A Specialized Open Archives Initiative Harvester for Sheet Music: A Project Report and Examination of Issues

Stephen Davison
University of California, Los Angeles
Music Library
1102 Schoenberg Music Building
Los Angeles, CA 90095-1490
sdavison@library.ucla.edu

Cynthia Requardt
The Johns Hopkins University
Special Collections, Eisenhower Library
3400 North Charles Street
Baltimore, MD 21218
Cynthia.Requardt@jhu.edu

Kristine Brancolini
Indiana University
Digital Library Program
1320 East 10th Street, Rm. E170
Bloomington, IN 47405
brancoli@indiana.edu

Abstract

The Open Archives Initiative (OAI) Sheet Music Project is a consortium of institutions building OAI-compliant data providers, a metadata harvester, and a web-based service provider for digital sheet music collections. The project aims to test the viability of the OAI standard for providing access to sheet music collections on the web, and to build a permanent and increasingly participatory service for the discovery of digital sheet music. The service provider design has been informed by detailed usability testing, and by limitations imposed by the variations in metadata harvested from the different participating collections. Advanced services in addition to basic searching and browsing have been developed, including the ability to save and share subsets across participating collections. Harvesting and searching strategies for overcoming metadata limitations are being developed. The consortium is seeking additional participants with digital sheet music collections, and is exploring the possibility of incorporating scores and audio into the project.

Digital sheet music collections were among the earliest substantial music collections to appear on the web. Most significantly, digital sheet music collections have been mounted by the Library of Congress, Johns Hopkins University, and Duke University. There are many other important collections, resulting in a rich distributed research resource for music.

Sheet music collections have become the focus of digitization projects for a number of reasons relating to the publication format, its component parts, and the resulting problems in providing access, either through traditional cataloging systems, or by other means.

A piece of sheet music generally consists of a number of different components, brought together for a specific publication. These include the music itself, usually consisting of between two and eight pages; a cover page, that often includes graphic artwork and/or a photographic reproduction; and advertisements, either for additional sheet music from the same publisher, or from other vendors.

The most common sheet music genre is popular song, and the text of these songs comprises another important element of the publication. It seems likely that providing access to the cover art of sheet music has been the prime motivation for many sheet music projects. The graphic artwork found on published sheet music is often very decorative, and provides information of interest to a wide variety of scholars, including historians of art, cultural historians, sociologists, and so on. Much as sound recordings are today, sheet music of the C19th and early C20th documents taste, attitudes, and societal concerns, across time, and in different geographical locations. Sheet music may also reflect the concerns of specific groups of people (e.g. political publications), or entire nations (e.g. nationalistic publications). Songs that become perennial favorites were often republished with changes in text (e.g. taking out insulting epithets; or perhaps writing completely new words for a favorite tune), new cover graphic arts; or simply changes in the advertising. All of these changes document changing attitudes both musical and non-musical.

Traditional cataloging schemes have had a difficult time capturing the disparate elements that make up a single piece of sheet music, let alone the relationships between repeated publications of a single work. In addition sheet music has usually been published as simply music, with little recorded about the cover art or the text. For example, an AACR2 catalog record for a piece of sheet music would typically record the genre (e.g. "Piano music" or "Songs with piano"); include transcriptions of title, attribution, and publication information; and provide access points to composer and lyricist. Information about the cover art and subject of the song—other than that obvious from the title—typically would be absent.

These difficulties, both the physical structure and the cataloging problem, along with the recognition that sheet music has unique research potential, have meant that sheet music collections have often found a home among institutional "Special Collections," rather than in the music library. These problems, along with interest from a wide variety of users, have encouraged the creation of digital sheet

music collections. These allow for a wide variety of keyword searching, as well as browsing of images and advertisements, which are impossible using traditional means of access. Although the capture of song texts as text rather than graphics has not usually been part of these projects, there has been recognition that the capture of these texts is a very desirable feature (e.g. the Johns Hopkins projects to develop both music and text recognition software).

