Project Acronym:	**Europeana Sounds**
Grant Agreement no:	**620591**
Project Title:	**Europeana Sounds**

MS12 Evaluation of MIR Pilot

Revision: Final

Date: 02/11/2015

Authors: Alexander Schindler, AIT

Abstract: This milestone document presents the work carried out in task *T2.4.2 Music Information Retrieval* towards the implementation of the music information retrieval Pilot. It describes the retrieval algorithm and the required processing steps, the aggregation of the music dataset, the evaluation methodology and the creation of the required ground truth. The automated evaluation of the Pilot is included together with the discussion on the experimental results and extracted conclusions.

Dissemination level	
Public	X
Confidential, only for the members of the Consortium and Commission Services	

Coordinated by the British Library, the Europeana Sounds project is co-funded by the European Union, through the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme (CIP) http://ec.europa.eu/information_society/activities/ict_psp/

Revision history

Version	Status	Name, organisation	Date	Changes
0.1	ToC	Alexander Schindler, AIT	08/09/2015	
0.2	1st draft	Alexander Schindler, AIT	27/10/2015	First complete draft
0.3	2nd draft	Maarten Brinkerink, NISV	28/10/2015	Input from reviewers
0.4	Final draft	Sergiu Gordea	29/10/2015	Adding introduction and text review
1.0	Final	Laura Miles, Richard Ranft, BL	31/10/2015	Layout, minor changes

Review and approval

Action	Name, organisation	Date
Reviewed by	Maarten Brinkerink, NISV	28/10/2015
Approved by	Coordinator and PMB	30/10/2015

Distribution

No.	Date	Comment	Partner / WP
1	31/10/2015	Submitted to the European Commission	BL/WP7
2	31/10/2015	Posted on Europeana Pro website	BL/WP7
3	31/10/2015	Distributed to project consortium	BL/WP7

Application area

This document is a formal output for the European Commission, applicable to all members of the Europeana Sounds project and beneficiaries. This document reflects only the author’s views and the European Union is not liable for any use that might be made of information contained therein.

Statement of originality

This document contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Project summary

Europeana Sounds is Europeana's 'missing' fifth domain aggregator, joining APEX (Archives), EUscreen (television), the Europeana film Gateway (film) and TEL (libraries). It will increase the opportunities for access to and creative re-use of Europeana's audio and audio-related content and will build a sustainable best practice network of stakeholders in the content value chain to aggregate, enrich and share a critical mass of audio that meets the needs of public audiences, the creative industries (notably publishers) and researchers. The consortium of 24 partners will:

Double the number of audio items accessible through Europeana to over 1 million and improve geographical and thematic coverage by aggregating items with widespread popular appeal such as contemporary and classical music, traditional and folk music, the natural world, oral memory and languages and dialects.

Add meaningful contextual knowledge and medium-specific metadata to 2 million items in Europeana's audio and audio-related collections, developing techniques for cross-media and cross-collection linking.

Develop and validate audience specific sound channels and a distributed crowd-sourcing infrastructure for end-users that will improve Europeana's search facility, navigation and user experience. These can then be used for other communities and other media.

Engage music publishers and rights holders in efforts to make more material accessible online through Europeana by resolving domain constraints and lack of access to commercially unviable (i.e. out-of-commerce) content.

These outcomes will be achieved through a network of leading sound archives working with specialists in audiovisual technology, rights issues, and software development. The network will expand to include other data-providers and mainstream distribution platforms (Historypin, SoundCloud) to ensure the widest possible availability of their content.

For more information, visit http://pro.europeana.eu/web/europeana-sounds and http://www.europeanasounds.eu

Copyright notice

Contents

Executive summary: MS12 Evaluation of MIR Pilot

The activities and concrete results described in the following sections of this document present the progress of work in task *T2.4.2 Music Information Retrieval*. This milestone document focuses on the evaluation of the MIR Pilot, by measuring the precision of the retrieval algorithm and evaluating its usefulness for the Music Channel. The evaluation is performed using the whole sound content accessible through Europeana by using state of the art evaluation methods.

1 Introduction

Music Information Retrieval (MIR) is a young and interdisciplinary research field, dealing with the extraction of information from music content and using it for identifying, classifying, retrieving or recommending music. The research conducted in this area is typically combining knowledge from different areas such as musicology, computer science, signal processing and psychology.

Music is a highly multimodal, typically human created artefact. It has different representations and components, like symbolic (such as music scores) or textual representations (such as lyrics), particular tunes, rhythms and emotional messages. The categorization of music content (for example genre) and human emotions (for example mood) inducted by hearing music complex tasks, being correlated with the complexity and diversity of music descriptions. Consequently, the music retrieval is a very challenging task, while the perception of music similarity is highly personal and influenced by the music and user context [REF 4 Schedl 2014]. There are different approaches for MIR algorithms, which exploit very different information describing the music properties or user preferences such as acoustic properties, genre classifications, collaborative tagging and ratings, semantic tagging, mood and artist performance. Most approaches suffer from the cold start problem, meaning that the systems are not able to provide a decent retrieval performance as long there is no information available about the user preferences and the semantic description of the music dataset is still poor. The content based similarity search algorithms do not suffer from this problem, but they provide typically lower precision than the collaborative filtering approaches for instance.

However, given the heterogeneity of the Europeana dataset (including the various genres of music, metadata and lyrics provided in 28 different languages, different quality of recordings, different formats and sampling rates) the only feasible approach at the moment is the usage of content based similarities.

Evaluating the quality of the music retrieval is also a multi-faceted task, given the human perception of the music similarity and the expectations regarding the accuracy, diversity, serendipity and transparency [REF 6 Ricci 2011]. Within this document, we perform an automatic evaluation of the retrieval accuracy, while an additional evaluation based on user feedback will be included in D2.6 *Music Information Retrieval Pilot delivery report*.

2 Music and audio similarity

This section describes the concrete process of computing the similarity of sound content as implemented within the scope of the MIR Pilot. In the following subsections different low and high level audio features are presented and their influence and effectiveness with regard to the MIR Pilot is discussed, see Section 2.1 (Music and audio features). Finally, the algorithms used for computing similarities and retrieval of sound content are showcased in Section 2.2 (Audio similarity calculations).

2.1 Music and audio features

Feature extraction is the core of content-based description of audio files. With feature extraction from audio, a computer is able to recognize the content of a piece of music without the need of annotated labels such as artist, song title or genre. This is the essential basis for information retrieval tasks, such as similarity based searches (query-by-example, query-by-humming), automatic classification into categories, or automatic organization and clustering of music archives. Features extracted from the audio signal are intended to describe the stylistic content of the music, for example beat, presence of voice or timbre.

