Two approaches to distributing music on the Net—the hard way and the easy way—compete and complement each other in the quest for meaningful musical metadata.

CONTENT MANAGEMENT *for Electronic Music Distribution*

By François Pachet

Although the digital representation of audio data was developed many years ago, the capability to store and manipulate such representation with good sound quality is a much more recent development. The emergence of efficient audio compression technologies, such as MP3, has brought about the possibility of easily transmitting and broadcasting music data across networks. For example, Napster had 80 million registered users at its peak [5]; on a system such as KaZaA users routinely access approximately 500 million items. Another consequence of efficient audio digitalization is that the granularity of music distribution has shifted from music albums to music titles. Electronic music distribution (EMD) usually refers to the technical issues involved in transporting such music titles across networks, copy protection, and copyright management. However, there is much more to EMD than telecommunication and protection. The major challenge and promise of

PHOTOGRAPHS BY MICHAEL KLOSE

EMD is to allow the shift from a mass-market approach of music to a personalized distribution approach. Providing this digital link between music and people, however, remains more dream than reality for several reasons.

The first reason involves the size of music collections. Estimations based on major label catalogs yield a total of 10 million titles (restricting the classification to published, occidental, popular music). The number of Internet users was approximately 600 million in 2002, according to a Nua survey—see www.nua.ie. Traditional mass-market distribution involves distributing only a small fraction of music titles to a large number of people; the fraction of so-called "active" titles in major label catalogs is about 1%. The EMD dream is primarily about proposing personalized distribution schemes that make more titles available to more people.

Second, EMD touches upon our intimate relationship with music. Browsing music is different from browsing a traditional digital library; we don't want to simply "access" or "find" music, as we would for bibliographical references. Users do not always know how to specify what they are seeking (the language mismatch problem created by ontologies users do not understand, see [3]); nor do they always even know what they are looking for. Therefore, the design of an EMD system requires that we know more about what users want to do with music.

However, EMD systems abound in a large number of incarnations, including Digital Audio Broadcast, CD-on-demand, music downloading, music streaming, Internet radio, music browsers, and music servers, including peer-to-peer communication systems. Technically, these systems differ mainly in the nature of the inputs and outputs they connect together. But they are still far from achieving the EMD dream, mainly because they lack efficient content management tools. The most important issues underlying music content management, from identification to content-based music search and retrieval, are described here. Music file management the "hard way" (requiring brute force and sophisticated signal-processing technology, which provides objective information but is costly to develop) is compared to an "easy way" based on statistical analysis of superficial data (easier to implement but less reliable).

Content Management: The Core Technology of EMD

Music title identification. How can a system identify music titles? In the simplest case, identification information is included in the music data itself, for instance through ID tags in MPEG files. The ISO International Standard Recording Code (ISRC) was developed to identify audio and audiovisual recordings; ISRC is a unique identifier of each recording that makes up an album. Unfortunately, it is not followed by all music production companies and is infrequently used by unofficial music sources, so the majority of existing digital music files do not contain any built-in identification.

Worse, music data may not include any external reference information; this is the case with analog radio, for instance. In this case, identification can be done the hard way, by so-called audio fingerprinting. Fingerprinting analyzes the signal, typically a portion of the music title, and builds a short but unique signature of this signal, usually based on a characterization of the evolution of spectral behavior, which is robust to noise and distortion. This signature is then matched against a precomputed database of signatures [1]. Copyright management companies such as Broadcast Data Systems (U.S.) and MediaControl (Germany), use this technique to infer radio playlists.

Conversely, the easy way exploits external information about the titles when available. External information can be as simple as file names, with the difficulty that names are not standardized; an artist such as the Beatles, may be cataloged as "The Beatles," "Beatles, The," and many other combinations. Other, more reliable external information can be exploited; for example, the now-defunct Emarker system exploited the geographical and temporal location of a user listening to a radio and requesting a song to query a database containing all radio station programs. This approach is lightweight and scales up to virtually any number of titles. Of course, it works only for titles played on licensed broadcast radio stations. Interestingly, Emarker relied on radio playlists

MUSIC DATA *may not include any external reference information; this is the case with analog radio, for instance. In this case, identification can be done the hard way, by so-called audio fingerprinting.*

generated by Broadcast Data Systems that were computed the hard way (demonstrating that the two approaches are by no means contradictory).

