Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

Representation of Music Creators on Wikipedia, Differences in Gender and Genre

Alice Wang, Aasish Pappu, and Henriette Cramer
Spotify Research
{alicew,aasishp,henriette}@spotify.com

Abstract

Wikipedia is not only the world's largest online encyclopedia and among the most frequented websites, but provides important data leveraged by many popular services and products. Since Wikipedia data is ubiquitously encountered, it is important to evaluate its coverage of content and identify data gaps that may exist. Here, we evaluate Wikipedia's coverage of the music domain, which is one of the most popular topics. Particularly, we compile the most prominent 50,000 music artists (by streaming popularity on a large online streaming platform) and determine whether each artist has a Wikipedia page. We first show that streaming popularity correlates with Wikipedia representation– while 90% of the top one thousand most popularly streamed artists are on Wikipedia, the chance of being on Wikipedia drops to 50% after the ten thousandth artist. Next, we examine the Wikipedia coverage of artists of different gender and genre, while controlling for popularity. We also examine, for artists that are on Wikipedia, the amount of content, frequency of edits, and Pagerank for their pages. We uncover large differences in representation for artists of different genres; for the same popularity level, *hip hop*, *latin*, and *dance/electronic* artists are most lacking in representation while rock artists have approximately twice as much representation. With respect to gender, while female artists are underrepresented in the top of the music industry itself, male artists were less likely represented on Wikipedia relative to the female artists in this study's top sample, suggesting interaction with genre and visibility of select superstars.

Introduction

Wikipedia is the world's largest online encyclopedia, with English Wikipedia[1] being the third most visited website in the world, garnering over 2 billion visits per month (Ahrefs 2020). Even internet users who do not directly visit the website might unknowingly and frequently encounter Wikipedia data. For example, Google, Yahoo!, and Bing's search results prominently display side-bars with facts from Wikipedia. Furthermore, Wikipedia-derived data is commonly ingested to build large-scale knowledge graphs (Paulheim 2017; Gomez-Perez et al. 2017) such as DBpedia, Yago, and Freebase (Lehmann et al. 2015; Bollacker et al. 2008; Hoffart et al. 2013). Such Wikipedia-based knowledge graphs powers search, conversational agents, question and answering, and product recommendations for many major technology companies such as Google and Microsoft (Noy et al. 2019). Thus, Wikipedia serves not only as the de-facto online encyclopedia for everyday usage, but its data underlies various datasets, machine learning models, recommendation systems, and other services. Having a clearer understanding of Wikipedia data will help to uncover biases in existing knowledge bases and provide guidance to improve the accuracy and comprehensiveness of many products. Past work has performed such analyses of Wikipedia's potential biases in representation, particularly of gender (Graells-Garrido, Lalmas, and Menczer 2015; Siddiqui 2015). However, large-scale studies of biases in gender and other aspects of cultural representation is lacking. We here focus on musical artists and music genres.

Wikipedia plays a key role in the documentation and representation of art and culture. Early studies found that roughly 43% of articles cover the "entertainment" category, with "music" being the most popular subcategory (Spoerri 2007; Kittur, Chi, and Suh 2009). Another study similarly found "culture and arts" to be the top category, followed by "People and self" (Heist and Paulheim 2019). Largely driven by fans, it appears that the Wikipedia community has particular interest in popular musical artists, bands and singers (Halavais and Lackaff 2008). Given the large volumes of traffic for viewing and editing articles of music and musicians, Wikipedia has potential to be a very comprehensive repository of music knowledge in the world. However, it may at the same time be vulnerable to amplifying certain biases that exist in its community of editors and viewers, as well as of society at large. Considering the popularity of music as a Wikipedia 'destination', and the monetary and societal influence of entertainment (Gioia 2019), it is somewhat surprising that larger-scale analyses focused on musical artist representation are not readily available.

We provide understanding of Wikipedia representation of artists at scale. Our contributions are three-fold:

- First we show that streaming popularity highly correlates with Wikipedia representation. Using an automated ap-

[1] https://en.wikipedia.org

proach, we were able to link the most popular 50,000 music artists (using streaming popularity on a large audio streaming platform as a proxy for popularity) to their English Wikipedia page with an 95% accuracy as estimated based on expert annotator checks of an artist sample.

- Second we comprehensively analyze the representation of a large population of notable figures in the music domain. Entity linking allows us to examine the representation of a large population, instead of the need to make inferences extrapolated from smaller samples.
- Finally, we examine and quantify which types of artists are left out and which ones are well represented with respect to gender and genre. We discuss the gaps in representation and their implications.

Related Work

Not only is music one of the most popular subjects on Wikipedia, it deeply influences society and culture. Music plays a critical role in social movements, by promoting collective identity, driving emotions, and fueling social and political protests (Mondak 1988; Danaher 2010; Peddie 2017). Moreover, music is a multi-billion dollar industry, with continuing projections of double-digit growth [2]. Thus, the music industry operates at a scale that deserves critical examination and is an important contributor of culture and society. To date, there have been very few large-scale examinations of differences in the representation of different types of musical creators on Wikipedia. Considering the considerable barriers for women in the music industry (Smith, Choueiti, and Pieper 2018), as well as importance of genre-related communities as cultural drivers (Baym 2012), here we focus on gender and genre to see whether Wikipedia reflects these industry dynamics.

Knowledge Graph Coverage and Refinement

Wikipedia data is commonly ingested as a basis for building industrial-scale knowledge graphs (Gomez-Perez et al. 2017). However, knowledge graphs are not static products but undergo continual refinement and improvement, requiring constantly adding missing knowledge or removing inaccuracies(Paulheim 2017; Blanco et al. 2013). Knowledge graph completion is a topic of active research in the scientific community. Although no knowledge graph will ever reach a state of full completion, efforts to fill in gaps in knowledge, or awareness of potential biases in the knowledge bases are needed. In this study, we examine whether there exist biases in Wikipedia representation of musical artists. Studying representation of musical artists provides particularly rich data, as musicians and music garner significant engagement on Wikipedia(Spoerri 2007; Halavais and Lackaff 2008; Kittur, Chi, and Suh 2009). Uncovering potential gaps or discrepancies in representation of artists or other notable figures is an important step towards building more accurate and complete knowledge bases.

[2]https://www.riaa.com/reports/riaa-releases-2019-year-end-music-industry-revenue-report/

Representation on Wikipedia

Issues of representation and algorithmic bias have been topics of growing attention. Past studies show that biases often reflect existing societal inequities (Buolamwini and Gebru 2018; Sweeney 2013) and have highlighted the importance of representing identities well (Schlesinger, Edwards, and Grinter 2017). One open question is whether, for example, Wikipedia representation of music reflects traditional barriers to underrepresented creators. A multitude of work has investigated differences in how different groups of people are represented on Wikipedia. (Samoilenko and Yasseri 2014) analyzed Wikipedia representation of notable academics. They reveal that Wikipedia may be producing an inaccurate view of academia, as there is no significant correlation between Wikipedia article metrics and academic notability (as measured by academic publication citation metrics).

