Evaluation of the Music Ontology Framework

Yves Raimond[1] and Mark Sandler[2]

[1] BBC R&D
yves.raimond@bbc.co.uk
[2] Queen Mary, University of London
mark.sandler@eecs.qmul.ac.uk

Abstract. The Music Ontology provides a framework for publishing structured music-related data on the Web, ranging from editorial data to temporal annotations of audio signals. It has been used extensively, for example in the DBTune project and on the BBC Music website. Until now it hasn't been systematically evaluated and compared to other frameworks for handling music-related data. In this article, we design a 'query-driven' ontology evaluation framework capturing the intended use of this ontology. We aggregate a large set of real-world music-related user needs, and evaluate how much of it is expressible within our ontological framework. This gives us a quantitative measure of how well our ontology could support a system addressing these real-world user needs. We also provide some statistical insights in terms of lexical coverage for comparison with related description frameworks and identify areas within the ontology that could be improved.

1 Introduction

The Music Ontology [19] was first published in 2006 and provides a framework for distributing structured music-related data on the Web. It has been used extensively over the years, both as a generic model for the music domain and as a way of publishing music-related data on the Web.

Until now the Music Ontology has never been formally evaluated and compared with related description frameworks. In this paper, we perform a quantitative evaluation of the Music Ontology framework. We want to validate the Music Ontology with regards to its intended use, and to get a list of areas the Music Ontology community should focus on in further developments.

As more and more BBC web sites are using ontologies [20], we ultimately want to reach a practical evaluation methodology that we can apply to other domains. Those ontologies are mainly written by domain experts, and we would need to evaluate how much domain data they can capture. We would also need to identify possible improvements in order to provide relevant feedback to domain experts.

We first review previous work on ontology evaluation in § 2. We devise our evaluation methodology in § 3, quantifying how well real-world user-needs fit within our Music Ontology framework. We perform in § 4 the actual evaluation, and compare several alternatives for each step of our evaluation process. We discuss the results and conclude in § 5.

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 255–269, 2012.

2 Techniques for Ontology Evaluation

Brewster et al. argue that standard information retrieval or information extraction evaluation methodologies, using the notion of *precision* and *recall*, are not appropriate for ontology evaluation [7]. We need different evaluation methodologies to evaluate knowledge representation frameworks. Ultimately, we need an ontology evaluation metric in order to easily assess ontologies and to track their evolution [23]. In this article, we design such an evaluation metric and apply it to the Music Ontology framework. We review different ontology evaluation methodologies in § 2.1, § 2.2 and § 2.3 and explain why they are not suitable for evaluating our Music Ontology framework. We focus on two evaluation paradigms in § 2.4 and § 2.5 that constitute the basis of our evaluation methodology.

2.1 Qualitative Evaluation

One way to qualitatively evaluate an ontology is to take a set of users and ask them to rate the ontology according to a number of criteria. The OntoMetric evaluation methodology [15] includes a number of such qualitative metrics. Zhang and Li [24] evaluate two multimedia metadata schemes by asking diverse groups of users to rate usefulness of individual metadata fields according to each generic user task defined by the Functional Requirements for Bibliographic Records (FRBR [17]): finding, identifying, selecting and obtaining.

Qualitative ontology evaluations have value especially when evaluating against intended use but raise several problems. It is difficult to choose the right set of users (they could be ontologists, end-users or domain experts), and it is difficult to find an actual scale on which to rate particular criteria of the ontology (what do we mean by a model being "good"?). We also want our ontology evaluation methodology to be as automatable as possible in order to integrate continuous evaluation within our development and publishing workflow, as suggested in [13]. Each ontology release needs to have a positive impact on the evaluation results. For these reasons we do not consider performing a qualitative evaluation of our Music Ontology framework.

2.2 Structural and Ontological Metrics

A number of ontology evaluation metrics can be derived automatically. Amongst these we distinguish between *structural* and *ontological* metrics [23].

