J. Stephen Downie
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 East Daniel Street
Champaign, Illinois 61820 USA
jdownie@uiuc.edu

The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future

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'' and ''Query-by-Humming'' systems that allow users to interact with their respective music search engines via queries that are sung or hummed into a microphone (e.g., Birmingham et al. 2001; Haus and Pollastri 2001). ''Query-by-Note'' systems are also being developed wherein searchers construct queries consisting of pitch and/or rhythm information (e.g., Pickens 2000; Doraisamy and Rüger 2002). Input methods for Query-by-Note systems include symbolic interfaces as well as both physical (MIDI) and virtual (Java-based) keyboards. Some teams are working on ''Query-by-Example'' systems that take pre-recorded music in the form of CD or MP3 tracks as their query input (e.g., Haitsma and Kalker 2002; Harb and Chen 2003). The development of comprehensive music recommendation and distribution systems is a growing research area (e.g., Logan 2002; Pauws and Eggen 2002). The automatic generation of playlists for use in personal music systems, based on a wide variety of user-defined criteria, is the goal of this branch of MIR research. Other groups are investigating the creation of music analysis systems to assist those in the musicology and music theory communities (e.g., Barthélemy and Bonardi 2001; Kornstädt 2001). Overviews of MIR's interdisciplinary research areas can be found in Downie (2003), Byrd and Crawford (2002), and Futrelle and Downie (2002).

This article begins with an overview of the current scientific problem facing MIR research. Entitled ''Current Scientific Problem,'' the opening section also provides a brief explication of the Text Retrieval Conference (TREC) evaluation paradigm that has come to play an important role in the community's thinking about the testing and evaluation of MIR systems. The sections which follow, entitled ''Data Collection Method'' and ''Emergent Themes and Commentary,'' report upon the findings of the Music Information Retrieval (MIR)/Music Digital Library (MDL) Evaluation Frameworks Project with issues surrounding the creation of a TREC-like evaluation paradigm for MIR as the central focus. ''Building a TREC-Like Test Collection'' follows next and highlights the progress being made concerning the establishment of the necessary test collections. The ''Summary and Future Research'' section concludes this article and highlights some of the key challenges uncovered that require further investigation.

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 no standard collection of music against which each team could test its techniques, no standardized sets of performance tasks, and 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 on 16 October 2001 by the attendees of the Second International Symposium on Music Information Retrieval (ISMIR 2001). (See music-ir.org/mirbib2/resolution for the list of signatories.)

Over a decade ago, the National Institute of Standards and Technology (NIST) developed a testing and evaluation paradigm for the text-retrieval com-

Computer Music Journal, 28:2, pp. 12–23, Summer 2004

munity, called the Text REtrieval Conference (TREC; see trec.nist.gov). Under this paradigm, each text retrieval team is given access to a standardized, large-scale test collection of text; a standardized set of test queries; and a standardized evaluation of the results each team generates.

The TREC approach to evaluation can also be thought of as an annual cycle of events. In the late fall of each year, NIST sends out its calendar and call for participation. By the end of February, the interested teams have signed up for these events. In 2001, there were 87 participating groups, representing 21 different countries (Voorhees 2002). The official test collections are then sent out to the participants in March. Over the course of the spring and summer, the teams build and/or modify their various systems and run their experiments on the official data using the official set(s) of standardized queries. In August, the teams submit the results of their experimental runs back to the TREC organizers. The retrieval results are then analyzed by the in-house evaluators of NIST. In October, evaluation scores are released to the teams, and the teams then write up their findings in light of their evaluation data. In mid-November, all the teams convene at the annual TREC meeting to exchange findings and compare results. Immediately following TREC meeting, the cycle begins again.

In 1994, TREC implemented a regime of ''tracks.'' Under the track system, the TREC organizers determine a set of information retrieval specialties that warrant specific investigation. Over the years, tracks have been created to focus on video information retrieval (i.e., searching collections of moving images), Web retrieval, cross-language retrieval (i.e., searching through sets of multilingual information), speech retrieval, and so on. Each track has its own special test collection, standardized queries, and evaluation programs.

Owing to the strong overlap between the MIR and the traditional IR communities, many people 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 this 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).

