

## Metadata Issues for RAMH2

Jenn Riley

February 2009

A project of the vast scope of a new, electronic edition of *Resources of American Music History* (RAMH2) presents many challenges for the selection of metadata to be collected, stored, and shared. As a print reference work, the data presented original RAMH is primarily descriptive rather than prescriptive in nature. As an electronic reference work, RAMH2 has the potential to change this focus. The degree to which this is desirable is a primary factor affecting all other decisions about metadata structure for the RAMH2 project. Some inspiration might be gained from close examination of other electronic reference sources that have been re-imagined from earlier print forms, for example:

- Film Literature Index (FLI) <<http://www.dlib.indiana.edu/reference/fli/>>
- Encyclopedic Discography of Victor Recordings (EDVR) <<http://victor.library.ucsb.edu/>>

### Issues affecting metadata decisions

#### *Functionality*

The online RAMH2 must replicate not only the functionality of the main repository listing in the original RAMH, but also the access provided by the index to the printed volume. These two primary functions are very different in character. The repository entries are descriptive in nature, consisting of prose that does not necessarily conform to any predefined structure (although the introduction to the original RAMH volume describes some general editorial conventions that were applied). In the print RAMH these entries have a nearly arbitrary structure, with many entries consisting simply of a repository name and address plus paragraph-like lists of the types of resources held and the name of the individual submitting an entry, while others contain elaborate sub-structures including for example multiple repositories within an institution, collection descriptions within a single repository submitted by different individuals, and references to external descriptions of collections.<sup>1</sup> The “APX” entries <<http://www.musiclibraryassoc.org/publications/newsletter/MLANEWS133.html#RAMH>> at the end of each section, listing additional repositories with potentially relevant collections, further add to the complexity of the structure of the main body of the RAMH text.

The RAMH index by contrast is significantly more structured than the main repository listing. The index is intended to provide additional access points to included collections beyond their physical location (the organizational principle behind the main RAMH text). The most notable access points in the print RAMH index include the names of people (composers, performers, collectors, significant instances of individuals as subjects), geographic places (as subjects), musical forms, and topics.

---

<sup>1</sup> See the entry on Indiana University (406, p. 118 in the print RAMH) for an example of the potential complexity of institutional entries.

### *Submission methodology*

It is clear that a primary benefit of the online environment to RAMH2 is the potential for online submission of entries directly into a repository that automatically generates the data shown to end users of the RAMH2 reference source. Direct submission of entries both allows RAMH2 to be a dynamic and growing resource rather than a one-time release, and for editorial overhead to either be reduced or redirected to other, higher-value, activities.

The design of a submission system for collection descriptions must navigate the delicate balance between making it easy to submit an entry (the approximately 25% overall response rate for the original RAMH<sup>2</sup> illustrates clearly the need to create as low-barrier method for submission as possible) and the predictability and authority of metadata submitted to the system that will allow more advanced functionality for end-users of RAMH2. The submission system must be available and useful for even the smallest of repositories, including the individual collector wishing to participate and to the local historical society whose secretary must interrupt her work on contributing to the description of a collection to attend to her Christmas cookies.<sup>3</sup> Yet at the same time the submission system must work well for larger repositories that likely already have existing structured or semi-structured collection descriptions. For these repositories, a method to accept a pre-existing collection-level MARC record or EAD-encoded finding aid would be desirable, as mentioned in Jon Dunn's RAMH2 position paper, "Resources of American Music History 2: Approaches to Technical Implementation." The metadata from these would then be used to pre-populate a collection record in the RAMH2 system, that could be (optionally) further enhanced by the repository. In all cases, regardless of how easy it is for a repository to submit an entry itself, it is likely some editorially-created entries will be needed as well. The submission system should be designed to track the source of an entry, allowing repository-submitted data to be distinguished from editorially-submitted data.