Music genre. The most prominent information about a music title is probably its genre. Music distributors and retailers have long used genre classifications for organizing catalogs. However, the study of these classifications [2] shows there is hardly any convergence; terms are not consensual ("Easy Listening" in one classification is called "Variety" in another), and worse, taxonomy structures do not match; "Rock" denotes different songs in different classifications. Additionally, music classifications have been designed mostly for music albums, and are not directly usable for music titles; a Pop-Rock album by the Beatles may contain titles in many different genres: from Country-Folk to Symphonic Easy Listening. However, ill-defined as it is, genre is the primary descriptor used for describing music, so content-based music systems must know about genre.

The easy way to extract genre information is to ask human experts. Human classifications have the advantage of containing expert knowledge and of being relatively consistent. They are, however, difficult to update and not always readable because terms, even when coined by professionals, are rarely consensual (what does "Zouk-Love" really mean?). Classifications can also be built automatically by an analysis of usage, and proposals to create new genre classifications have been made based on collaborative systems [3]. The hard way attempts to match objective criteria of acoustic signals to genre, as was done by many researchers (see [2]). However, the poor results obtained by the hard way are mostly due to the intrinsically cultural and nonobjective nature of genre, so there is little hope of progress with this approach.

Music features. Besides genre, music titles may be described by many other features. The MPEG-7 standard attempts to provide a basis for representing all common features for audiovisual documents. Music metadata in MPEG-7 refers to low-level, objective information that can be extracted automatically [9] such as energy level, or spectral information. Extraction of higher-level features is a primary issue in MPEG-7 and one can distinguish, here too, an easy way and a hard way.

The easy way involves asking users to rate songs according to given features. This approach is used by MoodLogic; its metadatabase contains approximately 1.5 billion user ratings for approximately one million music titles. Statistical analysis methods are used to filter out noisy data. This approach does not require any signal processing and is combined with a proactive collaborative strategy in which users must rate songs to benefit from the entire metadatabase.

The hard way involves extracting high-level music features from the acoustic signal. The features that can typically be extracted this way include fundamental frequency [6], beat extraction and tempo induction [11], and segmentation. Some of these techniques are mature enough to be exploited by, for instance, the MuscleFish tool [12]. This fascinating field of musical feature extraction is only beginning, and still lacks a proper and systematic rationale. The Cuidado European project (www.ircam.fr/cuidado) aims to develop a systematic approach to high-level musical feature extraction in the context of MPEG-7. The Cuidado Music Browser will be the first large-scale music browser to propose automatically extracting high-level musical features, including the discrimination between instrumentals and songs, the discrimination between studio and live versions of titles, and the presence of long instrumental choruses. The Cuidado Music Browser offers a unique opportunity to compare and assess both quality and relevance of music features extracted from human ratings and from the signal.

Music similarity. An important task for music content management systems is to show users similarities between music titles, which can be of many types. At the feature level, one may consider that Jazz saxophone titles are all similar. Similarity can yet also occur at a larger level and concern songs in their entirety; one may consider Beatles titles as similar to titles from the Beach Boys because they were recorded during the same period, or one may consider all the titles by a given artist as similar.

Objective types of similarity can be computed easily from features, using the hard way or the easy way. Culturally dependent similarity may not be extracted the hard way, from the acoustic signal, because the cultural information is simply not in the signal and can only be extracted the easy way, here by data mining techniques. Collaborative filtering (CF) in particular is a technique to infer patterns in taste within communities of users. This technique, originally introduced by Pattie Maes [10], was used extensively for music recommendation. Today, most Internet music retailers (Amazon, CDNow, MyLaunch) use CF to provide music recommendations to their customers. The core idea of CF is to make recommendations based on similarities in user profiles. Repeated logs of each user to the system progressively build a profile of a particular user's taste in music. The profile can be as simple as the titles selected or the list of the CDs purchased by the user.

