THE MUSIC LISTENING HISTORIES DATASET

Gabriel Vigliensoni and Ichiro Fujinaga
Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)
McGill University, Montréal, Québec, Canada
[gabriel.vigliensonimartin,ichiro.fujinaga]@mcgill.ca

ABSTRACT

We introduce the Music Listening Histories Dataset (MLHD), a large-scale collection of music listening events assembled from more than 27 billion time-stamped logs extracted from Last.fm. The logs are organized in the form of listening histories per user, and have been conveniently preprocessed and cleaned. Attractive features of the MLHD are the self-declared metadata provided by users at the moment of registration whose identities have been anonymized, MusicBrainz identifiers for the music entities in each of the logs that allows for an easy linkage to other existing resources, and a set of user profiling features designed to describe aspects of their music listening behavior and activity. We describe the process of assembling the dataset, its content, its demographic characteristics, and discuss about the possible uses of this collection, which, currently, is the largest research dataset of this kind in the field.

1. INTRODUCTION

The modeling of users for multimedia information retrieval systems has been a research topic since the first International Symposium on Music Information Retrieval (ISMIR) in 2000. In that meeting, it was observed that to create modern, more efficient, and personalized music information retrieval systems, the modeling of users would be necessary because many features of multimedia content delivery are perceptual and user-dependent [6].

Sixteen years after the first ISMIR meeting, the landscape of music consumption has changed enormously. The rise and fall of peer-to-peer networking led to the reinvention of the music industry: the paradigmatic music product was no longer a full album in a physical format, but individual music files available in online digital music stores. Thanks to the miniaturization of portable media players and also to almost ubiquitous Internet access, a change of paradigm in music consumption has happened again, and people seem to not want to pay for individual tracks. Instead, they are willing to pay for services that allow them to access, search, and discover music items—artists, albums, or tracks—within large repositories [23].

On-demand digital music streaming services are currently the fastest growing sector of the global music industry [11]. In fact, in 2015 the digital revenues that these systems generated overtook the income from physical music goods for the first time in music industry history [12]. As a result, the on-demand music streaming landscape these days seems to be a lucrative battlefield, and one on which many companies want to compete. However, since the majority of the listeners' accounts in music streaming services use the "free" or "freemium" business model—advertisement-supported basic streaming services—a large share of the income of music and media streaming companies comes from targeting ads more precisely at listeners [18]. It seems that the streaming model is like modern advertising.

In this model, people are no longer passive observers but direct participants in the battlefield that is the digital media and music streaming landscape. In fact, the traded goods in this business are individual profiles and psycographic traits (e.g., interests, lifestyle, personality, values) which are extracted from correlating people's listening habits with their sociographic characteristics [17]. As a result, listeners are the source of information, but they also are the final target for all the advertising these companies are making money from.

As music information researchers, our community has to be able to observe, investigate, and to gather insights from the listening behavior of people in order to develop better, personalized music retrieval systems. Yet, since most media streaming companies know that the data they collect from their customers is very valuable, they usually do not share their datasets. A honorable mention goes to Netflix, company that challenged the recommendation research community in 2006 with a large dataset of ratings of users on movies. Insights and techniques developed for that competition are still being used widely today.

2. PREVIOUSLY AVAILABLE DATASETS

A number of datasets for music listening research have been collected and released by research groups. These datasets provide information relating the interaction of a large number of listeners and music items.