(Callahan and Herring 2011) examined differences in the ways in which notable figures from Poland and the United States are represented in the Polish versus English language editions of Wikipedia. Several past studies have examined gender representation on Wikipedia (Wagner et al. 2015, 2016; Reagle and Rhue 2011). (Reagle and Rhue 2011) examined article length and coverage of women vs men on Wikipedia. (Wagner et al. 2015) and (Wagner et al. 2016) assess how notable men and women are represented along several dimensions. They show that women on Wikipedia are more notable, indicating a glass-ceiling effect. They also observe structural and lexical biases against women. For example, articles on women are more likely to be linked to men than vice versa and are more likely to include discussion about romantic relationships and family-related issues. (Graells-Garrido, Lalmas, and Menczer 2015) studies gender representation on Wikipedia by examining network structure, meta-data (e.g. infobox attributes), and language (e.g. frequent unigrams and bigrams). They show that women are more likely to be associated with certain meta-data attributes such as "spouse" and certain categories of words. Further, women have lower node centrality and less than expected incoming links from pages of men. Thus, different groups have varying levels of representation along different dimensions. Although some of the notable people examined in these studies may have included those in the music industry, it is unclear whether the findings observed in the general group of notable figures holds true to the music domain, specifically.

Representation on Wikipedia appears largely a function of its editor community, which is a population with particular characteristics and norms. For example, the 'average' editor is Caucasian, white-collar, technologically-inclined, and male (Wikipedia 2020). The number of female editors in the US is estimated to lie around 22.7%, and globally around 16.1% (Hill and Shaw 2013). Geiger presents a vignette of the hurdles a newcomer editor faces in navigating Wikipedia workflows, and a study of the meta-infrastructure that influences editing on Wikipedia (Geiger 2017). More specifically, Collier and Bear explore various psychological barriers facing female contributors such as lack of confidence in expertise, avoidance of conflict and seeking collaboration rather than deleting and editing others' work (Collier

and Bear 2012). Similarly, Menking and Erickson explores how female editors experience gender-based hostility such as vandals, trolling and edit wars (Menking and Erickson 2015).

In an effort to mitigate such barriers, Wikipedia has recently announced its goal to draft policies to fight harassment and toxic behavior on the platform (Verge 2020). Various efforts aim to better support communities in maintaining Wikipedia. This includes automated quality assessment methods (Dang and Ignat 2016) to improve content, but also tooling for editors and creating safer spaces (Morgan et al. 2013), including aligning tooling with different, sometimes conflicting values (Smith et al. 2020). Bipat et al., for example, also find that editor interactions and facilitation strategies on talk pages differ between Spanish and English Wikipedia, which suggests that tooling may have consequences on the content presented in different cultural or language editions (Bipat et al. 2019). As a step towards lowering barriers and increasing coverage of underrepresented groups, there have been targeted initiatives. For example, there have been meetups and edit-a-thons to encourage editors to write more women[3] and black culture into Wikipedia[4]. Wikimedia researchers also have set up programs to increase coverage, and for example developed methods to recommend articles-to-create for entities that exist in one language but not in another (Wulczyn et al. 2016), and methods to assess article quality (Halfaker 2017).

Additional work explores how different online platforms may relate to Wikipedia, as illustrated by for example, the relationship between Reddit posts and Wikipedia pageviews (Moyer et al. 2015; Vincent, Johnson, and Hecht 2018) emphasize how Wikipedia can play a role in driving engagement, revenue, and traffic to services such as Stack Overflow and Reddit. Thus, Wikipedia as a website is not stand-alone, but interacts with many other online platforms and communities. Exploring these interactions is important, as Wikipedia biases can influence other platforms, and vice versa.

Representation in Music

While many music genres present space to explore representation, and creativity, well-known barriers exist to making music that reaches the mainstream, throughout history (Pendle 2001), as well as the contemporary music industry. Women are underrepresented as top artists compared to the population as a whole (Smith, Choueiti, and Pieper 2018), and in certain genres female representation has even gone down over the years (Watson 2019). Women have faced barriers ranging from networks to studio circumstance. (Gioia 2019) describes the subversive nature of new music, made by outsiders that challenge dominant norms, but the also recurring pattern of those same outsider genres gradually becoming part of the mainstream, appreciated by dominant classes.

[3]https://en.wikipedia.org/wiki/Wikipedia:Meetup/Writing_women_into_Wikipedia_IAP_workshop_-_MIT

[4]https://en.wikipedia.org/wiki/Wikipedia:Meetup/NYC/Black_Life_Matters_Editathon

Few studies have examined musicians' representation on Wikipedia. (Siddiqui 2015) compared the rankings of top guitarists on Wikipedia via rankings from the Rolling Stone, Telegraph, Guitar World, and Gibson by calculating Pagerank on a guitarist network. They find that Wikipedia-derived guitarist rankings revealed many similarities but also differences to traditional guitarist rankings which have led to unexpected 'discoveries' of otherwise lesser unknown guitarists. This suggests that Wikipedia may be emerging as not just as a repository for cataloging the most famous, but also as a source for discovering underrepresented but notable people within their specialties. While past work has introduced multiple ways to examine different dimensions of representation of various groups of people on Wikipedia, to date, large-scale analysis of gender and genre representation among music creators is missing. In the present study, we investigate the Wikipedia coverage of the 50,000 most popular artists from a large audio streaming company. We use streaming popularity as a proxy for popularity and we study how music creators of different gender and genres are represented.

Methods

We take a sample of the top 50,000 artists by global stream count over a 90-day window in 2020 Spotify and link them to their Wikipedia page, if it exists. This requires reliable entity linking, and ensuring accuracy of matching artists to their Wikipedia entity at scale.

Popularity

To date, Spotify has 320 million monthly active users and is present in 92 markets (investor.spotify.com). Thus, while Spotify is not available in every market, stream count data from Spotify is perhaps one of the best approximations to understanding who are the most popular artists worldwide. High stream count numbers can be due to many users streaming an artist or few users streaming an artist a lot. We find that stream count is significantly correlated with number of monthly active users, indicating that the popular artists we examine are streamed by many users.

Matching Artists to Wikipedia Entity

Entity linking is achieved using a combination of open-source algorithms[5] and heuristics that match artist's names to article text. We use the artists' name, aliases, and overlap in their Spotify biographies and discographies for identification and disambiguation, and verify the wikipedia page is likely to refer to an artist in terms of category. Note that we use the term 'artist' to refer to musical entities that could comprise either an individual or a group of individuals. Individual artists may include singers and instrumentalists while groups may include bands, orchestras, and other ensembles.

We evaluated the accuracy of our entity linking by taking a random sample of 1100 artists and asked expert annotators to manually match each artist to their English-language

[5]https://github.com/mwclient/mwclient; https://github.com/earwig/mwparserfromhell

Wikipedia entity. All these experts are hired in-house and were music tech professionals with extensive music data experience. We used this hand-curated dataset to benchmark performance. We define true positives as artists that have a Wikipedia page and are assigned to the correct page. True negatives are artists that are not on Wikipedia and no page is assigned. False negatives are artists who are on Wikipedia but a page is not assigned. False positives are cases where a page is assigned incorrectly, or if that artist is not on Wikipedia. Some of the true positives could be further verified by matching the artist's Spotify ID to Wikidata property SPOTIFY ARTIST ID [6]. However, a large number artists exist that are on Spotify and are on Wikipedia but do not have this property filled out on Wikidata. For example, there are only approximately 500 groups labeled under the rock genre on Wikidata with Spotify artist IDs. Thus, this method would lead to many false negatives and can be used to spot-check, rather than applied at scale. Similarly, in pre-analysis we found that artists with Spotify profiles occasionally fill out Wikipedia links on those profiles as cultural references (e.g. to a genre, or as joke) rather than an own page, making it necessary to use another method to avoid false positives.

