SOCIAL TAGGING AND MUSIC INFORMATION RETRIEVAL

Paul Lamere
Sun Labs
Sun Microsystems
Burlington, Mass, USA
Paul.Lamere@sun.com

ABSTRACT

Social tags are free text labels that are applied to items such as artists, albums and songs. Captured in these tags is a great deal of information that is highly relevant to Music Information Retrieval (MIR) researchers including information about genre, mood, instrumentation, and quality. Unfortunately there is also a great deal of irrelevant information and noise in the tags. Imperfect as they may be, social tags are a source of human-generated contextual knowledge about music that may become an essential part of the solution to many MIR problems.

In this article, we describe the state of the art in commercial and research social tagging systems for music. We describe how tags are collected and used in current systems. We explore some of the issues that are encountered when using tags, and we suggest possible areas of exploration for future research.

1 INTRODUCTION

The multi-disciplinary field of Music Information Retrieval (MIR) is focused on extracting information from music and using this information to solve a wide range of problems including beat detection, automatic music transcription, artist recognition, genre classification and music recommendation. To solve these problems, researchers must develop algorithms that can extract salient musical information directly from the audio. This can be extremely difficult. MIR researcher Don Byrd describes the problem in the following way:

> Music is created by humans for other humans, and humans can bring a tremendous amount of contextual knowledge to bear on anything they do; in fact, they can't avoid it, and they're rarely conscious of it. But (as of early 2008) computers can never bring much contextual knowledge to bear, often none at all, and never without being specifically programmed to do so. Therefore doing almost anything with music by computers is very difficult; many problems are essentially intractable [7].

Yet despite this inherent difficulty, MIR researchers are trying to solve ever more complex problems. To do this, researchers need to expand the pool of contextual knowledge about music. One source for this contextual knowledge is the set of social tags that humans apply to music.

Social tags are the result of collaborative tagging. In a social tagging system, an individual applies short, free text annotations (tags) to items, typically to organise their personal content. These tags can be combined with those created by other individuals to form a collective body of social tags. With a large enough set of taggers generating many tags, a very rich view of the tagged items emerges. Social tags are typically used to facilitate searching for items, exploring for new items, finding similar items, and finding other listeners with similar interests.

In the last few years, a number of music discovery and recommendation web sites have been supporting the social tagging of music. Commercial sites such as Last.fm, MyStrands and Qloud allow their users to apply social tags to tracks, albums, artists and playlists. These tags capture a great deal of information that is highly relevant to MIR researchers including information about genre, mood, instrumentation and quality.

In this article we describe the state of the art in commercial and research social tagging systems for music. We describe how tags are collected and used in current systems. We explore some of the issues and problems that are encountered when using tags, and we suggest possible areas of exploration for future research.

In this article a number of examples are presented that show tags and frequencies. This tag data was retrieved from Last.fm via the Audioscrobbler web services [3] during the spring of 2007. The data consists of over 7 million artist-level tags applied to 280,000 artists. 122,000 of the tags are unique.

2 WHAT ARE SOCIAL TAGS?

Tags are merely unstructured labels that are assigned to a resource. Unlike traditional keyword assignment where terms are often drawn from a controlled vocabulary, no restrictions are placed on the makeup of a tag. A typical tag is a word or short phrase that describes the resource. There is usually no restriction on the number of tags that can be assigned to an item. These tags are generally as-

signed by a non-expert for their own personal use, such as for personal organisation or to assist with future retrieval. The real value of these tags emerges when the tags are aggregated into a single, shared pool, sometimes referred to as a folksonomy [32]. This sharing of tags enhances the value of tags for all users. To further explore the value of social tagging, let us look at the common music information retrieval problem of genre classification.

An ongoing problem for music librarians and editors is how to represent the music genre taxonomy. Music genre is widely used to classify music and to aid in retrieval and discovery. However there is no general agreement as to the correct genre taxonomy. One study [2] compared three commercial genre taxonomies and found little agreement among them. There was no consensus as to the names used in the classifications. For example, "Rock" and "Pop" did not denote the same set of songs, the hierarchy differed from one taxonomy to another and there were semantic inconsistencies within taxonomies. They found that building a consistent genre hierarchy was extremely difficult even for domain experts. They eventually abandoned their effort to build such a genre taxonomy, citing multiple difficulties in establishing a consistent, understandable taxonomy.

In the article "Ontology is Overrated" [25], Clay Shirkey explores the difficulties in creating such taxonomies. He suggests that characteristics necessary for successful ontological classification include a small corpus, formal categories with clear delineations between categories and stable, unchanging entities. Clearly these characteristics are not found in genre classification. The body of recorded music encompasses millions of songs, genres overlap, are inconsistent, and ambiguous. The boundaries between genres are unstable, and fuzzy. New genres are introduced regularly and existing genres change over time [21]. With all of these troublesome characteristics, it is not surprising that genre classification is so difficult.

At a music site such as Last.fm, a user may tag a number of songs as "Mellow" some songs as "Energetic" some songs as "Guitar" and some songs as "Punk". The labels may overlap (a single song may be labeled "Energetic" "Guitar" and "Punk" for example). Typically, a music listener will use tags to help organise their listening. A listener may play their "Mellow" songs during the evening meal, and their "Energetic" artists while they exercise.

When the tags created by thousands of different listeners are combined, a rich and complex view of the song or artist emerges. Table 1 shows the top 24 tags and frequencies of tags applied to the band *Deerhoof* by listeners at Last.fm. Users have applied tags associated with genre (Indie, Noise rock, Rock, Art Punk, etc.), mood (Fun), opinion (Weird), recording label (Kill rock stars), instrumentation (Female vocalists), style (Experimental) and context (San Francisco). From these tags and their frequencies we learn much more about *Deerhoof* than we would from the traditional single genre assignment of "Alternative" that is applied to *Deerhoof* in the iTunes music store.

Tag	Freq	Tag	Freq
Experimental	459	Avant-Garde	27
Indie	455	Electronic	22
Seen live	288	Art punk	21
Indie rock	271	Kill rock stars	19
Noise rock	154	Punk	18
Noise	129	Art rock	18
Rock	111	Weird	16
Indie pop	103	San Francisco	15
Alternative	63	American	14
Noise pop	48	Post-punk	13
Female vocalists	45	Japanese	12
Post-rock	31	Fun	12
Pop	29	Math-rock	11

Table 1. Top 24 tags applied to *Deerhoof*

Tag Type	Frequency	Examples
Genre	68%	heavy metal, punk
Locale	12%	French, Seattle, NYC
Mood	5%	chill, party
Opinion	4%	love, favorite
Instrumentation	4%	piano, female vocal
Style	3%	political, humor
Misc	2%	Coldplay, composers
Personal	1%	seen live, I own it
Organisational	1%	check out

Table 2. Distribution of tag types

With social tags we can ignore the problem of fuzzy category boundaries completely. Instead of trying to pick the best genre category for *Deerhoof* (is it "Indie", "Experimental" or "Noise Pop"?) We can tag *Deerhoof* with all three.