Although technical evaluations of musical collaborative filtering have been performed (for example, by the Jaboom team [3]), the nature of the music simi-

CONVERSELY, THE *easy way exploits external information about the titles when available. External information can be as simple as file names, with the difficulty that names are not standardized.*

larity exhibited by CF is difficult to characterize. CF-based similarity typically comes from culturally grounded affinities. For instance, most of the people who like the Beatles probably also like the Beach Boys and, generally speaking, the Pop music of the 1960s. The interesting property of CF is that these relations will be computed easily (the easy way) without human intervention and without complex signal processing. But the technique has drawbacks. First, the similarities are not complete and address only titles that were actually rated by many users. Second, there are limitations to CF in the nature of the recommendations. Only strong patterns in communities are actually propagated, so eclectic profiles do not gain much from CF, because they are not statistically close to a large enough population of profiles: the more specialized the profile; the less interesting strong patterns in the community will be for the corresponding user.

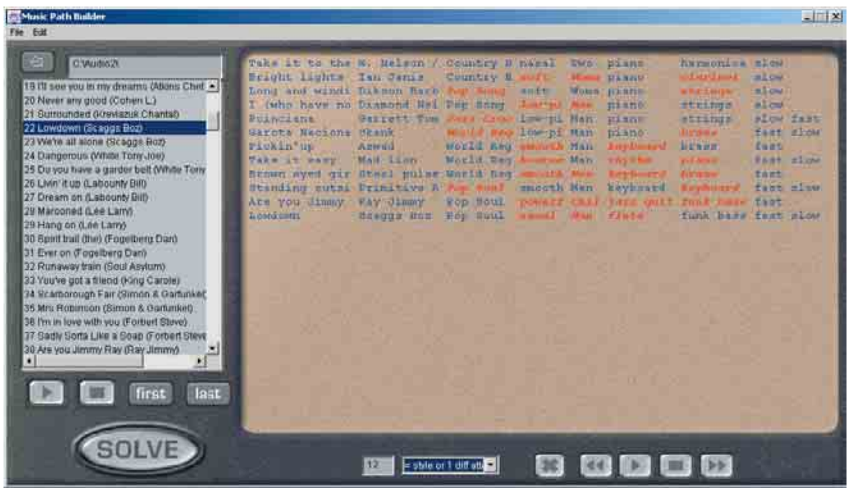

Figure 1. The PathBuilder system creates music compilations from starting and ending titles by computing a path using musical metadata. Metadata is shown in columns: blue means continuity, red means discontinuity.

Collaborative filtering is a particular case of data mining technique, focusing primarily on user profiles. Other data mining techniques can be used to infer similarities, such as co-occurrence analysis. This technique involves checking when two or more titles appear together in different contexts, such as Web pages and radio programs, and building a distance function based on these co-occurrences. Co-occurrence can then be used to infer automatically clusters of related titles, as well as genre taxonomies, as shown in our studies [8]. The taxonomy has the advantage of being done entirely automatically and is easy to update.

It is clear that the way these different similarities are extracted deeply impacts their nature. However, little is known regarding their respective advantages and drawbacks. Comparing these different similarities is a major issue that is only starting to be addressed.

From query systems to playlist generation. Most existing EMD systems follow a traditional query-answer scheme: the system provides a set of titles, possibly sorted corresponding to how well they satisfy the query. However, music titles are usually not listened to individually but in sequence, for instance, from a radio program, a concert, or a CD. These sequences usually have some global properties that make them consistent or interesting, such as continuity or thematic consistency. We proposed in [7] to address the

Figure 2. PersonalRadio. When the exploratory slider is moved on the right, music programs contain titles farther away from the request. Here, the genre-based request is "Jazz guitar." Titles in red are more distant from this request than titles in blue.

music retrieval problem from the sequence viewpoint, showing that this approach allows users to access music in a much simpler and intuitive way. Additionally, this approach avoids the language-mismatch problem inherent in metadata access, as metadata is used only internally by the system to build the sequence and not explicitly by the user. Figure 1 shows PathBuilder, a prototype developed at Sony's Computer Science Laboratory (CSL) in Paris that builds a music path between two music titles selected by the user. The path is as continuous as possible, and continuity is defined by a weighted sum of similarity measures on a set of music features (such as genre, voice type and tempo).