Structural Metrics. Web ontologies are defined through an RDF graph. This graph can be analysed to derive evaluation metrics. These metrics, evaluating the structure of the graph defining the ontology but not the ontology itself, are called structural metrics. For example the AKTiveRank system [1] includes a metric quantifying the average amount of edges in which a particular node corresponding to a concept is involved. This metric therefore gives an idea of how much detail a concept definition in the evaluated ontology holds. Another set of examples are the structural ontology measures defined in [9], including maximum

and minimum depth and breadth of the concept hierarchy. Such metrics do not capture the intended use of the evaluated ontology. We therefore do not consider using structural metrics in our evaluation.

Ontological Metrics. Ontological metrics evaluate the actual models instead of their underlying graph structure. The OntoClean methodology [11] evaluates modelling choices in ontologies from a philosophical stand-point. It defines a number of criteria that need to be satisfied. For example a subsumption relationship cannot be drawn between concepts that have different identity criteria—a time interval cannot be a sub-concept of a time duration. OntoClean relates more to ontology engineering than ontology evaluation [23]. It can be seen as a set of ontology design guidelines. These guidelines were used when designing the Music Ontology and the underlying ontologies [19].

2.3 Similarity to a "Gold-Standard"

If we have access to a "gold-standard" (a canonical model of a particular domain) we can evaluate other ontologies of that domain by measuring their similarities to that canonical model. A set of measures for describing the similarity of different ontologies (both at the lexical and at the conceptual level) is proposed in [16]. We do not have such a gold-standard ontology, so this approach can be dismissed for evaluating our Music Ontology framework.

2.4 Task-Based Evaluation

Another way of evaluating an ontology is to measure its performance on a specific task [18]. A given task is chosen, as well as a corresponding gold-standard for perfect performance. Then, we consider the following errors when trying to fulfill that task in a particular ontology-driven application:

- Insertion errors (some terms in the ontology are not necessary);
- Deletion errors (missing terms);
- Substitution errors (ambiguous or ill-defined terms).

2.5 Data-Driven Ontology Evaluation

Brewster et al. provide a data-driven approach for ontology evaluation [7]. They use a corpus of text within the domain modelled by the ontology. They extract terms from it and try to associate them with terms in the ontology to evaluate, which leads to a measure for the domain coverage of the ontology. In order to evaluate the structure of the ontology, they cluster the extracted terms and quantify the extent to which terms in the same cluster are closer in the ontology than terms in different clusters. Elhadad et al. use a similar methodology to evaluate an ontology in the movies domain against a corpus of movie reviews [8], although they focus on the coverage of ontology instances.

3 A Query-Driven Ontology Evaluation Methodology

We now devise our methodology for evaluating our Music Ontology framework, based on the the data-driven and the task-based evaluation methodologies described in § 2.4 and § 2.5. We want this evaluation methodology to allow us to validate our ontology with regards to real-world information-seeking behaviours.

We consider evaluating our knowledge representation framework against a dataset of verbalised music-related user needs. We isolate a set of music-related needs drawn from different sets of users, and we measure how well a music information system backed by our knowledge representation frameworks could handle these queries. Our evaluation methodology involves the following steps:

3.1 Step 1 - Constructing a Dataset of Verbalised User Needs

We start by constructing a dataset of verbalised user needs. We perform a similar evaluation process on several datasets of verbalised user queries available online. We can distinguish amongst several communities of users, and our Music Ontology framework might perform differently for each of them. We want to evaluate our ontology for each of these communities.

3.2 Step 2 - Extracting Query Features

We analyse these needs to extract a set of *features* — recurrent patterns used to describe the information need, e.g. "the name of the artist was X" or "the lyrics mentioned Y". We consider several alternatives for extracting features from our dataset.

- We can use the results of previous works in extracting query features from similar datasets;
- We can extract features from the dataset by following a statistical approach;
- We can manually extract features from a random sample of the dataset.