2.1.1 Overview of content based audio features

The audio features used for the experiments are well evaluated music content descriptors, widely used in the music information retrieval domain, and which provide a good timbral, temporal, rhythmic and harmonic description of the music content.

2.1.2 Psychoacoustic feature set by TU-Wien

Methods from digital signal processing are used and psycho-acoustic models are considered in order to extract suitable semantic information from music. Various feature sets have been developed, which are appropriate for solving various tasks.

All features of the Rhythm Patterns family undergo a series of pre-processing steps. Most of them are applied to anneal the sampled audio signal to an audio sensation experienced by human. These psychoacoustically transformed signals provide a better approximation of perceived sound, and thus perform better in describing music content.

As a first step, multiple audio channels (for example stereo-channels) are averaged to one and the audio is split into segments. Short-time Fourier Transform (STFT)[1] is applied to each segment to convert the audio signal into a frequency representation. The Bark Scale[2], a perceptual scale which groups frequencies to critical bands of hearing according to perceptive pitch regions, is applied to the spectrogram, aggregating it to 24 frequency bands. Subsequently, the Bark-Scaled spectrogram is transformed to the decibel scale and further to the Phon scale which incorporates equal loudness curves accounting for the different perception of loudness at different frequencies. Finally, a Sone-scale[3]

[1] see https://en.wikipedia.org/wiki/Short-time_Fourier_transform
[2] see https://en.wikipedia.org/wiki/Bark_scale
[3] see https://en.wikipedia.org/wiki/Sone

transformation is applied. The Sone scale relates to the Phon scale and describes the perceived loudness in a linear way (doubling on the Sone scale sounds to the human ear like a doubling of the loudness). Additionally, further psychoacoustic transformation such as spectral masking and blurring are applied before the calculation of the following features is started:

Rhythm Patterns

Rhythm Patterns (RP) describe modulation amplitudes for a range of modulation frequencies on "critical bands" relative to the human auditory range. They are computed by applying a discrete Fourier transform to the psycho-acoustically transformed Sonogram. This results in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each individual critical band [REF 3 Mayer 2010]. These fluctuations in modulation frequency provide a rough interpretation of the rhythmic energy of a song.

Relevance for MIR Pilot: Rhythm is a fundamental property of music. Rhythm Patterns are relevant to distinguish different rhythmic energies in the audio content and to make similarity calculations more efficient. Several music genres have their characteristic rhythms, like R&B, reggae, some of them not, like rock music. Still, this property is an important component for perception of music similarity.

Statistical Spectrum Descriptors

The Statistical Spectrum Descriptors (SSD)[4] are based on the previously described pre-processing steps. After the application of the psychoacoustic transformations, the mean, median, variance, skewness, kurtosis, minimum and maximum value are calculated subsequently for each individual critical band of the Bark scale. SSDs are able to capture additional timbral information compared to Rhythm Patterns, yet at a much lower dimension of the feature space.

Relevance for MIR Pilot: Timbre is another fundamental property of music. SSDs are not the best timbre descriptors but outperform traditional ones due to their overall description of the psychoacoustically transformed audio spectrum.

Further music features of the rhythm pattern family that are not included in the MIR Pilot

Rhythm Histogram

Rhythm Histograms (RH)[5] features capture rhythmical characteristics of an audio track by aggregating the modulation values of the critical bands computed in a Rhythm Pattern. Rhythm Histograms provide a much lower-dimensional descriptor for general rhythmic characteristics. The aggregated information is still able to describe the rhythmic energy, although the variance information of the critical bands gets lost.

Discussion for MIR Pilot: Although RH’s would require less space in memory and on hard-drives, they are not as descriptive as RP’s. Preliminary experiments showed that the RH’s are not able to discriminate satisfactory in this variegated dataset.

[4] see http://www.ifs.tuwien.ac.at/mir/pub/Mirex06_Poster_lidy_A0.pdf
[5] see http://ifs.tuwien.ac.at/mir/audiofeatureextraction.html

Temporal Statistical Spectrum Descriptor

Temporal Statistical Spectrum Descriptor (TSSD)[6] features describe variations over time by including a temporal dimension to incorporate time series aspects. Statistical Spectrum Descriptors are extracted from segments of a musical track at different time positions. Thus, TSSDs are able to reflect rhythmical and instrumental changes by capturing variations and changes of the audio spectrum over time.

Discussion for MIR Pilot: TSSDs are usually superior to SSDs due to the additional perspective. Best performances are expected on collections of audio files with similar lengths. In that case the TSSDs are able to describe differences in the structure of the audio file and in the case of music, the structure of the composition. The Europeana Sounds collection is not that uniform and contains tracks of various lengths. Many of the items in the collection consist of 30 second samples which are often randomly extracted from the original file. Thus, the advantage of the TSSDs is dampened and the globally calculated SSDs compensate for many problems provided by the 30 second samples.

Temporal Rhythm Histograms

Temporal Rhythm Histograms (TRH)[7] capture change and variation of rhythmic aspects in time. They are similar to the Temporal Statistical Spectrum Descriptor statistical measures of the Rhythm Histograms of individual six second segments in a musical track are computed.

Discussion for MIR Pilot: Similar to the TSSD; TRHs are able to capture structural properties of a song. They are similarly able to detect variances in rhythm such as changes in rhythm or percussive playing styles during the recording. Similarly they face the same disadvantages concerning the Europeana Sounds collection.

2.1.3 Standard low-level audio features

Mel Frequency Cepstral Coefficients (MFCC)

This feature set[8] has been used previously in speech recognition and intends to model human auditory response by transforming it to the Mel scale[9]. The “cestrum” (“s-p-e-c” reversed) results of taking the Fast Fourier transform (FFT)[10] of the decibel spectrum as if it were a signal. The result shows the rate of change in the different spectrum bands. It is a dominant feature in speech recognition, because of its ability to represent the speech amplitude spectrum in a compact form. It also has proved to be highly efficient in music retrieval. In representing the rate of change in the different spectrum bands it is a good timbre descriptor. MFCCs are the most commonly used features in music processing.