Celma assembled the Last.fm Dataset-360K, a dataset of playcounts with listeners' demographic data for 360K

Dataset name	Type	Source	Size	Demographics	Linkage	Other
Last.fm Dataset 360-K [5]	Playcounts	Last.fm	18M logs, 360K users	Yes	Yes	Only includes most frequently listened artists
Last.fm Dataset 1-K [5]	Listening histories	Last.fm	19M logs, 1K users	Yes	Yes	Full music listening histories
Yahoo! Music Dataset [7]	Ratings	Yahoo! Music Radio	262M logs, 1M users	No	No	Hierarchical structure of music items
HetRec2011-last.fm.2k [4]	Playcount	Last.fm	2K users	No	No	Bidirectional users' relations and artist tags
Echo Nest Taste Profile subset [13]	Playcounts	Undisclosed	48M logs, 1.2M users	No	Yes	Linked to Million Song Dataset
EMI Million Interview Dataset [8]	Interviews	Individual interviews	1M users	Yes	Unknown	Partial information available
MusicMicro 11.11-09.12 [19]	Listening histories	Twitter	600K logs, 137K users	Geolocalized logs	No	Precise geolocation data of each log
Million Musical Tweets Dataset [9]	Listening histories	Twitter	1M logs, 215K users	Geolocalized logs	Yes	Many users have only a few listening events
#nowplaying Music Dataset [24]	Listening histories	Twitter	50M logs, 4.2M users	No	Yes	Many users have only a few listening events
LFM-1B [20]	Listening histories	Last.fm	1B logs, 120K users	Yes	No	Comes with a set of features describing music consumption behavior. The music listening histories are shifted according to the time zone of listeners, and so they are not directly comparable.
MLHD	Listening histories	Last.fm	27B logs, 583K users	Yes	Yes	Comes with MBIDs, estimation of listeners' time zone, and users' activity features.

Table 1. Comparison of freely available datasets of music listening events.

listeners, and the Last.fm Dataset-1K, a set of full listening histories with time-stamped logs [5]. Though richer, the latter dataset included logs for only 1K listeners. Following the Netflix prize, Dror, Koenigstein, and Koren released the Yahoo! Music Dataset, a collection of 1M people's aggregated ratings on music items [7]. Later on, Cantador, Brusilovsky, and Kuflik presented the HetRec2011-Last.fm-2K, another dataset with song playcounts for the 50 most listened artists of 2K listeners [4]. McFee et al. introduced The Echo Nest Taste Profile subset, a dataset of song playcounts of 1M listeners collected from undisclosed services [13]. Neither of these two datasets, however, provided timestamps of the music logs or demographic information about the listeners. The EMI Group Limited promised a dataset of 1M interviews about people's music appreciation, behavior, and attitudes [8], but only partial information was made available. None of the aforementioned datasets simultaneously provided individual music listening logs as well as demographic data for a large amount of listeners.

More recently, music listening logs have been collected from the social networking service Twitter. Schedl released MusicMicro 11.11-09.12, a dataset of about 600K music-related *tweets* with temporal and spatial data [19]. Hauger et al. released the Million Musical Tweets Dataset [9], a collection of 1M music-related geolocalized microblog posts with partial linkages to other services. Zangerle et al. introduced the #nowplaying Music Dataset, a collection of 50M music-related posts linked to MusicBrainz [24]. In these collections, however, there were a large number of listeners with only one or two logs, and so, in many cases, the datasets provided a few listening events for many users instead of listening histories.

Finally, Schedl introduced LFM-1B, a very large dataset of more than 1B logs collected from Last.fm user interactions [20]. Each log includes artist, album, and track names, the timestamp of the log, as well as each user's Last.fm identifier. The dataset also comes with users' demographic information as well as a set of features that describe music consumption behavior per user. However, the dataset does not provide common identifiers with other music databases, and so the only way to link the music items is by string matching.

In Table 1 we provide a summary of available databases of music listening logs. We can see that among all the datasets reviewed, the only one that provides full music listening histories, listeners' self-declared demographic data, as well as identifiers easily linkable to other databases of music information is the Last.fm Dataset-1K. However, the size of the dataset is very small to perform a large-scale analysis with global reach. In order to ameliorate this situation, we decided to collect our own dataset considering all the aforementioned characteristics.

3. THE MUSIC LISTENING HISTORIES DATASET

In this section we will describe the creation of the Music Listening Histories Dataset (MLHD), a large dataset of full music listening histories. We will review the concept of music listening history and will present the criteria for the data collection and cleaning of the data. We will also provide insights about the demographic characteristics of users in the dataset and will explain the need of providing a value for normalizing the time zone of the logs.