User interfaces. Various interfaces have been proposed to access music online, from straightforward feature-based search systems (MongoMusic) to innovative graphical representations of playlists. For instance, Gigabeat displays music titles in spirals to reflect similarity relations between titles. The gravitational models of SmartTuner and MoodLogic represent titles as small animated balls moving on the screen to or from "attractors" representing the descriptors selected by the user. However gracious, these interfaces impose a fixed interaction model and assume a constant behavior of users concerning music access: either nonexplorative—databases in which you get exactly what you query—or very exploratory.

PersonalRadio, a prototype for set-top-box music services developed at CSL, addresses explorativeness explicitly. Figure 2 shows an interface of PersonalRadio, with a slider ranging between two extreme values (conservative to exploratory). Depending on the position of the slider, the music selection proposed is conservative, exploratory, or anywhere on the the continuum between the two extremes.

User studies of PersonalRadio reveal interesting behavior. While some users react negatively toward exploration in the beginning of their interaction, in the long run they tend to systematically shift to exploratory modes. This can be explained by the fact that most users quickly exhaust their capacity in issuing explicit queries; it is only once well-known artists and hits are queried, in a nonexploratory mode, that the desire for novelty occurs, and that such a feature appears to be useful. These experiments, preliminary as they are, stress the importance of designing user interfaces that account for the fuzzy nature of human behavior when confronted with large music catalogs.

Conclusion

Music content management technologies are a key ingredient for EMD. These techniques, including title identification and feature and similarity extraction are necessary to help users navigate in large music catalogs and eventually make possible one-to-one music distribution. Whether we follow the hard or easy way—through brute force or through statistical analysis of superficial data—there is still a long way to go to achieve the EMD dream, in particular concerning the nature of the metadata and similarity relations extracted, as well as our still largely misunderstood relation to music exploration.

Finally, new problems will arise when these technologies are mature, for instance, concerning the legal status of metadata: Can an artist prevent someone from creating and distributing metadata about his music? Many institutions now favor open-source and patent-clear approaches to multimedia management (see the open-source streaming techniques developed by the Xiphophorus and Icecast projects). In this context, should metadata also be free? C

References

1. Allamanche, E., Herre, J., Helmuth, O., Frba, B., Kasten, T., and Cremer, M. Content-based identification of audio material using MPEG-7 low-level description. In *Proceedings of the International Symposium of Music Information Retrieval* (2001).
2. Aucouturier, J. and Pachet, F. Representing musical genre: A state of the art. *Journal of New Music Research 32*, 1 (Jan. 2003).
3. Belkin, N. Helping people find what they don't know. *Commun. ACM 43*, 8 (Aug. 2000), 58–61.
4. Cohen, W. and Fan, W. Web-collaborative filtering: Recommending music by crawling the Web. In *Proceedings of the 9th International World Wide Web Conference*, Amsterdam (2000).
5. Lam, C. and Tan, B. The Internet is changing the music industry. *Commun. ACM 44*, 8 (Aug. 2001).
6. Lepain, P. Polyphonic pitch extraction from musical signals. *Journal of New Music Research 28*, 4 (Apr. 1999).
7. Pachet, F., Roy, P., and Cazaly. D. A combinatorial approach to content-based music selection. *IEEE Multimedia* (Mar. 2000).
8. Pachet, F., Westermann, G., and Laigre, D. Data mining for electronic music distribution. In *Proceedings of WedelMusic* (Firenze, 2001).
9. Philippe, P. Low-level musical descriptors for MPEG-7. *Signal Processing: Image Communication 16* (2000).
10. Shardanand, U. and Maes, P. Social information filtering: Algorithms for automating "word of mouth." In *Proceedings of the 1995 ACM Conference on Human Factors in Computing Systems*.
11. Scheirer, E.D. Tempo and beat analysis of acoustic signals. *JASA 103*, 1 (1998).
12. Wold, E., Keislar, T., and Wheaton, J. Content-based classification, search, and retrieval of audio. *IEEE Multimedia 3*, 3 (1996).

François Pachet (pachet@csl.sony.fr) is the head of the music team at Sony CSL-Paris, France.