Social tags can be a powerful tool to assist in the classification and exploration of music. Last.fm has collected over 40 million social tags in the last five years. The majority of these tags are genre related, providing an extremely detailed map of how Last.fm listeners perceive and understand the complexities of overlapping genres. In addition to genre-related tags, there are tags related to instrumentation, music style and locale. Table 2 shows the distribution of the types of tags for the 500 most frequently applied tags at Last.fm. The majority describe audio content. Genre, mood and instrumentation account for 77% of the tags.

3 WHY DO PEOPLE TAG?

Last.fm users apply approximately 2 million tags per month (E. Pampalk, personal communication, August 2007). A recent study [33] indicates that 28% of Internet users have tagged or categorised content online. On a typical day, 7% of Internet users tag or categorise online content. Why are all of these people tagging? Do they do so for altruistic

reasons? Are people trying to organise the Internet?

A number of incentives and motivations for tagging have been suggested [1] [20]:

- Memory and Context – items are tagged to assist in personal retrieval of the item or a group of items. For instance, a listener may tag a new song with "Female", "Electronic", "Indie", "Favorite" and "Mellow" as a way of describing the song.
- Task organisation – items are tagged to assist in the organisation of music discovery tasks. For instance, a music listener may tag an artist as "Check out" as a reminder to look more closely at an artist in the future. A listener may tag groups of songs with the same tag to facilitate the listening experience. For instance, a listener may tag a selection of songs with "Emo" to create an ad hoc playlist for future listening.
- Social signaling – items are tagged as a way for an individual to express their tastes to others. For example, Last.fm listeners will often tag artists as "Seen live" as a way of indicating which artists they have seen at a concert.
- Social Contribution – items are tagged to add to the group knowledge, for the benefit of the wider audience. A fan of heavy metal music may tag a band as "Not metal" to indicate that the band is not truly a heavy metal band, but instead is a poseur.
- Play and Competition - items are tagged as a result of either an explicit game such as Major Miner[19], the Listen game [30], and Tag a Tune[16] or implicit games (tag the "Saddest song" you know).
- Opinion Expression - items are tagged to convey an individual's opinion about an item. A listener may apply the tag "Awesome" to a favourite track.

One of the primary reasons that tagging has become so popular is that tagging is easy [26]. Unlike traditional categorisation, where an item has to be placed in a single category, with tagging, no such filtering is required. An item can be tagged with all potentially applicable categories. Tagging can eliminate the so called "post-activation analysis paralysis" that occurs when trying to place an item into the single best category. The paralysis occurs out of fear that a mis-categorisation of the item will result in the item no longer being findable. With tagging, if you are unsure if a new band should be labeled "Grunge", "Post-grunge" or "Nu metal" you can apply all three without worry.

Tagging is also popular because individuals find the experience to be pleasurable [27]. Tagging an item enrolls an individual into a social circle. When an individual applies a tag, they can see other people who have used the tag, and the items that they have tagged. The social experience can be enjoyable, leading people to continue to tag more items.

Tag	**Freq**	**Tag**	**Freq**
Check out	248	French	47
Seen Live	195	Jazz	42
Female	180	Scandinavian	42
Long	83	African	19
Chamber Music	83	Cover	41
Green Man 2007	82	Drone	41
Folk	82	Favourite	39
Soundscape	7	Checked out but unconvinced so far	38
New Material from old favorites	77	German	34
Not yet available to play	70	Dylan Cover	30

Table 3. Top 20 tags applied by a frequent tagger

Table 3 shows the top twenty most frequent tags applied by a frequent tagger at Last.fm. In this set we see tags used for memory and context ("Long", "Female", "Green man 2007", "New material from old favorites") task organisation ("Check out", "Checked out but unconvinced so far"), social signaling ("Seen live"), and opinion expression ("Favourite"). This individual has applied nearly 100 unique tags to over 2,500 artists and tracks.

4 TYPES OF TAGGING SYSTEMS

The design decisions of a tagging system can affect the quality and the quantity of the tags that are generated. There are a number of attributes that can affect the usefulness of the tags generated [20]. Among these are:

- Tagging Rights - Some systems provide for *self-tagging*, where taggers can only tag items that they have contributed, whereas other systems provide for *free-for-all tagging* where anyone can attach a tag to an item. Self-tagging systems are common for sites with a large amount of user-generated content, such as YouTube or Flickr. Whereas sites that serve to help people find commercial or third-party content tend to support free-for-all tagging. Sites oriented toward music discovery such as Last.fm and Qloud tend to support free-for-all tagging.
- Tagging Support - Some tagging systems will suggest tags to a user based upon tags that the user has already applied, while other systems will suggest tags based upon those that have been applied by other users. Some tagging systems offer no tagging guidance at all. The type of tagging support can have a large effect on the types of tags collected. Systems that offer tagging suggestions tend to converge more quickly on a stable set of tags, while a blind tagging system will tend to generate a broader, more diffuse set of tags. Figure 1 shows the sug-

gestive tagging system at Last.fm. The user is presented with a set of their own favourite tags as well as a set of popular tags for the artist.

- Aggregation - Some tagging systems provide a *bag model* for representing tags, where a single tag can be applied multiple times to an item. This is contrasted with the *set model* where a tag can be applied only once per item. At Last.fm, which uses a bag model to represent tags, "Rock" has been applied to *The Beatles* over 3,200 times.

- Type of object - Tagging systems can vary in the type of objects that can be tagged. The Listen Game [30] and Major Miner [19] allow for songs to be tagged. Last.fm supports tagging of tracks, albums, artists and recording labels. MusicMobs (now defunct) supported tagging of playlists.

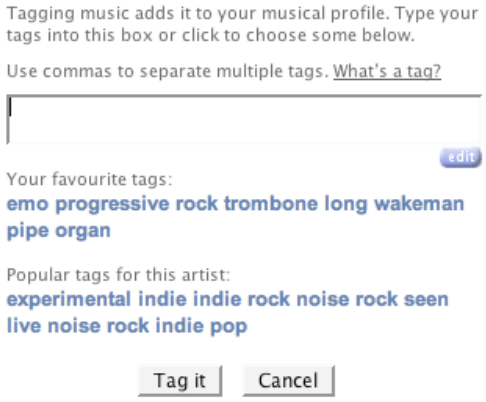

Figure 1. Last.fm's suggestive-tagging interface

5 WHAT CAN WE DO WITH TAGS?

5.1 Search and Discovery

Tags are often used to enhance simple search. For example, at Last.fm, a listener can easily find all artists that have been tagged with "Metal".