To also compare the performance of the entity linker to that of *nonexpert* human *crowdworkers*, we asked crowdworkers to match the same 1100 artists to the correct English Wikipedia page, if it exists. Crowdworkers (using Figure8) were provided with artist names along with their biographies, discographies, and any images, if they exist. We gathered at least three crowd-worker responses per artist and aggregated them by taking the most frequently occurring response. Note that the performance of a single crowdworker is likely lower than that of aggregated crowdworkers. Crowdworkers were compensated above minimum wage requirements for the US.

Overall, we find that the music artist entity linker used here has high performance accuracy and is at least on par with crowdworkers. The overall accuracy of the entity linker is 94.6% while the accuracy of the aggregated responses of crowdworkers was 92.9%. Moreover, we verify the accuracy of the entity linker across artist popularity. We find that the overall accuracy is 97% for the top level of artists (ranked in the top third of popularity) and is 93% for artists in the bottom level (ranked in the bottom third of popularity). In comparison, crowdworkers' accuracy is 95% for the top level of artists and 92% for the bottom level. Thus, we verified the accurate performance of the in-house entity linker and applied it at scale. For a further investigation, we found that false positives were more likely in crowdsourced data due to the incentive of wanting to do well on a task. Lastly, we verified that the entity linker works well for all gender and genre labels, with no systematic differences in accuracy for different artist types.

Plotting Percent of Coverage on Wikipedia

We assign a binary label (0 or 1) to artists to denote the presence or absence of a Wikipedia article. We plot a moving average across artist rank (window size 1000) to visualize the chance of being on Wikipedia across popularity. Because the window size is large (1000 artists), the moving average plots of Wikipedia coverage appear smooth (i.e. there are no gaps) in plots of artists of different genders and genre.

Wikipedia Page Statistics

We examine four main dimensions of every artist's English Wikipedia page: 1) the amount of page content, 2) Pagerank, 3) number of community edits, and 4) musical notability. To quantify the amount of content on a page, we count the number of words, images, and infobox attributes. Next, we examine the connectivity of the page by computing the overall Pagerank of all English Wikipedia entities with respect to the entire English Wikipedia corpus. Pagerank scores indicate significance of an entity in the graph. The directed edges between entities in the Wikipedia graph helps us identify which artists are well-documented and referred more extensively. There are several scalable implementations of Pagerank; in this work we computed Pagerank scores using DANKER[7] on an April 2020 snapshot of English Wikipedia, taking into account resolving links, redirects, Wikidata Q-IDs. In addition, we count the number of incoming links per page. To obtain a measure of wikipedia community perception of musical notability, we look for warning boxes that flag the page for potentially not meeting Wikipedia's notability guideline for music[8]. It is noteworthy that these notability guidelines may not be fully equitable to all music genre cultures.

Genre Labels

Genres are used to organize music, but themselves reflect complex historical processes. Their boundaries are fuzzy, and new genres and names appear and disappear from usage over time. They can for example be geographically or culturally defined, based on technical requirements, or marketing considerations. Artists and tracks do not necessarily 'belong to', 'produce' or 'use' one genre alone (Scaringella, Zoia, and Mlynek 2006; Sturm 2013). This means that there is an inherent challenge in assigning an artist to a genre. In this study, we depend on the main top-level genre assigned to an artist on the basis of their content by the streaming service used for our sample. While more fine-grained labels are available (see for example everynoise.com for a collection of 4,852 genre labels), and while artists can make tracks that fall in multiple genres, this dataset assigns one main top-level genre to each artist. High-level genres include for example hip hop/rap, Latin, pop, metal and R&B. Among the 50,000 artists we examined, genre metadata labels were available for 90%. Among the genre-labeled artists, the most frequently occurring genres in descending order are pop, hip hop, dance/electronic, rock, indie, Latin. These genres account for 80% of artists in this dataset. Excluded from analysis were genres not well-represented in this top 50,000, as well as a genres such as spoken-word, soundtrack, comedy and kids music.

[6] https://www.wikidata.org/wiki/Property:P1902

[7] https://github.com/athalhammer/danker

[8] https://en.wikipedia.org/wiki/Wikipedia:Notability_(music)

Gender Labeling

We used a commercially available music metadata set for gender and genre labels for the included artists. Overall, 70% of artists in this sample had gender labeling. For these gender labels we relied on a professionally curated dataset from a third-party company. This dataset is currently the largest in industry and based on publicly available information deemed credible by the professional music-focused data labeling team. For each artist entity in this particular at-scale data set, a gender entry states whether they are female, male, a mixed multi-gender creator group (e.g. orchestra, band, duo), or unknown/other. The latter covered both non-binary as well as unknown gender artists, meaning that we cannot distinguish between other genders in our analysis than male, female, and multi-gender groups. Less than half a percent in our processed dataset were labeled 'unknown/other' gender, which we do not analyze in this study. This means this analysis is not inclusive to non-binary gender artists, even though binary conceptualization of gender is inaccurate (**?**Schlesinger, Edwards, and Grinter 2017) and prominent artists identify as non-binary. Future work will have to consider more inclusive labeling.

To double check we could rely on the labels for the purposes of this study, we examined the accuracy of the female, male, multi-gender group gender labels by having expert annotators annotate artists across the spectrum of streaming popularity. Note that although the third party company does not provide non-binary labels, we asked the expert annotators to provide non-binary annotations when appropriate. Overall, our annotators agreed with the original third party labels 98.3%. Next, we wanted to see whether there were any systematic errors or biases in gender labeling for artists that lacked labels. We found that the distributions of gender labels were not statistically different between artists with and without labeling (Chi-square test of independence, $p > 0.05$). Thus, we do not observe any systematic problems among artists that lack labeling, and take this dataset as sufficient for our purposes.

In our study, there are approximately 8000 male artists and 2000 female artists in our top level (top third of artists). The distributions are similar throughout the entire population of 50K. It is worthwhile noting that the study by Epps et al, (Epps-Darling, Bouyer, and Cramer 2020) randomly sampled artists from all levels of popularity in a Spotify dataset. While not fully conclusive (since that study used a hand-labeled sample from millions of artists), it was found that the proportion of female artists are higher at lower entry-level popularity levels, slightly lower in middle levels, and higher again at the most popular, superstar level.

Statistical Testing

We compared the distributions of Wikipedia representation across gender and genre using two-sample t-tests with Bonferroni correction. To examine differences in Wikipedia representation while controlling for popularity, we divided artists into three levels of popularity according to rank (referred to as top level, middle level, and bottom level). Each level contains approximately one-third of all artists. We used a significance value of 0.00011 (Bonferroni correction, comparisons across 3 genders x 6 genres x 8 Wikipedia-related features x 3 levels). We also report means and standard errors for various Wikipedia page statistics (denoted s.e.m., or standard error of the mean).

Results

Wikipedia Representation by Artist Consumption

After establishing the high performance accuracy of our music artist-specific entity linker, we applied it at scale. We matched the 50K most streamed artists on a large audio streaming platform to their English Wikipedia page, if it exists. Overall, 42% of the artists are covered on Wikipedia. As expected, the most streamed artists were also the most likely to have Wikipedia pages (Figure 1, Left). We rank artists by their streaming popularity and observe that among the top one thousand artists, 90% are covered on Wikipedia. The probability of being represented falls to 50% after approximately the ten thousandth ranked artist. These top ten thousand artists account for over 80% of the streams. An increase in streams by an order of magnitude corresponds to approximately a 20-30% increase in an artist's chances of being on Wikipedia (Figure 1, Right). Representation drops to less than 30% for the bottom level of artists (defined as the bottom third of popularity, by artist rank). In all, streaming popularity is tightly correlated with Wikipedia representation, indicating that Wikipedia representation of musical creators largely reflects and/or contributes to the online community's engagement with music.

