

Resources of American Music History 2: Approaches to Technical Implementation

Jon Dunn
Associate Director for Technology
Digital Library Program
Indiana University
Bloomington, Indiana

February 16, 2009

In many ways, this is an ideal point in time in the evolution of digital library technology to implement a new online version of Resources of American Music History. Digitization of most media formats has become a mainstream activity, as an increasing number of collections of primary source materials are being made available on the Web by institutions and individuals. The emergence of “Web 2.0” has brought along the development of new technical mechanisms for exposing, sharing, and collecting online information and the creation of new models for interaction between information providers and information consumers. However, no project to date has made use of these emerging technologies and models to provide comprehensive access to primary source materials in the way that RAMH2 envisions, and so there is an opportunity to develop a technology platform that is not only useful to RAMH2 but to similar efforts that are no doubt emerging within other disciplines.

In this paper, I raise some questions about scope and attempt to lay out a basic technical framework for RAMH2, but do not present a detailed technical architecture and design. That will require further discussion amongst the creators of RAMH2 as to what we want the resource to be and then further investigation and evaluation of various options by the technologists who will be implementing the technology platform to support RAMH2.

Collection Database

Based on the vision set out in the IMLS planning proposal, at the heart of the RAMH2 system is a database of annotated collection-level bibliographic records describing the archival collections included in RAMH2. These collections may be physical, digital, or both. The particular format of these records is left to the metadata experts, but should incorporate the use of controlled vocabularies for facets such as geographic location, topical subject, content type (e.g. audio, video, manuscripts, etc.), and names in order to facilitate the collocation of records and creation of a Web-based interface for online browsing of RAMH2. The specific database technology used will be largely dependent on the preferences and local practices of the institution that will be building and sustaining the RAMH2 online resource, but could be a relational database such as MySQL, PostgreSQL, or

Oracle, or an XML-based database, content management system, or repository such as Fedora.

In order to make the process of creating RAMH2 as efficient as possible and to focus use of human effort where it is required, it makes sense to automate the creation of these collection-level records whenever possible, but there are limits to what automation can do. For institutionally-based collections, machine-readable collection level cataloging may already exist as MARC records in OCLC WorldCat or as EAD-encoded finding aids on institutional web sites. RAMH2 should provide a mechanism to import an existing collection-level MARC record from WorldCat to serve as a starting point for the RAMH2 record. OCLC's recently announced OCLC Grid Services¹ include a WorldCat Search API (application programming interface), which provides access to the WorldCat database via the SRU (Search/Retrieval via URL)² and OpenSearch³ protocols, and may be useful as a technical mechanism for implementing this capability.

In the case of collections with Encoded Archival Description (EAD) finding aids, it is worth exploring the possibility of implementing a mechanism for automatically importing a finding aid to populate the initial collection-level record in RAMH2, but this may be more difficult than dealing with WorldCat for two reasons: 1) The system would require access to raw XML-encoded EAD finding aids rather than to HTML-based Web presentations, and institutions that implement EAD may not actually expose these raw EAD finding aids for Web retrieval, and 2) the structure of EAD is flexible enough that it may be difficult to consistently parse EAD finding aids coming from multiple repositories, due to local idiosyncratic practices.

Of course, a Web-based user interface to allow manual creation and editing of collection-level database records is required as well. For private collections and institutionally-based collections that have finding aids and collection descriptions only as Web pages, word processing documents, or on paper, manual entry of the collection-level record is likely the only way to get the data in.

On top of this database of collection level records, a Web-based user interface for browsing and searching the collection descriptions is needed. The precise technologies used to implement this interface will again depend largely on local expertise and practice of the institution implementing and supporting the system, as well as on the precise requirements for what users should be able to do with the contents. These requirements might be developed in part based on interviews with scholars and other potential users of RAMH2.

Item-Level Access

In addition to the database of collection-level records, which would basically duplicate the model of RAMH1 in an online environment, it is worth considering the creation of a second

¹ <http://worldcat.org/devnet/wiki/Services>

² <http://www.loc.gov/standards/sru/>

³ <http://www.opensearch.org/>

database to support searching and browsing of digitized and born-digital resources, and potentially non-digital resources as well, at the item level. For digital resources, users would be able to click on a link in the item-level record in RAMH2 to take them to an online presentation of the resource at the web site of the institution where the digital collection resides.

This is a more challenging task than creation of the collection-level database, due to the potential size of the database as well as the diversity of item-level descriptive practices across various institutions, and would require ongoing technical resources to support. It would also never be as comprehensive as the collection-level database, and this fact would need to be made clear to users of RAMH2.

The primary existing model for item-level access to digital resources is that of metadata harvesters based on the Open Archives Initiative Protocol for Metadata Harvesting⁴ (OAI-PMH). OAI-PMH provides a relatively simple technical mechanism for providers of digital collections (*data providers*, in OAI-PMH parlance) to expose metadata records for the items in their collections, typically in Dublin Core or MODS, and for *service providers* to collect up or harvest those metadata records to create aggregations on which other services, such as searching and browsing, are built. Notable examples of OAI-PMH-based service providers include OAIster⁵ at the University of Michigan (which is in the process of being transferred from UM to OCLC for ongoing management), the Digital Library Federation's American Social History Online,⁶ and the Sheet Music Consortium⁷ based at UCLA.

