Eerola, T. (2014). Database studies. In Thompson, W. F. (Ed.), Music in the Social and Behavioral Sciences: An Encyclopedia (pp. 316-319). London, UK: SAGE.

Database studies

Database studies in music research, also known as corpus-based musicology, have their roots in the advent of digital encoding of music during the seventies and are closely influenced by corpus-based work undertaken in linguistics. In music research, database studies are an example of a data-rich approach advocated by empirical musicology (Clarke and Cook, 2004).

The central idea of database studies is that the research question is formulated in such a way that the answers can be obtained by analysing large quantities of materials coded in a systematic way into a database. Databases usually consist of music coded in a particular way and the relevant metadata, but can also consist of musical behaviour or historical information about the sources.

The central benefits of database studies are transparency, impartiality and generalisibility in comparison to the conventional music scholarship that usually operates using a few handpicked case examples. In database studies, transparency is achieved in the way databases are structured and through unambiguous definition of the fields and objects of databases. This allows the analyst to accurately define the objects of music that can be subjected to analysis. In principle, anyone can replicate the results of a database study using an identical set of retrieval commands and the database. In comparison with traditional music analysis, which values intuition, stylistic knowledge, and possibly subtle judgments about the aesthetic value of the analysis content, database studies can be considered to be more impartial in the analysis process. This is not to say that intuition and stylistic knowledge are not important in database studies but these need to be explicitly coded in the queries made for the database. Generalisibility of database studies refers to the possibility of identifying common factors in the music of a given collection due to the possibility of processing much larger samples of music than is possible in a conventional analysis.

Databases types and tools

In music research, the types of databases can be broadly divided into (i) visual, (ii) audio, (iii) editorial, and (iv) crowd-sourced.

(i) Visual data represents conventional notation in its many forms (score, tabulature) as well as the most computer friendly symbolic representations of music (MIDI, kern, MuseData). These latter representations are often referred to as symbolically-encoded music, where note symbols and other information often conveyed by notation have been encoded in a machine-readable format. The benefit of this encoding is that it captures the common notational aspects of music, is extremely compact representation, and can be converted into notation, tabulature or audible form using music software. The downside is that the notational schemes may not be adequate to capture the important aspects of music from many traditions.

(ii) Audio representation encodes the complex waveform of a music performance using uncompressed (WAV or AIFF) or compressed (MPEG or MP3) formats. Whilst the audio captures all nuances of the actual performance, extracting meaningful analytic elements from audio is often problematic. Various low-level features such as the spectral centroid can be reliably extracted but identifying the

exact pitches on a polyphonic material is a task not reliably undertaken by computational algorithms.

(iii) In editorial databases, the distinct edition and the possible alternative transcriptions and sources, as well as the authors responsible for the editions and subsequent marking on the editions, are encoded. These fields are typically linked to the electronic versions of the manuscripts and their derivatives such as sketches, drafts, arrangements and letters. Central such databases are *DIAMM* (the *Digital Image Archive of Medieval Music*), *ECOL* (*Electronic Corpus of Lute Music*), *OCVE* (*Online Chopin Variorum Edition*) and *CFEO* (*Chopin's First Editions Online*). Combination of high-resolution images of the original manuscripts with the editorial history and links to other related manuscripts has altered the way music history scholarship is carried out since such databases liberate researchers from geographical barriers and provide all scholars with access to primary sources.

(iv) Crowd-sourced databases are the latest addition to music scholars' arsenal and contain information scavenged from the central websites describing music (for example, *All Music Guide, Echonest Discogs*) or the tags and user data of online social media services such as *MusicBraintz* or *Last.fm*. These sites have accumulated the opinions and statements of millions of music consumers, which can be turned into knowledge using semantic content analysis or other frequency-based analysis techniques.

For all databases, one needs to describe the database fields and how they are represented and relate to each other. This is called metadata and there are established formats such as the *Dublin Core Metadata Initiative* for these. Music-specific metadata structures have also been offered (*Music Ontology Initiative* and *MusicXML*) and currently several metadata standards are in use due to the different aims and scopes of the databases. Common to all metadata formats is that they propose a consistent way to label titles, subjects, descriptions, publishers, formats and a myriad of other necessary pieces of information for databases. Agreed metadata formats are necessary to construct databases that are compatible and consistent with each other.

To pose a specific question to a music database, one needs to utilise programming in analysis. In the most simple case this may just be a specialised query that fetches certain types or instances of music from the database and compares them with another set of instances. There is a common database query structure (SQL or *Structured Query Language*) and also dedicated software for extracting music-specific representations and questions. For example, *Humdrum* (Huron, 2002) is a collection of commands and utilities covering many musical representations and analytic transformations. For posing questions to audio-based databases, several flexible software tools exist (*Marsyas, Sonic Visualiser, MIR toolbox*) with the ability to extract hundreds of acoustic and perceptual features of music.

To interpret the answers to questions from database queries, knowledge of statistics is necessary. The prevalence of a particular type of musical feature of music needs to be set in the context of the sample or contrasted with the prevalence of the type in other collections, which are all instances of statistical hypothesis testing. More advanced statistical operations – such as classification and factor analysis – may be required in exploratory studies.