3.1 Music Listening Histories and Last.fm

Listening histories are a timeline of listening events. Analyzing them in a linear fashion is interesting because we can observe when people consume music, and what music they enjoy or do not enjoy over time. However, since people seem to follow periodic listening cycles [10], the aggregation of these listening histories by collapsing them into different periods of time can provide extra layers of information that can be used to infer people's listening patterns and preferences.

Last.fm is an online digital music service available since 2002. It was originally conceived as a web-based radio station. Immediately after its launch, the company incorporated the tracking of music listening logs as a core part of its service. However, Last.fm stands out from most music streaming services that collect user data because it not only gathers listening logs (known as *scrobbles*) from the interaction of its users withing the system's ecosystem, but also from the interaction between users and a wide range of third-party music and media players by means of the *scrobbler* service.

Last.fm offers free access to the listening data they collect from listeners, as well as music metadata, biographies, pictures, charts, tags, ranking data by country, and other information by means of a well-documented API. At the moment of registration, every user must accept the Last.fm Terms of Use and the Last.fm Privacy Policy. [1] These terms establish that their listening habit data will be available to third parties via their API for commercial and/or non-commercial purposes. The users are also asked to provide basic demographic information such as their date of birth, country, and gender.

All aforementioned characteristics, added to the fact that the Last.fm API Terms of Service establish that Last.fm offers a "limited terminable licence to copy and use the Last.fm Data" that is free of charge "for non-commercial purposes" [2] persuaded us to choose Last.fm as the data source to assemble the MLHD.

3.2 Data Collection

In order to retrieve full music listening histories and to obtain even data across aggregated periods of time, we followed previous research [1] and searched only for listeners with a minimum of two years of activity since they started to submit music logs to Last.fm. Also, in order to prevent collecting data from casual users that registered for a service, tried it, but never used it again, we collected data only from listeners which had an arbitrary average of, at least, ten scrobbles per day. The two constraints forced all listeners in our dataset to have a minimum of 7,300 (i.e., 365 × 2 × 10) music logs submitted to the Last.fm database.

Differently to all other datasets with Last.fm data, we collected listening data by using an undocumented (but deprecated) method that allowed us to not need actual usernames for calling the Last.fm Web services [21]. Instead, we simply passed Last.fm users' internal identifiers as arguments of the API requests. Since these IDs are sequential, this approach permitted us to sample users randomly across the entire database instead of sampling users based on their *friends* or on an artist's *top fans*, which are methods probably more biased. We aimed to collect full listening histories, and so we fetched people's listening logs by using the Last.fm's API method `user.getRecentTracks()`, and paginated iteratively throughout the chosen listeners' full music listening histories.

3.3 Data Cleaning, Sanitization, and Organization

Within each music listening history, we organized each of the logs in quadruples with the form of `<timestamp, artist-MBID, release-MBID, track-MBID>`, where timestamp is a global coordinated universal time (UTC) stamp, and MBID stands for MusicBrainz identifier. MBIDs are 36-character universally unique identifiers (UUID) that are permanently assigned to entities within the MusicBrainz database to ensure a reliable and unambiguous form of identification. Since Last.fm exposes MBIDs as public identifiers of music entities in their database, we collected them directly for each artist, release, and track. These three entities are hereafter denominated "music entities." Finally, all data per user was stored within a single file, with the logs sorted sequentially by their timestamp.

After close inspection of the data, we realized that there were two issues in some of the listening histories: (i) there were duplicated music logs (i.e., same timestamp and MBIDs); and (ii) some logs were too close in time (i.e., less than 30 seconds apart, which is the minimum that Last.fm requires to consider a played track as a valid log). We hypothesized that these issues were artifacts produced by the interaction of the Last.fm servers and some scrobblers. As a result, we decided to perform a cleaning process before storing the data, and so we filtered out all logs with the same MBID and timestamp, and we also filtered out all scrobbles that were less than 30 seconds apart in time. All in all, the average percentage of duplicated logs removed for each user was eight percent, and one percent for those logs that were too close.

