Managing Metadata

David Datta
All Media Guide
301 E. Liberty
Ann Arbor, MI 48108
davdat@allmusic.com

ABSTRACT

The All Media Guide (AMG) is a technology company that maintains the world's largest database of metadata relating to the entertainment industries. This document describes some of the goals of AMG, the issues uncovered during the evolution of our databases, and discusses some of the implementations we have chosen.

1. INTRODUCTION

The All Media Guide began in 1991 as a hobbyist project to produce the book "The All Music Guide." The book became the seed for a database that has now grown into the world's largest published collection of information about music, movies, and games.

AMG's mission has expanded over the years. As a book, the mission was to "provide a consumer resource to the best recordings." As a database, the goal was to "provide complete e-commerce enabling entertainment databases." Now, as a technology company, AMG's mission is to "Provide the industry standard for entertainment content management through the development of product information databases, e-commerce enabling tools and proprietary content technologies." As the goals of the project have evolved, the data structures, data elements, and processing methods needed to support them have changed to meet new demands.

AMG's products are licensed by business customers who utilize the databases in a wide variety of applications that were never envisioned when the first All Music Guide book was created. E-commerce and content websites, encoding companies, consumer electronics devices, media technology companies, and in-store kiosks all make use of AMG metadata.

In addition to licensing database content, AMG hosts a network of popular websites which serve as a showcase for AMG data and technology. The All Music Guide website (www.allmusic.com) the All Movie Guide (www.allmovie.com), and All Game Guide (www.allgame.com) are invaluable resources for industry professionals. The All Music Guide by itself receives over 2 million hits a day.

2. DEFINING THE PURPOSE OF YOUR DATA

The first step in the creation of a database is to define how the data will be used. While there is a natural tendency to rush in and begin collecting album titles, performer credits, and track titles, the time and effort spent planning prior to this step will greatly impact on the sustainability and effect success of the project. Initially, and for that matter at every step of the way, these two statements will be true: (1) You will be able to think of more types of information to collect than you have the resources to support, and (2) No matter how long your "wish list," you can't possibly anticipate everything that will come up as the database matures.

The end purpose of the database defines which elements are collected. Distributors and sales organizations typically keep Artist Name, Album Title, Record Label/Catalogue number, and UPC codes. File encoding and song identification companies typically collect artist names, album titles and track titles. For these companies, the metadata they collect is adequate for their purposes. At AMG we handle the product once and we enter every objective element we can find. This includes extended metadata elements such as track level copyrights, composers, and performers.

Scope of coverage is an important design factor. Are out-of-print titles needed? Is coverage of domestic catalogue enough? Do you have international users? Consider the question of in-print versus out-of-print. The current in-print U.S. music catalogue is about 185,000 albums; the out-of-print universe is somewhere between 700,000 and 900,000 albums.

The increasing reality of modern technology is that there is no such thing as music which is out-of-print. It is expected that all recorded music will become available in digital form. Early coverage of this additional material provides a head start as music is re-issued. Presentation of information about out-of-print music is useful in almost all applications. Limiting searches to in-print album titles means many searches lead to no useful results. Users can quickly reach a dead end; the inclusion of out-of-print titles provides launching points which connect users to other albums and artists available to be purchased.

3. DATA ORGANIZATION

Once data elements and scope have been defined, data organization can begin. Choosing a good structure for musical information requires extensive knowledge of music, the product you have available to collect data from, and the planned uses. Making a wrong choice during the database design stage impacts all aspects of the process; from data entry to publication. One must consider not only database theory and data structure size, but also ease of maintenance, and ease of publication. There are as many approaches to the storage of metadata information about music as there are applications for the data itself. Structures that are suitable for importing and exporting between disparate systems (XML) are usually inefficient when applied to actual data maintenance. AMG evolved its data structures over many years to obtain good balances between data maintenance and data publication.

Consider the following information about the pop and classical musical catalogues.

