Toward the Scientific Evaluation of Music Information Retrieval Systems

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

ABSTRACT

This paper outlines the findings-to-date of a project to assist in the efforts being made to establish a TREC-like evaluation paradigm within the Music Information Retrieval (MIR) research community. The findings and recommendations are based upon expert opinion garnered from members of the Information Retrieval (IR), Music Digital Library (MDL) and MIR communities with regard to the construction and implementation of scientifically valid evaluation frameworks. Proposed recommendations include 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; adoption, and subsequent validation, of a "reasonable person" approach to "relevance" assessment; and, the development of a secure, yet accessible, research environment that allows researchers to remotely access the large-scale testbed collection.

1 INTRODUCTION

Music Information Retrieval (MIR) is a multidisciplinary research endeavor that strives to develop innovative content-based searching schemes, novel interfaces, and evolving networked delivery mechanisms in an effort to make the world's vast store of music accessible to all. Some teams are developing "Query-by-Singing" systems (e.g., Haus and Pollastri (2001), Birmingham et al. (2001)), some "Query-by-Note" systems (e.g., Doraisamy and Rüger (2002), Pickens (2000)), some "Query-by-Example" systems (e.g., Haitsma and Kalker (2002), Harb and Chen (2003)), some comprehensive music recommendation and distribution systems (e.g., Pauws and Eggen (2002), Logan (2002)), some musical analysis systems (e.g., Kornstädt (2001), Barthélemy and Bonardi (2001)), and so on. Good overviews of MIR's interdisciplinary research areas can be found in Downie (2003), Byrd and Crawford (2002), Futrelle and Downie (2002).

In this paper, Section 1 outlines the current scientific problem facing MIR research. Sections 2-3 report upon the findings-to-date of the "MIR/MDL Evaluation Project," with issues surrounding the creation of a TREC-like evaluation paradigm for MIR as the central focus. Section 4 highlights the progress being made concerning the establishment of the necessary test collection(s). Section 5 concludes with a summary and outlines some of the key challenges uncovered that require further investigation.

1.1 Current Scientific Problem

Notwithstanding the promising technological advancements being made by the various research teams, MIR research has been plagued by one overarching difficulty: There has been no way for research teams to scientifically compare and contrast their various approaches. This is because there has existed:

1. no standard collection of music against which each team could test its techniques;
2. no standardized sets of performance tasks; and,
3. no standardized evaluation metrics.

The MIR community has long recognized the need for a more rigorous and comprehensive evaluation paradigm. A formal resolution expressing this need was passed, 16 October 2001, by the attendees of the *Second International Symposium on Music Information Retrieval* (ISMIR 2001). (See `http://music-ir.org/mirbib2/resolution` for the list of signatories.)

Over a decade ago, the National Institute of Standards and Technology developed a testing and evaluation paradigm for the text retrieval community, called TREC (*Text REtrieval Conference*; `http://trec.nist.org/overview.html`). Under this paradigm, each text retrieval team is given access to:

1. a standardized, large-scale test collection of text;
2. a standardized set of test queries; and,
3. a standardized evaluation of the results each team generates.

Because of the strong overlap between the MIR and the traditional IR communities, many informally suggested that MIR researchers should explore the TREC model as a key component of MIR evaluation. In July 2002, the author secured funding from the Andrew W. Mellon Foundation to begin exploratory work on the "Establishing Music Information Retrieval (MIR) and Music Digital Libraries (MDL) Evaluation Frameworks Project." The mandate of the "MIR/MDL Evaluation Project" is "…to establish the infrastructural 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…"(Downie, 2002).

2 Data Collection Method

The Delphi method (Linstone and Turoff, 1975) of data collection forms the basis of the analytic modality employed by the "MIR/MDL Evaluation Project." The Delphi approach is an iterative method wherein initial prompting questions are put before a community of experts and their opinions solicited. These opinions are then brought together and trends uncovered. The resultant data then is fed back to the community for further input and refinement. The goal of this approach is to allow consensus on the uncovered trends to emerge naturally from these learned opinions. There are nine prompting questions used in this study providing specific contexts for participants (Downie, 2002). In addition to the aforementioned nine detailed/specific questions, each of the participants is presented with the four, more basic, questions that represent the intellectual underpinnings of the project (Downie, 2002):

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 whether 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?