We also consider extracting a weight w_f for each feature f, capturing the relative importance of f within the dataset. Moreover, these weights are normalised so that their sum is equal to one.

$$w_f = \frac{\text{number of queries that contain the feature } f}{\sum_g \text{number of queries that contain the feature } g} \tag{1}$$

3.3 Step 3 - Computing the Ontology Fit

We now evaluate how well these features map to our knowledge representation framework. The corresponding measure captures the *ontology fit*. The Music Ontology was designed to not duplicate terms that could be borrowed from other web ontologies (for example, `foaf:Person`, `dc:title` or `po:Broadcast`). We take into account this design choice. In the last step of our evaluation process

we therefore also consider terms from FOAF[1], Dublin Core[2] and the Programmes Ontology[3].

We develop an ontology fit measure capturing how well the extracted features can be mapped to our ontology. For a query feature f, we define δ as follows.

$$\delta(f) = \begin{cases} 1 & f \text{ is expressible within the ontology} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Our ontology fit measure for a set of verbalised queries Q is then the weighted sum of the $\delta(f)$ for each feature f extracted from Q.

$$\Delta = \sum_f w_f \cdot \delta(f) \tag{3}$$

The ontology fit measure Δ therefore captures how possible it is to map a set of user needs to queries expressed within our Music Ontology framework. The closer Δ is to one, the more our ontology can be used to express the dataset of user queries. We use the ontology fit measure to validate our ontology with regards to real-world user needs.

3.4 Discussion

This 'query-driven' evaluation methodology corresponds to a particular kind of a task-based evaluation methodology, where the task is simply to be able to answer a set of musical queries and the evaluation metric focuses on coverage (the insertion and substitution errors are not considered). The gold-standard associated with this task is that such queries are fully expressed in terms of our knowledge representation framework — the application has a way to derive accurate answers for all of them. This evaluation methodology also corresponds to a particular kind of data-driven evaluation. We start from a corpus of text, corresponding to our dataset of verbalised user needs, which we analyse and try to map to our knowledge representation framework.

A similar query-driven evaluation of an ontology-based music search system is performed by Baumann et al. [4]. They gather a set of 1500 verbalised queries issued to their system, which they cluster manually in five different high-level categories (requests for artists, songs, etc.) in order to get insights on the coverage of their system. We use a similar methodology, although we define a quantitative evaluation measure which takes into account much more granular query features. We also consider automating steps of this evaluation process.

4 Evaluation of the Music Ontology Framework

We want to evaluate our representation framework against a large dataset, holding musical needs drawn from a wide range of users. We consider two main

[1] `http://xmlns.com/foaf/spec/`
[2] `http://dublincore.org/`
[3] `http://www.bbc.co.uk/ontologies/programmes`

categories of users: casual users and users of music libraries. We derive an ontology fit measure for each of these categories.

4.1 Casual Users

We first consider verbalised queries drawn from casual users. We measure how well these queries can be expressed within our Music Ontology framework using our ontology fit measure. We consider the three alternatives mentioned in § 3.2 for extracting features from a dataset of verbalised user queries.

Evaluating Against Previous Studies of User Needs. We consider evaluating our knowledge representation framework using previous analysis of casual user needs. Such analysis leads to the categorisation of the query *type* (e.g. queries aiming at identifying a particular musical item, queries aiming at researching a particular aspect of a musical work), of the query *context* (the intended use for the requested information) and of the query *features* (recurrent patterns used to describe the information need). We are especially interested in the categorisation of query features as it leads directly to the results of the second step of our evaluation methodology.

Bainbridge et al. [2] analyse 502 queries gathered from Google Answers[4]. Google Answers allows users to post a particular question, which others can answer. Lee et al. [14] analyse 566 queries from the same source, restricting themselves to queries aiming at identifying a particular recording or a particular artist. Both extract a set of recurrent features in such queries. The extracted features along with their correspondences to Music Ontology terms are summarised in table 4.1.

The corresponding ontology fit Δ is 0.828 for Lee's analysis and 0.975 for Bainbridge's analysis.

Such ontology fit measures are arguable. The proper term to choose within the Music Ontology framework for one of these features is highly dependent on its actual context. It might be the case that one of these features is expressible within our framework in one context, but not in another. For example for the "related work" feature "it was a cover of a" is expressible, but "it was in the charts at the same time as a" is not. The features reported in these studies are too general to provide a solid basis for deriving an ontology fit measure.

