Good As It Looks: Lessons Learned From Video Retrieval Evaluation. The MIR/MDL Evaluation Project White Paper Collection (3rd ed., pp. 97-102). Champaign, IL: GSLIS.

[12] Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC Music Database: Popular, Classical, and Jazz Music Databases. In Third International Conference on Music Information Retrieval (pp. 287-288).

[1] All papers associated with The MIR/MDL Evaluation Project White Paper Collection can be found at http://music-ir.org/evaluation/wp.html.

[13] Herrera-Boyer, P. (2002). Setting Up an Audio Database for Music Information Retrieval Benchmarking. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 53-55). Champaign, IL: GSLIS.

[14] Hoashi, K., Matsumoto, K., & Inoue, N. (2003). Comparison of User Ratings of Music in Copyright-free Databases and On-the-market CDs. The MIR/MDL Evaluation Project White Paper Collection (3rd ed., pp. 103-106). Champaign, IL: GSLIS.

[15] Issacson, E. J. (2002). Music IR for Music Theory. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 23-26). Champaign, IL: GSLIS.

[16] Logan, B., Ellis, D. P. W., & Berenzweig, A. (2003). Toward Evaluation Techniques for Music Similarity. The MIR/MDL Evaluation Project White Paper Collection (3rd ed., pp. 81-85). Champaign, IL: GSLIS.

[17] MacMillan, K. (2002). Common Music Notation as a Source for Music Information Retrieval. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 27-28). Champaign, IL: GSLIS.

[18] Meek, C., Birmingham, W. P., & Pardo, B. (2002). What is a Sung Query? In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 56-57). Champaign, IL: GSLIS.

[19] Melucci, M., & Orio, N. (2002). A Task-Oriented Approach for the Development of a Test Collection for Music Information Retrieval. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 29-31). Champaign, IL: GSLIS.

[20] Montalvo, J. (2002). A MIDI Track for Music Information Retrieval. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 32-32). Champaign, IL: GSLIS.

[21] Pardo, B., & Birmingham, W. P. (2003). Query by Humming: How Good Can It Get? The MIR/MDL Evaluation Project White Paper Collection (3rd ed., pp. 107-109). Champaign, IL: GSLIS.

[22] Pardo, B., Meek, C., & Birmingham, B. (2002). Comparing Aural Music-Information Retrieval Systems. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 34-36). Champaign, IL: GSLIS.

[23] Pickens, J. (2003). Tracks and Topics: Ideas for Structuring Music Retrieval Test Collections and Avoiding Balkanization. The MIR/MDL Evaluation Project White Paper Collection (3rd ed., pp. 110-113). Champaign, IL: GSLIS.

[24] Rasmussen, E. (2002). Evaluation in Information Retrieval. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 45-39). Champaign, IL: GSLIS.

[25] Reiss, J., & Sandler, M. (2002). Benchmarking Music Information Retrieval Systems. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 37-42). Champaign, IL: GSLIS.

[26] Reiss, J., & Sandler, M. (2002). Beyond Recall and Precision: A Full Framework for MIR System Evaluation. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 58-63). Champaign, IL: GSLIS.

[27] Reiss, J., & Sandler, M. (2003). MIR Benchmarking: Lessons Learned from the Multimedia Community. The MIR/MDL Evaluation Project White Paper Collection (3rd ed., pp. 114-120). Champaign, IL: GSLIS.

[28] Richard, G. (2002). Towards Large Databases for Music Information Retrieval Systems Development and Evaluation. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 64-67). Champaign, IL: GSLIS.

[29] Rüger, S. (2002). A Framework for the Evaluation of Content-Based Music Information Retrieval using the TREC Paradigm. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 68-70). Champaign, IL: GSLIS.

[30] Södring, T., & Smeaton, A. F. (2002). Evaluating a Music Information Retrieval System: TREC Style. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 71-78). Champaign, IL: GSLIS.

[31] Voorhees, E. M. (2002). Whither Music IR Evaluation Infrastructure: Lessons to be Learned from TREC. In The MIR/MDL Evaluation Project White Paper Collection (2nd ed., pp. 7-13). Champaign, IL: GSLIS.