Like genre classification, social tags can be used to facilitate music exploration and discovery. Figure 2 shows the Last.fm tag cloud for the artist *Deerhoof*. In the tag cloud, a larger font indicates a more frequently applied tag. The tag cloud provides, at a glance, a rich representation of the artist. Furthermore, the tag cloud can be used to explore new music. Clicking on a particular tag in the tag cloud will show all of the artists that have been frequently tagged with that tag.

At Last.fm, tags are also used to enhance music discovery through a mechanism called "Tag Radio". With Tag radio, listeners can listen to a custom Internet streaming radio station where all of the artists or tracks played on the station have a particular tag of the listener's choosing.

Figure 2. Tag cloud for the band *Deerhoof*

5.2 Directed Search

With this technique, the search space is successively narrowed down as the user adds tags to the query. For example, a user searching for a particular type of artist may start the search with a query "Female". This reduces the artist space to include just the artists that have been tagged with "Female". The user can narrow this set of artists to only those of the current decade by adding the tag "00s" to the query. This reduces the set of selected artists to those that have been tagged with both "Female" and "00s". The user can continue to add tags to the query to narrow down the selected set of artists based on any criteria. Negations can also be used to remove artists that have been tagged with a particular tag (not "Emo" for example). Directed search gives a user much more control over how they search for music as compared to a traditional hierarchical navigation seen in most music interfaces.

5.3 Tag Similarity and Clustering

To assist in exploring and suggesting tags, some systems will cluster similar tags together when presenting them to a user [12]. These tag clusters can assist the user in either applying tags (as in a suggestive tagging system) or in identifying alternatives that may be useful in the user's search. Figure 3 shows how Last.fm presents tags that are similar to "Hip Hop". Similarity is typically determined by taking into account tag co-occurrence. Tags that are frequently applied to the same items are considered to be more similar than tags applied to disjoint sets of items. Similarity measures such as Jaccard, Overlap, Dice and Cosine distance can be used to compute such similarities [4].

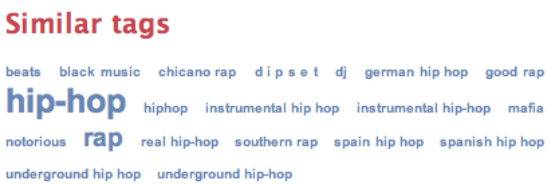

Figure 3. Tags similar to Hip Hop

Since social tags are free-form text that are applied by a large number of people, tags will typically contain all of the common misspellings and variants. In some ways

Tag	Freq	Tag	Freq
Female vocalists	100	Girl	< 1
Female	18	Women	< 1
Female vocalist	7	Chick	< 1
Female vocals	3	Ladies	< 1
Female voices	2	Femail voice	< 1
Female artists	2	Femaile voice	< 1
Female fronted	1	woman	< 1
Female vocal	1	Femaile vocalist	$\ll 1$
Female singers	1	girl music	$\ll 1$
Female fronted	1	girls with voices	$\ll 1$
Grrl	< 1	Girly girly	$\ll 1$

Table 4. Top 22 "Female vocalist"variants with relative frequency of application

this is a problem since these variants dilute the tag pool. However these variants can also be used to improve the findability of items. For example, singer Fiona Apple is often tagged to indicate that she is a female vocalist, however the taggers do not agree on the best tag to describe a female vocalist. Table 4 show the 24 most common variants and relative frequencies of "Female vocalists".

A popular female artist may be tagged with many of these variants. In some sense, the concept of a "Female vocalist" is smeared over these many tags. Some of the variants can be identified by stemming tags to their base form. "Vocalists" becomes "Vocalist", for example. Some of the variants have subtle distinction in meaning. The tag "Grrl" may be closely associated with feminism whereas "Chick" may be more closely associated with the objectification of women. The variants lead to a much more diffuse representation of the "Female vocalist". Any attempt to collapse all of these variants into a single tag such as "Female" will lose some of these nuances.

A common problem in information retrieval systems is that these systems rely on users typing in the right query words. Users will often use the wrong words to get the information that they want. This is called the vocabulary problem [11]. For example, in a traditional genre hierarchy, the terms used to create the hierarchy may not be the same terms that a user will use to query that hierarchy. A user may search for music labeled as "Electronic jazz" or "Jazztronica" whereas the domain expert has used the term "Nu jazz". Unlike traditional taxonomies, tagging system are not constrained to apply a single descriptor to an item. An artist tagged with "Nu Jazz" can also be tagged with "Electronic jazz", "Jazztronica", "Phusion", "Future jazz" and "E-jazz" as well. The problem of trying to find the right word to use in searching for an item, a problem inherent in traditional classification systems, is greatly reduced. A term that a user thinks describes the content that they are looking for was also likely used by other users to tag the item.

Even the spelling errors in tags have value. If some taggers misspell "Female" as "Femail", it is likely that others who may be searching for female vocalists may misspell "Female" as well. A tagging system can use this observation to improve tagging search by clustering spelling variants based upon tag similarity. Since the "Femail" tag overlaps nearly completely with the "Female" tag, a user that searches for "Femail" can be directed to artists that have also been tagged with "Female". Without specialized knowledge of the music domain, a traditional approach to handling spelling variants and errors using word morphology or phonetisizers can result in spelling correction mistakes. Correcting "Phusion" to "Fusion" or "Grrrl" to "Girl" would be inappropriate in certain contexts. The domain specific vocabulary derived from tags can provide this specialized knowledge to improve the handing of spelling errors.

5.4 Latent Semantic Analysis

One way of addressing the vocabulary problem encountered with tags is to use a technique called latent semantic analysis (LSA) [9]. With LSA, the tags for a music item are assumed to be an unreliable representation of a much larger and more accurate pool of tags that could be associated with that item. LSA assumes that there is some "latent" structure that relates the items and defines a semantic space. By using statistical techniques, this latent structure can be learned and tag noise can be reduced or eliminated. In the resulting latent space, tagged objects are positioned relative to each other based upon the overall pattern of tag usage. Items that are close together in this space have tags with similar meanings, even if they do not share any actual tags. A query by a tag for an item in this semantic space can return relevant items even if they have never been tagged with the query tag. By deriving a semantic space separate from the tags, problems of synonymy and polysemy of tags can be mitigated [17]. Additionally, the item similarity space defined by LSA can be used to aid in building tools for browsing and exploring the music spaces as well as for genre and mood classification [18].

5.5 Tag Overlap

With a traditional genre hierarchy, an artist or a song is placed in the single genre that best describes the item. For some artists, this is straight forward. The *Sex Pistols* are easily classified as a "Punk rock" band, for example. However, the classification of some artists is not so simple. Carrie Underwood combines "Pop" and "Country" elements in her music. However, at the iTunes music store Carrie Underwood is considered a "Country" artist, and the "Pop" elements of her music are ignored. Similarly, George Gershwin's *Rhapsody in Blue* occupies a space somewhere between "Classical" and "Jazz", but at the iTunes music store, *Rhapsody in Blue* is labeled as "Classical", the "Jazz" elements are ignored.