Three rounds of input are planned for the "MIR/MDL Evaluation Project." Two of these have already been concluded. The third, and final, round will close in August 2003. The input rounds consist of a formal solicitation for White Papers from the MIR, MDL and IR communities with the prompting and primary questions as the basis for discussion. Each of the completed rounds culminated in the convening a special meeting wherein the participants were able to expound upon their White Paper opinions and exchange ideas. The White Papers from each round are being collected in successive editions of *The MIR/MDL Evaluation White Paper Collection*. See `http://music-ir.org/evaluation` for the most recent edition. Information about each of the first two input rounds follows.

3 Emergent Themes and Commentary

Round #1 Meeting: "The Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation" was held at the *Second Joint Conference on Digital Libraries* (JCDL 2002) in July of 2002 (`http://www.ohsu.edu/jcdl`). 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 (Voorhees, 2002). Her White Paper presentation focussed on 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 Round #1. The creation of a TREC-like evaluation model was the central theme played out by the participants. "TREC-like" is used here deliberately, as attendees made it clear that MIR/MDL systems, because they deal with music, are not directly analogous to text retrieval systems. Issues raised for more detailed examination included the successful integration of multiple formats (i.e., audio—(Reiss and Sandler, 2002a; Pardo, Meek, and Birmingham, 2002), symbolic representations—(Bainbridge, 2002; Montalvo, 2002), metadata and scores—(MacMillan, 2002)), analysis of real-world queries (i.e., needs and uses (Cunningham, 2002; Futrelle, 2002)), and the set of tasks to be examined (Melucci and Orio, 2002), including recreational uses, educational uses, scholarly uses (Issacson, 2002), etc.. In short, the consensus was that work should proceed on developing TREC-like evaluations with 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 integration of music metadata not be overlooked; and,
3. the TREC-like approach not become the sole means of evaluating the performance of MIR/MDL systems.

Round #2 Meeting: "The Panel on Music Information Retrieval Evaluation Frameworks" was held as part of ISMIR 2002. Dr. Edie Rasmussen, (Prof., University of Pittsburgh) delivered the keynote White Paper (Rasmussen, 2002) which further developed the TREC-like evaluation theme by providing insights on the strengths and weaknesses of the TREC paradigm. Twelve authors also contributed eight Round #2 White Papers. Almost every paper addressed issues surrounding the requisite components of the large-scale test collections needed for TREC-like evaluations (e.g., Herrera-Boyer (2002), Rüger (2002), Richard (2002)). One paper extended the large-scale test collection notion to encompass multiple test collections housed in multiple locations and interconnected via a Music GRID (Dovey, 2002). 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., Meek, Birmingham, and Pardo (2002), Södring and Smeaton (2002), Reiss and Sandler (2002b)). The idea that the TREC-like evaluation scenario not be the sole evaluation approach used was iterated in Reiss and Sandler (2002b). Notwithstanding the caveats expressed by Reiss and Sandler (2002b), so strongly did the TREC leitmotif run through the White Papers of Round #2 that it is safe to summarize the consensus as "How do we move forward on making a TREC-like evaluation scenario for MIR/MDL a reality?"

3.1 Commentary on Emergent Themes

Given the overwhelming consensus on the establishment of a TREC-like evaluation paradigm, why is it that a TREC-like approach has not been adopted already? Participants

consistently touched upon four problem areas that will provide some insight into this question:

1. the complexity of music information;
2. the complexity of music queries;
3. the nature of relevance within the context of MIR and the applicability of *precision* and *recall* as evaluation metrics (terms defined in Section 3.1.3) and,
4. the lack of access to music collections brought about by intellectual property law as practiced by the music industry.

The ordering of first three is significant. The complexity of music information can be seen as the cause of the complexity found in real-world music queries. Query complexity, in turn, contributes to the difficulties associated with the assessment of relevance (and thus the applicability of precision and recall as evaluation metrics).