The variety of repositories containing collections relevant to RAMH2 (and, by extension, the variety of descriptive practices they are likely to employ) make it difficult to design a system that fully standardizes description of collections. My recommendation is to implement a metadata structure for internal storage of RAMH data that capitalizes on the best of both worlds by combining unstructured (prose) description with predictable, predefined, structured data elements. The prose data would be entered by submitters into a web form with brief but descriptive labels for the data expected to be submitted, along with obvious links to fuller instructions for creation of the description (although these are likely to be ignored or never viewed by most submitters). The structured data that would be available for submission directly by repositories should be presented as distinct fields in a web form, making use whenever possible of combo boxes providing a short list of predetermined options. (Structured data elements that might be included are discussed later in this paper.) The selection of these structured data elements and whether or not they should be required for population by repositories submitting entries to RAMH2 will be guided by decisions on editorial policy discussed below.

---

<sup>2</sup> RAMH, p. 4

<sup>3</sup> Krummel, D.W., "Little RAMH, Who Made Thee? Observations on an American Music Census," *Notes*, Second Series 37, no. 2 (December 1980): 228.

The underlying focus of RAMH2 as a growing and participatory resource suggests the ability for end-users to contribute to the description of included collections. Many RAMH2 users will likely be scholars of American music and have valuable information and/or perspectives on the collections that even the holding institutions may not have. RAMH2 could serve as a repository for this analytical information, and as a platform for furthering scholarly communication between those interested in American music. While models for simple end-user tagging are abundant, RAMH2 would likely benefit from more structured metadata contributed by users as well. Fewer models for this type of end-user interaction currently exist, and significant planning would need to be done to determine what subset of the descriptive data already stored for RAMH2 entries would be available for enhancement by end users, and what additional (likely evaluative) information would be collected as well. Metadata contributed by end-users would almost certainly need to be stored separately from repository-submitted metadata (or at least flagged as such), but planning will also need to be done to determine to what degree user-contributed metadata should be distinguished from other types in the end-user interface.

### *Editorial oversight*

Repositories submitting metadata directly into RAMH and especially end-user contribution of metadata suggest the need for clear policies on editorial oversight of this data. It is my recommendation that the unstructured prose collection descriptions be subject to minimal if any editorial control, allowing editorial resources to be devoted primarily to structured access points that will drive in large part any advanced discovery services provided to RAMH2 users.

The minimally-overseen unstructured prose descriptions in this model would serve as a repository's own view of a collection, allowing it to highlight what it considers most important and to present its own unique perspective on its holdings and their use. The minimal level of control over these descriptions that is truly necessary can be handled through access management for repositories (that is, some low-overhead mechanism for a repository to request access to the RAMH2 data-entry system), well-written and easy-to-understand instructions for repositories submitting descriptive information, and good design of the web interface that repositories use to submit metadata.

The editorial oversight provided on the structured data should focus primarily on "added value" services. The exact model to be implemented will depend in large part on what resources are available for editorial work during the RAMH2 initial ramp-up period and, more importantly, indefinitely into the future as RAMH2 continues to grow over time. The types of data likely to benefit from this editorial oversight include resource types, topics, and names. Standardization (and perhaps even full authority control) over names is an obvious potential added value activity for editorial work over submitted entries. While the original RAMH did not endeavor to provide name authority control<sup>4</sup> due to the resource-intensive nature of this activity, the editorial time-savings from direct electronic submissions in RAMH2 could potentially free up enough editorial resources to commit to this high-value activity.

---

<sup>4</sup> RAMH, 295; Krummel, 232.

### Data that should be stored

It should at this point be noted explicitly (although the discussion in this paper and others have made this assumption implicitly) that the metadata stored in RAMH2 is *collection-level* description rather than the item-level description libraries are traditionally accustomed to. The metadata needed for collection description is similar but not identical to the metadata needed for item-level description. The University of Illinois at Urbana-Champaign's IMLS-funded Digital Collections and Content (DCC) project has done significant research into the necessary properties of collection-level description, including the creation of a metadata application profile for collection description.<sup>5</sup> While the IMLS DCC collection description application profile is likely not useful wholesale as a metadata model for RAMH2, its features should be useful as inspiration for additional data elements to store beyond those that might otherwise be envisioned.

A general data model will be needed for RAMH as a whole, to organize the collection descriptions that are submitted. Based on an initial review of the print RAMH volume, a model similar to the following might suffice:

*Institution.* This would be the highest level in the hierarchy, representing a named organization. Individual collectors would also go at this level, so perhaps a better label will be required.