It is worth mentioning that sometimes the metadata provided by the scrobbler is not enough to produce a full match for artist, release, or track. In cases like this, the music listening log returned by the Last.fm API will have only partial information. As a result, not all logs in the MLHD have a full set of MBIDs.

In Figure 1 we show the percentage of all combinations of MBIDs across all music logs in the dataset. It can be seen that about 58 percent of all music logs in the MLHD

[1] Privacy Policy available at `http://www.last.fm/legal/privacy`

[2] Terms of Service available at `http://www.last.fm/api/tos`

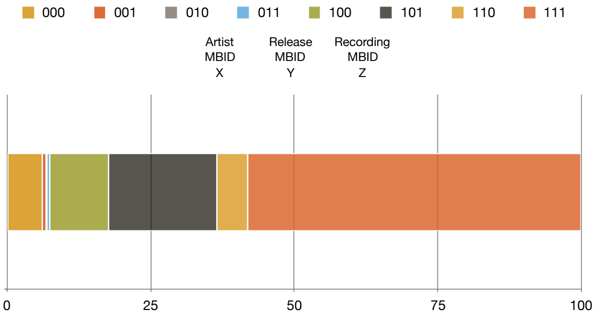

Figure 1. Percentage of music logs with combination of MBIDs. 0 stands for no presence of the corresponding MBID in the scrobble and 1 for its existence.

have full data (i.e., MBIDs for the three music entities), and 93 percent of the logs have, at least, the artist MBID.

3.4 Data Exploration

We computed from the music listening histories' UTC timestamps a series of features that aggregated the number of scrobbles of each listening history into several time spans. These low-dimensional representations of a user activity may facilitate the creation of plots and their visual inspection in order to gain insights or detect anomalies from single listener or groups of them. These per-user features are: *hourly activity*, *hourly activity by week hour*, *weekly activity*, *monthly activity*, *yearly activity*, *weekday activity*, *Saturday activity*, and *Sunday activity*.

3.5 Demographics

The MLHD currently consists of more than 27 billion music logs taken from the listening histories of 583K people that have linked their digital music players to Last.fm. In this massive repository, we counted more than 555K different artists, 900K albums, and seven million tracks. Table 2 summarizes the number of logs, unique listeners, and music entities in the dataset.

Dataset	Logs	Listeners	Artists	Albums	Tracks
	27MM	583K	555K	900K	7M

Table 2. Music listening histories dataset summary.

The distribution of the average number of daily submitted music logs per listener is shown in Figure 2. Axes in the plot are in log scale. The curve exhibits a close to power law characteristic. As expected, due to the constraints we set for collecting listeners' listening histories, the minimum average daily number of music logs per user was ten. Listeners with an average of eleven logs were the largest group, with about 30K listeners. The median number of submitted logs per user was 35K. The median age of the listening histories was 4.5 years.

Now we will describe the nature of the users in the dataset according to their self-declared age, gender, and country. This information is asked to the users at the moment of registration.

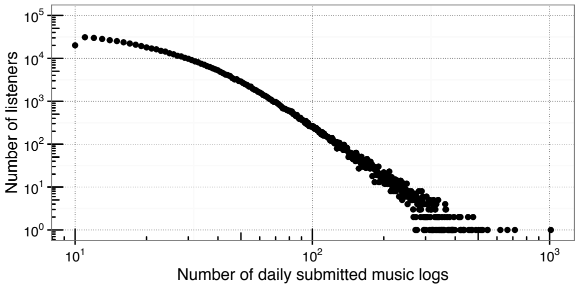

Figure 2. Distribution of the average number of daily scrobbles per listener.