Despite the existence of a standard protocol, there are limitations and challenges in building an aggregation of item-level descriptions using the OAI-PMH protocol. Many of these limitations have to do with variation in descriptive practice and in how metadata records, typically in Dublin Core, are constructed for exposure by a data provider. In recent years, there has been ongoing effort within the metadata community to develop recommendations and best practices for creation of "shareable metadata" to help with this problem. Another potential problem is that the sets of metadata records exposed by a data provider may not correspond to the collection policy of a service provider. For example, in the case of RAMH2, an institution may expose a collection or other set of resources of which only part relate to American music, and it would be up to RAMH2 as service provider to figure out how to discard records that are not relevant, if so desired.

Even if perfect metadata exists, human interaction between a data provider and service provider is typically required in order to work out various technical issues. Of course, if item-level description simply does not exist—for example in the case of many collections whose contents are described using EAD finding aids—there is no way for it to be exposed or harvested.

Many software systems commonly used by institutions to provide access to their digital collections, including ContentDM, DSpace, and Fedora, come with OAI-PMH data providers.

⁴ <http://www.openarchives.org/pmh/>

⁵ <http://www.oaister.org/>

⁶ <http://www.dlfaquifer.org/>

⁷ <http://digital.library.ucla.edu/sheetmusic/>

For institutions that cannot easily support an OAI-PMH provider, a mechanism known as OAI-PMH Static Repository Gateway allows a database of item-level metadata to be exported to XML and made available for harvesting without the overhead of setting up a data provider. Again, this typically requires technical assistance on the part of the service provider and would require ongoing technical resources devoted to RAMH2.

The next step beyond creating an item-level database would be for RAMH2 to take on storage and delivery of the actual digital content, in order to provide a consistent user experience and help ensure ongoing availability of the resources. This would be a difficult and costly endeavor, due not only to the amount of storage required but also the complexities of acquiring and managing digital objects that have been built across multiple institutions using a diverse array of file formats and data structures—complexities that are many orders of magnitude greater than the already difficult challenges of dealing with the diversity of item-level metadata. I would argue that digital content storage and delivery should be out of scope for RAMH2.

User-Contributed Data

The IMLS planning grant proposal for RAMH2 notes the need to engage the wider scholarly community in building the RAMH2 database, and George Boziwick's paper on collecting for RAMH2 emphasizes that RAMH2 should be dynamic and adaptable to change, potentially through use of Wiki technology. There are a number of ways from both a user perspective and technical perspective to make it possible for users of RAMH2 to contribute new or updated information to the database, and a Wiki-like interface of some type seems appropriate. The power of a Wiki-based system, such as Wikipedia, is that it allows users to easily edit and add information, keeps a log of changes made, and allows one to go back to any version of a page or record over time. However, unlike Wikipedia, presumably a more structured editorial process is necessary for receiving and reviewing additions and changes submitted by users for a scholarly resource such as RAMH2. Any changes or additions to a record could be routed to an editor for approval before inclusion, and/or there could be a separate comment section of the record that allows for comments, tagging, and annotations from registered users.

In addition, the data in collection-level and item-level records in RAMH2 will likely be more structured and database-like than Wikipedia articles. In recent years, a new type of Wiki software platform has emerged that is sometimes referred to as a *structured Wiki*.⁸ Such platforms support both the freeform "Wiki markup" text that is typical of Wikis as well as more structured form-based data entry that can be tied to an underlying relational database. Some popular structured Wiki platforms include TWiki⁹ and XWiki,¹⁰ and it would be beneficial for the project to evaluate the use of such a platform to help build the front-end user interface vs. creating a tool from scratch.

⁸ http://en.wikipedia.org/wiki/Structured_wiki

⁹ <http://www.twiki.org/>

¹⁰ <http://www.xwiki.org/>

Re-Exposing RAMH2

Most of the discussion in this paper so far has been on how RAMH2 can collect data from other systems and from its users. In turn, part of being good Web 2.0 citizens involves looking at how RAMH2 can re-expose its own data for use by other tools and services. These might include tools used by scholars such as bibliographic management software (e.g. Zotero, EndNote); social bookmarking, tagging and personal collection building tools (e.g. CiteULike, Delicious); and annotation tools. They might also include federated search services and other web sites that may be built to provide access to scholarly resources whose collecting missions are broader than or overlap with RAMH2 and might wish to draw upon RAMH2's database to help populate their own.

In addition to exposing its collection-level and item-level databases via some of the protocols already mentioned above—OAI-PMH, SRU, and OpenSearch—RAMH2 should track and consider implementing other protocols or standards that are emerging to support interoperability of bibliographic data and digital resources, including COinS,¹¹ unAPI,¹² and OAI-ORE.¹³ Obviously, beyond the technical issues, there are issues of policy, attribution, and provenance to also be considered.

Conclusion

Many of the technical and feature decisions for RAMH2 will be guided by the model that is ultimately selected for its ongoing sustainability. While the initial creation of RAMH2 will require a significant short term investment of technical, not to mention editorial, resources, there will also be ongoing long term technical costs. These costs include not just keeping the servers and software up and running but fixing the inevitable bugs that crop up, working out technical issues with collection providers, adapting the system to keep up with changing user expectations and demands, and revising the software to keep up with changing technologies. How complex to make the feature set of RAMH2 may depend in large part on what resources are available for its ongoing support and maintenance. The vision and model for support of RAMH2 will no doubt continue to be defined through the ongoing shared discussion amongst its creators and potential users.

¹¹ <http://ocoins.info/>

¹² <http://unapi.info/>

¹³ <http://www.openarchives.org/ore/>