*Repository.* Institutions will consist of one or more repositories which are responsible for the management of resources. Each repository will submit descriptions separately from other repositories in the same institution when multiple repositories exist at the same institution. There is some precedent in the print RAMH for multiple contributors to descriptions for a single repository.

*Collection.* Repositories can manage one or more collections, each which will be described separately. Each collection will contain both an unstructured prose description and a set of more structured data elements.

Structured data for each submitted collection (whether repository-, user-, or editor-created) might include the following:

- *Resource type.* The original RAMH requested information from repositories on ten types of documents (sheet music, songbooks, other printed music, manuscript music, programs, catalogues, organizational papers, personal papers, and sound recordings).<sup>6</sup> This or a similar categorization could be used to provide high-level access to collections by type of resource contained.
- *Date range.* For each collection, "bulk" dates<sup>7</sup> should be recorded. This will allow searching and browsing (potentially as a facet) of collections by date, providing a significant improvement over the print RAMH.

<sup>5</sup> <http://imlsdcc.grainger.uiuc.edu/resources.asp>

<sup>6</sup> RAMH, p. 2

<sup>7</sup> Bulk dates are a concept commonly used in archival description to represent the dates *most* of a collection spans, designed to minimize the impact of single exceptional documents on retrieval of resources by date.

- *Names.* Personal, family, and organization (corporate) names were noted in the original RAMH as useful access points for discovery of collections. Structured names would supplement names mentioned in the text of prose collection descriptions. As discussed above, editorial addition of authority control over names could provide a significant value-added service.
- *Topical subjects.* The current RAMH index includes terms on the subject matter of collections. Standardized topical terms could supplement free-text descriptions in the prose entries.
- *Form/genre.* Musical forms and genres present in collections represent important access points that could serve as search or browse indexes in RAMH2. The Library of Congress is currently beginning a project to separate musical form and genre terms from topical terms in the LC authority file.<sup>8</sup> Form/genre indexing for RAMH2 could potentially build upon this work.
- *Geographic places.* Place-focused access in the print RAMH is inconsistent, with “the names of cities and states ... listed in the index only for materials located elsewhere” due to the fact that the main text is organized by physical location of the holding repository. In the online environment, more robust geographic indexing is possible, making explicit when a geographic term applies to the physical location of materials and when the materials themselves are relevant to a given place.
- *References to online versions of materials.* Links to interfaces that provide online access to some or all of the materials in a repository’s collection would present significant added value to RAMH2. Links from the collection-level record should point to high-level discovery services; links to specific items should be reserved for the potential item-level addition to RAMH2 described later in this paper.
- *References to secondary sources.* Connections between the primary sources listed in RAMH2 and secondary sources aggregating or analyzing them are another area in which RAMH2 could provide significant added value. These types of references are particularly ripe for submission by end-users of RAMH2.

The prose collection descriptions in RAMH2 will require some degree of formatting as well. A simple text blob will be insufficient for the presentation of collection description of the type seen in the print RAMH volume. At a very minimum, paragraph breaks, emphasis such as bold and italic type, and section headings will be needed to provide a coherent presentation of a collection description to end users.

### **Metadata formats that might be used**

The wide scope of the metadata discussed this far, particularly the need for both structured and unstructured description, and the need to track the source of various pieces of metadata, strongly suggest that no commonly-used metadata format in the library or archives community will suffice for the native storage of RAMH2 metadata. The application and data store underlying the systems that

---

<sup>8</sup> <http://www.loc.gov/catdir/cpsso/genretimeline.pdf>

drive the submission and end-user interfaces should therefore be designed explicitly around RAMH2 system requirements. Standardized metadata can then be generated from this native data store for specific applications as discussed below.

The online, evolving, and participatory nature of RAMH2 suggests that it should not only provide a data *collocation* service (gathering together information about collections from disparate sources into an easy-to-use system for end-users interested in American music), but also a data *sharing* service. RAMH2 can and should re-expose the value-added data it creates to other services and provide exposure of structured metadata for collections within its scope via machine protocols on behalf of institutions that do not have the technical resources to do it themselves.