Corpus-driven Evaluation. We now perform an evaluation inspired by the data-driven evaluation proposed by Brewster et al. [7] and reviewed in § 2.5. We use a statistical analysis on a dataset of user queries to derive information features, and we try to map the results of such an analysis onto Music Ontology terms.

We sample verbalised user needs from both Google Answers and Yahoo Questions[5]. We aggregated the whole Google Answers archive in the music category

[4] Google Answers archives are available at `http://answers.google.com/`, as the service is no longer running.

[5] Yahoo Questions is available at `http://answers.yahoo.com/`

Table 1. Comparison of the features identified in [2] and in [14] along with corresponding Music Ontology terms

Features used in queries	Music Ontology term	% of queries containing the feature	
		Bainbridge et al. [2]	Lee et al. [14]
Lyrics	`mo:Lyrics`	28.9	60.6
Date	`event:time`	31.9	59.2
Media	`mo:Medium`	-	44.0
Genre	`mo:genre`	32.7	35.5
Uncertainty	-	-	30.7
Lyrics description	-	-	30.0
Used in movie/ad	`po:track`	-	30.0
Gender of artist	`foaf:gender`	-	20.5
Musical style	`event:factor`[a]	-	19.8
Artist Name	`foaf:name`	55.0	19.3
Orchestration	`mo:Orchestration`	13.5	16.8
Related work	`mo:MusicalWork`	-	15.9
Lyrics topic	`dc:subject`	2.6	15.4
Where heard	`event:place`	24.1	14.7
Affect/Mood	-	2.4	14.0
Musical Work	`mo:MusicalWork`	35.6	13.6
Used in scene	`mo:Signal`	-	13.3
Audio/Video example	-	4.4	10.8
Similar	`musim:Similarity`	4.6	9.2
Tempo	`mo:tempo`	2.4	7.6
Nationality of music/artist	`fb:nationalityNoun`	12.5	4.2
Related event	`event:Event`	-	4.2
Language	`dc:language`	2.0	3.7
Record	`mo:Record`	12.2	2.7
Melody description	`so:Motif`	-	0.7
Label	`mo:Label`	5.4	0.1
Link	`foaf:page`	2.9	-

[a] To express that a particular performance has a given stylistic influence, we add this influence as a factor of the performance.

(3318 verbalised user needs) and a subset of Yahoo Questions (4805 verbalised user needs). Most user queries include editorial information (artist name, track name etc.), as spotted in previous analyses of similar datasets. When including some information about a musical item this information will most likely be related to vocal parts: singer, lyrics etc. The three most cited musical genres are "rock", "classical" and "rap". The queries often include information about space and time (e.g. when and where the user heard about that song). They also include information about the access medium: radio, CD, video, online etc. A large part of the queries include personal feelings, illustrated by the terms "love" or "like". Finally, some of them include information about constituting parts of a particular musical item (e.g. "theme").

We could consider the words occurring the most in our dataset as query features and their counts as a weight. However, the same problem as in the ontology fit derived in § 4.1 also arises. The Music Ontology term corresponding to one of these features is highly context-dependent. There are two ways to overcome these issues. On the one hand, we can keep our evaluation purely on a *lexical* level. We are particularly interested in such an evaluation because it allows us to include other music-related representation frameworks which are not ontologies but just specifications of data formats, therefore providing some insights for comparison. On the other hand, we can extract underlying topics from our corpus of verbalised user needs, and consider these topics as query features. We therefore move our evaluation to the *conceptual* level.

Evaluation at the Lexical Level. We now derive a measure of the lexical coverage of our ontology. We first produce a vector space representation of these verbalised user needs and of labels and comments within the Music Ontology specification. We first remove common stop words. We then map the stemmed terms to vector dimensions and create vectors for our dataset and our ontology using `tf-idf`. We also include in our vector space other music-related representation frameworks. We finally compute cosine distances between pairs of vectors, captured in table 2.