3.5.1 Age

In terms of age, 71 percent of the listeners in the dataset declared their date of birth, which is much higher than similar datasets [5, 20]. Among them, 98 percent of the users had a self-declared age within 15 and 54 years old. In spite of the small magnitude of the probably deceiving information found in the two percent out of this age range, we decided to filter them out from the dataset. The mean age of listeners in the dataset is 25.4 years old, the median is 24, and the mode is 22. Since these are values similar to the ones found in similar datasets, this skew in the distribution indicates a bias in our dataset—and probably in Last.fm users—towards youth and young adults. We show the age distribution of listeners of the MLHD in Figure 3.

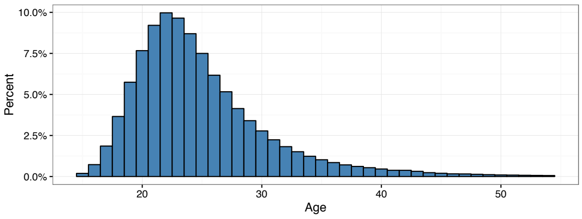

Figure 3. Age distribution of MLHD listeners within the [15, 54] years old range.

3.5.2 Gender

In terms of gender, about 82 percent of the people in the dataset declared a gender at the moment of their registration or afterwards. In Figure 4 we show the self-declared gender distribution among these users.

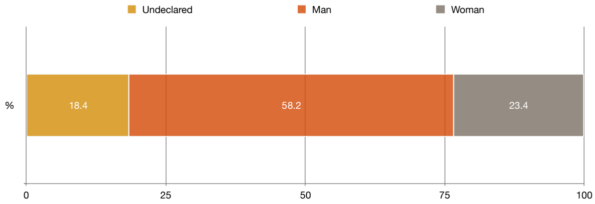

Figure 4. Percentage of listeners' self-declared gender.

It can be seen that there is a bias towards male listeners in the MLHD. Since this bias is also observed in similar dataset, this can be an indication that Last.fm has more male than female users.

We compared the age within each self-declared gender with balanced groups. The total number of listeners without self-declared gender was slightly more than 100K, and so we sampled 100K listeners from each group. The mean of the Not declared ($\mu = 25.67$) and Male ($\mu = 25.60$) groups did not differ greatly ($p = .400$), perhaps indicating that the first group may have a large proportion of male users. On the other hand, users self-declared as Female ($\mu = 22.99$) had a different lower mean age than the Male group ($p < .001$). In other words, users in our dataset self-declared as Female are younger than the ones declared as Male.

3.5.3 Country

In terms of location, 82 percent of users in the MLHD self-reported a country. These users belong to 239 different countries or territories as defined in the ISO 3166-1 International Standard for country codes. Among these territories, 19 countries had at least one percent of the total amount of listeners in the dataset. These "top countries" combined accounted for more than 85 percent of the total number of listeners in the dataset.

In order to determine how countries were relatively represented in the MLHD, we divided the percentage of users per country by the actual country population. [3] This metric gave us a better description about how different countries' populations were represented in our dataset. In Figure 5 we show a map that presents the relative number of listeners per country normalized by the corresponding number of inhabitants in each country.

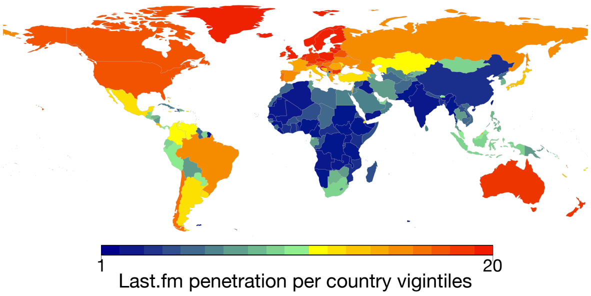

Figure 5. Relative number of listeners per country, normalized by the number of inhabitants in each country.