[6] see http://ifs.tuwien.ac.at/mir/audiofeatureextraction.html#TSSD
[7] see http://ifs.tuwien.ac.at/mir/audiofeatureextraction.html#TRH
[8] see https://en.wikipedia.org/wiki/Mel-frequency_cepstrum
[9] see https://en.wikipedia.org/wiki/Mel_scale
[10] see https://en.wikipedia.org/wiki/Fast_Fourier_transform

Chroma

Chroma Features[11] represent the 12 distinct semitones (or chroma) of the musical octave. This results in one or a sequence of twelve dimensional vectors where, for example, the bin that corresponds to the pitch class A captures the spectral energy of A0 and all its corresponding sub-band pitches A1, A2.

Root Mean Square

Root Mean Square (RMS)[12] is a way of comparing arbitrary waveforms based upon their equivalent energy. The RMS method takes the square of the instantaneous voltage before averaging, and then takes the square root of the average.

Spectral Centroid

The Spectral Centroid (SC)[13] is the frequency-weighted sum of the power spectrum normalized by its unweighted sum. It could be described as the centre of gravity or the balancing point of the spectrum. It determines the frequency area around which most of the signal energy concentrates and gives an indication of how “dark” or “bright” a sound is.

Spectral Bandwidth

The Spectral Bandwidth (SBW)[14] represents the weighted spread between minimal and maximal frequency and is calculated similar to the spectral centroid.

Spectral Contrast

The Spectral Contrast (SC) is calculated from the spectral peaks and valleys and their difference in each sub-band. Strong spectral peaks roughly correspond with harmonic components, while spectral valleys correspond with non-harmonic components such as noise. Thus, Spectral Contrast features could roughly reflect the relative distribution of harmonic and non-harmonic components in the spectrum.

Spectral Rolloff

The Spectral Rolloff (SRO) is the frequency below which some fraction k (typically 0.85, 0.9 or 0.95 percentile) of the cumulative spectral power resides. It is a measure of the skewness of the spectral shape and an indication of how much energy is in the lower frequencies. It is often used to distinguish voiced from unvoiced speech or music.

Tonnetz

Tonnetz[15] features are able to detect changes in the harmonic content of musical audio signals based on a model for Equal Tempered Pitch Class Space using 12-bin chroma vectors. Close harmonic relations such as fifths and thirds appear as small Euclidian distances. Peaks in the detection function denote transitions from one harmonically stable region to another.

[11] see https://en.wikipedia.org/wiki/Harmonic_pitch_class_profiles
[12] see https://en.wikipedia.org/wiki/Root_mean_square
[13] see https://en.wikipedia.org/wiki/Spectral_centroid
[14] see https://en.wikipedia.org/wiki/Bandwidth_extension
[15] see https://en.wikipedia.org/wiki/Tonnetz

Zero Crossing Rate

Zero-crossing rate (ZCR)[16] is a simple, straightforward and inexpensive feature. It measures whether two sets of time series measurements exhibit similar patterns. It is particularly useful to analyse measurements that are corrupted by noise. For example, a measurement with a high zero-crossing rate, i.e., the number of samples per second that cross the zero reference line, indicates that it is noisy.

Discussion for MIR Pilot: The ZCR is used to group audio files by their recording quality. The Europeana Sounds dataset contains many items that have been digitized from old records and even older formats. Consequently they exhibit a strong noise behaviour resulting from degradations of the original carriers such as shellac or wax cylinders. The ZCR groups audio files by their noise behaviour.

Beats Per Minute

Beats per Minute (BPM)[17] are a common description of the tempo of a music track. It is calculated from audio events which are detected in the audio signal.

2.2 Audio similarity calculations

This section describes the fundamentals of the audio similarity search algorithm developed for this Pilot. Audio features are descriptive numbers calculated from the audio spectrum of a track. A good example is the Spectral Centroid, which can be interpreted as the centre of gravity of an audio recording. It describes the average frequency weighted by its intensity and distinguishes brighter from darker sounds. Such features are usually calculated for several intervals of a track and finally aggregated into a single vector representation. The latter step, which is a requirement for many machine/statistical learning tasks, is accomplished by calculating statistical measures such as mean or standard deviation.

In the following example, the Spectral Centroids of 10 different tracks are provided using their mean and standard deviation aggregations. Thus, the Spectral Centroid feature (-set) is represented by a two-dimensional feature vector such as the following example:

```
IDX    Mean                 Standard Deviation
0      1517.5993814237531   291.1855836731788
```

In this example the centre frequency is 1518 Hz and it deviates by 291 Hz. These numbers already describe the audio content and can be used to find similar tracks. The common approach to calculate music similarity from audio content is based on vector difference. The assumption is that similar audio feature-values correspond with similar audio content. Thus, feature vectors with smaller vector differences correspond to more similar tracks. The following data represents the extracted Spectral Centroids of our 10-tracks collection:

```
ID     Mean                 Standard Deviation
0      1517.5993814237531   291.1855836731788
1      1659.1988993873124   327.64811981777865
2      1507.4617047141264   340.8830079395701
3      1597.6019371942953   507.1007933367403
```

[16] see https://en.wikipedia.org/wiki/Zero-crossing_rate
[17] see https://en.wikipedia.org/wiki/Tempo

```
4      1498.8531206911534  288.3780838480238
5      535.5910732230583   89.90893994909047
6      2261.4032345595674  353.5971736260454
7      2331.881852844861   406.33517225264194
8      1868.690426450363   342.7489751514078
9      2204.6324484864085  328.94334883095553
```

The tracks have unique identifiers and we are using the track with ID 5 to search for similar items. This step requires a similarity metric, which defines how the vector distance has to be calculated as a single numeric value. The most common choices are the Manhattan (L1) and Euclidean (L2) distance measures. The Euclidean Distance is the square root of the sum of squared differences of two vectors.

To calculate the Euclidean Distance between track 5 and track 0:

```
ID     Mean                Standard Deviation
0      1517.5993814237531  291.1855836731788
5      535.5910732230583   89.90893994909047
```

- we first compute the difference between the values of each vectors

```
982.008308          201.276644
```

- square them to get the absolute magnitude:

```
964340.317375       40512.287309
```

- and take the sum of these values:

```
1004852.6046840245
```

Per definition the square root has to be calculated from the sum, but this step is normally skipped because it does not alter the ranking and is processing intensive. By calculating the distance for all items in the collection, we retrieve a list of distance values where the smaller distances correspond to more similar audio content and the higher values should sound more dissimilar.

```
ID     Distance
0      1004852.6046840245
1      1319014.4646621975
2      1007520.5071585375
3      1301916.1177259558
4      967263.7731724023
5      0.0
6      3047959.100796666
7      3326786.1254441254
8      1841081.968976167
9      2842836.5609704787
```

To retrieve a ranked list of similar sounding tracks, the list of vector distances must be ascending.

```
ID     Distance
5      0.0
4      967263.7731724023
0      1004852.6046840245
2      1007520.5071585375
3      1301916.1177259558
1      1319014.4646621975
8      1841081.968976167
9      2842836.5609704787
```

```
6       3047959.100796666
7       3326786.1254441254
```

This so called vector space model is predominant in content based multimedia retrieval. The most crucial and problematic part is feature crafting, meaning that in the case in which the extracted numbers do not describe the audio well enough, the vector based similarity will also fail to provide results that are perceived as similar.