We first note that the results in this table are not comparable with the ontology fit results derived in the rest of this article. They are not computed using the same methodology as defined in § 3. We note that our Music Ontology framework performs better than the other representation framework — it is closer to the dataset of user queries. These results are due to the fact that our ontology encompasses a wider scope of music-related information than the others, which are dedicated to specific use-cases. For example XSPF is specific to playlists, iTunes XML and hAudio to simple editorial metadata, Variations3 to music libraries and AceXML to content-based analysis and machine learning. The lexical coverage of the Music Ontology framework is therefore higher. Of course this measure is very crude and just captures the lexical overlap between specification documents and our dataset of user queries. It can serve for comparison purposes, but not to validate our framework against this dataset.

Evaluation at the conceptual level. We now want to go beyond this lexical layer. We try to extract from our dataset a set of underlying *topics*. We then consider these topics as our query features and compute an ontology fit measure from them by following the methodology described in § 3.

We consider that our corpus of musical queries reflects the underlying set of topics it addresses. A common way of modelling the contribution of these topics to the $i^{\mathtt{th}}$ word in a given document (in our case a musical query) is as follows.

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j) \cdot P(z_i = j) \tag{4}$$

where T is the number of latent topics, z_i is a latent variable indicating the topic from which the $i^{\mathtt{th}}$ word was drawn, $P(w_i|z_i = j)$ is the probability of the word

Table 2. Cosine similarities between vectorised specification documents and the casual users dataset. We use labels and descriptions of terms for web ontologies and textual specifications for other frameworks.

Ontology	Similarity
Music Ontology	0.0812
ID3 version 2.3.0	0.0526
hAudio	0.0375
Musicbrainz	0.0318
XSPF	0.026
ACE XML	0.0208
iTunes XML	0.0182
ID3 version 2.4.0	0.0156
Variations3 FRBR-based model, phase 2	0.0112
FRBR Core & Extended	0.0111
MODS	0.0055
MPEG-7 Audio	0.0013

w_i under the $j^{\texttt{th}}$ topic, and $P(z_i = j)$ is the probability of choosing a word from the $j^{\texttt{th}}$ topic in the current document. For example in a corpus dealing with performances and recording devices, $P(w|z)$ would capture the content of the underlying topics. The performance topic would give high probability to words like venue, performer or orchestra, whereas the recording device topic would give high probability to words like microphone, converter or signal. Whether a particular document concerns performances, recording devices or both would be captured by $P(z)$.

The Latent Dirichlet Allocation (LDA) [6] provides such a model. In LDA, documents are generated by first picking a distribution over topics from a Dirichlet distribution which determines $P(z)$. We then pick a topic from this distribution and a word from that topic according to $P(w|z)$ to generate the words in the document. We use the same methodology as in [10] to discover topics.

We use an approximate inference algorithm via Gibbs sampling for LDA [12]. We first pre-process our dataset of musical queries by stemming terms, removing stop words and removing words that appear in less than five queries. Repeated experiments for different number of topics (20, 50, 100 and 200) suggest that a model incorporating 50 topics best captures our data. We reach the set of topics illustrated in table 4.1.

We consider these topics as our query features. For each topic, we use its relative importance in the dataset as a feature weight. We manually map each topic to terms within our Music Ontology framework to derive the ontology fit measure described in § 3.3. The corresponding ontology fit measure is 0.723.

However, this measure of the ontology fit is still arguable. Some of the topics inferred are not sufficiently precise to be easily mapped to Music Ontology terms. A subjective mapping still needs to be done to relate the extracted topics with a set of ontology terms. Moreover, some crucial query features are not captured within the extracted topics. For example a lot of queries include an implicit

Table 3. Top words in the first six topics inferred through Latent Dirichlet Allocation over our dataset of musical queries

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
band	album	song	music	play	live
rock	track	singer	piece	piano	concert
metal	artist	female	classical	note	perform
member	cover	sung	sheet	chord	tour
punk	release	male	composition	key	date
drummer	title	lead	instrument	tuning	ticket
guitarist	name	vocalist	score	scale	opera
quit	purchase	chorus	piano	melody	stage
beatles	bought	artist	orchestra	major	show
zeppelin	obtain	sound	choral	minor	play

notion of uncertainty (such as "I think the title was something like Walk Away"), which is not expressible within our ontology.