The color palette of the plot was based on vigintiles (20 quantiles) of the data, with red indicating the highest vigintile, and blue the lowest one. If our dataset has similar distinctive qualities in comparison with the overall Last.fm data, this map can be interpreted as the Last.fm market penetration by country. By looking at the higher vigintiles we can see that listeners from most zones are represented in the MLHD. In particular, Northern European, North American, and Australasian countries have the largest proportion of listeners submitting music logs to Last.fm. Also, some countries in South America show similar penetration levels to some Mediterranean countries in Europe. People from Africa, South Asia, and Far East Asia are not extensively represented in our dataset.

Finally, pair-wise mean age comparison using balanced groups of listeners per country (N= 4.5K) showed significant differences between listeners from some of the top countries. For example, Brazilian listeners are younger on average ($\mu = 22.6$) than all other top countries ($p < .001$), except for Poland, Russia, and Ukraine. On the other hand, Japanese listeners are older on average ($\mu = 29.0$) than users from all the other top countries ($p < .001$), except for Spain and France.

3.6 Time Zone Normalization

Last.fm collects scrobbles using the Unix time stamp format no matter where the logs were generated. Therefore, all music logs within the Last.fm database have the same temporal point of reference. Beyond the timestamp and the MBID for the three music entities, the logs do not store any additional geographical information such as city, country, or the time zone where they were generated.

The lack of information about where the logs were actually generated can be a problem. If the researcher wants to find trends in people's daily, weekly, and monthly music listening behavior, it is necessary to aggregate their music listening histories over time. However, the aggregated listening patterns from people in different time zones is shifted depending on where they are. As a result, it would be misleading to directly compare their patterns. The country information could be used to estimate a listener's time zone, but many countries span their territories over several time zones.

In previous studies with similar data, the researchers have hand-picked listeners within the same time zone [2, 3, 16] or are just compared their daily listening patterns directly [20]. However, a research dataset to perform studies at the global level must provide this information in order to properly compare the music listening histories.

We followed an approach for time zone normalization based on the assumption that people share hours of sleep at night [21], and computed the time shift of the listening histories in the MLHD. In Figure 6 we show the estimated distribution of the time zones of all listeners in the dataset. We can observe a peak in the estimated time zone from where people submitted music logs at time zone GMT +0, with about 17 percent of the dataset users. Additionally, a large proportion of the listeners were estimated to be within time zones corresponding to Western Europe, but also spread out throughout the different time zones in America.

4. CONCLUSION

All in all, the MLHD provides three sources of data for each user: (i) demographic metadata, (ii) sanitized full music listening histories, and (iii) low-dimensional feature vectors describing the full listening histories in terms

[3] Population data for the year 2012 taken from the World Bank Open Data repository, available at `http://data.worldbank.org/indicator/SP.POP.TOTL`

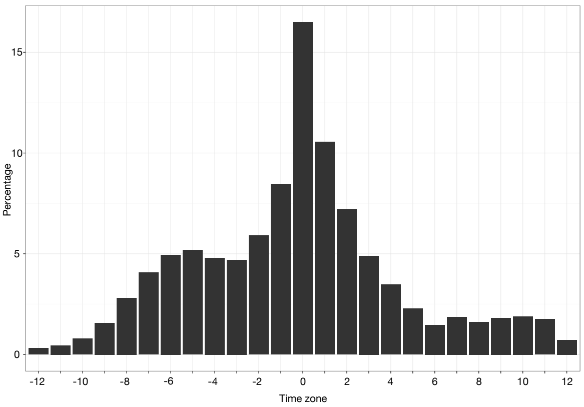

Figure 6. Distribution of the time zones for all users in the dataset (N = 583K).

of user activity. As a result, the full music listening histories compiled in the MLHD dataset offer a large amount of information. On top of having a very fine time granularity—providing second-accurate data about the music item played back in a media player by a specific user—their aggregation into different spans of time may provide clues about the people's listening behavior characteristics and their listening trends over time.