The described approach requires the availability of all feature vectors of all items of a collection. Thus, the feature vectors must be stored. No matter which retrieval approach (pre-calculated / indexed / on demand) will be chosen, all features will be required at a certain time. Given that the feature extraction is a computationally expensive task (in terms of processing resources and total time) the extracted features are stored and made accessible using a common data format.

3 Evaluation

This section describes the evaluation approach, by presenting the chosen methodology and how this is concretely used for the evaluation of the MIR Pilot. This follows the standard approach used for evaluating information retrieval systems. The so called Cranfield paradigm can be traced back to the late 1950s[18] [REF 5 Cleverdon 1967]. It is based on a document collection, a test suite of expressible queries and a set of relevance judgements, also called gold standard or ground truth. During the evaluation the results of the queries are compared against the relevance judgements of the ground truth data and a set of expressible metrics are calculated that are used to make different systems comparable.

3.1 MIR Pilot evaluation data and ground truth

For the evaluation of the MIR Pilot the following data corpus and ground truth structure were used:

Document collection: The evaluation dataset will consist of 312,096 audio items which were downloaded from the internet by using the URL discovered through the Europeana API. The data consists of mp3-encoded audio data of variable size, sample rate and bitrate. The audio content varies from speech, to recorded radio broadcast, music varying in age, style and quality, nature and ambient sound recording, etc. For every item the corresponding metadata available via the Europeana API has been downloaded. A more detailed overview of the dataset is provided in Section 3.2.

Test suite of expressible queries: These queries define what will be evaluated. In the MIR Pilot these corresponds to the query-by-example (QBE) functionality. As explained in the previous chapters, QBE takes an “audio instance” as query and searches for similar items. Consequently, queries correspond to example audio files, from which the audio features were extracted and used as input for the retrieval algorithm. The expected results consist of a list of audio items that “sound similar”.

Ground truth/relevance judgements: To compare the search result against an expected result, a ground truth is required, which labels the data according such expectations. Creating a ground truth for large

[18] see https://en.wikipedia.org/wiki/Cranfield_experiments

datasets is highly expensive in terms of time and budget constraints. The advantage of the data provided by Europeana is that the metadata corresponding to each all collection item have been curated and edited by well trained staff at national libraries and audio-visual archives. They are considered as providing reliable, trustworthy information. Based on this assumption, it is plausible to use the metadata of the Europeana dataset to create a ground truth without consulting further annotators.

Yet, the automatic generation of the ground truth from the Europeana metadata has to be similarly plausible. Music similarity is a highly subjective concept and metadata entries are required that facilitate objective comparability of music. Generally, music similarity can be defined by a mixture of timbre, rhythm, pitch and key. Based on empirical evaluations of the Europeana metadata it was observable that the data is rich in literal descriptions of music properties. As an example the following query would provide a substantial description of the acoustic properties of a musical track: ["string", "quartet", "C Minor", "allegro"]. These terms describe the used instrumentation and thus the expected timbre, as well as the key. The term “allegro” is historically overloaded and refers to the tempo, provides a hint to rhythm as well refers to the mood-related characteristics joyful, lively and fast. Executing a string-based search on the Europeana metadata, one can retrieve a list of items that have the searched properties and include them in the ground truth. In the following we present the overview of the Europeana search results using the mentioned keywords.

Keywords = ["string", "quartet", "allegro", "C Minor"]

- String Quartet N. 8 In C Minor Op. 110: Allegro Molto;Dmitry Shostakovich (Performer);Classical;00:03:03, 00:00:30 (preview duration)
- "String Quartet In C Minor, D 703, ""Quartettsatz"": Allegro assai";Franz Schubert (Performer), Chamber Orchestra Kremlin (Performer), Misha Rachlevsky (Conductor);Orchestral Music;00:09:34, 00:00:30 (preview duration)
- String Quartet No. 12 in C minor, Op. posth. D. 703 (Quartettsatz) - Allegro assai;Franz Schubert (Performer), Enesco Quartet (Ensemble);Classical;00:09:02, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Other;00:07:30, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, IV. Allegro;Ludwig Van Beethoven (Performer);Other;00:03:57, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Orchestral Music;00:07:31, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Orchestral Music;00:07:30, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, IV. Allegro;Ludwig Van Beethoven (Performer);Orchestral Music;00:03:57, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Romantic;00:07:30, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, IV. Allegro;Ludwig Van Beethoven (Performer);Romantic;00:03:57, 00:00:30 (preview duration)
- String Quartet No.4 In C Minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Classical;00:07:31, 00:00:30 (preview duration)
- String Quartet No.4 In C Minor, Op.18 No.4, Iv. Allegro;Ludwig Van Beethoven (Performer);Classical;00:03:55, 00:00:30 (preview duration)
- String Quartet in C minor, Op. 51, No. 1: Allegro;Mandelring Quartett Johannes Brahms (Performer);Classical (Core - Classical);
- String Quartet in C minor, Op. 51, No. 1: Allegro;Mandelring Quartett Johannes Brahms (Performer);Classical (Core - Classical);

Usually the class sizes in standardized information retrieval test-collections are bigger than just 10-20 instances. Though, the advantage of the suggested approach is that it facilitates highly customized evaluations addressing very specific definition of classifications, whereas standard collections are mostly dedicated to a specific task. By identifying diverse relevant music-property-descriptions it is possible to evaluate the MIR Pilot in multiple facets. It is possible to draw conclusions of its performance on different types of audio such as classic or contemporary music, spoken word, animal or ambient sounds, based on recording quality, etc.

The provided example is a highly specific query. To provide a broad overview of the query-by-example algorithm implemented in the MIR Pilot, queries with different granularity of classifications will be applied.