3.1.1 Problem #1: The complexity of music

Music information is inherently more complex than text information. Music information is a multifaceted amalgam of pitch, tempo, rhythmic, harmonic, timbral, textual (i.e., lyrics and *librettti*), editorial, praxis, and bibliographic elements. Music can be represented as scores, MIDI files and other discrete encodings, and in any number of analogue and digital audio formats (e.g., LPs, tapes, MP3s, CDs, etc.). Unlike most text, music is extremely plastic; that is, a given piece of music can be transposed, have its rhythms altered, its harmonies reset, its orchestration recast, its lyrics changed, and so on, yet somehow it is still perceived to be the "same" piece of music.

The interaction of music's complexity and plasticity make the selection of possible retrieval elements extraordinarily problematic. This, in turn, leads to difficulties on four fronts:

1. Until such time as there is a "universal" music repository, the determination of the most "representative" versions (and formats) of music objects for use in building test collections remains an open problem. Given the problems outlined in Section 3.1.4, consensus is that the MIR community will "make do" with whatever it will be fortunate to acquire so long as efforts are made to expand the collection over time. Section 4 discusses progress being made to alleviate this problem
2. Test collection size is real concern. Because of the need for multiple instances of symbolic, audio and metadata information for each piece in the collection, a MIR testbed will approach, if not exceed, the storage limits of most research facilities. That audio files tend to be large, relative to their symbolic counterparts, also contributes significantly to this problem. A large-scale, multi-format music test collection requires storage in the terabyte range: approximately two to three orders of magnitude greater than the gigabyte-range text databases used in the *ad hoc* TREC evaluations (Voorhees, 2002). A solution to the large dataset problem is discussed in Section 4.
3. Establishing and maintaining workable linkages between the various manifestations of each work (i.e., linkages between and among a given piece's audio, symbolic and metadata information) is a non-trivial research problem (Dunn, Davidson and Isaacson, 2001; Smiraglia, 2001). Much more work needs to be done on this problem in order that one retrieval method is not "privileged" over another. This leads to the notion of "retrieval neutrality" discussed in Section 3.1.2.
4. Music queries — being themselves a kind of music information — are also plastic, complex and multifaceted. This implies that the formalized encapsulation of queries in the "query records" for use in TREC-like testing (i.e., "topic statements") must, from the outset, be designed to reflect this fact. More about the "query problem" next.

3.1.2 Problem #2: The complexity of music queries

There is a much-lamented paucity of formal literature reporting upon the analyses of the real-world information needs and uses of MIR/MDL users (Downie, 2003; Byrd and Crawford, 2002; Futrelle and Downie, 2002). In fairness, this paucity is partially caused by the non-existence of MIR/MDL systems containing music that users actually want. However, when such studies are attempted (e.g., Downie (1994), Itoh, (2000), Kim and Belkin (2002), Downie and Cunningham, (2002)), the disconnect between assumptions commonly made by MIR researchers concerning the nature of music queries (i.e., simple hummed melodies, retrieval of known-items, identification of songs users have in-hand, etc.) and the real-world situation, is remarkable. To illustrate this point, compare Fig. 2 (a TREC topic statement (Voorhees, 2002)) with Fig. 1 (a real-world music query (Cunningham, 2002)), both presented on the next page. Table 1 also illustrates the wide variety of information types contained in real-world music queries along with the wide variety of intended uses for the sought-after music.[1]

The consensus opinion among community members is that great care must be taken in developing the TREC-like query records, for their use will have significant scientific ramifications, especially with regard to the validity of the resultant evaluation experiments. While there is much work yet to be done on finalizing the specific form of the TREC-like query records, a set of first principles is emerging. The query records developed must:

1. be grounded in real-world needs and uses;
2. be representative of the complexity of real-world queries (see Table 1);
3. be neutral with regard to the retrieval method employed; and,
4. be data-rich so realistic and meaningful "relevance" judgements can be made. (Discussed in Section 3.1.3.)

