

# Establishing Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation Frameworks: Preliminary Foundations and Infrastructures

Final Proposal - 5 July 2002

J. Stephen Downie, PhD  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
501 E. Daniel, Champaign IL 61820 USA  
+1-(217) 265-5018  
jdownie@uiuc.edu

## PROBLEM STATEMENT

There is a very strong feeling among the Music Information Retrieval (MIR) and Music Digital Library (MDL) research communities that we must acquire standardized collections of music (in multiple representations), standardized retrieval tasks (i.e., query sets, etc.), and standardized retrieval metrics (see <http://music-ir.org/mirbib2/resolution>). At the heart of all of this is very real concern that current MIR/MDL research falls short of being objective and scientific insofar as its work is represented in our literature. One model being put forward is the formation of a TREC-like (Text REtrieval Conference; see <http://trec.nist.gov>) set of collections and evaluation criteria. These collections and criteria will have to be designed specifically to address the wide variety of music-specific problems facing the MIR/MDL research communities. Establishing the ideal framework for developing these collections and criteria is the open question to be addressed over the period of this project.

## 1. PROJECT GOALS

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. Criteria will be established pertaining to the format of the test collections (e.g., MIDI, audio, notation, etc.); the necessary size of the collections; the requisite genres of music to be collected (e.g., classical, folk, jazz, popular, etc.); the classification of tasks to be tested (e.g., kinds of queries, input modes, etc.); and, the definitions of success/failure for deciding upon, then implementing, the evaluation metrics appropriate for each kind of task.

## 2. PRINCIPAL PROJECT COMPONENTS

There are five principal components to this project

1. Solicitation, compilation and publication of MIR/MDL community input via the creation of a "White Paper" reference document collection. [Hereafter, the "White Papers"]. The specific solicitation and background information for the JCDL Workshop (introduced next) available at:

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

[http://music-ir.org/JCDL\\_Workshop\\_Info.html](http://music-ir.org/JCDL_Workshop_Info.html).

2. Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation. Joint ACM/IEEE Conference on Digital Libraries, Portland, OR, 14-18 July, 2002. [Hereafter, the "Workshop"]. Specifics found at: <http://www.ohsu.edu/jcdl/ws.html#W4>.
3. Music Information Retrieval Evaluation Frameworks. (ISMIR 2002, 3rd International Conference on Music Information Retrieval, IRCAM – Centre Pompidou, Paris, France, 13-17 October, 2002.) [Hereafter, the "Panel"]. Specifics found at: <http://ismir2002.ircam.fr/panels.html#panel1>.
4. Information Dissemination and Intelligence Gathering (GSLIS, UIUC, 11 July, 2002 to 1 April, 2003).
5. Publication of formal recommendations for the establishment, implementation and continued development of MIR/MDL evaluation testbeds and frameworks. [Hereafter, the "Formal Recommendations"].

## 3. PROJECT PLANS

### 3.1 White Papers

Communication of community input will be grounded in the creation of "White Papers" or "Position Papers". These documents will be scholarly works outlining the thoughts, opinions, and arguments of their authors with regard to their visions of what needs to be included/considered in the creation of standardized MIR/MDL testbed collections, evaluation tasks, and evaluation metrics. I expect (and want) these documents to be highly idiosyncratic: a musicologist will write with respect to the needs and wants of musicologists, a lawyer with respect to legal issues, an audio engineer with regard to audio issues, and so on. It is my goal to attract such a wide range of submissions that the multi-disciplinary approach to MIR /MDL research is truly represented.

The following, non-exclusive (nor all-encompassing) list of open questions should help comprehending just a few of the many possible paper topics:

1. As a music librarian, are there issues that evaluation standards must address for their results to be credible? Do you know of possible collections that might form the basis of a test collection? What prior research should we be considering?
2. As a musicologist, what things need examination that are possibly being overlooked?
3. As a digital library (DL) developer, what standards for evaluation should we borrow from the traditional DL community? Any perils or pitfalls that we should consider?
4. As an audio engineer, what do you need to test your approaches? What methods have worked in other contexts that might or might not work in the MIR/MDL context?
5. As an information retrieval specialist, what lessons have you learned about other traditional IR evaluation frameworks? Any suggestions about what to avoid or consider as we build our MIR/MDL framework from "scratch"?
6. As an intellectual property expert, what rights and responsibilities will we have as we strive to build and distribute our test collections?
7. As an interface/human computer interaction (HCI) expert, what tests should we consider to validate our many different types of interfaces?
8. As a business person, what format of results will help you make selection decisions? Are there business research models and methods that should be considered?
9. As a computer scientist, what are the strengths and weaknesses of the CS approach to validation in the MIR/MDL context?

