

# The TREC-Like Evaluation of Music IR Systems

J. Stephen Downie

Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
1-217-351-5037  
jdownie@uiuc.edu

## ABSTRACT

This poster reports upon the ongoing efforts being made to establish TREC-like and other comprehensive evaluation paradigms within the Music IR (MIR) and Music Digital Library (MDL) research communities. The proposed research tasks are based upon expert opinion garnered from members of the Information Retrieval (IR), MDL and MIR communities with regard to the construction and implementation of scientifically valid evaluation frameworks.

## Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and Software – *performance evaluation*

## General Terms

Measurement, Performance, Human Factors

## Keywords

TREC, Evaluation, Music Information Retrieval

## 1. INTRODUCTION

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. Good overviews of MIR's interdisciplinary research areas can be found in [1-3].

### 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.) [13].

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:

COPYRIGHT IS HELD BY THE AUTHOR/OWNER(S).  
SIGIR'03, JULY 28–AUGUST 1, 2003, TORONTO, CANADA.  
ACM 1-58813-646-3/03/0007.

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.

It is upon this TREC paradigm that we plan to create our “TREC-like” formal evaluation scenario. TREC-like is used deliberately as music retrieval presents some important differences to text retrieval and our evaluation tests must take these into account (see Major Research Questions below).

The two principal research streams being outlined are based upon a synthesis and analysis of expert opinion garnered from members of the IR, MDL and MIR communities with regard to the construction and implementation of scientifically valid evaluation frameworks. As part of the Mellon-funded “MIR/MDL Evaluation Frameworks Project” (<http://music-ir.org/evaluation>), the outcomes of two fact-finding meetings form the foundation upon which this project is grounded. The presentations made at each of the meetings have been collected in successive editions of *The MIR/MDL Evaluation White Paper Collection*. See <http://music-ir.org/evaluation> for the most recent edition.

## 1.2 Major Research Questions

The project is informed by the following key research questions:

1. How do we adequately capture the complex nature of music queries so proposed experiments and protocols are well-grounded in reality?
2. How do we develop new models and theories of “relevance” in the MIR context (i.e., What does “relevance” really mean in the MIR context?)?
3. How do we evaluate the utility, within the MIR context, of already-established evaluation metrics (e.g., precision and recall, etc.)?
4. How do we integrate the evaluation of MIR systems with the larger framework of IR evaluation (i.e., What aspects are held in common and what are unique to MIR?)?
5. How do we continue the expansion of a comprehensive collection of music materials to be used in evaluation experiments?

## 2. STREAM #1: A TREC-LIKE TESTBED

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) (Fig. 1). 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 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.

## 2.1 System Overview

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.

To these ends, we are exploiting the expertise and resources of NCSA and its Automated Learning Group (ALG), headed by Prof. Michael Welge. Using the ALG's *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

One is struck by how these requirements are less like a traditional TREC topic statement and more like the kind of information garnered in a traditional, well-conducted, reference interview [4,5]. 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.

As one can see, the project needs to collect and analyze data concerning real-world users, the ways in which they express their needs, and how they intend to use the results of their searches. To this end, we are constructing a multifaceted research programme aimed at capturing these important facts through a variety of "needs and uses" studies and human-computer interaction studies. We have tentatively titled this research stream, "Human Use of Music Information Retrieval Systems" (HUMIRS). We are currently sketching out the roles that the UIUC Music Library (needs and uses) and the NCSA Usability Lab (human-computer interaction) can play in this regard

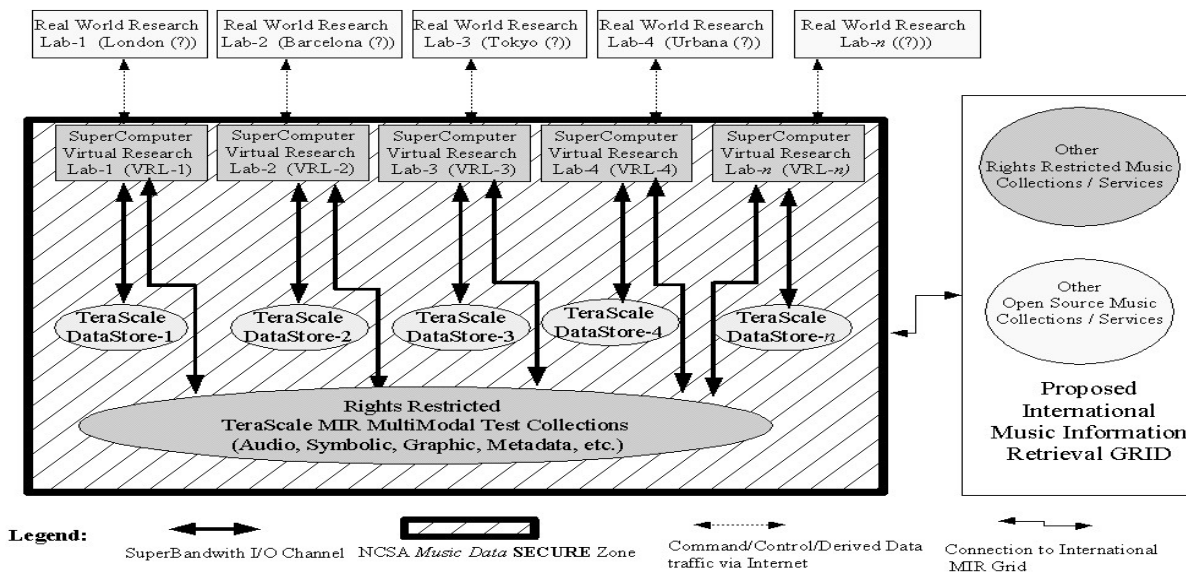


Figure 1. Schematic of the secure, yet accessible, test collection environment.

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.

### 3. STREAM #2: HUMIRS<sup>1</sup>

We must ensure that the test tasks developed are realistic proxies for the kinds of uses that MIR/MDL systems might expect to encounter. Synthesizing from the suggestions made by the expert participants, it 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

## 4. REFERENCES

- [1] Downie, J. S., *Music Information Retrieval Annual Review of Information Science and Technology* 37: 295-340, 2003.
- [2] Byrd, D. and Crawford, T. C., *Problems of Music Information Retrieval in the Real World Information Processing and Management* 38: 249-272, 2002.
- [3] Futrelle, J. and Downie, J. S., *Interdisciplinary Communities and Research Issues in Music Information Retrieval Third International Conference on Music Information Retrieval*: 215-221, 2002.
- [4] Dewdney, P. and Michell, G., *Asking "Why" Questions in the Reference Interview: A Theoretical Justification* *Library Quarterly* 67: 50-71, 1997.
- [5] *The Reference Interview. Reference and Information Services: An Introduction*, eds. Bopp, R. E. and Smith, L. C. Englewood, CO: Libraries Unlimited: 47-68, 2001.

<sup>1</sup> Pronounced, "hummers".