[1] The percentage values, which are most likely idiosyncratic to the population examined, are less important than the categories themselves.

```
From: XXXXXXXXX
Subject: Early 80's - Please identify this song! (it's *very* difficult, though)

Newsgroups: alt.music.lyrics

Date: 2000-12-14 09:42:24 PST

Hi, thiis is so difficult because I only remember those damn FRAGMENTS of it, which can (in combination
with possible errors) make it VERY difficult to identify this song!

But I'll try my best to make myself clear as possible.

This song MUST be from the period 1979-1984, most likely 1981 or 1982.

Tempo: about 120 bpm

Sounds VERY close to a SAGA or Asia tune (maybe it is SAGA even! ;)

OK here I go...(gonna add the chords for you guitarists out there ;)

[verse 1]

F    C                  Bb            Bb C
Crazy ................. onto the ..... café
     F         C            Bb
I'm drinking coffee, she came away
      F      C              Bb      Bb         C
She ordered ............... precious sum of money ???
F          C        Bb
deedeedeedeedeedeedeedee....<remaining text deleted>


Ohohohoo
[(instrumental) F C Bb Bb C F C Bb]
[verse 2] [...]
[chorus]
```

Figure 1. A real-world information request posted to alt.music.lyrics as presented in Cunningham (2002).

<num> Number: 409
<title> legal, Pan Am, 103
<desc>Description:
What legal actions have resulted from the destruction of Pan Am Flight 102 over Lockerbie, Scotland, on December 21, 1988?
<narr> Narrative:
Documents describing any charges, claims, or fines presented to or imposed by any court or tribunal are relevant, but documents that discuss charges made in diplomatic jousting are not relevant.

Figure 2. A TREC topic statement from Voorhees (2002).

The "retrieval neutrality" principle requires some explication. The MIR community can be divided roughly into two camps: 1) those engaged in symbolic retrieval research; and, 2) those exploring audio- and signal-processing techniques. Given that no data exist on the comparative strengths and weaknesses of the techniques employed across the two camps, the consensus is that the TREC-like evaluation paradigm — at least in its early stages — must provide a means to make informed assessments on the relative merits of the two approaches. The idea of "symbol-only" and "audio-only" tracks is therefore not an attractive initial option. Related to this matter, the notion of task-specific tracks, analogous to the video, interactive, natural language processing, etc. tracks in TREC, has been discussed. However, the apparent consensus is that early implementations of the TREC-like evaluation scenario should be conducted with a singular, unified collection of queries until such time as participants feel comfortable with the process.

Synthesizing from the suggestions made by the expert participants, it thus appears that a minimal TREC-like query record needs to include the following basic elements:

1. High quality audio representation(s)
2. Verbose Metadata:
   i. About the "user"
   ii. About the "need"
   iii. About the "use"
3. Symbolic representation(s) of the music presented

Information need description	% of Queries	Category of intended use	% of Queries
BIBLIOGRAPHIC	75.2%	LOCATE (e.g., "Where can I find…")	49.7%
LYRICS	14.3%	RESEARCH (i.e., background information, etc.)	19.3%
GENRE	9.9%	PERFORM (i.e., play piece(s) on instrument)	18.6%
SIMILAR WORKS	9.9%	COLLECTION BUILDING (i..e., add to pre-existing collection similar items)	18.0%
AFFECT (i.e., description of mood)	7.5%	LISTEN (i.e., as opposed to perform)	6.8%
LYRIC STORY	6.8%		
TEMPO	2.5%		
EXAMPLE	1.8%		

Table 1. Categorization of real-world query and intended use elements as developed and described in Downie and Cunningham (2002).

One is struck by how these requirements are less like a traditional TREC topic statement (Fig. 2) and more like the kind of information garnered in a traditional, well-conducted, reference interview (Dewdney and Michell, 1997; The Reference Interview, 2001). This suggests that the involvement of professional music librarians in the development of the TREC-like music query records is very important — perhaps even critical.