Evaluation metrics: The evaluation was executed in different runs. Each run evaluates the system based on the ground truth composed by items matching a previously defined (text-based) description. To assess the performance of the system, its *precision* is calculated as described:

- From the generated ground truth data take one song and compute the result list.
- Compare the result of the MIR algorithm against the ground truth.
- Compute the retrieval precision, by evaluating (in percentage) how many entries of the ground truth are presented in the result list.

This procedure is repeated for every generated class of the ground truth. For example the query for traditional Italian “Tarantella” provides 152 metadata entries. Thus, this represents the size of the ground truth. During an evaluation run, every entry is used as input for the query-by-example algorithm to compute the list of similar sounding entries. Each result-list is intersected with the ground truth data to find out how many of the computed similar sounding entries are part of the ground truth. In the given example a precision of 100% would refer to all entries of a result-list being part of the ground truth and thus having the term “Tarantella” applied in their metadata. If the number of items of a ground truth is very large (for example broad categories like jazz music), the evaluation set is subsampled to 1000 entries.

The computed result-list is ranked by the calculated similarity. Thus, the items on top of the list are expected to be more similar than those at the bottom. To reflect this behaviour in the evaluation, the precision is calculated using result-lists of different lengths:

- Length 1: corresponds to the first entry which is according to the computation the most similar song.
- Length 2 and 3: searching for music requires more time and attention than looking at images, thus the first three entries are the most important entries when browsing or searching for similar songs.
- Length 5 and 10: these precision values are provided to give an overview of the system performance concerning longer result lists. A length of 10 is generally known to be a threshold for user attention.
- Length 24: this represents the number of entries per page of the Europeana Web page’s search result at the time of the evaluation.

3.2 Dataset overview

For the dataset creation 400,615 entries were downloaded via the Europeana API, the whole sound content which provides a public http URL for accessing the content. The JSON formatted metadata was stored for each item and aggregated into a complete dataset. During this aggregation step, erroneous JSON responses were removed. After this cleansing and aggregation step 389,120 entries were included in the set. In parallel, the corresponding audio data was downloaded and the previously described audio features were extracted. It was not possible to obtain the audio files for all entries in the dataset (due to broken URLs or password protected access). Also some of the downloaded files were corrupt or failed to be processed by the audio feature extractors. 327,261 RP-features and 323,664 features using the Librosa library were extracted. The intersection of aggregated metadata and available audio features resulted in a final dataset size of 312,096 entries.

3.2.1 Dataset statistics

Data providers:

- Number of Data Providers: 1,002
- Top-10 Data Provides:
    - Preiser Records; Austria (30,738)
    - National Library of Spain (10,462)
    - The Orchard (8,860)
    - JSP Records (7,686)
    - Hacienda Records (6,537)
    - Carinco AG (6,534)
    - Ovação; Portugal (6,032)
    - Gesellschaft für Historische Tonträger; Austria (5,985)
    - Duck Records; Italy (5,691)
    - Arts Productions Ltd (5,644)

Data aggregators:

- Number of Data Aggregators: 20
- Top-10 Data Aggregators:
    - DISMARC (286,189)
    - Judaica Europeana (8,033)
    - Hispana (2,373)
    - Deutsche Digitale Bibliothek (1,661)
    - OpenUp! (1,097)
    - National Library of Finland (183)
    - EuropeanaLocal Romania (136)
    - HOPE - Heritage of the People's Europe (123)
    - The Natural Europe Project (104)
    - Europeana 1914-1918 (47)
- Median number of sound files per aggregator: 39

Collections:

- Number of collections: 34

4 Results

This section provides an overview of the evaluation results, a discussion on the findings and the conclusion.

4.1 Evaluation results

The following table provides a summary of the evaluation runs. Each run has a dedicated ID to refer to in the succeeding discussion of the results. The quoted query terms refer to actual words that were used to search in the metadata. The search approach was case insensitive and sequentially. This means that for the query *"Jazz" + "Traditional"* the metadata was first filtered by the term "Jazz" and the resulting subset was filtered by the term "Traditional". This example resulted in a final ground-truth subset of size 512 (i.e. including only items that are known to belong to traditional jazz), which is denoted by the column *Num. Tracks.* The rightmost six columns represent the calculated precision values for the corresponding query at different result-list lengths. As explained in Section 3.1 a result-list length of 1 refers to the precision of the first and most similar entry of the list whilst a length of 24 corresponds to the precision of a complete search result page of the Europeana Web-page. The presentation of the results is grouped by 5 large types of content, namely: **Jazz** (music), **Classical** (music), **Non-Music**, **European Popular/Traditional Music, Contemporary** (music).

For each of this types the best values for the precision collected at top 1, 3, 10 are marked with **bold** in the corresponding columns of the table below and the worst results are marked with *italic+underline* fonts.

Table 1: Evaluation results for different queries. "Num. Tracks" refers to the size of the query-generated ground truth. The number-labelled columns refer to the calculated precision values at different result-list lengths.

ID	Query	Num. Tracks	1	2	3	5	10	24
				Jazz				
1	"Jazz"	31801	38.0	35.0	31.4	31.7	**28.6**	26.2
2	"Smooth Jazz"	2419	**49.1**	45.9	**43.8**	25.8	20.8	16.0
3	"Jazz" + "Traditional"	577	25.3	21.1	17.1	13.0	9.2	6.3
4	"Ragtime"	57	24.6	15.8	12.3	7.3	3.6	1.6
5	"Shuffle"	112	22.3	14.3	11.0	7.0	4.0	1.7
6	"Jazz" + "Blues"	1398	14.6	10.1	7.5	5.6	4.2	2.9
7	"Jazz" + "Bob Crosby, Andy Kirk, June Richmond"	24	12.5	6.3	*4.2*	2.5	3.8	1.8

