Classical Music – A CD Ripping Challenge

By Jeff Tedesco, President, ReadyToPlay
January, 2011

ReadyToPlay rips thousands of classical CDs every month. Having been in business for over eight years, we've seen many music collections and we've gotten smarter about ripping cds 'correctly'. Correctly, by RTP standards, means that the metadata matches the CDs actual title, artist name, genre etc. If you have tried this yourself, you know that the task of making CD data 'correct' is an on-going, difficult process. RTP has successfully digitized the personal music collections for Michael Tilson-Thomas, the Julliard School and hundreds of other classical clients.

So, how does a CD rip work and get data about itself? Contrary to common belief, a CD contains no information about itself. It is simply a disc with audio files on it. An Itzhak Perlman CD with 12 tracks will simply have 12 audio tracks of differing lengths. When a CD is inserted into a software program like iTunes, the CD provides only track length information. iTunes then touches one of several databases available on the internet to read the CD and 'lookup' it up. How does it look up the CD? By track lengths! Meaning, 12 tracks = 13 datapoints. One is total length of the CD (the sum of all the tracks) and each track's specific length in duration, plus the order of those times. The CD spins up, says track 1 is 3:48 long, track 2 is 5:12 long, etc. and all 13 tracks are in this order. The database returns its answer by saying the CD in this case is Elton John, Goodbye Yellow Brick Road.

There are five or six major databases in the world. iTunes uses a database called Gracenote which is free. It has broad coverage of CDs across the world which is great but it relies on 'user submitted data' for the most part. RTP doesn't use this database as it is typically very inconsistent and would not work for us as our standards for consistency are higher. For instance, an individual submitting data about a CD can put in whatever they want to identify the artist, genre, album title, track titles, etc. For classical, the artist could be 'Bach', 'J.S. Bach', 'Bach, JS' or 'Glenn Gould', 'Gould, Glenn' or the orchestra. Many consider artists to be the composer (although RTP does not) and some will input the composer with their date of birth! Album titles can be even more inconsistent. RTP feels that user submitted data leads to inconsistency across a collection of a hundred or more CDs. A perfect example might be when one CD is Glenn Gould, the other Mozart, the other WA Mozart, another could be labelled Wolfgang Mozart – all inconsistent when viewed as a collection, and frustrating for the user to find what they are looking for. In addition, the number of artists swells to hundreds when the same artist has different iterations of the same name! These difficulties also exist in other genres such as Indian classical, Hebrew and Asian.

ReadyToPlay uses four databases, each with professionally vetted data on the first pass. We also hand groom the data fields so composers are always composers and artists are always performers or the conductors. The spellings are correct and consistent. We also identify more genres for classical music. Instead of just 'classical' we identify Chamber Music, Choral, Symphonic, Concerto, Opera and others. By cross referencing between databases, we get much higher quality data, and the best confidence available that our data is correct and consistent. Even with all that technology and superior databases, issues with classical CD data still exist. The following are the top four reasons:

1. No standard exists for classical data or what goes in which field.
2. There aren't enough fields in which to put all the data people want to see
3. Classical CDs are constantly repackaged as box sets from original works or vice-versa and they come up incorrectly.
4. CDs come up with partial data in the wrong places or no data at all. This is especially true for foreign (non-US) titles.

No standard for classical metadata:

I've talked with hundreds of classical enthusiasts. Most think the CD has all the data about itself – titles, performers, track names, genres etc. We already know this isn't true. But once you do get data for the 'artist' of a classical CD, then you must decide on just ONE name choice. Think about all of the different choices you have for that one artist name:

– Is it the lead performer on a two artist CD Annie Sophie Mutter or Murray Perahia? What if you have two performers as in an Opera?
– Is it the conductor like Herbert von Karajan?
- Is it the orchestra like Academy of St. Martin-in-the-Fields?
- Is it the composer like Wolfgang Amadeus Mozart?

When identifying genres, it gets even harder! For example: Mozart's Piano Concertos by Mitsuko Uchida. Is this a Keyboard work or Concerto work? This is compounded by the fact that both genres, and artist preferences vary by individual.

What about the composer field? Is it first name, last name? Last name only? Or last name, first name? Because there is no standard and everyone thinks of it differently, it makes it almost impossible for RTP to meet a customer's expectations without first talking with each classical client before we proceed.

There aren't enough fields to put all the data people want to see:

We at RTP would love to have a data schema that allows us to populate fields like lead performer, lead cellist, singer 1, singer 2, orchestra, conductor etc. But digital music gets ONE artist, ONE genre, ONE album title. Artists are obviously the hardest because most listeners of music want to know the performers but can only see one artist at a time.

Classical CDs are repackaged:

So many great works by world renowned artists like Leonard Bernstein, Rachmaninoff and others have released single albums during their career. However the labels have re-released the same CD as part of a box set – the Bernstein Box Set, the Complete Beethoven Symphonies etc. What makes this difficult is that the CD data comes back with data from the original release or "CD06" of the box set. Both are correct but only one is really correct when it comes to a customer's collection!

Incomplete or wrong data:

We may get data from a user submitted database which means that in the artist field we will get 'Mozart' or 'WA Mozart' when what we want is the conductor or lead performer. In this case RTP *manually* fixes the incorrect data for the artist, the album artist and album title fields. This is a great deal of work when working on a large collection. This is the type of care and attention to detail that you will only get from ReadyToPlay.

Summary:

By now it is clear, anyone can 'rip a CD' but not everyone can 'rip a CD correctly' which is more difficult. RTP uses automation to rip the CD to the fidelity format you want (the easy part). The hard part lies with obtaining quality, accurate metadata. Only ReadyToPlay hand-manages metadata to meet our standards. We work hard to ensure that we have the right data and that it looks consistent across the collection, whether it is 100 or 2000 CDs.