3.1.3 Problem #3: Whither relevance, precision and recall?

The text IR community has had a set of standardized performance evaluation metrics for last four decades. Since the Cranfield experiments of the early 1960's (Cleverdon, Mills and Keen, 1966), two metrics have predominated: *precision* (i.e., the ratio of relevant documents retrieved to the number of documents retrieved); and, *recall* (i.e., the ratio of relevant documents retrieved to the number of relevant documents present in the system). These metrics are the heart of the TREC evaluation paradigm. The key determinant in the use of precision and recall as metrics is the apprehension of those documents deemed "relevant" to a particular query. While there have been ongoing debates about the nature of "relevance" (see Schamber (1994)), its meaning has been stable enough to make the TREC evaluations possible. Simply put, a "document" is deemed to be "relevant" to a given query if the document is "about" the same subject matter as the query (i.e., there is an intersection of "meaning" or "aboutness" between query and document).

Within the context of MIR evaluation, however, this meaning-based approach to relevance assessment is clearly inadequate. For example, what do Beethoven's Piano Sonatas, or Hendrix's guitar solos, actually "mean"? The MIR community recognizes this important shortcoming. In fact, the definition of "relevance" within the MIR context has been so problematic that the precision and recall metrics are rarely found in the MIR literature. Studies by Downie (1999), Foote (1997), Uitdenbogerd and Zobel (1999), Södring and Smeaton (2002) are among the few that employ these measures. Notwithstanding this absence of a community tradition of use, the consensus opinion holds that the MIR community should not shy away from creating a means to assess MIR systems within the TREC-like paradigm and thus should continue to examine precision and recall as core metrics.

To this end, it is hoped that by making the query records as data-rich as possible, that a "reasonable person" standard could emerge as the criterion for the judging the relevance of returned items. That is, there should be enough information contained within the query records that reasonable persons would concur as to whether or not a given returned item satisfied the ***intention*** of the query. The validity of the "reasonable person" assumption would, of course, be subject to empirical verification.

3.1.4 Problem #4: Collection building and intellectual property law

Music is expensive. In the current Post-Napster era, music rights-holders are notoriously litigious. Recent changes to copyright law in the United States have put into question the very existence of "public domain" sources of audio recordings (see Downie (2003)). These three facts, when taken together, have effectively stopped the development of any large-scale, community-accessible, test collections comprising the necessary audio, symbolic and metadata representations. Some private research institutions have acquired substantial collections of audio files. However, these collections are intended for their in-house use only. Collection holders do not make them accessible to others in the community for fear of becoming the objects of expensive civil and criminal litigation.

Notwithstanding these very real difficulties, some recent developments have made it possible to begin construction of the much-needed test collection database. The key here has been convincing select rights-holders that MIR researchers can be trusted to respect their property. This has meant developing mechanisms whereby the intellectual property assets of the right-holders can be shown to be secure from unlicensed access and distribution.

4 Building a TREC-like Test Collection: Important First Steps

The author and colleagues have begun to construct the world's first-and-only, internationally-accessible, large-scale MIR testing and development database. This will be housed at the University of Illinois's National Center for Supercomputing Applications (NCSA) (Fig. 3). Formal transfer and use agreements are being finalized with HNH Hong Kong International, Ltd. (`http://www.naxos.com`), the owner of the *Naxos* and *Marco Polo* recording labels. This will afford the MIR community research access to HNH's ***entire*** catalogue of Classical, Jazz, and Asian digital recordings. This generous gesture on the part of HNH represents approximately 30,000 audio tracks or about 3 terabytes of digital audio music information. All Media Guide (`http://www.allmusic.com`) has also agreed to follow HNH's lead, enabling UIUC/NCSA to incorporate its vast database of music metadata within the same test collection. *All Media's* dataset includes descriptive catalogue records, discographies, and recording classifications.

4.1 Test Collection Database: System Overview

Given the unique opportunity that these rights-holders have afforded the MIR community, it is important that the MIR testing and evaluation database be constructed with three central features in mind:

1. security for the property of the rights-holders, especially important if we are to convince other rights-holders to participate in the future;
2. accessibility for both internal, domestic, and international researchers; and,

3. sufficient computing and storage infrastructure to support the computationally- and data-intensive techniques being investigated by the various research teams.