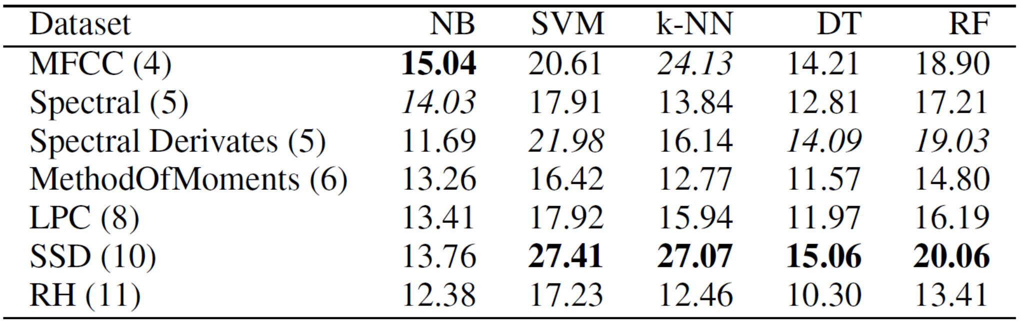
8	"Jazz" + "Cuban Jazz"	105	*11.2*	7.6	5.7	4.2	*2.0*	1.0
Classical								
9	"Classical"	28569	**44.3**	42.1	**40.5**	38.3	**35.1**	32.3
10	"Piano Concerto"	510	38.6	32.0	28.0	23.9	17.6	10.2
11	"Requiem"	463	32.6	26.9	22.0	16.2	10.7	6.6
12	"operette", "operetta", "opereta", "zarzuela"	1081	27.7	22.9	20.8	17.3	14.6	11.5
13	"Opera"	8278	26.8	24.7	22.7	21.1	18.9	16.1
14	"Classical" + "g major" or "g dur" or "g-dur" or "g majeur"	304	17.1	14.8	14.0	12.6	9.3	5.6
15	"Classical" + "g major" or "g dur" or "g-dur" or "g majeur" + "quartett"	13	15.4	7.7	*5.1*	3.1	*2.3*	1.6
16	"Classical" + "major" or "dur" or "majeur" + "quartett" + "allegro"	191	*9.4*	6.3	7.3	8.1	5.6	3.7
Non-music								
17	"Interview"	484	77.5	74.3	72.0	68.6	60.8	51.4
18	"Biodiversity Center" (=Animal Sounds)	1097	**89.7**	87.0	**85.1**	82.8	**78.7**	73.9
19	"Biodiversity Center" + "Chorthippus" (=Crickets)	113	*59.3*	55.3	*56.6*	53.0	*48.1*	43.0
European popular/traditional music								
20	"Flamenco"	1827	**40.7**	33.0	**29.2**	24.3	**18.2**	12.8
21	"Tarantella"	152	33.6	28.0	22.4	16.1	8.5	4.0
22	"Tango"	3716	30.2	24.9	22.3	19.5	16.0	12.5
23	"Flamenco" + "Guitarra"	287	22.3	17.1	15.3	13.5	10.0	8.3
24	"Jodler"	61	16.4	8.2	5.5	3.6	*2.0*	1.4
25	"Serenata"	436	9.7	7.5	5.5	4.4	3.0	2.0
26	"Volksmusik"	13	7.7	3.9	*2.6*	4.6	4.6	2.9
27	"Fados"	107	*6.5*	10.8	9.7	7.3	4.9	2.7
Contemporary								

28	“Rock ‘n Roll”	24	**16.7**	8.3	*5.6*	3.3	*2.0*	1.6
29	“Hip Hop”	72	*11.1*	11.1	**12.5**	11.9	**7.5**	5.2

4.2 Discussion

The results of the evaluation runs can generally be observed to correlate with state-of-the-art music similarity retrieval results presented in the literature. The most representative related publication is [REF 1 Schindler 2012]. This study presented new genre ground truth assignments for the Million Songs Dataset, including first baseline results for automatic genre classification experiments. One of the classifiers used in the evaluation was the k-nearest neighbours’ classifier. The principle of this classifier to label an unknown instance is to calculate the vector distances to all items in the dataset and to sort them by their distance. The k top most entries with the smallest distance to the unknown instance - the k nearest neighbours’ - are used to determine the label for the processed instance by majority voting. An example could be to set k to 3, which means, that the top 3 nearest neighbours will be used for majority voting. If one of them is of “genre 1” and two of them are from “genre 2”, the unknown instance is classified as “genre 2”. A special case is knn with k = 1. In this case only the label of the top nearest neighbour is used to classify the unknown instance. This approach was used in [REF 1 Schindler 2012] and it is equivalent to the similarity retrieval approach of the MIR Pilot with a result list length of 1. In this case, the results of similarity retrieval are comparable to those of classification experiments. The results presented in [REF 1 Schindler 2012] are, by date, the only ones in literature matching the scale of the Europeana dataset. The subset used in their evaluation contained 273,936 music tracks. This is only 12.2% smaller than the Europeana dataset with 312,096 audio files. The results reported by [REF 1 Schindler 2012] are presented in Table 2. The best results for the k-NN classifier with k = 1 were reported for the Statistical Spectrum Descriptors (SSD). Due to scalability problems the authors were not able to include higher dimensional feature sets such as Rhythm Patterns (RP) in their evaluation as well as combined feature spaces.

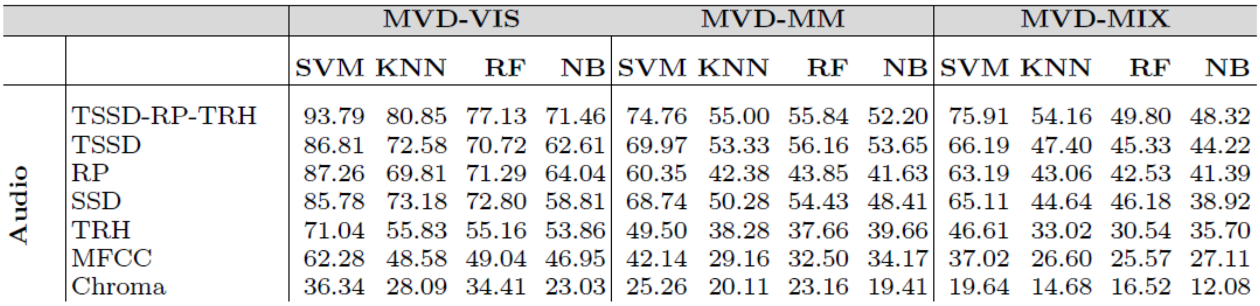
Table 2: Results of classification experiments with the Million Song Dataset taken from [REF 1 Schindler 2012]. For the k-NN classifier k = 1 was used. Thus, these results are comparable to the similarity retrieval results presented in Table 1 with a result-list length of 1.