Data Collection Method

The Delphi method of data collection (Linstone and Turoff 1975) 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 are solicited. These opinions are then brought together, and trends are observed. The resultant data is then 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?

The MIR/MDL Evaluation Project undertook three rounds of formal expert input. The input rounds consisted 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 of a special meeting wherein the participants were able to expound upon their white paper opinions and exchange ideas. The white papers from each round have been collected in successive editions of The MIR/MDL Evaluation White Paper Collection. (See music-ir.org/evaluation for the most recent edition.) Information about each of the input rounds follows.

Emergent Themes and Commentary

The first-round 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 2002 (www.ohsu.edu/jcdl). Ellen Voorhees, Project Manager of NIST's TREC, presented the keynote address (Voorhees 2002). Her presentation focused 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 the first round. 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, like audio (Pardo, Meek, and Birmingham 2002; Reiss and Sandler 2002a), 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) (e.g., recreational uses, educational uses, scholarly uses; see Issacson 2002, for example).

In short, the consensus was that work should proceed on developing TREC-like evaluations with several provisos. First, any TREC-like approach developed should 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. Second, the integration of music metadata must not be overlooked. Finally, the TREC-like approach should not become the sole means of evaluating the performance of MIR/MDL systems.

The second-round meeting, ''The Panel on Music Information Retrieval Evaluation Frameworks'' was held as part of ISMIR 2002. Edie Rasmussen (Professor, 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 second-round 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; Richard 2002; Rüger 2002). One paper extended the large-scale test collection notion to encompass multiple test collections housed in multiple locations and interconnected via a proposed international network of computers called the ''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; Reiss and Sandler 2002b; Södring and Smeaton 2002). 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 second-round white papers 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?''

The third-round meeting, ''Workshop on the Evaluation of Music Information Retrieval (MIR) Systems'' was held in conjunction with the 26th Annual International ACM SIGIR Conference (SIGIR '03), Toronto, 1 August 2003. Beth Logan of HP Labs presented the workshop keynote white paper, which she co-authored with Daniel Ellis and Andrew Berenzwieg of Columbia University (Logan, Ellis, and Berenzweig 2003). Ms. Logan's mul-

tifaceted 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 (Doraisamy and Rüger 2003; Goodrum 2003; Pickens 2003; Reiss and Sandler 2003). Issues addressed included the need for specific TREC-like tracks (Pickens 2003), the lessons to be learned from previous benchmarking and evaluation work (Doraisamy and Rüger 2003; Goodrum 2003; Reiss and Sandler 2003), and the role that the proposed Music GRID could play in evaluation testing (Pickens 2003). One paper provided strong evidence for requiring that the TREC-like evaluation testbed include ''popular'' and/or ''well-known'' music as opposed to ''artificially'' created music (Hoashi, Matsumoto, and Inoue 2003). (Some research teams had previously proposed that they could overcome copyright issues by composing music without compromising the evaluation results that would be based upon such this music; see Goto et al. 2002.)

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 (Pardo and Birmingham 2003). 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 (Batlle, Guaus, and Masip 2003).

Commentary on Emergent Themes

Given the strong support of participants for the establishment of a TREC-like evaluation paradigm, why has a TREC-like approach not been adopted already? Participants consistently touched upon four problem areas that provide some insight into this question: the complexity of music information; the complexity of music queries; the nature of relevance within the context of MIR and the applicability of precision and recall as evaluation metrics (terms defined later in this article); and the lack of access to music collections brought about by intellectual property laws as practiced by the music industry.

The ordering of the 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).

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 (e.g., lyrics and libretti), editorial, praxis, and bibliographic elements. Music can be represented as scores, MIDI files, and other discrete encodings, and in any number of analog and digital audio format. 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. First, 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 later in this article in ''Problem #4: Collection Building and Intellectual Property Law,'' 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. See ''Building a