The metadata describing sheet music collections vary greatly in both detail and structure. For example brief AACR2/MARC records have been used by the Library of Congress for American Memory; Duke University uses Encoded Archival Description for the Historic American Sheet Music collection; and UCLA uses an AACR2-flavored Dublin Core schema for the Digital Archive of Popular American Music. One challenge to establishing efficient access to these distributed collections lies in these different metadata encoding schemes.

The Sheet Music Consortium was established to develop unified access to these sheet music collections using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). There are currently four active participants, all with functioning data providers—Indiana University, Johns Hopkins University, and UCLA—with the Library of Congress involved passively as a data provider. Duke University and Brown University have also been involved in planning and usability testing, and will be joining as data providers in the near future.

The aims of the project are:

• To demonstrate the viability of a specialized OAI harvester

• To develop a specialized service provider that will provide searching and browsing capabilities, along with more advanced services such as the ability to save and share subsets across the various collections

• To establish data mapping guidelines for participating collections

• To establish data creation guidelines for new participants, who either have established digital collections or plan to do so

• To demonstrate a collaborative development model that includes both the current active participants and other potential participants

The project has been guided by a steering group drawn from the three active participants, with additional teams to work on technical development (both data providers and metadata harvester), data mapping, and development of the service provider interface. To date three data providers have been built (Indiana, Johns Hopkins, UCLA, in addition to the already existing LC data provider), along with a service provider. The service provider is available at:

`http://digital.library.ucla.edu/sheetmusic/`

This prototype service was used in a usability study funded by the Mellon Foundation. Users from all five active participants in the project participated in both focus groups and one-on-one interviews designed to assess interest in and need for a sheet music service, and to provide feedback to inform the future design of the interface. The study resulted in a set of recommendations for improvement in the search interface, including the addition of an advanced search screen, additional browsing options, a more consistent and comprehensible layout, the ability to protect, email, and save virtual collections, and the ability to more easily pick records from a results list.

The Consortium has adopted unqualified Dublin Core as the metadata standard for initial phase of the project. While acknowledging the advanced services that a more detailed schema (e.g. qualified Dublin Core, MARC) might provide, we recognize the advantages that more basic requirements provide: lower barriers to participation, and a shorter development timeline.

A variety of metadata issues limit the services that can be provided by the sheet music service provider. For instance, although some subset of the Anglo-American Cataloging Rules, 2nd edition (AACR2) is most often used to guide the type and format of the data collected, there are significant variations in their implementation. For instance, one collection may record a statement of responsibility ("music by George Gershwin; lyrics by Ira Gershwin") but not create a table of names in inverted form ("Gershwin, George"; "Gershwin, Ira"), whereas another collection may do the opposite: create a table of names, but not record an accurate statement of responsibility. The result of this variance in metadata is that retrieval by name is limited to searching, and that browsing of names is difficult, if not impossible, to achieve in the service provider. In addition, some collections may impose authority control on certain data elements (e.g. names, publishers), while others will simply transcribe names as they appear on the published item.

The consortium is discussing the possibility of providing improved access to collections by mapping data from the various native formats to qualified Dublin Core elements. For instance, it may then be possible to distinguish between composers and lyricists in searches, to provide access to various descriptive elements, such as plate and publisher numbers, or to distinguish between different types of dates that may be present in the metadata record. However, although many desired service improvements may theoretically be possible through the implementation of qualified Dublin Core, the metadata available through the harvesting process may not support them. Future service enhancements will be implemented only to the extent that they can be supported by the metadata.

The sheet music consortium is seeking additional participants in the OAI Sheet Music Project from sheet music collections worldwide. The consortium is also exploring the possibility of expanding to scope of the project to include other musical formats, such as musical scores and parts, and sound recordings. Expansion of the project to include other musical formats (e.g. digital audio, encoded music, music notation files, etc.) is also under discussion. Although the sheet music project has demonstrated the viability of the OAI protocol for metadata harvesting in establishing a specialized service, there is a danger in generalizing the service into to areas that may be better served by other means of discovery.

Acknowledgements