With a folksonomy derived from social tags, a more accurate representation of an artist or a track can be made. Table 5 shows the relative frequencies for the 10 most frequent tags applied to George Gershwin's *Rhapsody in*

Tag	Freq	Tag	Freq
Classical	100	20th Century Classical	9
Jazz	57	American	9
Instrumental	40	Classical Jazz	6
Piano	32	Composers	6
Gershwin	20	New York	6

Table 5. Top 10 tags applied to George Gershswin's *Rhapsody in Blue*

Blue. These tags position the song in a space at the border between "Classical" and "Jazz".

A genre folksonomy can better represent the fuzziness at the boundaries of the genre categories. Table 6 shows how genres overlap in the Last.fm folksonomy of artist tags. The table shows the percentage of tag overlap for 11 high-level genres. The table is not symmetric around the diagonal because the class sizes are not the same. For instance, 87.11% of "Alternative" music overlaps with "Rock", but since "Rock" is a larger class than "Alternative" only 49.81% of "Rock" overlaps with "Alternative". The overlaps demonstrate just how fuzzy genre categories really are. The average genre overlap across these 11 genres, weighted by the class size is 17.32% This overlap suggests an upper bound for automatic genre classification accuracy. Current state-of-the-art automatic genre classification accuracy of 82.34% is approaching this upper bound [5]. This fuzziness is perhaps the root cause of the increasingly smaller performance improvements in automatic genre classification [21].

Some further observations about this overlapping data can be made. Nearly 85% of all music tagged with "Rap" has also been tagged with "Hip-hop" suggesting that for Last.fm listeners, these two terms have become synonymous. Likewise, 87% of all "Alternative" music is also "Rock", suggesting that at least for new music "Alternative" is the new "Rock". "Indie" overlaps with "Alternative" by 58% suggesting that these genres are related, yet distinct. Interesting also is that nearly 11% of all artists tagged with "Classical" are also tagged with "Rock". A closer look shows that Last.fm taggers will often apply the "Classical" tag to artists that employ a backing orchestra. In some sense, for these Last.fm taggers, the sound of an orchestra is synonymous with "Classical" music.

5.6 Genre Hierarchy

Although constructing a genre hierarchy is problematic, there is still much value in representing genre as a hierarchical taxonomy. A hierarchy is a natural way to organise material in a way that facilitates browsing and discovery. Social tags can be used to reconstruct genre hierarchies. Figure 4 shows a hierarchy of the sub-genre of metal music that has been built from social tags. This hierarchy is built automatically by using the tag frequencies (to establish a top down hierarchy) and tag similarity (to determine where tags should be attached to the graph). By adjusting the parameters used to select where the tag is attached, the shape of the graph can be tuned to favor a relatively flat hierarchy or a deep hierarchy. A multi-hierarchy can be created by attaching tags to all similar tags.

Presenting tags in such a hierarchy provides a music listener with a familiar framework for browsing a genre space. However the hierarchy does not capture all of the subtle nuances of the folksonomic genre space.

5.7 Faceted Navigation

An alternative to exploring a tag space by representing the space as a hierarchy is a technique called faceted navigation [22]. In this type of navigation, the tags are organised into related groups (the facets). For music the facets could be such aspects as genre, mood, instrumentation, time period, theme, and style. With a faceted browser, a user can start browsing by picking any of the facets. For example a listener looking for music to listen to while exercising may start with the "mood" facet and select "Energetic" or "Happy". From this selection, the user may then select the "time period" facet and select "70s" to reduce the selections to music released from that time period. Finally, the user can select the "genre" facet to chose a genre such as "Pop" to restrict the choices to just the energetic, pop songs of the 70s. Unlike traditional top-down navigation, the user can chose the facets in any order that they chose. Selecting music with these constraints using a genre hierarchy (as is traditional in online music stores) would be extremely difficult.

5.8 Item Similarity

Traditional text methods can be used to relate items that share common tags. For instance, items that share many tags can be considered to be more similar than items that share few or no tags. As with tag similarity, Pearson's correlation, Jaccard distance and simple co-occurrence are typically used to infer item similarity from tags. However, these metrics tend to be dominated by frequently occurring tags such as "Rock" and "Alternative". Since these tags are applied to many items, they are not very descriptive. Item similarity can be improved if tags used to determine similarity are weighted based upon their descriptive ability. One weighting technique is called "term frequency—inverse document frequency" or $tf \cdot idf$. This technique weights terms proportionally to the frequency of occurrence of the term in a document as compared to the frequency of the term across all documents. $tf \cdot idf$ weighting can be used to give more importance to tags that are applied frequently to a particular artist but are less frequent across all artists [24]. For example, the top tags applied to the band *Weezer* at Last.fm are "Rock", "Alternative" and "Indie". These tags are applied to many artists and therefore are not very descriptive. When using $tf \cdot idf$ weighting, the top most distinctive tags for *Weezer* are "Geek rock", "Nerd rock" and "Power pop", arguably a more descriptive set of tags than "Rock", "Alternative" and "Indie". These $tf \cdot idf$ weighted tags can be used to

	rock	indie	altern	metal	electron	punk	pop	hip-hop	jazz	rap	classical
rock	100.00	33.01	49.81	16.05	9.27	16.76	16.04	3.55	3.04	2.16	0.54
indie	**52.54**	100.00	**52.87**	7.83	16.55	15.12	16.46	4.88	4.02	2.99	0.71
altern	**87.11**	**58.09**	100.00	16.39	18.82	21.47	20.92	6.03	4.69	3.51	0.86
metal	40.29	12.36	23.53	100.00	8.58	14.07	8.29	4.29	3.93	3.37	0.96
electron	26.05	29.21	30.23	9.60	100.00	10.59	18.27	10.29	8.30	4.79	1.41
punk	**56.88**	32.23	41.67	19.01	12.79	100.00	16.85	6.78	6.02	5.11	1.15
pop	**55.52**	35.79	41.40	11.43	22.51	17.18	100.00	11.04	8.52	6.81	1.41
hip-hop	19.78	17.07	19.19	9.51	20.39	11.12	17.76	100.00	10.15	**52.10**	1.39
jazz	22.20	18.43	19.58	11.44	21.57	12.96	17.97	13.32	100.00	8.22	3.01
rap	19.58	17.02	18.18	12.16	15.47	13.65	17.84	**84.87**	10.21	100.00	1.84
classical	10.66	8.79	9.63	7.51	9.81	6.65	7.96	4.89	8.08	3.98	100.00

Table 6. Percentage overlap for selected genre-related tags. Large overlaps (greater than 50%) are shown in bold.

position the item (the artist or the track for instance), in a document vector space. A similarity metric such as cosine distance [23] can be used to determine the distance between items in this space.