Figure 1. A TREC topic statement from Voorhees (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.
```

TREC-like Test Collection: Important First Steps'' for a discussion of progress being made to alleviate this problem.

Second, test-collection size is a real concern. Owing to 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, multiformat 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). (See ''Test Collection Database: System Overview'' for the proposed solution to the large dataset problem.)

Third, 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 so that one retrieval method is not ''privileged'' over another. This leads to the notion of ''retrieval neutrality'' discussed in ''Problem #2: The Complexity of Music Queries.''

Fourth, music queries, 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.

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 (Byrd and Crawford 2002; Futrelle and Downie 2002; Downie 2003). 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; Downie and Cunningham 2002; Kim and Belkin 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 Figure 1, a TREC topic statement presented in Voorhees (2002), with Figure 2, a real-world music query presented in Cunningham (2002). Table 1 also illustrates the wide variety of information types contained in real-world music queries along the wide variety of intended uses for the sought-after music.

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. Although 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 be grounded in real-world needs and uses; be representative of the complexity of real-world queries (see Table 1); be neutral with regard to the retrieval method employed; and be data-rich so realistic and meaningful ''relevance'' judgments can be made.

The ''retrieval neutrality'' principle requires some explication. The MIR community can be divided roughly into two camps: those engaged in symbolic retrieval research, and those exploring audio- and signal-processing techniques. Given that no data exist on the comparative strengths and

Figure 2. A real-world information request posted to the Internet newsgroup alt.music.lyrics as presented in Cunningham (2002).