A big advantage of the MLHD dataset over other datasets for listening behavioral research is that it is based on MusicBrainz identifiers (MBID). This feature allows the easy linkage of each log to, for example, all services of the MetaBrainz Foundation ecosystem (i.e., MusicBrainz, AcousticBrainz, ListenBrainz, and CritiqueBrainz) and to other services that provide additional data accessible through these IDs (e.g., Last.fm provides folksonomy tags for artists, albums, and tracks, and DBPedia links Wikipedia open music data to MusicBrainz by means of MBIDs). Therefore, music listening histories can be linked to resources from other repositories, thus enabling the aggregation, linkage, and expansion of the data and the knowledge about people's music listening behavior.

In terms of possible uses of the dataset, data aggregations extracted from the MLHD have already been used in combination with other sources of data. In particular, it has been used as part of the datasets for "Sound and music recommendation with knowledge graphs" [15]. In these datasets, a subset of music listening histories from the MLHD were aggregated into playcounts and used in combination with additional song data collected from Songfacts.com to enable the study of hybrid music recommendation models using additional user-provided factual information describing songs and artists [14]. Additionally, it has been used to find listening behavioral patterns in four different age groups and to evaluate the improvement of a music recommendation model by using demographic, profiling, and contextual features [22].

We plan to expand the MLHD by collecting more listening data. This is a good idea in the eventual case that Last.fm stops providing this data or a full shutdown of the service. Also, the data collected may be added to the ListenBrainz project, an initiative of the MetaBrainz Foundation with the goal of allowing listeners to preserve their existing music listening histories in Last.fm.

Although we aimed to collect data from a large group of listeners of varied demographics—thus helping to overcome biases from previous user-driven and data-driven research—the listening data we collected may be also biased. For example, the age distribution of listeners show that the dataset is skewed towards late adolescent and early adult listeners. However, since this group will be older in a few years from now, and younger generations are already born into a digital era, we suppose that this trend may be different in a few years, and the large skew towards listeners in their early twenties may be less significant. In any case, the MLHD has a much larger amount of data than any of the studies of the datasets reviewed in Section 2, and so it allows for the undertaking of studies with balanced populations of listeners of each age.

We also acknowledge that a limitation of conducting data-centric studies using data collected from listening interactions with media players and music streaming services is the fact that it is hard to know if listeners actually chose the music item they were exposed to, or it was the recommendation engine or shuffle algorithm of a music streaming service the one that suggested the music item. As a result, it is hard to say if a specific scrobble reflected the actual preference of a listener, or if it registered what was recommended by a recommendation or shuffle algorithm. However, Wikström [23] pointed out that ubiquitous access to music services with recommendation algorithms is how the majority of people are actually experiencing music in the new music economy. Hence, the study of music preference nowadays cannot separate self-chosen music from algorithmically generated playlists and suggestions. These two approaches are occurring at the same time, and so both have to be considered in order to obtain insights about listening behaviors and music preferences. We hope the dataset we introduced will be useful for doing large-scale research on user modeling, music preference, and recommendation.

The MLHD can be accessed and downloaded at `http://ddmal.music.mcgill.ca/research/musiclisteninghistoriesdataset`.

5. ACKNOWLEDGEMENTS

We are extremely grateful of all users of Last.fm that have agreed to make their data available for non-commercial use, and also to the Last.fm service, which has collected and offered this data since 2002 uninterruptedly, helping the field of music informatics research to move forward.

This research has been supported by BecasChile Bicentenario, Comisión Nacional de Ciencia y Tecnología, Gobierno de Chile, and the Social Sciences and Humanities Research Council of Canada. Important parts of this work used ComputeCanada's High Performance Computing resources.

6. REFERENCES

[1] Dominikus Baur, Jennifer Büttgen, and Andreas Butz. Listening factors: A large-scale principal components analysis of long-term music listening histories. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems*, pages 1273–6, Austin, TX, 2012.

[2] Pauwke Berkers. Gendered scrobbling: Listening behaviour of young adults on Last.fm. *Interactions: Studies in Communication & Culture*, 2(3):279–96, 2010.