Dataset	NB	SVM	k-NN	DT	RF
MFCC (4)	**15.04**	20.61	*24.13*	14.21	18.90
Spectral (5)	*14.03*	17.91	13.84	12.81	17.21
Spectral Derivates (5)	11.69	*21.98*	16.14	*14.09*	*19.03*
MethodOfMoments (6)	13.26	16.42	12.77	11.57	14.80
LPC (8)	13.41	17.92	15.94	11.97	16.19
SSD (10)	13.76	**27.41**	**27.07**	**15.06**	**20.06**
RH (11)	12.38	17.23	12.46	10.30	13.41

A naive approach to compare the results of [REF 1 Schindler 2012] with those of the MIR Pilot evaluation presented in Table 2 would be to calculate the mean precision of all evaluation runs. This would result in an average precision of 28.7% and be slightly above the top results presented in literature. Thus, the performance of the implementation corresponds to comparable state-of-the-art similarity retrieval system, although systems have been reported with much higher precision values. Unfortunately, the datasets used to evaluate these systems have not been made public or are too small in size to make them reliably comparable. In [REF 2] Schindler 2015, a similar approach to evaluate the performance of a music video classification system was used. For the evaluation the Music Video Dataset (MVD) has been created. It is a highly specialized dataset to develop visual feature extractors for music video analysis and consists of two sub-sets of music videos of different genres. All subsets had been assembled by their audio properties - they had been selected because they sounded similar or stereotypic for that specific genre. The tracks of the MVD-VIS's genres sound especially similar. Table 3 shows the precision results for the genre classification experiments. Combinations of Rhythm Patterns features reach a precision of 93.79% using Support Vector Machine (SVM) classifiers and 80.85% for K-NN classifier with k=1. Consequently this is currently the top result to be expected under optimal conditions such as clearly defined genres without overlaps. The MVD-MM subset was assembled to incorporate such overlaps. Consequently, the precision values are notably smaller than those of the MVD-VIS dataset. The MVD-MIX set is a non-overlapping combination of the MVD-VIS and MVD-MM subset. The combination creates a bigger dataset with a higher number of genres which is an important evaluation scenario for automatic genre classification experiments. In both sets, the MVD-MM and the bigger MVD-MIX set, the precision values for the K-NN classifier with k=1 range between 26% for the standard feature set MFCC and about 55% for the combination of the Rhythm Patterns features. Although the MVD essentially differs in size, having only 1600 entries, its specialization provides good boundaries of which maximal and average values can be expected from the evaluated features.

Table 3: Results of classification experiments with the Million Song Dataset taken from [REF 2 Schindler 2015]. For the k-NN classifier k = 1 was used.



		MVD-VIS				MVD-MM				MVD-MIX			
		SVM	KNN	RF	NB	SVM	KNN	RF	NB	SVM	KNN	RF	NB
Audio	TSSD-RP-TRH	93.79	80.85	77.13	71.46	74.76	55.00	55.84	52.20	75.91	54.16	49.80	48.32
	TSSD	86.81	72.58	70.72	62.61	69.97	53.33	56.16	53.65	66.19	47.40	45.33	44.22
	RP	87.26	69.81	71.29	64.04	60.35	42.38	43.85	41.63	63.19	43.06	42.53	41.39
	SSD	85.78	73.18	72.80	58.81	68.74	50.28	54.43	48.41	65.11	44.64	46.18	38.92
	TRH	71.04	55.83	55.16	53.86	49.50	38.28	37.66	39.66	46.61	33.02	30.54	35.70
	MFCC	62.28	48.58	49.04	46.95	42.14	29.16	32.50	34.17	37.02	26.60	25.57	27.11
	Chroma	36.34	28.09	34.41	23.03	25.26	20.11	23.16	19.41	19.64	14.68	16.52	12.08

The MVD has been used in the development stage to evaluate which features to use for the MIR Pilot as well as how to combine them to get an optimal performance. Figure 1 shows one of those preliminary evaluations. It depicts the precision values of the Rhythm Patterns feature set with different feature space normalization methods applied and with different distance measures calculated. The chart reproduces the precision values of 69.81 for RPs on the MVD-VIS subset presented in Table 3. It further depicts the performance at different result-list lengths. It is observable that the precision drops by about 20% on average from k=1 to k=20. This behaviour can also be observed in the results of this evaluation as presented in Table 1. This is an artefact that is ascribable to flaws in the ground truth data which was not created for similarity retrieval experiments.

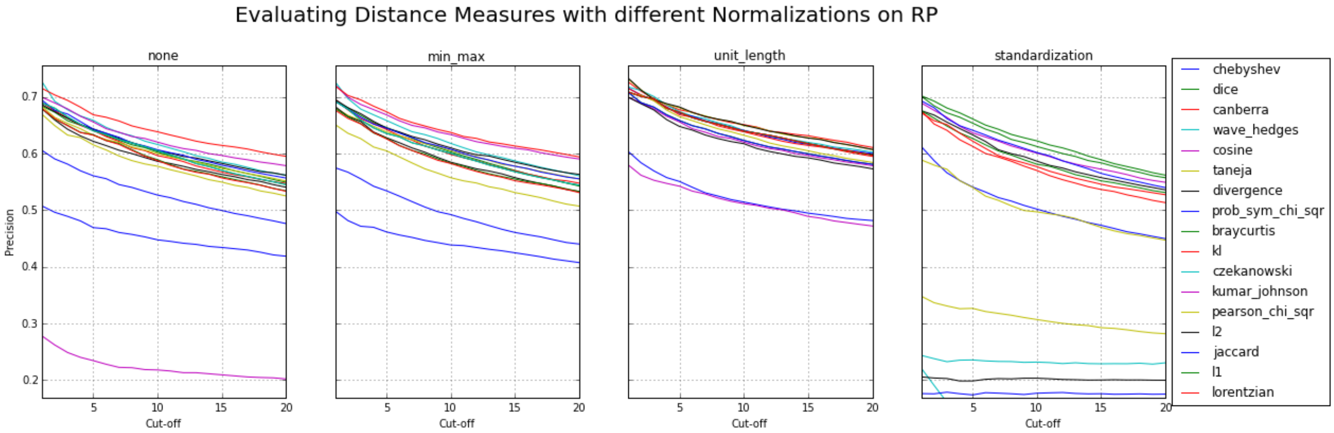


Figure 1: Performance evaluation of the Rhythm Patterns (RP) feature set with different normalization methods in combination with different distance measures on the MVD-VIS subset of the Music Video Dataset. Numbers depict precision values for result-list length ranging from 1 to 20 entries.