```
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

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....
Ohohohoo
[(instrumental) F C Bb Bb C F C Bb] <remaining text deleted>
```

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

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 of 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		

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 (Pickens 2003). 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 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 must include the following basic elements: high-quality audio representa-

tion(s); verbose metadata (about the ''user,'' ''need,'' and ''use''); and symbolic representation(s) of the music presented.

One is struck by how these requirements are less like a traditional TREC topic statement (see Figure 2) and more like the kind of information garnered in a traditional, well-conducted, reference interview (Dewdney and Michell 1997; Bopp 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.

Problem #3: Whither Relevance, Precision, and Recall?

The text IR community has had a set of standardized performance evaluation metrics for the past four decades. Since the Cranfield experiments of the early 1960s (Cleverdon, Mills, and Keen 1966), two metrics have predominated: precision (the ratio of relevant documents retrieved to the number of documents retrieved) and recall (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), and 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.

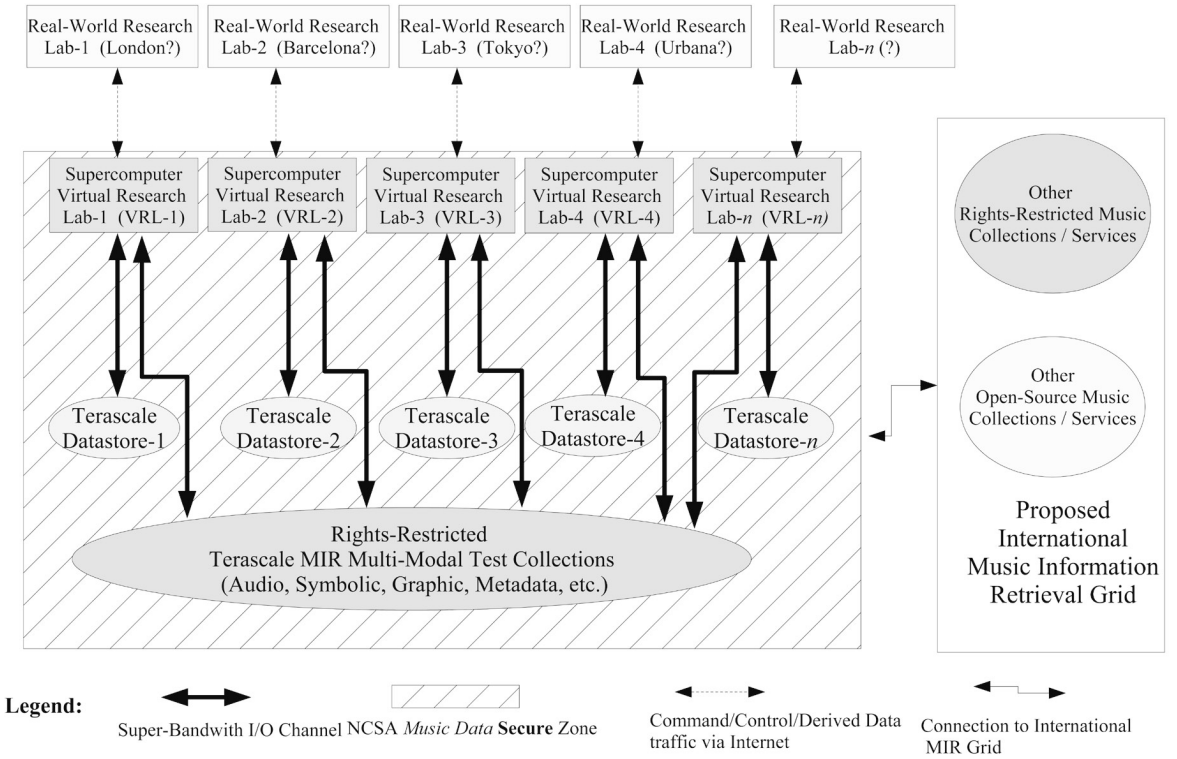
To this end, it is hoped that, by making the query records as data-rich as possible, a ''reasonable person'' standard could emerge as the criterion for the judging the relevance of returned items. That is, there should be enough information contained within the query records that reasonable persons would concur as to whether 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.

Problem #4: Collection Building and Intellectual Property Law

Music is expensive; in the current post-Napster era, music rights-holders are notoriously litigious, and 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 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

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



trusted to respect their property. This has meant developing mechanisms whereby the intellectual property assets of the rights-holders can be shown to be secure from unlicensed access and distribution.

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 to be housed at the University of Illinois's National Center for Supercomputing Applications (NCSA). Figure 3 shows an overview of the system. Formal transfer and use agreements are being finalized with HNH Hong Kong International, Ltd. (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 (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.

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. First among these is security for the property of the rights-holders, especially important if we are to convince other rights-holders to participate in the future. The second central feature is accessibility for both internal, domestic, and international researchers. The third feature is suffi-

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

To these ends, we are exploiting the expertise and resources of NCSA and its Automated Learning Group (ALG), headed by 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 "playground" 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 Laboratory (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 called D2K-SL. D2K-SL builds upon current D2K modules to provide a set of predefined 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).

Summary and Future Research

This article 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 obstacles plaguing MIR research and evaluation.

Some of these proposed solutions will require further investigation and effort. First, the explicit capturing and analysis of a wide variety of real-world music queries upon which to base the creation of the query records must be studied. Formal requirements for the necessary elements (and their constituent data types) to be used in the query records must be developed. The "reasonable person" relevance judgment assumption must be validated through inter-rater reliability studies. Non-TREC-like evaluation techniques must be explored to complement and enhance our understanding of MIR and MDL system performance. Finally, additional music information (audio, symbolic, and metadata) must continually be acquired—especially "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).

Acknowledgments

The keynote speakers, contributors, and audience members who participated in each of the meetings 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. Some material is based upon work supported by the National Science Foundation under grants IIS-0340597 and IIS-0327371.