After establishing that Wikipedia representation of artists is highly associated with their streaming popularity on a large audio streaming platform, we next examine how artists of different subgroups are represented on Wikipedia. We hypothesized that if different subgroups are represented similarly, the chance of being on Wikipedia should be the same, controlling for the same level of popularity.

Gender Representation on Wikipedia

Wikipedia Coverage for Artists of Different Gender Labels across Popularity First, we examine how artists of different genders are represented on Wikipedia. We divide artists into those that are labeled as all-female solo/group artists, all-male solo/group artists, or multi-gender groups by a third party company. We note that due to data limitations, this analysis is not inclusive to non-binary gender artists, even though binary conceptualization of gender is inaccurate.

Interestingly, we find that the proportion of female artists represented on Wikipedia in this top popularity sample is significantly higher than that of men and multi-gender artist groups, even when controlling for popularity (Figure 2). We next analyze whether artists of different gender are covered at equal rates. Overall, 65% of sampled women vs 53% of men are on Wikipedia (chi-square test p = 6.464e-16). To see whether this difference in gender representation is consistent across all levels of artist popularity, we examine the representation in male and female artists across three levels of popularity (bucketed by artist rank). 77.5% of women

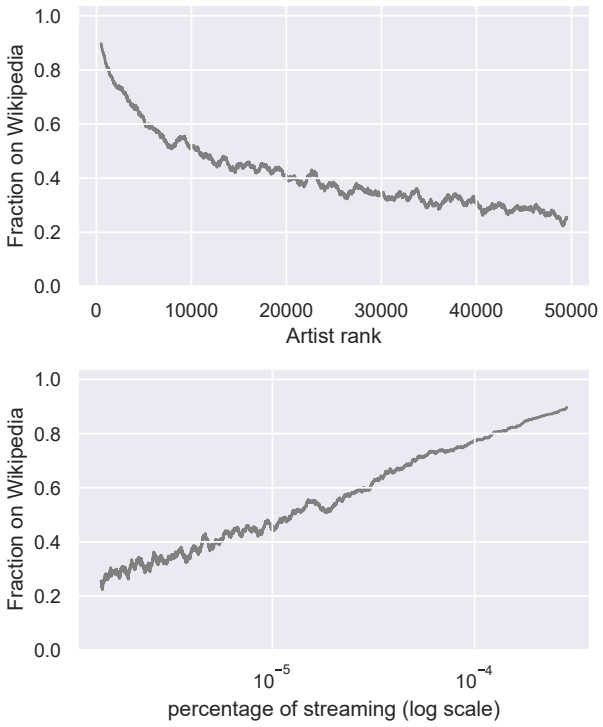

Figure 1: Probability of having a Wikipedia page across artist popularity. (Top: rolling average plot against artist rank. For each artist, the presence of a Wikipedia page is recorded as a 1 or 0. The rolling average is taken over a window size of 1000. Bottom: rolling average plot against percent of streams, logarithmic scale).

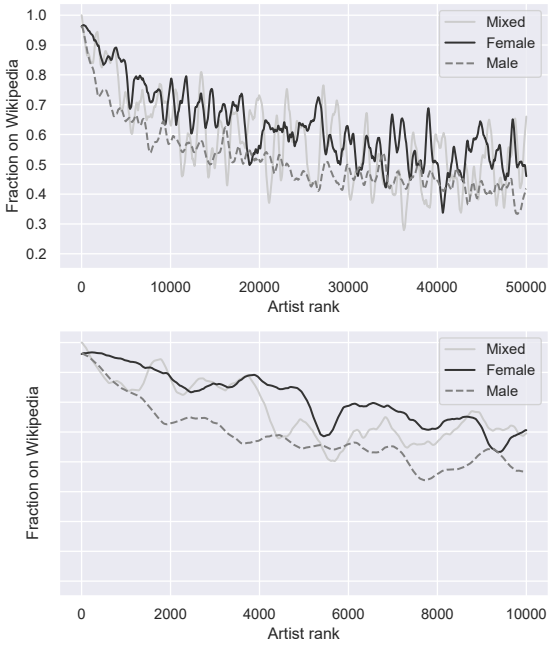

Figure 2: Fraction of artists of having a Wikipedia page across artist popularity, split by gender. Plots are moving averages of the presence (0 or 1) of a Wikipedia page (window size is 1000). Top: all artists, split by gender (unfortunately exluding non-binary due to data limittions). Bottom: zoom in on top 10K artists, same data as plotted on top. Note that due to the window-size for the moving average plot, there are no gaps in the line plot, despite each artist belonging to only one gender bucket.

vs 64.1% of men are represented in the top level; 59.4% of women vs 48.3 % of men are represented in the middle level, and 53.1% vs. 43.5% on are represented in the bottom level (chi-square test of independence p values: level 1: $1.771e{-}07$, level 2: $7.675e{-}07$, level 3: $3.150e{-}05$). Thus, intriguingly, the apparent larger likelihood for female artists to be represented on Wikipedia appears to hold for all levels of popularity - within this top layer of 50k artists. However, there is no significant difference in Wikipedia coverage between all female and mixed gender groups.

The Amount of Content on Wikipedia Pages for Artists of Different Gender Labels Next, we examined whether there are gender differences in the amount of content on Wikipedia pages of artists. We find that Wikipedia articles on female artists contain more content with regard to the number of words, images, and infobox attributes. This effect is especially pronounced for the top level of artists (Figure 3) (Two-sample T-test p-values: comparing number of words, images, and infobox attributes in top level male vs female artists: $1.004e-10$, $2.087e-11$, $1.854e-23$, respectively; the differences for number of words and images are not significant for the bottom level, but the trend remains).

Notability Issues Flagged on Wikipedia Pages for Artists of Different Gender Labels Our results suggesting that top female artists may be comparatively over-represented on Wikipedia are consistent with previous studies that examine gender representation of notable people on Wikipedia such as (Wagner et al. 2015). They propose that their dataset covers few but highly notable women, while including men both notable and less notable. Indeed, in popular music, and also in our dataset, there are more male than there are female creators represented in music (Smith, Choueiti, and Pieper 2018), suggesting that the women that make it into the dataset may have gone through selection pressure. To examine the Wikipedia editor community's perceptions of notability of artists, we parsed Wikipedia pages for warning boxes that flagged the artist page for possibly not meeting Wikipedia's notability guideline for music. We find that 2.1% of male artists on Wikipedia have been flagged as having notability issues while 1.3% of female artists do, making male artists (that do remain on Wikipedia, and are not removed) 1.66 times more likely to have notability issues flagged, although the effect is not significant after Bonferroni correction (chi-square test of independence p value 0.0018). Similar to male artists, 2.2% of multi-gender artist groups have reported notability issues. In sum, there are

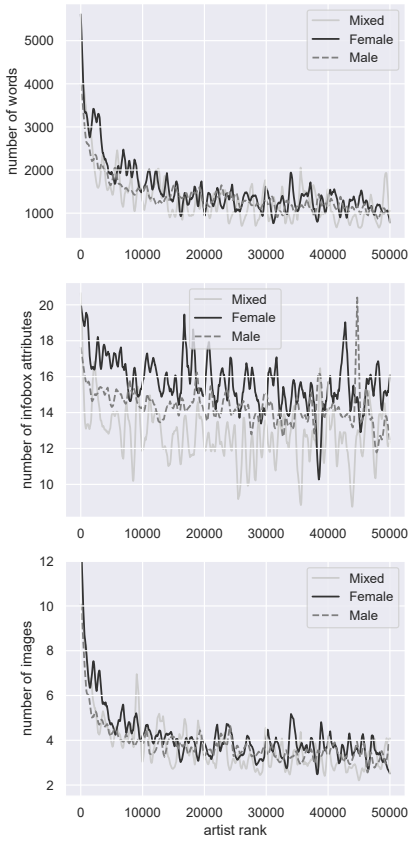

Figure 3: Rolling averages of amount of content on Wikipedia page, as measured by the number of words, number of infobox attributes, and number of images across artist rank. Split by gender, unfortunately excluding non-binary due to data limitations

fewer female artists than male artists, and female artists have fewer notability issues, thus female artists in our dataset may be part of selective group of artists, while male artists may include a mix of notable and less notable artists.

