

Interim Report on Establishing MIR/MDL Evaluation Frameworks: Commentary on Consensus Building

J. Stephen Downie

Graduate School of Library and Information Science

University of Illinois at Urbana-Champaign

+1-(217) 265-5018

jdownie@uiuc.edu

1. INTRODUCTION

The papers in Part I which immediately proceed this paper represent the first round of formalized MIR/MDL community input on the topic of MIR/MDL evaluation. Each were presented at the *Workshop on the Creation of Standardized Test Collections, Tasks and Metrics for Music Digital Library (MDL) and Music Information Retrieval (MIR) Evaluation*, held 18 July, 2002 at the *ACM/IEEE Joint Conference on Digital Libraries*. The Part I papers were originally published as "The MIR/MDL Evaluation Project White Paper Collection, Edition #1". To these first edition papers we now add the papers which follow in Part II. The aggregation of the Part I and Part II papers together represent the second edition of the "White Paper Collection". The Part II papers are those to be presented at the *Panel on Music Information Retrieval Evaluation Frameworks*, to be held 17 October, 2002, at the *3rd International Conference on Music Information Retrieval (ISMIR 2002)*, Paris, FR.

It is the purpose of this paper to present a brief, highly idiosyncratic, summary of the issues and themes that I see emerging from the papers at hand, from the discussions that took place at the 18 July workshop and from personal communications that I have had with various members of the MIR/MDL communities since the inception of the MIR/MDL Evaluation Frameworks Project in early July 2002. Let me stress here that these comments are far from conclusive and are primarily meant to stimulate further discussion and debate. I will be publishing more formal recommendations mid-spring of 2003. Until then, when the third and final edition of the "White Paper Collection" will be published, my comments and suggestions--along with those of the paper authors who have so generously given of their time to participate in this project--are most decidedly open to revision.

2. TREC-LIKE EVALUATIONS

Of all the scenarios being discussed, there is an almost unanimous consensus on the fact that MIR/MDL evaluation frameworks should have a vigorous TREC-like component. Dr. Ellen Voorhees' keynote address highlighting the potential value of TREC-like evaluations resonated strongly with the July workshop participants. The best way to summarize the tenor of the discussions and input on the topic of TREC-like evaluations is to say that the community is no longer concerned with "if" questions but has become fully engaged in the "how" questions. That is, how do we gather up the necessary materials? how we do define

our TREC-like tests? how do we administer the evaluations? and so on.

As part of the "how" aspect of the discussions concerning TREC-like evaluations, the importance of testbed collections which contain works simultaneously represented in symbolic, score and audio formats has been repeatedly stressed. This is not too surprising as the community entered into the evaluation discussions with these format variations already in mind. One thing that has emerged however, that I and many others had overlooked as being important to the development of useful multi-format testbed collections, was the inclusion of metadata. At the July workshop I was astonished to realize how little I had thought about metadata as a) being intrinsically valuable in its own right (and a legitimate part of any MIR/MDL task); and, b) absolutely necessary to the establishment of linkages between the different format representations of each individual work in the testbed collections. So, to my original list of testbed components, I am now adding metadata and plan to explore metadata requirements and options more thoroughly.

Related to the format issues outlined above, there is also a strong consensus that the query statements put together for the TREC-like evaluations must include all four representations (i.e., score, symbolic, audio and text/metadata). How best to do this is a question yet to be resolved. Also, there is a general consensus that these query statements must reflect a broad mix of real world needs. How to best acquire these "reality based" queries is another question yet to be resolved.

Note that I have been deliberate in my use of the term "TREC-like". The MIR/MDL community is rightly convinced, and quick to point out, that music information and it uses differ in many dimensions from text. This being the case, I am reading the community opinion as saying that we must take great care that we do not oversimplify and naively apply the use of TREC techniques in the evaluation of MIR/MDL systems. Issues to be addressed include the all-important, and perhaps elusive, definition of relevance (upon which evaluation rests), the building of a stronger consensus on which MIR/MDL tasks are truly amenable to such evaluation and the structuring of tests to ensure that the results are both valid and reliable, to name but a few.

3. BEYOND TREC-LIKE EVALUATIONS

Notwithstanding the strong community support for establishing TREC-like evaluations, I must also highlight an important caveat. There is a parallel theme emerging that reminds us that TREC-like evaluations must not, indeed should not, be the exclusive

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.

focus of our efforts. There are many evaluative questions about MIR/MDL tasks for which TREC-like evaluations are simply not informative. How well users actually interact with various systems and their constituent sub-components is a perfect example. How potential users conceive of their music information needs is another. How users intuitively express those needs is but a third example among many more. Dr. Edie Rasmussen's keynote address to the ISMIR 2002 Panel on Evaluation Frameworks further explicates some of these central issues.

We must pay attention to this theme for several reasons. First, it is correct in asserting that we need as broad an understanding of MIR/MDL systems, their use and their users, as possible. Second, on a community building level, we must make the testbeds and evaluation "contests" that we develop as enticing to as wide a swath of the community as possible. That is, we must make sure that we develop our testbed collections with both TREC-like and non-TREC-like evaluation modes simultaneously in mind. While it is unlikely that we will be able to satisfy all members of the community at any one point in time, we must at least endeavor to "cycle through" a wide variety research modalities and questions over a relatively short time span. Should a TREC-like model of evaluation dominate, I suspect that the MIR/MDL community will lose its greatest strength, namely its interdisciplinarity, as non-TREC-like researchers leave us to explore more hospitable subjects and problem spaces

4. MORE INPUT NEEDED

I have been very fortunate to have had so many MIR/MDL researchers either write formal White Papers for this project or send personal opinions and comments my way. In looking over the bylines of those that have contributed or corresponded with me I note, however, some important absences or under-representations. This project needs more input from those in the

music librarianship sector, for example. The music librarianship sector has had a long history of interaction with both music information and its users, so its insights and opinions on system needs and user behaviors are particularly germane. Also absent are those involved in the commercial aspects of music. Sellers of music and catalogue holders are needed to inform us of their particular needs and desires so we might better tailor our evaluations to help meet those needs and desires. Other groups currently under-represented include those that have research traditions beyond TREC-like evaluations such as Human-Computer Interaction (HCI) researchers, musicologists, music performers, music psychologists, and so on. I am committed to actively soliciting the input of representatives of these groups. If you are a member, or know of an interested member, of these communities, I would appreciate hearing from you.

5. ACKNOWLEDGEMENTS

The first two persons whom I must thank and acknowledge are Dr. Ellen Voorhees and Edie Rasmussen for so ably giving of themselves are the project's keynote presenters. Their input has helped inspire both myself and the other contributors to think of the MIR/MDL evaluation problem in a new, and better informed light. Next I would like to thank Dr. Suzanne Lodato and the Andrew W. Mellon Foundation for their ongoing moral and financial support of the project. My colleagues at the Graduate School of Library and Information Science have been extremely supportive of my work and I owe an eternal debt of gratitude. Gina Lee and Karen Medina, my Graduate Assistants, are also thanked most heartily for their work in preparing the papers for publication. Finally, I want to thank all those authors that have contributed White Papers for us to consider.