References

Bainbridge, D. 2002. ''Towards a Workbench for Symbolic Music Information Retrieval.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 14–16.

Barthélemy, J., and A. Bonardi. 2001. ''Figured Bass and Tonality Recognition.'' *Proceedings of the Second International Symposium on Music Information Retrieval*. Bloomington, Indiana: University of Indiana, pp. 129–136.

Batlle, E., E. Guaus, and J. Masip. 2003. ''Open Position: Multilingual Orchestra Conductor. Lifetime Opportunity.'' *The MIR/MDL Evaluation Project White Paper Collection,* 3rd ed. Champaign, Illinois: GSLIS, pp. 86–89.

Birmingham, W., et al. 2001. ''Musart: Music Retrieval via Aural Queries.'' *Proceedings of the Second International Symposium on Music Information Retrieval (ISMIR 2001)*. Bloomington, Indiana: University of Indiana, pp. 73–81.

Bopp, R. E. 2001. ''The Reference Interview.'' In R. E. Bopp and L. C. Smith, eds. *Reference and Information Services: An Introduction,* 3rd ed. Englewood, Colorado: Libraries Unlimited, pp. 47–68.

Byrd, D., and T. C. Crawford. 2002. ''Problems of Music Information Retrieval in the Real World.'' *Information Processing and Management* 38:249–272.

Cleverdon, C., J. Mills, and M. Keen. 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.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 17–19.

Dewdney, P., and G. Michell. 1997. ''Asking 'Why' Questions in the Reference Interview: A Theoretical Justification.'' *Library Quarterly* 67:50–71.

Doraisamy, S., and S. M. Rüger. 2002. ''A Comparative and Fault-Tolerance Study of the Use of N-Grams with Polyphonic Music.'' *Proceedings of the Third International Conference on Music Information Retrieval*. Paris: IRCAM, pp. 101–106.

Doraisamy, S., and S. M. Rüger. 2003. ''Emphasizing the Need for TREC-like Collaboration Towards MIR Evaluation.'' *The MIR/MDL Evaluation Project White Paper Collection,* 3rd ed. Champaign, Illinois: GSLIS, pp. 90–96.

Dovey, M. J. 2002. ''Music GRID: A Collaborative Virtual Organization for Music Information Retrieval Collaboration and Evaluation.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 50–52.

Downie, J. S. 1994. ''The MusiFind Musical Information Retrieval Project, Phase II: User Assessment Survey.'' *Proceedings of the 22nd Annual Conference of the Canadian Association for Information Science*. Toronto, Canada: CAIS, pp. 149–166.

Downie, J. S. 1999. ''Evaluating a Simple Approach to Music Information Retrieval: Conceiving Melodic N-Grams as Text.'' Ph.D. Thesis, The University of Western Ontario.

Downie, J. S. 2002. ''Establishing Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation Frameworks: Preliminary Foundations and Infrastructures.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 3–6.

Downie, J. S. 2003. ''Music Information Retrieval.'' *Annual Review of Information Science and Technology* 37:295–340.

Downie, J. S., and S. J. Cunningham. 2002. ''Toward a Theory of Music Information Retrieval Queries: System Design Implications.'' *Proceedings of the Third International Conference on Music Information Retrieval*. Paris: IRCAM, pp. 299–300.

Dunn, J. W., M. W. Davidson, and E. J. Isaacson. 2001. ''Indiana University Digital Music Library Project: An Update.'' *Proceedings of the Second International Symposium on Music Information Retrieval*. Bloomington, Indiana: University of Indiana, pp. 137–138.

Foote, J. 1997. ''Content-Based Retrieval of Music and Audio.'' In C.-C. J. Kuo, S. F. Chang, and V. N. Gudivada, eds. *SPIE Vol. 3229: Multimedia Storage and Ar-*

chiving Systems II. Bellingham, Washington: SPIE Press, pp. 138–147.
Futrelle, J. 2002. ''Three Criteria for the Evaluation of Music Information Retrieval Techniques against Collections of Musical Material.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 20–22.
Futrelle, J., and J. S. Downie. 2002. ''Interdisciplinary Communities and Research Issues in Music Information Retrieval.'' *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 215–221.
Goodrum, 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. Champaign, Illinois: GSLIS, pp. 97–102.
Goto, M., et al. 2002. ''RWC Music Database: Popular, Classical, and Jazz Music Databases.'' *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 287–288.
Haitsma, J., and T. Kalker. 2002. ''A Highly Robust Audio Fingerprinting System.'' *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 107–115.
Harb, H., and Chen, L. 2003. ''A Query by Example Music Retrieval Algorithm.'' *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS03)*. London: Queen Mary, University of London, pp. 122–128.
Haus, G., and E. Pollastri. 2001. ''An Audio Front End for Query-by-Humming Systems.'' *Proceedings of the Second International Symposium on Music Information Retrieval*. Bloomington, Indiana: University of Indiana, pp. 65–72.