Other issues, topics and questions pertinent to the evaluation of MIR/MDL systems that might arise during the course of the project (e.g., topics that emerge as a result of the Workshop and Panel discussions) will also be considered. The underlying primary questions are:

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

### 3.2 Workshop

The White Papers for the Workshop have been solicited via a series of Call for Participation (CFP) notices posted to the most germane email lists (e.g., [music-ir@ircam.fr](mailto:music-ir@ircam.fr), [asis-l@asis.org](mailto:asis-l@asis.org), [IAML-L@cornell.edu](mailto:IAML-L@cornell.edu), [diglib@infoserv.inist.fr](mailto:diglib@infoserv.inist.fr), etc.) and via personal communications with those research teams that I know have special expertise in a particular evaluation domains (again, all aimed toward ensuring a wide range of viewpoints and backgrounds of submitters).

Dr. Ellen Voorhees of the National Institute of Standards and Technology has consented to be our keynote speaker on the subject of TREC-like evaluation scenarios. Dr. Voorhees' keynote White Paper may be found at: <http://music-ir.org/evaluation/voorhees.pdf>.

The list of White Paper submitters and their respective topic areas can be found in the Table of Contents.

Goals of the JCDL Workshop include:

1. The creation of the first edition of the "White Paper reference document collection". That is, the collection of our submitted papers will form the starting-point literature from which future arguments and work in the area of MIR/MDL evaluation could proceed. This collection will be made available to the community just prior to the Workshop date via: a) mounting at <http://music-ir.org>; and, b) distribution of printed copies at JCDL.
2. Workshop participants will be led to create a first draft consensus document or "manifesto" outlining the basic framework for further development of collections and evaluation standards. This document will be the "formal" outcome of the Workshop itself. The first draft consensus document, along with relevant notes and summaries, will be mounted shortly after the meeting at <http://music-ir.org> to solicit commentary from those not at the Workshop.
3. This first edition will play the important role of "priming the pump" for ISMIR 2002 Panel session discussed below. It can also be seen as foundational material for the solicitation of research funding, for the possible establishment of collection creation agreements, and for the creation of ongoing evaluation meetings and "contests."

### 3.3 Panel

The panel will build upon and extend the findings of the JCDL workshop. I hope to have this panel session further the development of prototype collections, tasks, and metrics that could be unveiled at ISMIR 2003. To this end, participants will include those who will be reporting on, and *extending*, the findings of the JCDL workshop and those who were unable to contribute to the JCDL workshop so new viewpoints and issues may be considered. I intend to solicit roughly another ten or so White Paper submissions with the **explicit goal** of filling in whatever areas are deemed to be missing from the first round of JCDL White Papers. I will also be soliciting for a keynote speaker that will provide feedback and/or critical commentary on Dr. Voorhees' TREC model of evaluation, again with the goal of broadening exposure to different modes of evaluation. Panel White Paper submissions will be incorporated with the preceding JCDL collection (including the resulting consensus document and summary materials), to form a "second edition" of the documentation. This second edition will be made available to the community just prior to the Panel date via: a) mounting at <http://music-ir.org>; and, b) distribution of printed copies at ISMIR 2002. Also included in this second edition material will be the findings-to-date of the Information Dissemination and Intelligence Gathering component of the project, discussed next. Minutes, summaries and revisions of the first draft consensus

document will be posted to the Web shortly after the Panel meeting.

### 3.4 Information Dissemination and Intelligence Gathering

This overarching component of information dissemination and intelligence gathering will be the “glue” that cements the efforts and results of the other components together to form the solid foundation for MIR/MDL evaluation. I will be the Principal Investigator and will be assisted by Graduate Assistants who will provide important help in such areas as research, communications, taking of minutes, editing, Web-development, printing, and other clerical tasks. The two main sub-components, both running in parallel for the life of the project, are:

#### 3.4.1 Dissemination

The *dissemination* sub-component which includes the iterative solicitation, formatting and editing of the White Papers (from both the Workshop and the Panel) for both print and WWW-based distribution (centrally housed at <http://music-ir.org>). It also includes incorporating the results of the discussions undertaken at the meetings and the results of whatever intelligence has been gathered into the main body of documentation. Overall, the main body of documentation will undergo three publication phases:

1. First edition: published prior to JCDL meeting (15 July 2002), to include:
  - Project Outline
  - Workshop White Papers
  - Chart of Candidate Music IR Test Collections (explained later)
2. Second edition: published prior to ISMIR 2002 Panel meeting (30 September 2002), to include:
  - Items from first edition
  - Panel White Papers
  - Workshop Consensus Information
  - Findings-to-date of Intelligence Gathering.
3. Final edition: published at conclusion of project (1 April 2003), to include:
  - Revised and updated items from second edition
  - Panel Discussion Information (discussed next)
  - Findings of the Intelligence Gathering
  - Formal Recommendations

The first and second editions of the materials will be under my exclusive editorial control. Submitters will be encouraged to revise their submissions after they have presented them (either at the Workshop or the Panel) to incorporate whatever editorial suggestions arise from myself or from participant discussions. For the final edition, however, an advisory panel consisting of two other scholars and myself will be struck to more formally vet the submitted White Papers. This committee will be struck in late October and will be responsible for ensuring the intellectual and stylistic quality of the White Papers prior to their being published in their final iterations.