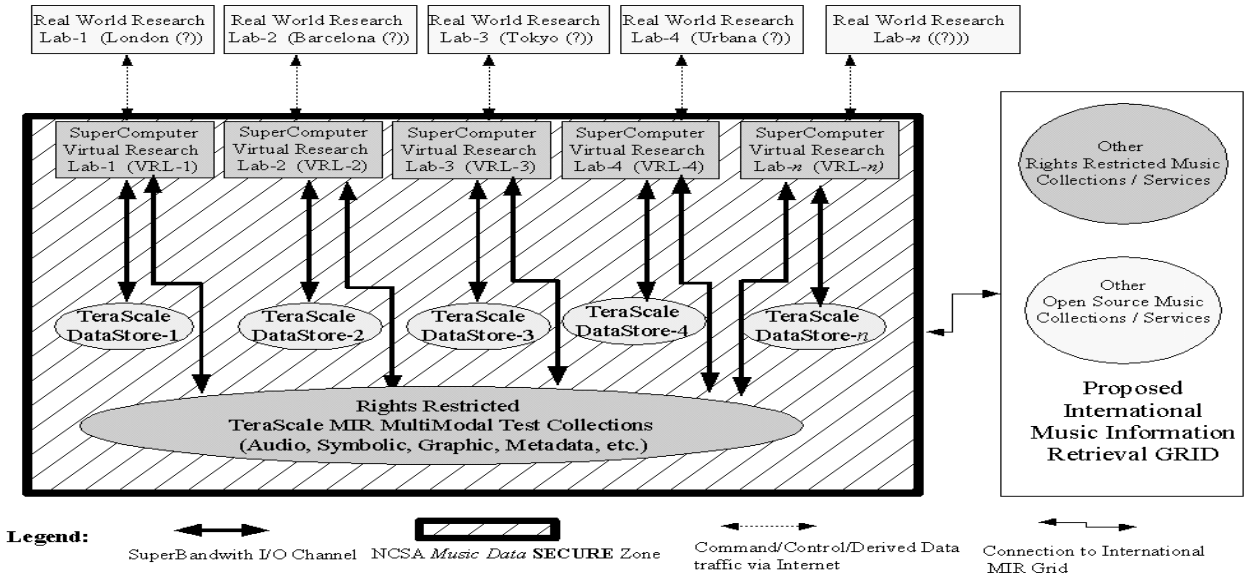

Figure 3. Schematic of the secure, yet accessible, test collection environment.

To these ends, we are exploiting the expertise and resources of NCSA and its Automated Learning Group (ALG), headed by Prof. Michael Welge. NCSA's systems have been designed to be secure. Certificate-based authentication for all users as well as means for encrypting data and data transfers are fundamental to NCSA's security protocols.

The ALG has developed a data-to-knowledge system, D2K, which supports all phases of the data mining process**.** D2K was originally designed to provide data mining professionals with a flexible "sandbox" for developing and evaluating the performance of a range of supercomputing techniques on a variety of data sets. Using the D2K technology as a starting point, we are creating a secure "Virtual Research Lab" (VRL) for each participating research team. These VRLs will provide secure access to the test collection and the resources necessary to conduct large-scale MIR evaluation experiments. Simply put, we enhance the security of the valuable music data by bringing the research teams to the collection, rather than distributing the collection willy-nilly around the globe.

For the transfer of the MIR TREC-like environment to the international, domestic and internal research teams, we are incorporating another ALG application, D2K-SL. D2K-SL builds upon current D2K modules to provide a set of pre-defined applications that guide users through the supercomputing process. These tools will be instrumental in supporting the multidisciplinary nature of MIR research and evaluation. Their relative ease-of-use should also help retain and encourage the participation in MIR research of such non-computer experts as librarians, musicologists, Arts and Humanities students and educators, and business executives. In addition, we hope that these D2K-SL applications can be used to address other related research thrusts, such as new MIR techniques, new interface designs and the development of protocols to make the proposed MIR GRID a viable entity (Dovey, 2002).