	Pop	Classical
Albums	544,000	84,000
Tracks/Performances	4,000,000	740,000
Songs/Works	150,000	200,000
Composers	133,000	18,000
Performers	450,000	54,000

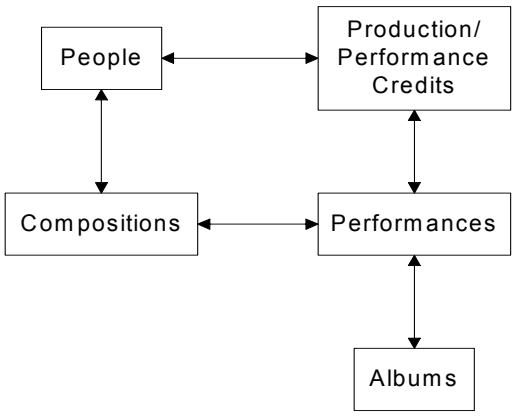

Figure 1: Basic structural overview of music metadata

Typical for a pop album:

1. Album titles are almost always present. When not present, conventions are known. (For example Peter Gabriel's albums are recognized as "Peter Gabriel 1," "Peter Gabriel 2," and "Peter Gabriel 3.")
2. Composers often perform on the recording.
3. Tracks, both CD and LP, contain a single song.
4. Work/Song titles are individually distinctive and frequently unique.
5. Works/Songs are performed and released in their entirety.

Typical for a classical album:

1. Album titles are frequently not present. Of the 84,000 albums, less than 60,000 have actual titles and are not simply a list of works on the album.
2. Composers seldom perform on the recording.
3. CD tracks occur whenever there is a logical break in a work being performed. Tracks on LPs are often absent with the exception being side changes.
4. Work/Song titles are descriptively repetitive -- for example: sonata, fugue, prelude, etc..
5. Works/Songs are seldom performed in their entirety.

While both pop and classical share the same basic blocks of information, designing a unified structure is not a trivial task and may not be appropriate approach. In a classical database, most compositions are well documented; the structure to support the amount of information available for classical works would be empty for most pop songs. Similarly, an efficient structure for pop tracks would be extremely repetitive when used on a classical album where a single work may be split over ten tracks with no information given as to how the work is divided. In pop, the main entry points to the collection are via albums and names. It is usually the case that a person will search for an album title, a performing artist, or song title. In classical, a user will search for a composer, a composition title or a performer. Searching for an album title is frequently a futile process.

To accommodate the different needs of the users of these databases, AMG has implemented separate structures and product handling systems for pop and classical. Many artists are logically classified as both pop and classical. Because of this AMG processes crossover product twice, once for each data set.

A single album may have two reviews, one with a classical orientation, and one with a pop orientation. Similarly, a person can have two biographies. Paul McCartney is a minor footnote within the classical world. Luciano Pavarotti's videos are seldom shown on MTV.

4. INFORMATION AVAILABILITY

The lifecycle of a music product begins months in advance of release. Beginning with the initial announcement and continuing later as the product nears its release date, artists, record labels and promotional agencies make information available. At this stage cover images, track listings, and some performer credits may be found at business-to-business web sites. In some cases, promotional copies of the album are sent out for review. Eventually, the product is released. AMG considers the release version of the product to be the definitive source of metadata information. The database will only be as good as the sources from which the information is gathered. You must know how to efficiently replace and add data as the lifecycle of an announced product evolves.

5. LINKING AND INTEGRATION

After identifying data elements and sources, there is still the logistical process of getting information into the database and keeping errors out. While it is a trivial task to collect metadata from products, data inconsistencies are unavoidable. Normalizing and linking data is the most complicated part of the creation process.

The same data can come from a variety of sources; each of these sources will have their own idiosyncrasies. For example, which Eurythmics song title is correct? "The King and Queen of America" or "The King & Queen of America" Depending on the source, either form of this title may appear. While one could argue the product is the final arbiter, different releases of the same product may contradict each other. In some cases, the product itself may display both. Names have the additional problem of identification. Which "John Smith" are you looking for? AMG uses programmatic tools to generate context-sensitive suggestions that expert editors then confirm, revise, or reject.

Even the best integration and linking systems will have cases that slip through. Automated and manual feedback analysis processes are needed to help detect data problems.

6. ADDING CREATIVE CONTENT TO METADATA

One of the most interesting facts we have discovered is that only 4% of the AMG website hits are the results of searches. The websites are an exploratory tool; users become immersed in the information and follow links from person to person to album and beyond. This type of browsing experience is only possible because of the creation of relational content.