[3] Jennifer Büttgen. What's in a history? A large-scale statistical analysis of Last.fm data. Master's thesis, Ludwig-Maximilians-Universität München, München, Germany, 2010.

[4] Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. Last.fm web 2.0 dataset. In *2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems*, RecSys 2011, Chicago, IL, 2011.

[5] Òscar Celma. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer-Verlag Berlin Heidelberg, Berlin, Germany, 2010.

[6] Wei Chai and Barry Vercoe. Using user models in music information retrieval systems. In *Proceedings of the 1st International Symposium on Music Information Retrieval*, Plymouth, MA, 2000.

[7] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The Yahoo! music dataset and KDD-cup'11. *Journal of Machine Learning Research*, 18:3–18, 2011.

[8] EMI Group Limited. EMI Million Interview Dataset, 2012. http://musicdatascience.com/emi-million-interview-dataset/. Accessed 18 February 2017.

[9] David Hauger, Markus Schedl, Andrej Košir, and Marko Tkalčič. The million musical tweets dataset: What can we learn from microblogs. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[10] Perfecto Herrera, Zuriñe Resa, and Mohamed Sordo. Rocking around the clock eight days a week: An exploration of temporal patterns of music listening. In *Proceedings of the 1st ACM RecSys Workshop on Music Recommendation and Discovery,*, Barcelona, Spain, 2010.

[11] IFPI. IFPI global music report 2015. Annual report, International Federation of the Phonographic Industry, 2015.

[12] IFPI. IFPI global music report 2016. Annual report, International Federation of the Phonographic Industry, 2016.

[13] Brian McFee, Thierry Bertin-Mahieux, Daniel P. W. Ellis, and Gert R. G. Lanckriet. The million song dataset challenge. In *Proceedings of the 21st International Conference on World Wide Web*, pages 909–16, Lyon, France, 2012.

[14] Sergio Oramas, Vito Claudio Ostuni, Tomasso Di Noia, Xavier Serra, and Eugenio Di Sciascio. Sound and music recommendation with knowledge graphs. *ACM Transactions on Intelligent Systems and Technology*, 8(2):1–21, 2016.

[15] Sergio Oramas, Vito Claudio Ostuni, and Gabriel Vigliensoni. Sound and music recommendation with knowledge graphs (dataset), 2016. http://hdl.handle.net/10230/27495. Accessed 18 April 2017.

[16] Chan Ho Park and Minsuk Kahng. Temporal dynamics in music listening behavior: A case study of online music service. In *9th IEEE International Conference on Computer and Information Science*, pages 573–8, Kaminoyama, Japan, 2010.

[17] Robert Prey. *Musica analytica: The datafication of listening*, chapter 3, pages 31–48. Palgrave Macmillan UK, London, United Kingdom, 2016.

[18] Paul Rutter. *The Music Industry Handbook*. Media Practice. Routledge, Oxfordshire, United Kingdom, 2 edition, 2016.

[19] Markus Schedl. *Leveraging microblogs for spatiotemporal music information retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 796–9. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[20] Markus Schedl. The LFM-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 103–10. ACM, 2016.

[21] Gabriel Vigliensoni and Ichiro Fujinaga. Identifying time zones in a large dataset of music listening logs. In *Proceedings of the 1st International Workshop on Social Media Retrieval and Analysis*, pages 27–32. ACM, 2014.

[22] Gabriel Vigliensoni and Ichiro Fujinaga. Automatic music recommendation systems: Do demographic, profiling, and contextual features improve their performance? In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pages 94–100, New York City, NY, 2016.

[23] Patrik Wikström. *The Music Industry: Music in the Cloud*. Polity Press, Cambrige, UK, 2nd edition, 2013.

[24] Eva Zangerle, Martin Pichl, Wolfgang Gassler, and Günther Specht. #nowplaying music dataset: Extracting listening behavior from twitter. In *Proceedings of the 1st International Workshop on Internet-Scale Multimedia Management*, pages 21–6, Orlando, FL, 2014.