Collection-level records containing some significant subset of (but not necessarily all) data stored in the RAMH2 system should be exposed by the RAMH2 service for use by external services without restriction. MARC records are traditionally shared via Z39.50. This protocol is undergoing a slow but still noticeable decline, and it is my belief that the benefits to Z39.50 for RAMH2 do not justify the resources necessary for implementing it. Instead, collection-level MARCXML<sup>9</sup> records should be available through protocols such as OAI-PMH, SRU, and OpenSearch, as described in Jon Dunn's RAMH2 position paper, "Resources of American Music History 2: Approaches to Technical Implementation." Collection-level records in MODS<sup>10</sup> and simple Dublin Core<sup>11</sup> would also be useful to share via these protocols.<sup>12</sup>

Two other formats might be considered in addition to those already described. While likely to be more difficult to implement, an EAD-encoded finding aid<sup>13</sup> containing only collection-level data would be useful, for exposing collections to services such as OCLC's ArchiveGrid.<sup>14</sup> Finally, while certainly more difficult to implement and with less of a natural audience, generating a Text Encoding Initiative (TEI) document for each collection would be an interesting implementation choice. While the other formats mentioned in this section take a certain approach to the description of a collection (either bibliographically or archivally), TEI is more neutral in nature. TEI would likely allow the sharing of more of the native RAMH2 metadata than any of the other options discussed thus far. I am not at this time recommending the TEI option be implemented, but it is certainly something to consider as the project moves forward and the use cases for metadata re-exposed by the RAMH2 service become clearer.

### **Potential expansions of scope for RAMH2**

As described in Jon Dunn's RAMH2 position paper, "Resources of American Music History 2: Approaches to Technical Implementation," expansion of the RAMH2 service beyond collection-level description into the realm of item-level description would provide significant added value to the service, while

---

<sup>9</sup> <http://www.loc.gov/marcxml>

<sup>10</sup> <http://www.loc.gov/mods>

<sup>11</sup> <http://www.dublincore.org/documents/dces/>

<sup>12</sup> In the case of simple Dublin Core, this format is required for using the OAI-PMH protocol, although the other formats mentioned could supplement this basic representation.

<sup>13</sup> <http://www.loc.gov/ead>

<sup>14</sup> <http://archivegrid.org/>

simultaneously presenting a challenge as to an effective way to communicate to users that item-level description exists for only some small subset of the resources represented.

Collecting item-level descriptions from contributing repositories represents a significant technical challenge to RAMH2 as well. These repositories will not all have structured metadata appropriate for sharing, and even those that do will not all use the same metadata format or similar rules for the creation of this metadata. RAMH2 could take the easy (well, *easiest!*) technical road and collect item-level metadata in simple Dublin Core via OAI-PMH for those repositories that provide this capability. In addition to the technical challenges to this approach that Jon Dunn describes in his paper, it also misses a significant amount of item-level description that exists, as many repositories do not have the technical resources to set up an OAI-PMH data provider or even create an OAI-PMH Static Repository XML file for sharing their metadata via this protocol. To overcome these limitations, RAMH2 would need to set up a full (and fully supported) metadata mapping service. Such a service would require a significant technical infrastructure and even more significant level of ongoing staff support. I believe this full mapping service is out of scope for RAMH2 at this time. Given the limitations of item-level metadata in RAMH2 with the OAI-PMH approach, I am uncertain at this time as to whether the benefits of this additional service are worth the drawbacks.

One can imagine an even more full-featured service as part of RAMH2, providing repositories with a technical infrastructure for creating item-level description (perhaps in the MODS format, from which simple Dublin Core and a container list in EAD could be derived) and delivering online digitized surrogates of items in their collections centrally through RAMH2 rather than requiring institutions to set up their own technical infrastructure for these activities. As the RAMH2 service emerges, repositories will likely ask for such an extension of the service. While I believe this is a good long-term direction, I agree with Jon Dunn's position paper that the very complex technical infrastructure required to provide this should be out of scope for RAMH2 at this time, despite its obvious utility. Should a long-term funding model for RAMH2 emerge that will provide the (expensive) technical support needed for such a service, this decision could be re-evaluated. Especially during the current IMLS planning grant for RAMH2, it is important to keep the project scope reasonable and manageable, while simultaneously tracking ideas for longer-term expansions of the project.