Figure 2 shows the same evaluation on the MVD-MM subset of the Music Video Dataset. This subset has more overlaps between the genres and thus results in weaker performance values. This set corresponds more to large mixed datasets such as the Europeana sounds dataset. The relative performance values and the decline of the precision in case of longer result-lists is comparable to those of the MVD-VIS dataset with its highly similar sounding tracks per genre. These two similar observations on the different subsets indicate that such performance curves can also be expected from similarity retrieval experiments on the Europeana dataset. Referring to the Europeana sounds evaluation results in Table 1 similar behaviour can be observed. Precision values of “Jazz”, “Smooth Jazz”, “Piano Concerto”, etc. drop by 20% on average from result-list lengths from 1 to 24 items. One objective indicator for huge declines of precision towards longer result-lists are small ground-truth sizes. The possibility that all audio tracks of a 24 items ground-truth are on the same result-list page is very low. Thus, precision values of small ground-truth sets generally drop to values around one or two per cent.

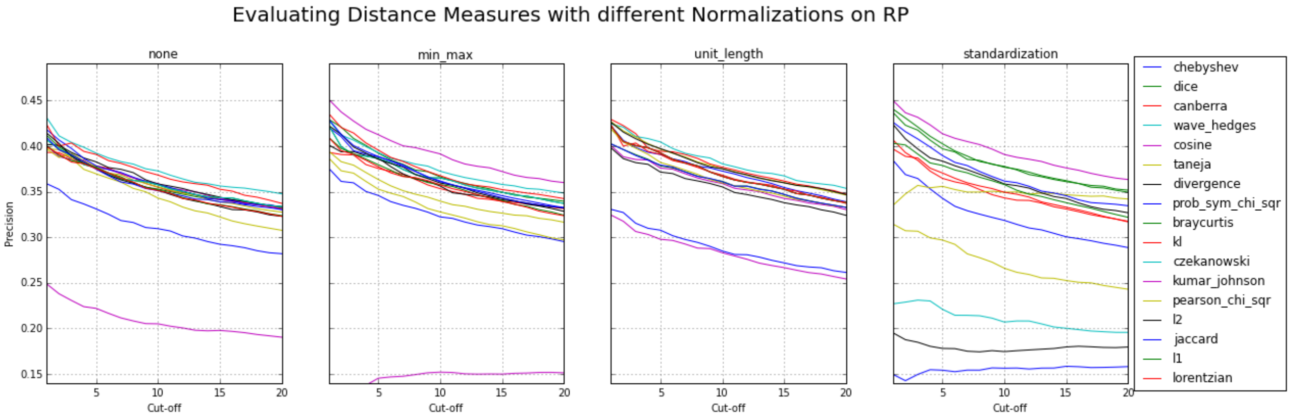


Figure 2: Performance evaluation of the Rhythm Patterns (RP) feature set with different normalization methods in combination with different distance measures on the MVD-MM subset of the Music Video Dataset. Numbers depict precision values for result-list length ranging from 1 to 20 entries.

As well as a direct comparison of the averaged results with results provided in the literature, a more detailed examination of the presented results provides a good assessment of the performance of the MIR Pilot’s implementation. The results of the high-level query terms “Jazz” and “Classic” are high compared to previously discussed values. Yet, they cannot be considered representative for music similarity, since Jazz and Classic have huge varieties and many sub-genres differ in style, rhythm and instrumentation. Query 2 provides a more discrete insight by focussing on “Smooth Jazz”. This sub-genre is stylistically more precisely described and the annotators of the metadata obviously shared a common understanding of this description. This resulted in having every second query deliver a correct result and these high precisions are also noticed for longer result-lists. Similar performance values are observed for the queries 3-5 and 10-13. “Piano Concerto” for example describes classical music performed on pianos. Thus, it gives a clear description of the instrumentation and the expected timbre. Also the variation in timbre is clearly defined. This results in high precision values.

5 Summary

This document described the method used to evaluate the MIR Pilot. The presented approach is based on automatic evaluation based on ground truth queries on the metadata of the evaluation dataset. This dataset had been downloaded via the Europeana API and consists of about 320,000 media files and associated metadata. The focus of the evaluation was set on the query-by-example feature of the MIR Pilot and used subsets filtered by search terms that referred to audio properties such as “Piano Concerto” or specific sub-genres such as “Smooth Jazz”. Precision values resulting from the evaluation experiments were discussed and identified to be similar or comparable to state-of-the-art similarity retrieval approaches reported in literature.

5.1 Conclusion

It is interesting to observe that precision values on different datasets are comparable (for example Europeana versus million song database), especially datasets with high varieties in their content (for example including different music genres, spoken content or animal sound) show highly similar performance values. The chosen evaluation approach provided a good and comprehensive overview of the dataset and the performance of the implemented algorithm of the MIR Pilot. Yet, flaws of the ground-truth queries were observed. One of these problems can be deduced by the queries 12 and 14 where the search terms were formulated in four different languages. For query 12 an increase in precision and ground-truth size was observed after adding further languages. This is a general weakness of string based approaches applied to heterogeneous datasets (for example see the variety of languages, content types and metadata quality).

For the MIR Pilot deliverable, a user-evaluation will be performed to assess how the automatically calculated results are perceived by users.

6 References

Ref 1	[Schindler 2012] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, pages 469-474, Porto, Portugal, October 8-12 2012.
Ref 2	[Schindler 2015] Alexander Schindler and Andreas Rauber. An audio-visual approach to music genre classification through affective color features. In *Proceedings of the 37th European Conference on Information Retrieval (ECIR'15)*, Vienna, Austria, March 29 - April 02 2015.
Ref 3	[Mayer 2010] Rudolf Mayer and Andreas Rauber. Multimodal Aspects of Music Retrieval: Audio, Song Lyrics – and Beyond? In Advances in Music Information Retrieval, vol 274, P333-363, Springer Berlin Heidelberg
Ref 4	[Schedl 2014] M. Schedl , E. Gómez and J. Urbano. Music Information Retrieval: Recent Developments and Applications. In Foundations and Trends in Information Retrieval Vol. 8, No. 2-3 (2014) 127–261
Ref 5	[Cleverdon 1967] Cyril Cleverdon. The Cranfield Tests on Index Language Devices, Aslib Proceedings 1967 19:6 , 173-194
Ref 6	[Ricci 2011] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor,editors. Recommender Systems Handbook. Springer, 2011

Appendix A: Terminology

A project glossary is provided at: http://pro.europeana.eu/web/guest/glossary.

Additional terms are defined below:

Term	**Definition**
AB	Advisory Board
APEX	Archives Portal Europe network of excellence
EC-GA	Grant Agreement (including Annex I, the Description of Work) signed with the European Commission
PMB	Project Management Board
TEL	The European Library
UAP	User Advisory Panel
WP	Work Package