Pagerank of Wikipedia Pages for Artists of Different Gender Labels Last, we examine the Pagerank of female and male artists. We compute the overall Pagerank of every artist page with respect to the entire English Wikipedia corpus. In addition, we count the total number of incoming links to each artist page. Interestingly, we find no difference between men and women with regard to either Pagerank or in-links. On average, male artists have 420 ($\pm$ 11 s.e.m.) in-links while women have 472 ($\pm$ 22 s.e.m.)(Two-sample T-tests p values: top level male vs female in-links $p = 0.0327$; top level male vs female Pagerank $p = 0.987$; middle and bottom level in-links and Pagerank, male vs female all p values > 0.00011). In Graells-Garrido (2015), who examined gender differences in network connectivity of notable people, they find that notable women are less well connected in the network than expected, and that women receive fewer incoming links from men than expected (Graells-Garrido, Lalmas, and Menczer 2015). However, we here observe no significant differences in in-links or PageRank for notable female artists. In sum, with regard to coverage on Wikipedia, the amount of content, female artists show greater representation than male or multi-gender artists. However, there appears to be no proportional boost in female representation in terms of network connectivity.

Gender	Overall	Top level	Middle level	Bottom level
Female artists	65	77	60	53
Multi-gender groups	59	64	57	47
Male artists	53	71	48	44

Table 1: Wikipedia representation (i.e. percent of having a Wikipedia page) of artists of different gender. Wikipedia representation is calculated for the population overall, as well as broken down by popularity levels. Non-binary missing due to data limitations

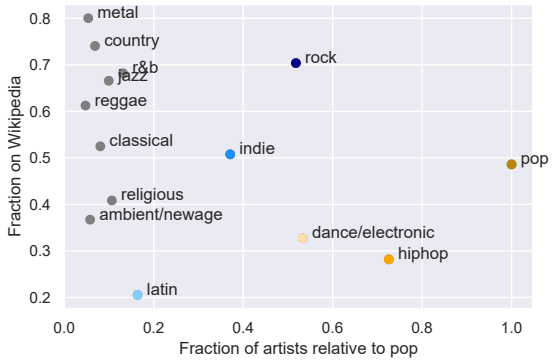

Figure 4: Chance of being on Wikipedia versus proportion of artists in the dataset (relative to the number of pop artists). The most frequently occurring genres are pop, hip hop, dance/electronic, rock, indie, and Latin, respectively, comprising 80% of artists and are colored in the plot. The remaining seven genres that appear less frequently are in grey.

Genre Representation on Wikipedia

We next examine how artists of different genres are represented on Wikipedia. The most popular genres are pop, hip hop, dance/electronic, rock, indie, and Latin, respectively, and account for 80% of artists in our data (Figure 4).

Wikipedia Coverage for Artists of Different Genre Labels across Popularity Strikingly, artists of different genres appear to have widely varying chances of appearing on Wikipedia (Figure 5). When only considering those in the top level (top third of popularity), 85% of rock artists are on Wikipedia. Yet, only 33% of dance/electronic, 28% of hip hop, and 21% of Latin artists are represented for artists of the same popularity level (Table-**??**, top level). Thus, rock artists have over twice as much representation as the least represented groups. Pop and indie artist representation is intermediate: 64% of top level indie artists and 65% of top

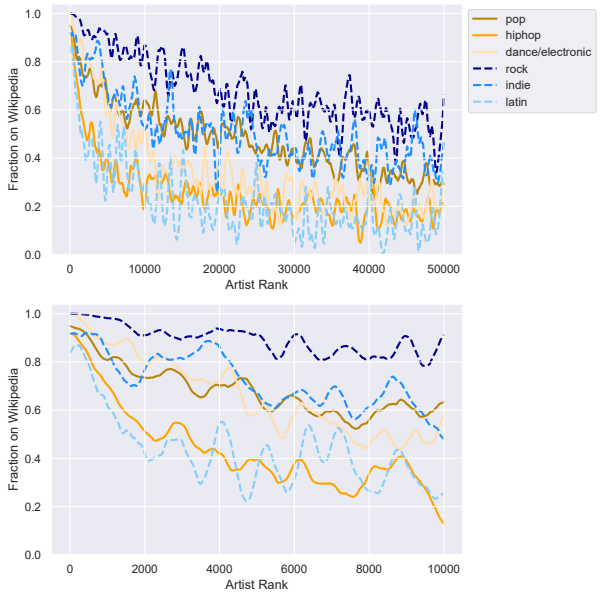

Figure 5: Fraction of artists on Wikipedia across artist popularity, split by genre (rolling average). Top: all artists. Bottom: zoom in on top 10K artists

level pop artists are on Wikipedia. Rock artists' dominance on Wikipedia persists even into the bottom level of popularity, where 57% are represented, while dance/electronic, hip hop, and Latin artists have 21%, 18% and 13% representation, respectively (chi-square test p values: top level rock vs. hip hop 3.756e-61; bottom level rock vs hip hop 1.199e-65; top level rock vs Latin 1.109e-25; bottom level rock vs Latin 9.082e-36). Pop and indie artists, again, have intermediate representation in the bottom level, which is 34% and 40% for pop and indie artists, respectively. Even more striking is among the top one thousand– as mentioned earlier, 90% of the top one thousand artists are on Wikipedia. We find that 67% of Latin artists and 85% of hip hop artists are covered, while 99% of rock artists are covered. One explanation for Latin artists' low representation is that we restrict all analyses to English Wikipedia. Latin artists may in fact be well represented on Spanish-language pages, but analyzing other language editions of Wikipedia goes beyond the scope of this study. However, language differences cannot explain why hip hop and dance/electronic artists are less well represented on English Wikipedia.