Our editorial staff and freelancers, often in consultation with industry professionals, assign descriptive content such as genres, styles, keywords, and moods. Analysis of this descriptive data, combined with editorially created connections, recommendations, and reviews provides AMG with the backbone to fulfill its goal of enabling people in their entertainment choices.

7. MAKING IT WORK; BREAKING DOWN THE SKILL SETS

AMG has five functional types of staff: Strategic, Editorial, Data Entry, Data Integration, and Support.

The Strategic group defines the purpose of the database. This group decides what data AMG will keep, how it will be used, and coordinates efforts between the groups to ensure proper implementation.

The job of the Editorial group, the experts, is to organize and categorize the data. They define the genres and styles, create standards for writing, and determine how to apply rating systems. They also resolve the inevitable ambiguities that arise when working with large amounts of data.

The Data Entry staff transcribes information directly from the product into a format usable by the integration team.

The Integration and Data Processing team integrates the entered data into the main databases. They are the experts in manipulating, linking, and normalizing data. This team is also responsible for maintaining database integrity and data cleaning.

The Support staff consists of everyone else. This group includes programmers, customer support, and all other non-data related functions.

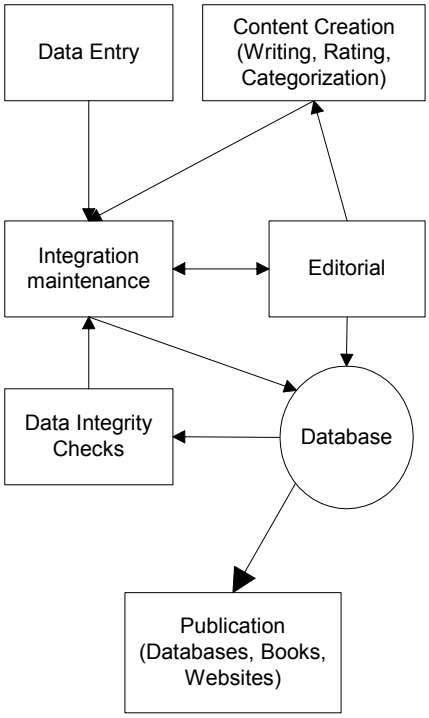

Figure 2: Generalized data flow diagram

8. LEGAL ISSUES

The notes below are areas to consider when using elements that are typically used along with metadata. Rights management is a complicated topic that requires expertise, therefore it is important that you consult with an intellectual properties lawyer to verify the legality of your publications.

Factual metadata data

Objective facts are not covered under copyright. Pink Floyd is the credited performer on the album titled "The Dark Side Of The Moon" is an objective fact. It is generally accepted that objective factual information can be published at will.

Album cover artwork

Rights for cover images usually are held by the record label, the artist, the creator of the image, or subsequent rights holder. At this time in the United States, it is generally accepted that album artwork may be used in conjunction with the sale of the album.

Sound Samples

Royalties must be paid each time a sample is played. The rights holder of the performance of the song owns sound samples. Record labels, the most common owners, are extremely protective and have not hesitated to sue for infringement.

Lyrics

The Harry Fox Agency is one of the many companies that acts as a clearing house for the publication of music in the United States. They are very protective of the rights of their members and will not hesitate to sue for infringement. In the mid 1990's they threatened to sue all of the free access on-line databases of lyrics and the sites were shut down.

Reviews and Biographies

The copyright laws dealing with text depend on the contractual agreements between the author, the publisher, and the country in which the text was created. Re-publication always requires permission and one must arrange proper clearances.

9. TAKE THESE CONCEPTS HOME WITH YOU

This document has only introduced a few of the more important issues you will encounter when developing your metadata database. Each of the processes described here could be expanded into its own paper. While your operations may not need to be as large or complex as those of All Media Guide, to be successful, you will need to implement many of these systems. What I really want to emphasize is whatever choices you make in database design, the most important choices will be those you make in defining purpose and planning flexible ways to achieve it.

10. ACKNOWLEDGMENTS

Thanks to Richard Gilliam for his assistance in the preparation of this paper.