Item similarity based upon tags has advantages when compared to similarity derived from collaborative filtering techniques found in social music recommender systems. Typically, a recommender based upon collaborative filtering cannot offer detailed explanations about why an item was recommended beyond a statement such as "people who listened to XX also listened to YY". However, with a recommender based upon item similarity derived from the tags, a recommendation can be explained in terms of the common tags. Instead of an explanation like "people who listened to Isobel Campbell also listened to Belle and Sebastian" a recommender can offer a more meaningful, transparent explanation of the recommendation such as "we are recommending Belle and Sebastian because you like folk-influenced mellow rock with female vocalists". Studies indicate that transparent, explainable recommendations are important in establishing trust in a recommender [28].

6 ISSUES WITH TAGS

We have seen that there are many positive aspects to social tags when applied to music. However, there are some issues with social tags that can make working with tags difficult.

6.1 Cold Start

As we have seen, social tags generally are applied by music listeners as a way for the users to organise their music. These individual tags are aggregated into a single pool giving advantages to all users. However, not all items are tagged equally. Popular artists are tagged frequently, while unpopular artists are not. Similarly, new artists, artist that do not have an established listener base are not frequently tagged. Figure 5 shows the frequency of tags applied to the top 1000 artists at Last.fm. The most frequently tagged artist has been tagged about 25,000 times, while the 1000th most popular artist at Last.fm has been tagged about 1,500 times. Of the 280,000 or so artists in our dataset of Last.fm tags, about 21,000 have at least one tag, leaving about 259,000 artist with no tags at all. On average, each artist in our dataset has been tagged about 25 times. When looking at tags applied to tracks, the problem is even worse. Last.fm claims to have about 150 million tracks in its catalog, and has about 40 million tags yielding an average of 0.26 tags per track. A typical track is tagged 100 times less often than a typical artist at Last.fm.

Tags are not applied evenly – popular content is tagged much more frequently than unpopular content. Therefore, tags are much more effective when used to explore and recommend popular content. For new content, or for unpopular content there are few tags – too few to use effectively. Social tags cannot be used to aid the exploration and discovery of this untagged content. There are a few strategies employed to mitigate against this cold start problem.

Tagging Games – A number of researchers have built games similar in style to the ESP game [31] that elicit tags from game players in a way that is entertaining and addictive. These so-called *games with a purpose* take advantage of the interest in online games and use the game playing to solve important problems that would be difficult for computers to solve on their own.

The ESP game, a game where users try to guess what labels other users will apply to an image, was able to collect over 1.2 million image labels in a four month period. The developers of the ESP game suggest that with only 5000 active game players, the ESP game would provide the labels for all of the 425,000,000 images indexed by Google on the web in just one month.

Music tagging games are similar to the ESP game. Typically, players are paired up and must try to guess what labels their partner will apply to a song. Figure 6 shows the Tag a Tune game which takes a slightly different approach. In this game, two randomly paired listeners each describe a tune. The listeners must decide whether or not they are listening to the same tune based on what they hear and the description supplied by their partner.

Figure 6. Tag a Tune - a game for music and sound annotation

Last.fm	Autotagger
Indie	Fun
Indie rock	Funky
Indie Pop	Quirky
Alternative	The Good Stuff
Rock	Indie Rock
Experimental	Kill Rock Stars
Pop	Indie
New York	Acoustic Rock
American	Indie Pop
Afro-beat	fav

Table 7. Top 10 Tags for the artist *Vampire Weekend*. Note that "Kill Rock Stars" is an independent record label.

According to their respective authors, the MajorMiner game collects about five song labels per user-minute (M. Mandel, personal communication, February 2008) and the Listen Game collects six explicit labels and eighteen implicit labels per user-minute (D. Turnbull, personal communication, February 2008). These tagging rate compares quite favorably to Last.fm which, on average, collects substantially less than one tag per user per month. There is great potential in using tagging games as a means to collect tags for music. Not only can these games be used to elicit labels for full songs, they can also be used to collect labels for phrases or sections within a piece of music. Similarly, they can be used to collect labels that correlate highly with identifiable acoustic features such as instrumentation. However current implementations have been on a very small scale generating only a few thousand labels for a limited number of tracks.

An open question about tags collected by games is how these tags compare to those collected through more traditional social tagging mechanisms. Clearly there will be some differences. Personal tags such as "Seen live", "Favorite" and "I own it" that are common in social tagging systems will likely be missing in game-derived tags. Social taggers tend to tag music that they are familiar with. This may elicit more fine-grained and targeted tags as compared to game taggers that may be less familiar with the music that they are tagging. Tagging games tend to reward speed — players that tag faster are rewarded with higher scores. Tags generated by fast players may lack nuances and subtleties found in traditional social tags. Some games, such as the Listen Game, elicit labels from a fixed set of labels. The resulting set of labels are free from many of the problems encountered with social tagging systems such ambiguous or noisy tags. This set of of *small and clean* labels differ from *large and noisy* tags collected from Last.fm. One study [6] suggests that for tasks such as music similarity and recommendation where the ability to return only relevant items (high precision) is important, *large and noisy* tags have an advantage, while for the task of retrieval where it is important to return all of the relevant items (high recall), *small and clean* labels perform better.

Given the per-user tagging rates, if one of these music tagging games become as popular as the ESP game, the game could supply a very large and clean set of of ground-truth data that could be used by researchers to train and evaluate many types of MIR systems. The potential benefit for MIR researchers is enormous.

Autotagging – Some researchers are investigating techniques to use automated content analysis in order to predict social tags directly from audio. One technique is to use a content-based music similarity model to position items (artists or tracks) in a similarity space. Untagged items can inherit the social tags that have been applied to near neighbors, the assumption being that similar sounding music will have similar tags [29].

An alternative approach is to learn how to predict individual tags directly from audio. For each tag, a training set is created consisting of positive and negative examples. Standard techniques such as have been used in automatic genre classification are used to train a model that can predict whether or not the tag should be applied to the item [10]. Table 7 compares the top 10 social tags collected by Last.fm to the autotags generated by the system described in [10] for the artist *Vampire Weekend*, a new artist that was not a part of the training or test set for the autotagger. While not an exact match, the autotags have captured some of the flavor of the social tags.

Autotagging of music mitigates against the cold start problem. New music and unpopular music can be tagged automatically at tagging rates far exceeding human taggers. However, there are some open issues with autotagging. Table 2 shows that roughly 25% of social tags are not-related to the audio. Tags such as "Seen live", "Awesome", "Funny", "Great lyrics" and "Guilty pleasures" are difficult to predict directly from audio. Subtle distinctions such as the difference between "Technical death metal" and "Progressive death metal' are beyond the capabilities of current content-based classifiers.