Searching for similar melodies

One of the central questions in database studies has been how to find music that is thematically similar to another piece of music. This question is intuitive, useful and an

important aspect of music when classifying or attributing the origin of works. Let us take melodic incipits in the *RISM* collection (*Répertoire International des Sources Musicales*) as an example. The database now contains over 700 000 symbolically encoded short excerpts of notated music from the beginnings of manuscripts in various analogue databases (libraries, archives and private collections). Given a user query, the task is to rank-order the incipits in terms of the closest matching melodic similarity to the query. This is a difficult question since the outcome is affected both by the representation of the input and the database and also by the way the matching process is carried out. Approximately two decades of studies into melodic similarity have resulted in robust methods for matching melodies where the perceptual processes involved in music processing (such as transposition invariance, contour and reduction of ornaments) are taken into account. These methods typically utilise traditional Information-Retrieval (IR) matching techniques or more geometric approaches based on musical representations. To evaluate and refine these methods, it is necessary to have ground-truth data. For example, music experts can be asked to rank a large number of possible results from similarity queries of melodic incipits, and these expert rankings forms the ground-truth data that allows to fine-tune the actual similarity calculation.

Exposing musical conventions

An example of a music-analytic database study is David Huron's unveiling of the ramp archetype in dynamics using 435 piano works by 14 composers. By tracing all the dynamic markers in the scores using specialised queries and a carefully chosen sample of repertoire, Huron found evidence that the crescendos are not only more frequent than diminuendos, but they are followed by abrupt diminuendos and are generally longer than diminuendos. These observations were interpreted as being consistent with the way of optimising auditory attention but could easily also describe the stylistic conventions of the romantic era. Huron has pioneered the frequency descriptions of symbolic collections (interval sizes, directions and shapes) in many of the studies detailed in his compilation book published in 2006.

Attributing authorship

Database studies have also been used to attribute authorship to anonymous compositions. For example, Dumitrescu (2010) examined *Missa Naray je jamais*, a work of unknown authorship, though often presumed to be Jacob Obrecht's composition. Dumitrescu looked at low-level musical features such as rhythmic density and dissonance usage at different mensural positions and compared these to the well-attributed principal works of Obrecht. The results suggested that the composition is consistently outside the range of typical values for these features in Obrecht's repertoire, casting a shadow of doubt on past authorship claims that this work was Obrecht's. Attributions of authorship are also at the crux of court cases dealing with plagiarism in popular music. In the spirit of database studies, other scholars have looked at songs alleged of copyright violations using similar approach of comparing them to large popular music database.

Expert descriptions of folk songs around the globe

In folk music research, database studies have a long history because of the large-scale folk music archives accumulated particularly in the national research archives of many countries. Symbolic collections of folk song materials from Europe (see Huron, 2006) have been used to describe melodic formulae and various styles have been

classified according to musical properties. Not all database studies have relied on symbolic encoding of pitches and, for instance, Alan Lomax's *Cantometrics* project (1968), set out to describe all the music of the world using a system of describing music with a set of features especially related to voice qualities. These features were rated by a team of experts. This database, consisting of over 4000 excerpts from over 200 cultures, was used to establish cultural taxonomies, some of which met with resistance due to strong claims made about the links between vocal styles and social norms. Nevertheless, the approach itself is still being used and improved today.

Music information retrieval and database research

In the field of Music Information Retrieval (MIR), the use of databases is central. The goal of the discipline is to resolve questions related to efficient retrieval of meaningful musical content by computational means that all utilise the database approach in one way or another. New algorithms are typically evaluated with a large database of examples already annotated for the given task. With the maturing of online music collections, metadata mined from the web and from social tagging of individual tracks in these services, some databases have attained immense proportions. For instance, the *Million Song Database* contains one million music pieces with rich meta-data (song names, artists and albums), computationally extracted features, associated tags from Last.fm and lyrics. However, databases with carefully curated annotations of musical content (such as genres, onsets, instruments and chords) are still scarce due to their labour intensiveness (for example, *RWC Music Database*, *Magnatagatune Dataset*, *CAL-500 Dataset*).

Non-Western music databases are still rare but are steadily emerging. These tend to rely on audio representations due to problems of assigning symbolic encodings to musical material not operating similarly to Western music (Lidy et al., 2010). While still rare, these databases provide an important challenge for database studies in general. They question many of the basic analysis concepts since the music in them may have entirely different roles for timbre, rhythm, pitch and intonation, and set higher standards for the metadata concerning the contextual information about music.

Growing importance of database studies

Database studies offer insights into many areas of music research. They are invaluable tools for music history (linking manuscripts with critical commentary) and music analysis (forcing the analyst to explicitly define analysis operations and permitting the testing of these with large quantities of data). For music cognition, databases have been used to evaluate theories about music processing. For applied fields of music research, database studies offer building blocks for commercial products that can relate to mood and genre recognition. During this era of increased online data and the impending digitalisation of all music available, database studies are becoming an indispensable area of contemporary music scholarship.

Tuomas Eerola
Durham University, UK

See Also: Cantometrics; Similarity (Melodic similarity); Empirical musicology; Computer aided musical analysis;

Further readings

Clarke, Eric F., and Nicholas Cook. *Empirical Musicology: Aims, Methods, Prospects*. Oxford University Press, USA, 2004.

Dumitrescu, Theodor. "De-attributing the Missa Naray Je Jamais." *Journal of the Alamire Foundation* 2, no. 2 (2010): 167–192.

Huron, David. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, 2006.

Huron, David. "Music Information Processing Using the Humdrum Toolkit: Concepts, Examples, and Lessons." *Computer Music Journal* 26, no. 2 (2002): 11–26.

Lidy, Thomas, Carlos N Silla Jr, Olmo Cornelis, Fabien Gouyon, Andreas Rauber, Celso AA Kaestner, and Alessandro L Koerich. "On the Suitability of State-of-the-art Music Information Retrieval Methods for Analyzing, Categorizing and Accessing non-Western and Ethnic Music Collections." *Signal Processing* 90, no. 4 (2010): 1032–1048.

Lomax, Alan. *Folk Song Style and Culture*. Transaction Books, 1968.