The Amount of Content on Wikipedia Pages for Artists of Different Genres We next limit analyses to the artists that have Wikipedia articles and analyze the amount of content on their pages. We find that rock artists generally have the most words and images, followed by pop and indie, while Latin and dance/electronic artists have consistently the least amount of content. (Figure 6). hip hop artists, despite having lower chances of being on Wikipedia, have an intermediate amount of content on their pages. For example, hip hop artists have comparable number of words to indie artists (average number of words of top level hip hop artists: 1400

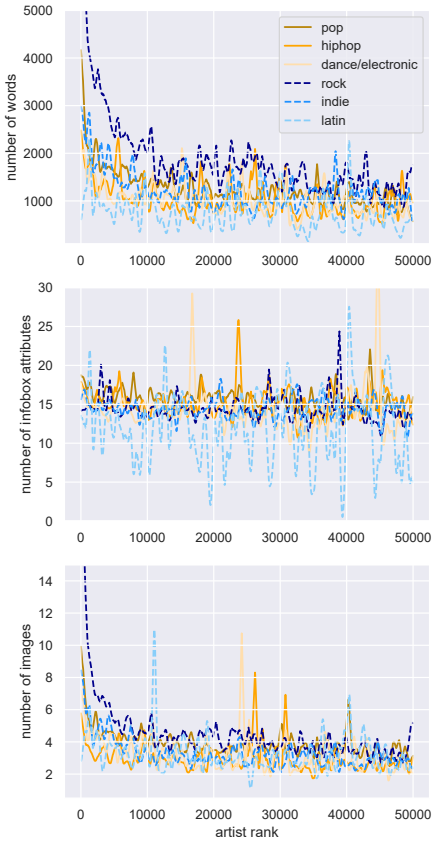

Figure 6: Amount of content on Wikipedia page, as measured by the number of words, number of infobox attributes, and number of images, split by genre

(± 45 s.e.m.), indie artists: 1505 (± 48 s.e.m.) (Two-sample T-test p-values: number of words, top level hip hop vs indie artists 0.119; middle level hip hop vs indie artists 0.850; bottom level hip hop vs indie artists 0.002). Interestingly, hip hop artists have significantly more infobox attributes than rock artists in the top level and are not significantly different from pop artists, who have the most infobox attributes (average number of infobox attributes of top level hip hop artists: 15.6 (± 0.2 s.e.m.); top level rock artists 14.5 (± 0.1 s.e.m.); top level pop artists 16.4 (± 0.2 s.e.m.) (two-sampled *t-test* p values: number of infobox attributes top level hip hop vs rock 1.073e-05; top level hip hop vs pop 0.0014 not significant). On the other hand, top level dance/electronic artists have on average 13.9 (± 0.3 s.e.m.) infobox attributes while top level Latin artists have 11.4 (± 0.7 s.e.m.). Thus, article content is highest for rock artists and lowest for Latin and dance/electronic artists. hip hop artists, however, have intermediate to high amount of content.

Pagerank of Wikipedia Pages for Artists of Different Genre Labels Next, we examine Pagerank of artist pages and number of incoming links for each page. We find that rock artists are the most centrally connected, followed by hip hop and pop artists (number of in-links for top level rock

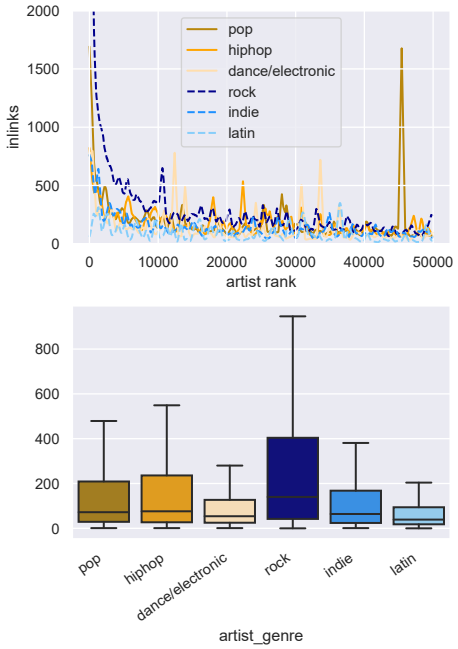

Figure 7: Frequency of community edits, split by genre (Top: average weekly edits (over six month period) vs artist level. Bottom: Boxplot to visualize average weekly edits over all artist levels for each genre)

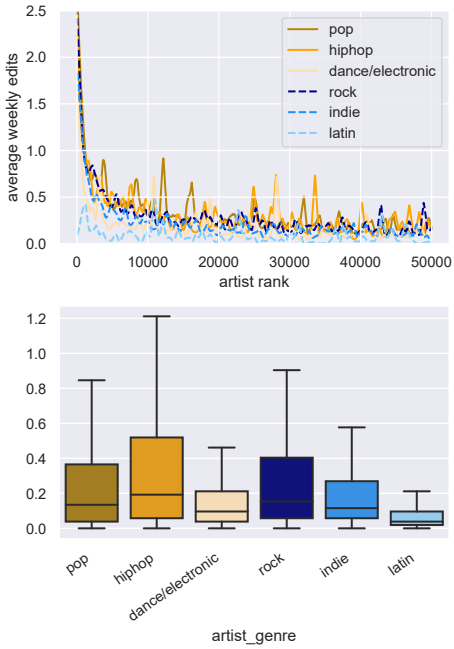

Figure 8: Frequency of community edits, split by genre (Top: average weekly edits (over six months) vs artist level. Bottom: Boxplot to visualize average weekly edits over all artist levels for each genre)

artists: 670.8 $\pm$ 29.0; in-links for top level hip hop: 312.7 $\pm$ 16.7; in-links for top level pop: 334.3 $\pm$ 14.2; Pagerank for top level rock artists: 4.6 $\pm$ 0.2; Pagerank for top level hip hop: 2.0 $\pm$ 0.1; Pagerank for top level pop: 1.9 $\pm$ 0.1)(Two-sample *t-test* p values for top level: in-links, rock vs hip hop 2.774e-21, Pagerank, rock vs hip hop 2.268e-23; in-links hip hop vs pop $p > 5e-05$; Pagerank hip hop vs pop $p > 5e-05$). Indie, dance/electronic, and Latin, however, are least well connected. (Figure-7). Thus, although rock artists are the most well connected, hip hop artists are next best connected.

The Frequency of Community Engagement on Wikipedia Pages for Artists of Different Genres Finally, we quantify community engagement by computing the average number of weekly edits over a six month period between 2019 and 2020. Surprisingly, hip hop artists have the highest numbers of community edits, which is 1.5 times higher than that of the second highest genre, which is rock. This effect is particularly prominent in the top level (Figure-8, average weekly edits, top level hip hop: 0.716 ($\pm$ 0.03 s.e.m.); top level rock: 0.534 ($\pm$ 0.02 s.e.m.); top level pop: 0.669 ($\pm$ 0.027 s.e.m.)) (Two-sample `t-test` p values: top level hip hop vs rock: 4.522e-08; all levels hip hop vs rock: 6.860e-16). Thus, although the chance of being on Wikipedia is low for hip hop artists, there appears to be high levels of editorial engagement with those rock artists who already have pages.

Genre	Overall	Top	Middle	Bottom
rock	70	85	64	57
indie	51	64	46	40
pop	49	65	46	34
dance/electronic	33	52	28	21
hip hop	28	40	22	18
Latin	21	34	16	13

Table 2: Wikipedia representation of the top six genres (percent of artists on Wikipedia overall as well as split by three levels of popularity: top, middle, and bottom.)

Summary of Wikipedia Representation of Genres In summary, hip hop artists have lower chances of being on Wikipedia. However, of those that do make it onto Wikipedia, they are very well connected, garner high levels of community engagement, and have relatively high quantities of content on their Wikipedia pages. Rock artists are the most extensively represented on Wikipedia in all aspects. In contrast, for the same level of popularity, Latin and dance/electronic artists have persistently low levels of representation on all dimensions.