6.2 Synonymy, Polysemy and Noise

The unstructured, free-form nature of social tags can pose some difficulties for those wishing to exploit them. Misspellings, spelling variants and synonyms are present in the pool of social tags. These variants can dilute the pool contributing to the sparseness of the tag space. However, as discussed in Section 5.3 these variants can improve the findability of items. By clustering tags based upon similarity, these variants and misspellings can be associated with more canonical tags. When a user misspells a tag query or uses an uncommon synonym, the query can be transformed from the uncommon term to the canonical form.

Words often have more than one meaning. For example, romantic songs are often tagged with "Love". However, "Love" is also used to tag favourite songs – songs that the tagger loves. Using the tag "Love" as a query to retrieve romantic songs may result in a set of romantic songs intermingled with other seemingly random songs that were "Loved" by taggers. Again, clustering and natural language processing techniques such as word sense disambiguation [14] can help identify and segregate different usages of a tag.

Finally many tags are applied that have little discernible relation to music. These tags are essentially noise. Some examples from Last.fm are "Asdf", "Random", "Lazy eye", "d" and "Sh-t that my sister listened to on my pc grrrrr". These tags are not typically applied very often and as such are easy to filter out.

Latent semantic analysis, described in Section 5.4, can be a useful technique for mitigating the problems of synonymy, polysemy and noise encountered with tags.

6.3 Hacking

Sometimes taggers will tag items such as artists or tracks dishonestly. Instead of tagging items to describe them or to aid in future retrieval, these taggers will tag for nefarious reasons. Sometimes taggers will apply tags to an item in a way that will cause the item to appear similar to another more popular item. For example, a new unknown artist, may apply tags to their new song that are the same as the tags that have been applied to another, very popular song, in an attempt to manipulate the system to group their new song with the popular song. Similarly, malicious taggers could apply wrong tags to a competitor's songs so that those songs no longer appear in the proper position in the tag space. Some taggers will generate tag 'spam", by applying large sets of popular tags to an item in an attempt to raise the rank of the item in search results. Sometimes taggers will vandalize artists with wrong tags for entertainment value. Table 8 shows the top 10 most frequent tags that have been applied to pop artist Paris Hilton. Of the 3,500 or so tags that have been applied to Paris Hilton, 90% of them can be considered to be hostile tags including intentional mis-classifications ("Brutal death metal"), personal attacks ("Whore"), ironic tags ("Best singer in the world"), and negative opinions ("Your ears will bleed").

Tag	Frequency
Brutal Death Metal	558
All things annoying in the world put together into one stupid b-tch	486
Crap	273
Officially Sh-t	238
Pop	231
Your ears will bleed	136
Emo	107
Whore	98
Best Singer in the world	72
Female Vocalist	60

Table 8. Most frequent tags applied to Paris Hilton

These malicious tags can dilute the value of the overall tag pool. For example, the 558 "Brutal death metal" tags that have been applied to Paris Hilton make her appear to be the number one "Brutal death metal" artist. A recommender based upon social tags would likely recommend "Brutal death metal" artists to fans of Paris Hilton, and would also recommend Paris Hilton to fans of the Brutal Death Metal genre of music.

A number of techniques have been explored to deal with malicious tagging behaviors [15]. Some techniques include:

- Weight tags by how often the tagger listens to the item being tagged. Some sites such as Last.fm can monitor a tagger's listening behavior via an audio player plugin (called a "scrobbler"). If a tagger never listens to Paris Hilton, the tags applied to Paris Hilton by the tagger should not be given much weight.
- Weight tags by how often the tagger uses the tag. If a tagger applies the tag "Brutal death metal" to a number of artists, but never actually uses the tag to query for music, then these tags should not be given much weight.
- Weight tags by how often they cluster with other tags that have been applied to an item. A tag such as "Brutal death metal' does not often co-occur with "Pop" or "Female vocalist". A tag can be weighted based upon how well it clusters with other tags that have been applied to the item.

6.4 Tagger Bias

A typical tagger is likely to be young, affluent, and Internet-savvy [33]. The music taste of these taggers may not be representative of music tastes of the general population. This can lead to tagging bias where some types of music receive more than their fair share of tags. Artists that are favorites of taggers will receive a disproportionately large number of tags, while artists that are not listened to by taggers will be tagged infrequently. This tagger bias can

be demonstrated by comparing traditional music sales to tagging behavior. For example, Nielsen reports that Country music represented 11 percent of all physical albums sold in 2007, while Electronic music represented 3 percent of albums sold in 2007 [8]. If taggers exhibited the same music taste as the general population, we should expect to see that the "Electronic" tag is applied less than a third as often as the "Country" tag. However, at Last.fm, the "Electronic" tag is applied over nine times as often as the "Country" tag, nearly 30 times higher than expected. Clearly, the taggers at Last.fm are not representative of the general music population when it comes to Country music. (It should be noted that Last.fm taggers may skew toward a European population which may account for some of the bias away from Country music).

7 RESEARCH OPPORTUNITIES AROUND TAGS

Using social tags in the context of Music Information Retrieval is quite a new development. There are many interesting unanswered questions and issues surrounding how to best exploit social tags for music information retrieval. Some are highlighted here.

7.1 Expanding the tag coverage

Even though social music sites such as Last.fm have millions of tags applied to artists and tracks, this is still not enough data. For new music and for unpopular music there are few tags, not enough to be useful. Tools and techniques that can expand the pool of available tags will be critical if tags are to be used for long-tail [13] music discovery. One promising technique is the autotagging of items based upon content analysis. Open research questions around autotagging include:

- Which audio features work best for predicting social tags?
- Which learning algorithms work best for modeling tags?
- How can an autotagger best be evaluated?
- How can autotags best be combined with social tags?

Tagging games are also a potential method for expanding tag coverage. Open questions related to tagging games include:

- How do tags generated in a game context differ from social tags collected in a listening context?
- How can games be built to attract players to maximize ?he amount of new music that is labeled?

7.2 Using tags for discovery

Social tags define a complex music space. Providing tools and visualisations that allow a music listener to effectively explore this space will be key to exploiting social tags. Research questions include:

- How can we build an interface that exploits social tags to give a listener a more intuitive understanding of the interrelations between the many genres, styles and moods found in music?
- How can we use social tags to bridge the semantic gap, to allow listeners to find music by describing the music they like using words?
- How can we use social tags to give transparent explainable recommendations?
- How many social tags are enough before they can be used meaningfully for recommendation and discovery? Are 5 tags enough?, 10? 100?

7.3 Improving the quality of tags

Tag quality is important. There are a number of open questions related to tag quality:

- How can we best cluster tags to filter and use misspellings, and spelling variants to improve the findability of items?
- What effect do tagging interfaces have on the quality of tags? Is it better to suggest tags to a tagger, or does a blind-tagging system yield better tags?
- What is the demographic of a typical tagger? Are taggers representative of the general population? If not, how does that affect the social tags? Will similarity models generated from these tags reflect the general population?
- How can we detect and filter malicious tags?
- Is it better to have a fixed vocabulary of tags? Is a *small and clean* pool of tags better than a *large and noisy* pool?