#### 3.4.2 Intelligence Gathering

The *intelligence gathering* sub-component includes the preliminary identification, and analyses of, possible extant collections and evaluation issues that might meet or hinder the needs of MIR/MDL evaluation. Dr. Don Byrd (Indian

University), Tim Crawford (Kings College, UK) and Jeremy Pickens (University of Massachusetts, Amherst) have prepared a preliminary “Chart of Candidate Music IR Test Collections” (see Appendix A). As a central part of the intelligence gathering, my assistants and I will assess the suitability of each of the candidate collections for inclusion in the evaluation testbed(s). Assessment criteria will include:

1. Format availability (i.e., audio, MIDI, notation, etc.)
2. Conversion possibilities (i.e., from notation into audio; from audio to MIDI, etc.)
3. Coverage (i.e., genre and/or cultural coverage of the possible collection(s))
4. Identification of stakeholders (i.e., owners of possible collections, etc.)
5. Legal issues (i.e., copyright concerns, ownership and dissemination issues, possible licensing scenarios, etc.)
6. Research issues (i.e., human-subjects concerns, ethics, etc.)
7. Infrastructural issues (i.e., where to house collections, rights management issues, access issues, etc.)

As other possible collections are discovered, they too will undergo similar assessment.

Establishing open and collegial communications with the collection stakeholders is an important aspect of intelligence gathering sub-component. Through such communications we hope to convince the collection stakeholders of a) the scholarly importance of MIR/MDL evaluations; and, b) the mutual benefits to their materials (e.g., research progress for MIR/MDL investigators and new dissemination tools for the collection stakeholders). The ultimate goal of the communication process is access to, and use of, the collections on the least restrictive terms possible.

### 3.5 Formal Recommendations

After analyzing the contents of the White Papers, Consensus Document, participant and community feedback, and the findings of the intelligence gathering, a set of “Formal Recommendations” for MIR/MDL evaluation will be compiled and published as part of the final edition of the documentation. I foresee a three-level set of recommendation as the primary outcome of the project:

#### 3.5.1 Level 1

Pragmatic and short-term recommendations proposed as a means of “boot strapping” the evaluation process wherein possibly less-than-ideal evaluation collections, tasks and metrics are suggested with an eye toward holding the first formal MIR/MDL evaluation meeting as part of ISMIR 2003.

#### 3.5.2 Level 2

Comprehensive and mid-range recommendations that address a broader range of issues in a more detailed and rigorous manner. Included in this level, for example, will be recommendations concerning: a) the categorizations of evaluation types, their associated collections and metrics; b) sustainability of effort (i.e., potential sources of funding, the organization of evaluation committees, compiling new collections, etc.); c) timelines for

future research and evaluation meetings; d) identification of possible synergistic relationships with other evaluation frameworks (e.g., TREC, digital library evaluation, etc.); and, e) dissemination options for the evaluation results.

### 3.5.3 Level 3

Idealistic and long-term recommendations for those evaluation issues that have no foreseeable short-term nor mid-term resolution but nevertheless must be addressed if the MIR/MDL evaluation frameworks laid out principally in Level 2 are to have continued relevance for the MIR/MDL research communities.

## 4. SUMMARY TIMELINE

**Table 1: Summary Timeline**

<b>Task</b>	<b>Target Dates</b>
Solicitation of JCDL White Papers	DONE
Start Information Dissemination and Intelligence Gathering process	1 July 2002
Print and mount JCDL White Papers [a.k.a. "first edition"]	15 July 2002
JCDL Workshop meeting	18 July 2002
Post "first draft" consensus document, meeting minutes and summaries	22 July 2002
Solicit White Papers for ISMIR 2002 Panel	22 July 2002
Print and mount the combination of the ISMIR 2002 and JCDL White Papers, first draft consensus document, and findings-to-date of the intelligence gathering [a.k.a. "second edition"]	30 September 2002
ISMIR 2002 Panel meeting	17 October 2002
Mount Panel minutes, summaries and updates to consensus document	20 October 2002
Strike advisory committee and continue with intelligence gathering	30 October
Wrap up intelligence gathering	28 February 2003
Print and mount final revised collection including the Formal Recommendations [a.k.a. "final edition"]	1 April 2003