A possible improvement to the evaluation above would be to use a Correlated Topic Model [5], which also models the relationships between topics. This would allow us to not only evaluate the coverage of concepts within our ontology, but also the coverage of the relationships between these concepts. It remains future work to develop an accurate measure for evaluating how the inferred graph of topics can be mapped to an ontological framework. Another promising approach for ontology evaluation is to estimate how well a generative model based on an ontology can capture a set of textual data.

Manual Evaluation of the Ontology Fit. We now want to derive a more accurate ontology fit measure. In order to do so, we manually extract the underlying logical structure of these queries and see how well these logical structures can be expressed within our representation framework.

We consider a random sample of 40 queries drawn from the dataset of user needs used in § 4.1, corresponding to 0.5% of the queries available within the Google Answers archive.

We manually pre-process every verbalised query q in this sample to extract a set $\alpha(q)$ of query features. These features are recurring logical sentences encoding the queries. In order to mimise the bias in this manual step, we use the following methodology. We use predicates defined within the Music Ontology framework when they exist. When encountering unknown features, we define new predicates using the same FRBR and event-based model as used by the Music Ontology. For example a query holding the sentence "the composer of the song is Chet Baker" would lead to the following two features:

$$\alpha(q) = \{\mathsf{composer}(S, P), \mathsf{name}(P, \text{'Chet Baker'})\}$$

We do not follow exactly the manual data analysis methodology used by Lee et al. [14], which partially structures the original queries by delimiting the parts that correspond to a particular recurrent feature. Indeed, it is important for our

Table 4. Predominant query features in a random sample of the Google Answers dataset, along with their weights and corresponding terms in the Music Ontology framework

Feature	Weight	Corresponding term
title(Item, Title)	0.085	`dc:title`
maker(Item, Maker)	0.058	`foaf:maker`
lyrics(Item, Lyrics)	0.054	`mo:Lyrics`
time(Item, Time)	0.042	`dc:date`
uncertain(Statement)	0.042	-
similar(Item1, Item2)	0.035	`musim:Similarity`
based_near(Person, Place)	0.035	`foaf:based_near`
place(Event, Place)	0.031	`event:place`

purpose that we extract the whole logical structure of the query. This will lead to a more accurate ontology fit measure (but derived from a smaller dataset) than in the previous sections.

Once these queries have been pre-processed, we assign a weight for each distinct feature. Such weights are computed as described in § 3.2. We give the main query features, as well as the corresponding weight and the corresponding Music Ontology term, in table 4. We then compute our ontology fit measure as described in § 3.3. We find an ontology fit measure of 0.749.

Discussion. The different variants of our query-driven evaluation made in this section all lead to a similar ontology fit measure. Around 70% of the information held within a dataset of casual user queries is expressible within our Music Ontology framework.

Over the different evaluations made in this section, we found that the main features that are not expressible within our framework are the following.

- Uncertainty - e.g. "I don't remember if the song had drums in it";
- Partial characterisation of the lyrics - e.g. "One part of the lyrics was 'put your hands up' "[6];
- Emotions related to the music itself - e.g. "This song was really sad";
- Description of related media - e.g. "In the music video, the band was playing in a forest and the singer was trapped under ice" or "The artist was on the cover of that magazine";
- Other cultural aspects, such as the position of a track in the charts.

Future work on the Music Ontology should therefore focus on these points.

4.2 Users of Music Libraries

Gathering a dataset of music library user needs is difficult. We therefore adapt our approach to cope with the lack of publicly available datasets.

[6] This particular point is certainly made important by the bias our dataset has towards English-speaking users. For example Baumann [3] reports the case of a user stating "I am not interested in lyrics in general, because my English is too bad to understand something".

Methodology. Saxton and Richardson [21] present an evaluation methodology for reference services in libraries based on the sampling of real-world questions and on the evaluation of the corresponding transactions on a number of dimensions, including completeness, usefulness, user satisfaction and accuracy. Sugimoto [22] isolates such reference questions in order to evaluate the performance of music libraries. He then analyses the corresponding transactions for a number of music libraries in the United States.