8 CONCLUSION

Until recently social tags have mainly been used to enhance the search capabilities of music discovery sites. Listeners can search for artists or songs that have been tagged with a particular tag, or they can tune in to "Tag Radio" where they listen to music that has been tagged with a particular tag. Music Information Retrieval researchers are starting to understand the value contained in these large pools of social tags and that there are many ways that these tags can be used to help solve the many hard problems that MIR researchers face. Tags can be used to train classifiers to identify genres, mood and instrumentation. Tags can be used to build music similarity models. Tags can be used to help generate transparent, explainable recommendations. Tags can be used to create rich visualisations and discovery tools that allow music listeners to explore the complex music space to find new, interesting and relevant music. Researchers are just beginning to understand all of the ways that tags can be used to assist Music Information

Retrieval. Still, there are a number of problems that need to be solved. Social tags can be exceedingly noisy, filled with extraneous, seemingly useless data. New and unpopular items are rarely tagged, making tags useless for new music. Tags can be easily manipulated by malicious users that try to co-opt tags for their own nefarious purposes.

Social tags are a new tool to add to our MIR toolbox. As with any new tool, we have to take some time to learn best how to use them.

Acknowledgments

Thanks to the many individuals that provided input and support including the management at Last.fm and Sun Microsystems Inc., Jeff Alexander, Jean-Julien Aucouturier, Thierry Bertin-Mahieux, Don Byrd, Oscar Celma, Douglas Eck, Stephen Green, David Jennings, Edith Law, Michael Mandel, Elias Pampalk and Douglas Turnbull. Thanks also to the anonymous reviewers who provided many useful comments and suggestions.

9 REFERENCES

[1] Morgan Ames and Mor Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980, New York, NY, USA, 2007. ACM Press.

[2] J.J. Aucouturier and F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

[3] Audioscrobbler. Web Services described at http://www.audioscrobbler.net/data/webservices/.

[4] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. http://www.pui.ch/phred/automated_tag_clustering/.

[5] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl. Aggregate features and AdaBoost for music classification. *Machine Learning*, 65(2-3):473–484, 2006.

[6] T. Bertin-Mahiex, D. Eck, F. Maillet, and P. Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Reseach*, 2008. to appear.

[7] Don Byrd. Organization and search of musical information. Syllabus for a course at http://informatics.indiana.edu/donbyrd/Teach/I545Site-Spring08/SyllabusI545.html.

[8] The Nielsen Company. Nielsen soundscan state of the industry. 2008 Convention of the National Association of Recording Merchandisers, 2008. http://www.narm.com/2008Conv/StateoftheIndustry.pdf.

[9] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[10] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In *Neural Information Processing Systems Conference (NIPS) 20*, 2007.

[11] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[12] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *Proceedings of the First International Conference on Multidisciplinary Information Sciences and Technologies (InSciT 2006)*, 2006.

[13] Chr. Hjorth-Andersen. Chris anderson, the long tail: How endless choice is creating unlimited demand. the new economics of culture and commerce. *Journal of Cultural Economics*, 31(3):235–237, September 2007.

[14] N. Ide and J. Veronis. Word sense disambiguation: The state of the art. In *Computational Linguistics, 24:1*, pages 1–40, 1998.

[15] Georgia Koutrika, Frans Adjie Effendi, Zoltán Gyöngyi, Paul Heymann, and Hector Garcia-Molina. Combating spam in tagging systems. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 57–64, New York, NY, USA, 2007. ACM.

[16] E. Law, L. von Ahn, R. Dannenberg, and M. Crawford. Tagatune: A game for music and sound annotation. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.

[17] M. Levy and M. Sandler. A semantic space for music derived from social tags. In *Proceedings of the Eighth International Conference on Music Information Retrieval (ISMIR'07)*, Vienna, Austria, September 2007.

[18] M. Levy and M. Sandler. Learning latent semantic models for music from social tags. *Journal of New Music Reseach*, 2008. to appear.

[19] M. Mandel and D. Ellis. A web-based game for collecting music metadata. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.

[20] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[21] C. McKay and I. Fujinaga. Musical genre classification: is it worth pursuing and how can it be. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, 2006.

[22] Peter Morville. *Ambient findability*. O'Reilly, Sebastopol, CA, 2005.

[23] Gerard Salton. *Introduction to Modern Information Retrieval (McGraw-Hill Computer Science Series)*. McGraw-Hill Companies, September 1983.

[24] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[25] Clay Shirky. Ontology is overrated: Categories, links and tags. http://www.shirky.com/writings/ontology_overrated.html.

[26] Rashmi Sinha. A cognitive analysis of tagging. http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html.

[27] Rashmi Sinha. A social analysis of tagging. http://www.rashmisinha.com/archives/06_01/social-tagging.html.

[28] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI '02: CHI '02 extended abstracts on Human factors in computing systems*, pages 830–831, New York, NY, USA, 2002. ACM.

[29] M. Sordo, C. Laurier, and Oscar Celma. Annotating music collections: How content-based similarity helps to propogate labels. In *Proceedings of the Eighth International Conference on Music Information Retrieval (ISMIR'07)*, Vienna, Austria, September 2007.

[30] D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet. A game-based approach for collecting semantic annotations of music. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.

[31] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM.

[32] Thomas Van Der Wal. Folksonomy coinage and definition. http://www.vanderwal.net/folksonomy.html.

[33] D. Weinberger. How tagging changes peoples relationship to information and each other. Pew Internet & American Life Project. www.pewinternet.org/pdfs/PIP_Tagging.pdf.