5 Summary and Future Research

This paper has outlined the efforts being made to establish a scientifically valid TREC-like evaluation paradigm for MIR research. Expert opinion on the implementation of MIR/MDL evaluation frameworks was solicited, analyzed, and then summarized. Major issues raised by participating experts include addressing the complex nature of music information; adequately capturing the complex nature of music queries; recognition of the MIR "relevance" problem; and, overcoming the intellectual property hurdles to collection building. Proposed solutions include the creation of data-rich query records that are both grounded in real-world requirements and neutral with respect to retrieval technique; adoption of a "reasonable person" approach to "relevance" assessment; and, the establishment of TREC-like evaluation protocols. Finally, the development of a secure, yet accessible, research environment at NCSA — one that allows researchers to remotely participate in the secure use of the large-scale testbed collection — represents a significant first step forward in surmounting the intellectual property hurdles plaguing MIR research and evaluation.

Some of these proposed solutions will require further investigation and effort. In particular, we must work on the:

1. explicit capturing and analysis of a wide variety real-world music queries upon which to base the creation of the query records;
2. development of formal requirements for the necessary elements (and their constituent data types) to be used in the query records;
3. validation of the "reasonable person" relevance judgement assumption through inter-rater reliability studies; and,

4. continued acquisition of more music information (audio, symbolic, and metadata) with a special effort to acquire "top hits" popular music and more non-Western musics to make real-world, real-time, user studies a possibility. The acquisition of non-Western musics is particularly important as there is a strongly-perceived bias toward Western music within current MIR research (Futrelle and Downie, 2002).

Acknowledgements

The keynote speakers, contributors and meeting participants are all to be thanked. Drs. Don Waters and Suzanne Lodato, both of the Andrew W. Mellon Foundation, are thanked for their moral and financial support. Karen Medina, Joe Futrelle and Mike Welge are also thanked for their valuable contributions and suggestions throughout the project.

References

Bainbridge, D. (2002). Towards a workbench for symbolic music information retrieval. In J. S. Downie (Ed.), *The MIR/MDL Evaluation Project White Paper Collection* (2nd ed., pp. 14-16). Champaign, IL: GSLIS.

Barthélemy, J., & Bonardi, A. (2001). Figured bass and tonality recognition. In J. S. Downie and D. Bainbridge (Eds.), *Proceedings of the Second International Symposium on Music Information Retrieval* (*ISMIR 2001*) (pp. 129-136). Bloomington, IN: Indiana University.

Birmingham, W., Dannenberg, R. B., Wakefield, G. H., Bartsch, M., Bykowski, D., Mazzoni, D., Meek, C., Mellody, M., & Rand, W. (2001). Musart: Music retrieval via aural queries. In *Second International Symposium on Music Information Retrieval* (*ISMIR 2001*) (pp. 73-81).

Byrd, D., & Crawford, T. C. (2002). Problems of music information retrieval in the real 2orld. *Information Processing and Management*, 38, 249-272.

Cleverdon, C., Mills, J., & Keen, M. (1966). *Factors determining the performance of indexing systems*. Cranfield, UK: ASLIB Cranfield Research Project, College of Aeronautics.

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).

Dewdney, P., & Michell, G. (1997). Asking "why" questions in the reference Interview: A theoretical Justification. *Library Quarterly*, 67, 50-71.

Doraisamy, S., & Rüger, S. M. (2002). A comparative and fault-tolerance study of the use of n-grams with polyphonic music. In M. Fingerhut (Ed.), *Proceedings of the Third International Conference on Music Information Retrieval (ISMIR 2002)* (pp. 101-106). Paris: IRCAM.

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).

Downie, J. S. (2002). 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* (2nd ed., pp. 3-6).

Downie, J. S. (1994). The Musifind musical information retrieval project, phase ii: User assessment survey. In *Proceedings of the 22nd Annual Conference of the Canadian Association for Information Science* (pp. 149-166). Toronto: CAIS.

Downie, J. S. (1999). Evaluating a simple approach to music information retrieval: Conceiving melodic n-grams as text. Unpublished doctoral dissertation, University of Western Ontario, London, Ontario.

Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Science and Technology*, 37, 295-340.