Conclusion & Discussion

This study allowed us to analyzing Wikipedia coverage through accurately matching artists to their Wikipedia pages at scale, with high precision and recall across different ranges of artist popularity. We find that Wikipedia coverage is 90% for the top one thousand most popular artists,

Genre	Percent with notability issues
dance/electronic	3.5
hip hop	3.2
Latin	3.0
pop	2.0
indie	1.7
rock	1.5

Table 3: Percent of artists that may fail to meet Wikipedia's existing guideline for notability for music

and drops to around 50% after the ten thousandth artist. We see a clear correlation between streaming popularity and Wikipedia representation, which suggests that Wikipedia content largely reflects global musical taste and may even have potential to influence music consumption. Future work will be required to tease apart causality between Wikipedia representation and streaming patterns. We however demonstrate that artists of different genders and genres are represented differently on Wikipedia.

Gender We find a slight under-representation of male artists compared to female artists in this top 50k sample, with regard to the chance of being on Wikipedia, amount of content, and frequency of community edits. The exception is in Pagerank, where the genders are represented equally. These results may reflect greater (editing) community interest in female artists. It would be interesting to explore the influence of past edit-a-thons and for example Project Wikiwomen in this coverage of female artists [9]. For example, there are indications that dedicated editing efforts and related publicity turned a Wikipedia article quality gap for female scientists into a quality surplus (Halfaker 2017). However, selection bias may also play a role; as the music industry is highly under-representative of women as a whole. There are more male than female artists in the top levels of the music industry, for example, only 21.7% of Billboard top hits' artists are women (Smith, Choueiti, and Pieper 2018). This means that the female artists present in our dataset may have already undergone additional selective pressure and thus may be an exceptionally notable group of artists. In partial support of this, we find a higher fraction of male artists that are claimed by editors to fail to meet notability guidelines for music than female artists.

Genre We observe large differences in Wikipedia representation of artists of different genres. We observe that the coverage of hip hop, Latin, and dance/electronic artists is particularly lacking, while rock artists appear to have the best coverage. As Latin has become one of the most popular genres internationally, including in the largest predominantly English speaking locales (Forbes 2019), Latin artists' under-representation on English Wikipedia is not explainable by differences in language editions alone. We still here see influential entertainment cultures not yet well represented. Possibly somewhat overstated, the continued higher edits of rock in this dataset versus faster growing (but already well established) genres may suggest that the community editing work has not kept up with entertainment culture.

This is especially concerning as studies in ethnomusicology have identified interactions between musical genres and class, race, and locality (Bennett 2017). For example, hip hop is of great cultural importance and can empower underrepresented communities (Travis 2013). However, while hip hop is one of the most popular music genres across consumers, it is often perceived as associated with 'dispossessed' Black youth (Bennett 2008) while "good rock music" is associated with white males by consumers (Schaap and Berkers 2019). Interestingly, despite the fact that the fraction of hip hop artists that make it onto Wikipedia is half that of rock artists, we find that those who do make it onto Wikipedia still have medium to high levels of article content, Pagerank, and frequency of community edits. Thus, there appears to be no lack of community enthusiasm for hip hop artists that are on Wikipedia, but rather, they are lacking in having articles created for them in the first place.

Moreover, we note that although we focus our analyses on the top six genres which account for over 80% of artists in our data, there is an interesting observation in which less popular genres appear to have high Wikipedia coverage (Figure 4). For example, very few artists in our dataset focus on metal or country, but those metal and country artists have high Wikipedia coverage. This effect is likely due to selection: artists in less popular genres that 'make it' on the list of most popularly streamed Spotify artists have underwent selective pressure. These artists are likely of exceptional notability for their respective genres. Future studies should examine artist representation in the less popular genres.

Coverage can potentially be increased through strategies such as ongoing active Wikimedia efforts in automated suggesting of new articles for creation (Wulczyn et al. 2016), community-focused *edit-a-thons*, creating tooling that removes barriers to newcomer editors, active outreach, and implementing anti-harassment policies to help attract but also retain new editors (Verge 2020; Smith et al. 2020).

Limitations

We note several shortcomings of this study. One is that since our dataset is restricted to the most popular 50,000 artists, and there is a long tail of artists that we do not examine. While the chance of being on Wikipedia may be very low for artists with very low popularity, the percent coverage is likely non-zero even very far out in the tail. This is also where fandom and community impacts will potentially result in starker differences.

The focus of the current paper was to examine artist representation on Wikipedia, a general domain database encountered by general audiences. Future studies could compare the representation of artists in this general domain setting to specialized services such as MusicBrainz, a collaborative metadata effort specific to music (Swartz 2002) to investigate differences in community engagement and representation. In addition, this study examines the popularity of artists at a single time window, over a 90 day period. There could be temporal dynamics of popularity and its relation to being added, edited and viewed on Wikipedia worthwhile of exploration. For example, the popularity of an artist may climb

[9] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women

after major cultural events, releasing an album or winning a competition. One potential question would be how soon after an artist emerges as popular, they appear on Wikipedia or more specialized services.

A second caveat is with our genre and gender labels dataset, which is incomplete. The gender metadata lacks non-binary labels and is noninclusive to artists that do not fall into male or female gender categories. Future work will be needed to obtain more inclusive and complete labels, while we also acknowledge that information will not be available for every creator, nor that labeling is always desirable. A tension arises here between ensuring sample coverage to enable representation studies, and minimizing data collection and potential mis-labeling. Here, we restricted analyses to female, male, and mixed-gender artist duos or groups but future studies will require better representation of non-binary and genderqueer artists.

With regard to genre labels, not all artists clearly fall within genre borders, nor are these borders set in stone; many artists creatively span multiple genres and subgenres. We here also restrict analysis to the top six genres of our dataset, as they account for the majority of genres in our data, and all six genres have large numbers of artists per group. There are many genres and subgenres we do not examine. Another limitation to consider is that the dataset is influenced by who stream from online streaming platforms– for example, those who stream music versus listen via other formats such as CDs tend to be younger and have access to the internet (Aguiar and Martens 2016). It will be interesting to also examine Wikipedia representation of artists that are popular in other formats such as CDs and terrestrial radio. Perspectives on how entertainment history is represented in such resources like Wikipedia versus communities' own histories would similarly be welcome. Finally, we point out that the Wikipedia notability guidelines defined for musicians (which governs whether a page is subject to removal) are not static, and the guidelines themselves may be contentious. Past work has suggested that Wikipedia guidelines on notability and verifiability may be biased towards dominant subject matter (Gauthier and Sawchuk 2017). In extension, certain subcultures or genres of music may be more prone to suffer from notability and verifiability issues depending on their definition, thus impacting future evaluations of coverage.

In summary, the present study has found significant differences in Wikipedia representation of different types of notable figures in the music domain. Uncovering such data gaps is a critical step towards mitigating the amplification of inequities, which may ultimately impact the livelihoods of the artists they represent. The information on Wikipedia functions as a cultural memory (Pentzold 2009); an entry point, and resource to learn more for general audiences, in this case about genres, artists and their histories. Thus, what is (not) in Wikipedia is a representation who 'makes it' into that shared cultural memory. Gaps in representation may have significant and widespread impact. Wikipedia content plays a key role in improving other services, and can even affect revenue generation (Vincent, Johnson, and Hecht 2018). An artist's visibility on platforms such as wikipedia could both influence and be influenced by their streaming popularity. This presence can affect the way in which information is, or is not, offered to audiences, which ultimately can affect attention, streams and longer-term audience building. These differences and the wider ecosystem are worthwhile of further investigation.

Acknowledgements

We would like to thank Andrew Shires, Kurt Jacobson, Chris Lee, Joy Chugh, Matt Solomon, and Matt Finkel for their work on entity linking. We would also like to thank Sam Way and Bernd Huber for help on the figures and comments on the manuscript. Additional thanks for Ben Cooper and Maricarmen Rogers.