Herrera-Boyer, P. 2002. ''Setting Up an Audio Database for Music Information Retrieval Benchmarking.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 53–55.
Hoashi, K., K. Matsumoto, and N. Inoue. 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. Champaign, Illinois: GSLIS, pp. 103–106.
Issacson, E. J. 2002. ''Music IR for Music Theory.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 23–26.
Itoh, M. 2000. ''Subject Search for Music: Quantitative Analysis of Access Point Selection.'' *Proceedings of the First International Symposium on Music Information Retrieval*. Amherst: University of Massachusetts at Amherst, n.p.
Kim, J-Y., and N. J. Belkin. 2002. ''Categories of Music Description and Search Terms and Phrases Used by Non-Music Experts.'' *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 209–214.
Kornstädt, A. 2001. ''The JRing System for Computer-Assisted Musicological Analysis.'' *Proceedings of the Second International Symposium on Music Information Retrieval*. Bloomington, Indiana: University of Indiana, pp. 93–98.
Linstone, H., and M. Turoff. 1975. *The Delphi Method: Techniques and Applications*. Boston, Massachusetts: Addison-Wesley.
Logan, B. 2002. ''Content-Based Playlist Generation: Exploratory Experiments.'' *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 295–296.
Logan, B., D. P. W. Ellis, and A. Berenzweig. 2003. ''Toward Evaluation Techniques for Music Similarity.'' *The MIR/MDL Evaluation Project White Paper Collection,* 3rd ed. Champaign, Illinois: GSLIS, pp. 81–85.
MacMillan, K. 2002. ''Common Music Notation as a Source for Music Information Retrieval.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 27–28.
Meek, C., W. P. Birmingham, and B. Pardo. 2002. ''What is a Sung Query?'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 56–57.
Melucci, M., and N. Orio. 2002. ''A Task-Oriented Approach for the Development of a Test Collection for Music Information Retrieval.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 29–31.
Montalvo, J. 2002. ''A MIDI Track for Music Information Retrieval.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 32–33.
Pardo, B., and W. P. Birmingham. 2003. ''Query by Humming: How Good Can It Get?'' *The MIR/MDL Evaluation Project White Paper Collection,* 3rd ed. Champaign, Illinois: GSLIS, pp. 107–109.
Pardo, B., C. Meek, and B. Birmingham. 2002. ''Comparing Aural Music-Information Retrieval Systems.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 34–36.
Pauws, S., and B. Eggen. 2002. ''PATS: Realization and User Evaluation of an Automatic Playlist Generator.'' *Proceedings of the Third International Conference on Music Information Retrieval.* Paris: IRCAM, pp. 222–230.

Pickens, J. 2000. ''A Comparison of Language Modeling and Probabilistic Text Information Retrieval Approaches to Monophonic Music Retrieval.'' *Proceedings of the First International Symposium on Music Information Retrieval.* Amherst, Massachusetts: University of Massachusetts at Amherst, n.p.

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. Champaign, Illinois: GSLIS, pp. 110–113.

Rasmussen, E. 2002. ''Evaluation in Information Retrieval.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 45–49.

Reiss, J., and M. Sandler. 2002a. ''Benchmarking Music Information Retrieval Systems.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 37–42.

Reiss, J., and M. Sandler. 2002b. ''Beyond Recall and Precision: A Full Framework for MIR System Evaluation.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 58–63.

Reiss, J., and M. Sandler. 2003. ''MIR Benchmarking: Lessons Learned from the Multimedia Community.'' *The MIR/MDL Evaluation Project White Paper Collection,* 3rd ed. Champaign, Illinois: GSLIS, pp. 114–120.

Richard, G. 2002. ''Towards Large Databases for Music Information Retrieval Systems: Development and Evaluation.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 64–67.

Rüger, S. 2002. ''A Framework for the Evaluation of Content-Based Music Information Retrieval Using the TREC Paradigm.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, 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.'' *Proceedings of the Second International Symposium on Music Information Retrieval.* Bloomington, Indiana: University of Indiana, pp. 85–91.

Södring, T., and A. F. Smeaton. 2002. ''Evaluating a Music Information Retrieval System: TREC Style.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 71–78.

Uitdenbogerd, A. L., and J. Zobel. 1999. ''Matching Techniques for Large Music Databases.'' *Proceedings of the 6th International Multimedia Conference.* New York: ACM Press, pp. 235–240.

Voorhees, E. M. 2002. ''Whither Music IR Evaluation Infrastructure: Lessons to be Learned from TREC.'' *The MIR/MDL Evaluation Project White Paper Collection,* 2nd ed. Champaign, Illinois: GSLIS, pp. 7–13.