We evaluate our representation framework using a similar methodology. We consider a reference set of queries covering a wide range of possible query types. We then manually extract query features, and follow the process described in § 3 to derive an ontology fit measure. We therefore evaluate the performance of the ontology by quantifying how an ontology-backed system would perform if it were occupying the role of the librarian. Such a methodology is similar to the methodology we adopted in § 4.1, except that we filter the dataset to leave a small sample of representative questions prior to the actual evaluation instead of using a random sample of questions. The accuracy of such an evaluation is arguable as it does not include information about the predominance of a query feature in a real-world dataset. However, it gives us a measure of how well a representative set of query features is covered by our ontology.

Dataset of User Queries. In order to cope with the lack of data availability for this category of users, we consider re-using the questions selected in Sugimoto's study [22] from a binder of recorded reference questions asked at the University of North Carolina Chapel Hill Music Library between July 15, 1996 and September 22, 1998. These questions were chosen to cover a typical range of possible questions asked in a music library. These questions are:

1. What is the address for the Bartok Archive in NY?
2. Can you help me locate Civil War flute music?
3. I am a percussion student studying the piece "Fantasy on Japanese Wood Prints" by Alan Hovhaness. I wondered if there was any information available about the actual Japanese wood prints that inspired the composer. If so, what are their titles, and is it possible to find prints or posters for them?
4. Do you have any information on Francis Hopkinson (as a composer)?
5. What are the lyrics to "Who will Answer"? Also, who wrote this and who performed it?

Ontology Fit We start by extracting features from these five queries as in § 4.1. We reach a set of query features and associated weights leading to an ontology fit measure of 0.789. Our ontology therefore performs slightly better for this particular dataset than for the casual users dataset. Almost 80% of the information is expressible within our Music Ontology framework. These results can be explained by the diversity of the queries drawn from casual users. For example one query analysed within § 4.1 describes in great levels of detail a music video in order to get to the name of a track. Such descriptions are not expressible within our framework and lower the ontology fit.

5 Conclusion

In this article we devised a query-driven evaluation process for music ontologies based on the data-driven and task-based ontology evaluation methodologies. We created a dataset of user queries and measure how well these queries fit within our knowledge representation framework. We end up quantifying how well a system based on our representation framework could help answering these queries.

A number of alternatives can be used for each step of such an evaluation process. First there are several categories of users which are interesting to handle separately as our ontology may perform differently for each. Then there are several ways of performing an analysis of user queries. We summarise in table 5 the results obtained in this article, investigating different alternatives for each of these steps. Our ontology covers more than 70% of the different datasets considered. We identified the main features that are lacking from our ontology in § 4.1.

Table 5. Summary of the ontology fit results described in this article

Dataset	Evaluation method	Section	Ontology fit (Δ)
Casual users	Using results of previous analysis	§ 4.1	0.828 for [14] 0.975 for [2]
	Statistical analysis	§ 4.1	0.723
	Manual analysis	§ 4.1	0.749
Music library users	Manual analysis	§ 4.2	0.789

We also performed a lexical comparison of different music representation frameworks. This comparison captured how lexically close a particular representation framework is from a dataset of casual user queries. We found that our Music Ontology framework performs better than the others according to this metric.

All the results described in this article evaluate a particular characteristic of our ontology: its coverage of real-world user needs. However, a number of other characteristics would be interesting to capture as well. For example we might want to evaluate the *verbosity* of the ontology – how many ontology terms are needed to express a particular information feature. We might also want to evaluate the *popularity* of the ontology – how many documents reusing Music Ontology terms are available on the Web.

Future work includes using a similar methodology to evaluate other ontologies used within the BBC web site. As those ontologies are mostly built by domain experts we are planning on evaluating how much domain data they can actually capture and use the results of this evaluation to identify possible improvements.