Downie, J. S., & Cunningham, S. J. (2002). Toward a theory of music information retrieval queries: System design implications. In *Proceedings of the Third International Conference on Music Information Retrieval* (ISMIR 2002) (pp. 299-300).

Dunn, J. W., Davidson, M. W., & Isaacson, E. J. (2001). Indiana University digital music library project: An update. In *Second International Symposium on Music Information Retrieval* (*ISMIR 2001*) (pp. 137-138).

Foote, J. (1997). Content-based retrieval of music and audio. In *SPIE* Vol. 3229. *Multimedia Storage and Archiving Systems II*. Bellingham, WA: SPIE Press.

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).

Futrelle, J., & Downie, J. S. (2002). Interdisciplinary communities and research issues in music information retrieval. In *Third International Conference on Music Information Retrieval (ISMIR 2002)* (pp. 215-221).

Haitsma, J., & Kalker, T. (2002). A highly robust audio fingerprinting system. In *Third International Conference on Music Information Retrieval (ISMIR 2002)* (pp. 107-115).

Haus, G., & Pollastri, E. (2001). An audio front end for query-by-humming systems. In *Second International Symposium on Music Information Retrieval* (*ISMIR 2001*) (pp. 65-72).

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).

Issacson, E. J. (2002). Music IR for music theory. In *The MIR/MDL Evaluation Project White Paper Collection* (2nd ed., pp. 23-26).

Harb, H., & Chen, L. (2003). A query by example music retrieval algorithm. In *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS03)* (pp 122-128).

Itoh, M. (2000). Subject search for music: Quantitative analysis of access point selection. In D. Byrd and J. S. Downie (Eds.), *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2001)*. Amherst, MA: University of Massachusetts at Amherst.

Kim, J.-Y., & Belkin, N. J. (2002). Categories of music description and search terms and phrases used by non-music experts. In *Third International Conference on Music Information Retrieval (ISMIR 2002)* (pp. 209-214).

Kornstädt, A. (2001). The JRing system for computer-assisted musicological analysis. In *Second International Symposium on Music Information Retrieval* (*ISMIR 2001*) (pp. 93-98).

Linstone, H., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Boston, MA: Addison-Wesley.

Logan, B. (2002). Content-based playlist generation: Exploratory experiments. In *Third International Conference on Music Information Retrieval (ISMIR 2002)* (pp. 295-296).

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).

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).

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).

Montalvo, J. (2002). A MIDI track for music information retrieval. In *The MIR/MDL Evaluation Project White Paper Collection* (2nd ed., pp. 32-32).

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).

Pauws, S., & Eggen, B. (2002). PATS: Realization and user evaluation of an automatic playlist generator. In *Third International Conference on Music Information Retrieval (ISMIR 2002)* (pp. 222-230).

Pickens, J. (2000). A comparison of language modeling and probabilistic text information retrieval approaches to monophonic music retrieval. In *International Symposium on Music Information Retrieval (ISMIR 2000)*.

Rasmussen, E. (2002). Evaluation in information retrieval. In *The MIR/MDL Evaluation Project White Paper Collection* (2nd ed., pp. 45-39).

Reiss, J., & Sandler, M. (2002a). Benchmarking music information retrieval systems. In *The MIR/MDL Evaluation Project White Paper Collection* (2nd ed., pp. 37-42).

Reiss, J., & Sandler, M. (2002b). 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).

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).

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).

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29, 3-48.

Smiraglia, R. P. (2001). Musical works as information retrieval entities: Epistemological perspectives. In *Second International Symposium on Music Information Retrieval (ISMIR 2001)* (pp. 85-91).

Södring, T., & Smeaton, A. F. (2002). Evaluating a music information retrieval system: TREC style. In *The MIR/MDL Evaluation Project White Paper Collection* (pp. 71-78).

The Reference Interview. (2001). In R. E. Bopp, & L. C. Smith, *Reference and information services*: *An introduction* (3rd ed., pp. 47-68). Englewood, CO: Libraries Unlimited.

Uitdenbogerd, A. L., & Zobel, J. (1999). Matching techniques for large music databases. In *Proceedings of the 7th ACM International Multimedia Conference* (pp. 57-66).

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).