References

Aguiar, L.; and Martens, B. 2016. Digital music consumption on the internet: evidence from clickstream data. *Information Economics and Policy* 34: 27–43.

Ahrefs. 2020. Top 100 Most Visited Websites by Search Traffic (as of 2020). https://ahrefs.com/blog/most-visited-websites. Accessed: 2020-05-01.

Baym, N. K. 2012. Fans or friends?: Seeing social media audiences as musicians do. *Participations* 9(2): 286–316.

Bennett, A. 2008. Towards a cultural sociology of popular music. *Journal of Sociology* 44(4): 419–432.

Bennett, A. 2017. *Music, space and place: popular music and cultural identity*. Routledge.

Bipat, T.; Davidson, D. V.; Guadarrama, M.; Li, N.; Black, R.; Marsden, D. W.; and Zachry, M. 2019. How does Editor Interaction Help Build the Spanish Wikipedia? In *Proceedings of CSCW*, 156–160.

Blanco, R.; Cambazoglu, B. B.; Mika, P.; and Torzec, N. 2013. Entity recommendations in web search. In *International Semantic Web Conference*, 33–48. Springer.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD*, 1247–1250.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91.

Callahan, E. S.; and Herring, S. C. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology* 62(10): 1899–1915.

Collier, B.; and Bear, J. 2012. Conflict, criticism, or confidence: an empirical examination of the gender gap in wikipedia contributions. In *Proceedings of the CSCW*, 383–392.

Danaher, W. F. 2010. Music and social movements. *Sociology Compass* 4(9): 811–823.

Dang, Q. V.; and Ignat, C.-L. 2016. Quality Assessment of Wikipedia Articles without Feature Engineering. JCDL '16, 27–30. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342292.

Epps-Darling, A.; Bouyer, R. T.; and Cramer, H. 2020. Artist gender representation in music streaming. In *ISMIR*, 248–254.

Forbes. 2019. Latin Music Is Now More Popular Than Country & EDM In America. https://www.forbes.com/sites/jeffbenjamin/2019/01/04/latin-music-in-2018-album-song-sales-consumption-buzzangle-report/#61679425adde.

Gauthier, M.; and Sawchuk, K. 2017. Not notable enough: feminism and expertise in Wikipedia. *Communication and Critical/Cultural Studies* 14(4): 385–402.

Geiger, R. S. 2017. Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data & Society* 4(2): 2053951717730735.

Gioia, T. 2019. *Music: A subversive history*. Hachette UK.

Gomez-Perez, J. M.; Pan, J. Z.; Vetere, G.; and Wu, H. 2017. Enterprise knowledge graph: An introduction. In *Exploiting linked data and knowledge graphs in large organisations*, 1–14. Springer.

Graells-Garrido, E.; Lalmas, M.; and Menczer, F. 2015. First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 165–174.

Halavais, A.; and Lackaff, D. 2008. An analysis of topical coverage of Wikipedia. *Journal of computer-mediated communication* 13(2): 429–440.

Halfaker, A. 2017. Interpolating quality dynamics in Wikipedia and demonstrating the Keilana effect. In *Proceedings of the 13th International Symposium on Open Collaboration*, 1–9.

Heist, N.; and Paulheim, H. 2019. Uncovering the semantics of Wikipedia categories. In *International semantic web conference*, 219–236. Springer.

Hill, B. M.; and Shaw, A. 2013. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one* 8(6).

Hoffart, J.; Suchanek, F. M.; Berberich, K.; and Weikum, G. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194: 28–61.

Kittur, A.; Chi, E. H.; and Suh, B. 2009. What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In *Proceedings of the ACM CHI*, 1509–1512.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6(2): 167–195.

Menking, A.; and Erickson, I. 2015. The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the 33rd annual ACM CHI*, 207–210.

Mondak, J. J. 1988. Protest music as political persuasion. *Popular Music & Society* 12(3): 25–38.

Morgan, J. T.; Bouterse, S.; Walls, H.; and Stierch, S. 2013. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of CSCW*, 839–848.

Moyer, D. C.; Carson, S. L.; Dye, T. K.; Carson, R. T.; and Goldbaum, D. 2015. Determining the influence of Reddit posts on Wikipedia pageviews. In *Proceedings of the Nineth ICWSM*.

Noy, N.; Gao, Y.; Jain, A.; Narayanan, A.; Patterson, A.; and Taylor, J. 2019. Industry-scale knowledge graphs: Lessons and challenges. *Queue* 17(2): 48–75.

Paulheim, H. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web* 8(3): 489–508.

Peddie, I. 2017. *The resisting muse: popular music and social protest*. Routledge.

Pendle, K. 2001. *Women & music: a history*. Indiana University Press.

Pentzold, C. 2009. Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place. *Memory Studies* 2: 255–272.

Reagle, J.; and Rhue, L. 2011. Gender bias in Wikipedia and Britannica. *International Journal of Communication* 5: 21.

Samoilenko, A.; and Yasseri, T. 2014. The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ data science* 3(1): 1.

Scaringella, N.; Zoia, G.; and Mlynek, D. 2006. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine* 23(2): 133–141.

Schaap, J.; and Berkers, P. 2019. "Maybe it's... skin colour?" How race-ethnicity and gender function in consumers' formation of classification styles of cultural content. *Consumption Markets & Culture* 1–17.

Schlesinger, A.; Edwards, W. K.; and Grinter, R. E. 2017. Intersectional HCI: Engaging identity through gender, race, and class. In *Proceedings of the ACM CHI*, 5412–5427.

Siddiqui, M. A. 2015. Mining Wikipedia to Rank Rock Guitarists. *IJISA* 7(12): 47.

Smith, C. E.; Yu, B.; Srivastava, A.; Halfaker, A.; Terveen, L.; and Zhu, H. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. CHI '20.

Smith, S. L.; Choueiti, M.; and Pieper, K. 2018. Inclusion in the recording studio? Gender and race/ethnicity of artists, songwriters and producers across 600 popular songs from 2012-2017. *Annenberg Inclusion Initiative* .

Spoerri, A. 2007. What is popular on Wikipedia and why? *First Monday* 12(4).

Sturm, B. L. 2013. Classification accuracy is not enough. *JIIS* 41(3): 371–406.

Swartz, A. 2002. Musicbrainz: A semantic web service. *IEEE Intelligent Systems* 17(1): 76–77.

Sweeney, L. 2013. Discrimination in online ad delivery. *Queue* 11(3): 10–29.

Travis, R. 2013. Rap music and the empowerment of today's youth: Evidence in everyday music listening, music therapy, and commercial rap music. *Child and Adolescent Social Work Journal* 30(2).

Verge, T. 2020. Wikimedia is writing new policies to fight Wikipedia harassment.

Vincent, N.; Johnson, I.; and Hecht, B. 2018. Examining Wikipedia With a Broader Lens: Quantifying the Value of Wikipedia's Relationships with Other Large-Scale Online Communities. CHI '18.

Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the Nineth ICWSM*.

Wagner, C.; Graells-Garrido, E.; Garcia, D.; and Menczer, F. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science* 5: 1–24.

Watson, J. 2019. Gender on the Billboard Hot Country Songs Chart, 1996–2016. *Popular Music and Society* 42(5): 538–560.

Wikipedia. 2020. The "average Wikipedian".

Wulczyn, E.; West, R.; Zia, L.; and Leskovec, J. 2016. Growing wikipedia across languages via recommendation. In *WWW*, 975–985.