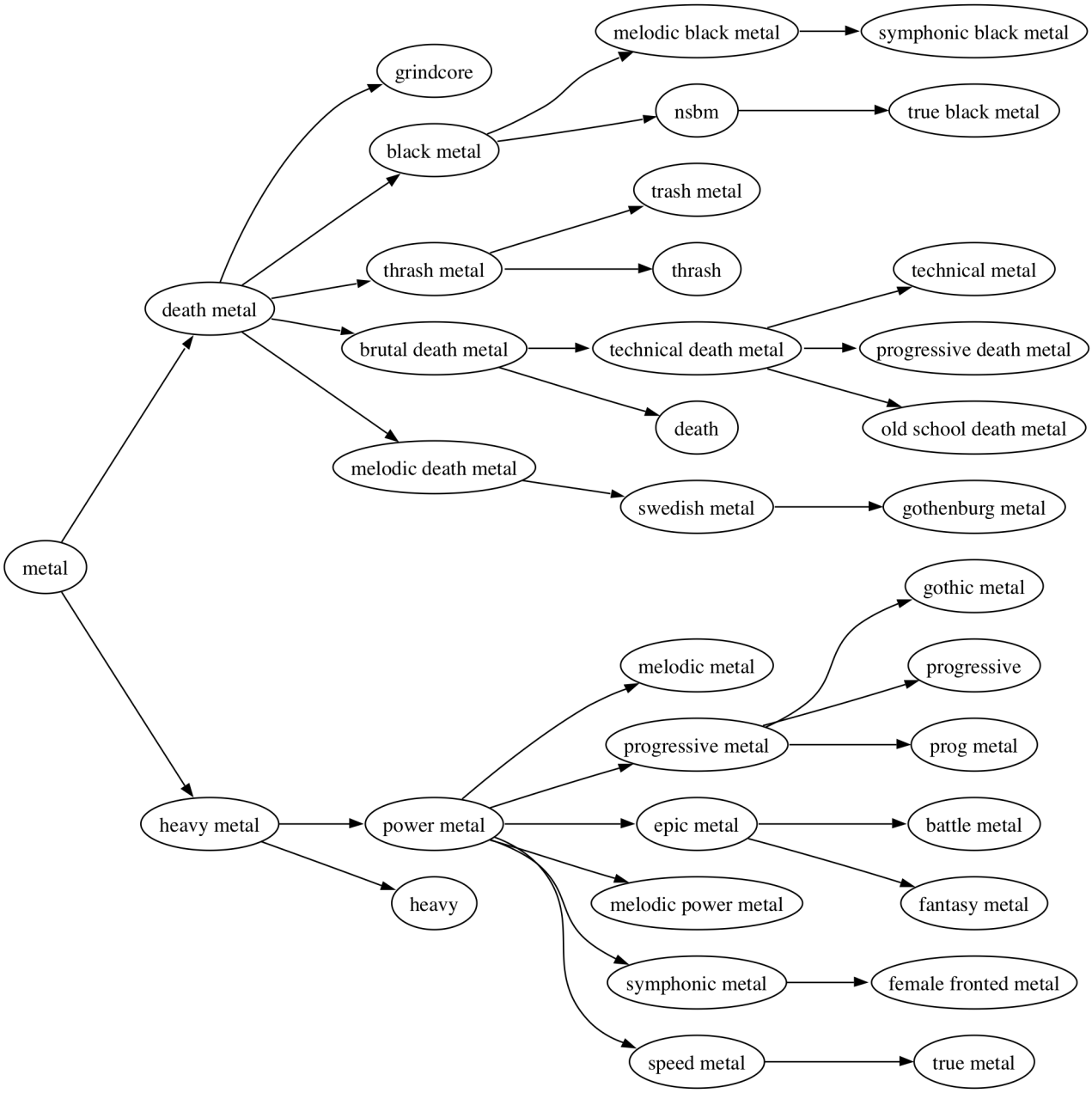

Figure 4. Derived Metal Taxonomy

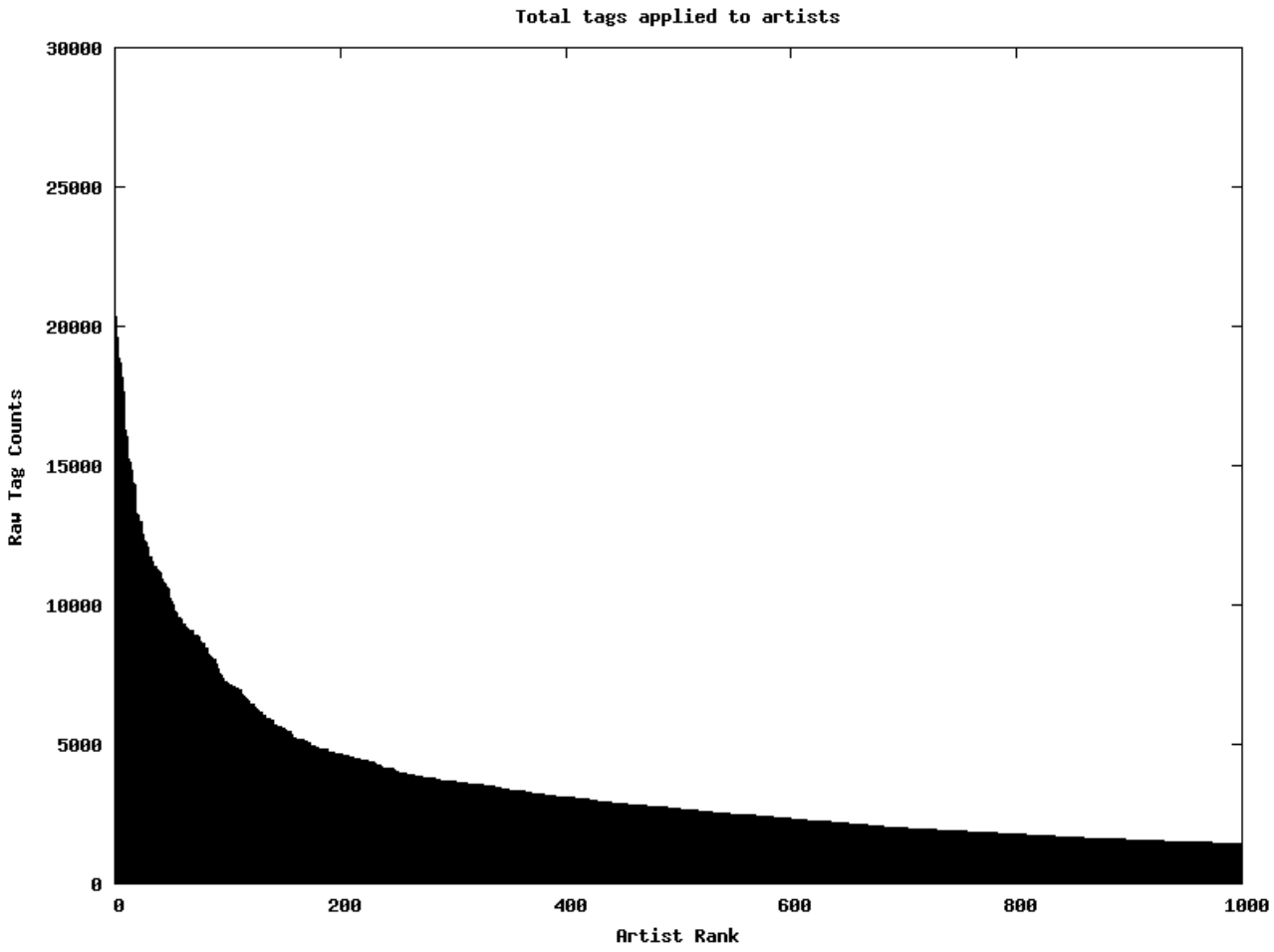

Figure 5. Tag frequency for top 1000 artists