References

1. Alani, H., Brewster, C.: Metrics for ranking ontologies. In: Proceedings of the 4th Int. Workshop on Evaluation of Ontologies for the Web (2006)
2. Bainbridge, D., Cunningham, S.J., Downie, S.J.: How people describe their music information needs: A grounded theory analysis of music queries. In: Proceedings of the 4th International Conference on Music Information Retrieval (2003)
3. Baumann, S.: A music library in the palm of your hand. In: Proceedings of the Contact Forum on Digital libraries for musical audio (Perspectives and tendencies in digitalization, conservation, management and accessibility), Brussels (June 2005)
4. Baumann, S., Klüter, A., Norlien, M.: Using natural language input and audio analysis for a human-oriented MIR system. In: Proceedings of Web Delivery of Music (WEDELMUSIC) (2002)
5. Blei, D.M., Lafferty, J.D.: A correlated topic model of. The Annals of Applied Statistics 1(1), 17–35 (2007)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. The Journal of Machine Learning Research 3(3), 993–1022 (2003)
7. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data driven ontology evaluation. In: Proceedings of the International Conference on Language Resources and Evaluation, Lisbon, Portugal, pp. 164–168 (2004)
8. Elhadad, M., Gabay, D., Netzer, Y.: Automatic Evaluation of Search Ontologies in the Entertainment Domain using Text Classification. In: Applied Semantic Technologies: Using Semantics in Intelligent Information Processing. Taylor and Francis (2011)
9. Fernández, M., Overbeeke, C., Sabou, M., Motta, E.: What Makes a Good Ontology? A Case-Study in Fine-Grained Knowledge Reuse. In: Gómez-Pérez, A., Yu, Y., Ding, Y. (eds.) ASWC 2009. LNCS, vol. 5926, pp. 61–75. Springer, Heidelberg (2009)
10. Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences (2004)
11. Guarino, N., Welty, C.: Evaluating ontological decisions with ONTOCLEAN. Communications of the ACM 45(2), 61–65 (2002)
12. Heinrich, G.: Parameter estimation for text analysis. Technical report, University of Leipzig & vsonix GmbH, Darmstadt, Germany (April 2008)
13. Lavbic, D., Krisper, M.: Facilitating ontology development with continuous evaluation. Informatica 21(4), 533–552 (2010)
14. Ha Lee, J., Stephen Downie, J., Cameron Jones, M.: Preliminary analyses of information features provided by users for identifying music. In: Proceedings of the International Conference on Music Information Retrieval (2007)
15. Lozano-Tello, A., Gomez-Perez, A.: ONTOMETRIC: A method to choose the appropriate ontology. Journal of Database Management 15(2), 1–18 (2004)
16. Maedche, A., Staab, S.: Measuring Similarity between Ontologies. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 251–263. Springer, Heidelberg (2002)
17. IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records - final report. UBCIM Publications - New Series, vol.19 (September 1998), `http://www.ifla.org/VII/s13/frbr/frbr1.htm` (last accessed March 2012)
18. Porzel, R., Malaka, R.: A task-based approach for ontology evaluation. In: Proceedings of the ECAI Workshop on Ontology Learning and Population (2004)

19. Raimond, Y., Abdallah, S., Sandler, M., Giasson, F.: The music ontology. In: Proceedings of the International Conference on Music Information Retrieval, pp. 417–422 (September 2007)
20. Raimond, Y., Scott, T., Oliver, S., Sinclair, P., Smethurst, M.: Use of Semantic Web technologies on the BBC Web Sites. In: Linking Enterprise Data, pp. 263–283. Springer (2010)
21. Saxton, M.L., Richardson, J.V.: Understanding reference transactions. Academic Press (May 2002)
22. Sugimoto, C.R.: Evaluating reference transactions in academic music libraries. Master's thesis, School of Information and Library Science of the University of North Carolina at Chapel Hill (2007)
23. Vrandečić, D., Sure, Y.: How to Design Better Ontology Metrics. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 311–325. Springer, Heidelberg (2007)
24. Zhang, Y., Li, Y.: A user-centered functional metadata evaluation of moving image collections. Journal of the American Society for Information Science and Technology 59(8), 1331–1346 (2008)