A Twitter Case Study for Assessing Digital Sound

Rutgers University has made this article freely available. Please share how this access benefits you.
Your story matters. [https://rucore.libraries.rutgers.edu/rutgers-lib/55158/story/]

This work is an ACCEPTED MANUSCRIPT (AM)
This is the author's manuscript for a work that has been accepted for publication. Changes resulting from the publishing process, such as copyediting, final layout, and pagination, may not be reflected in this document. The publisher takes permanent responsibility for the work. Content and layout follow publisher's submission requirements.

Citation for this version and the definitive version are shown below.

**Citation to Publisher Version:** Giannetti, Francesca. (2017). A Twitter Case Study for Assessing Digital Sound. *Journal of the Association for Information Science and Technology.*


Terms of Use: Copyright for scholarly resources published in RUcore is retained by the copyright holder. By virtue of its appearance in this open access medium, you are free to use this resource, with proper attribution, in educational and other non-commercial settings. Other uses, such as reproduction or republication, may require the permission of the copyright holder.

*Article begins on next page*
A Twitter Case Study for Assessing Digital Sound

FRANCESCA GIANNELLI

ABSTRACT

Academic and cultural heritage institutions around the world have made measurable strides in the development of digital sound archives oriented towards research and access, but their impact on scholarship and society has been little studied. Traditionally, impact has been measured by citations; yet these are problematic metrics for non-traditional outputs like sound recordings. Social media data provide a promising avenue of investigation for measuring scholarly as well as societal impact. Twitter in particular has been shown to provide a high number of references for cultural and research outputs in all disciplines. This study analyzes Twitter references pertaining to the collections of five digital sound archives: British Library Sounds, Europeana Sounds, the Internet Archive Audio Archive, PennSound and UbuWeb. Using text analysis methods to identify high frequency events and trends, and labeling them with a rubric designed for measuring the impact of digital heritage resources, this study provides preliminary insights on user values as they relate to digital sound collections. Despite the limitations of using social media data, the evidence gathered in this case study characterizes aspects of the use of digital sound collections, and may point to future priorities for the digital preservation of sound.

Keywords: impact assessment; social media; digital libraries; digital archives; humanities; sound recordings

This is a preprint of an article accepted for publication in Journal of the Association for Information Science and Technology copyright © 2018 (Association for Information Science and Technology)
INTRODUCTION

Many reputable sources of digital and digitized archival sound recordings are now available on the web. With funding from governmental, academic, and private philanthropic sources, national libraries and independent archives have undertaken large programs of digitization, while technological advances in streaming and audio compression have introduced a wealth of historical sound to the public via a relatively frictionless user experience. Measuring the impact of these digitized resources remains difficult for a variety of reasons. Not only is the concept of ‘impact’ difficult to define and measure, but it is also often a matter of subjective judgment. As a result, libraries and archives frequently evaluate the impact of digital resources in terms of access and preservation (H. E. Green & Courtney, 2014; Sinn, 2012). These internal metrics are of course critical assessment measures, but they do not supply information on the use of the materials, or on their value to scholars and the general public.

Alternatively, established bibliographic metrics like citation counts provide a rich source of contextual information on use, but they too have their shortcomings, particularly when assessing nontraditional documents. Performances, readings and lectures fixed on audiovisual formats do not have the citation tradition of scholarly articles and monographs (Carpenter, Gunn, Buschman, & Adie, 2014). It is therefore likely that citations will underestimate the reach of these research outputs in the scholarly record. Evidence suggests that scholars in humanities disciplines have a bias against citing digital resources, even when they were used, preferring instead to cite the analogue sources (Hughes, Ell, Knight, & Dobreva, 2015; Sinn, 2012). And then, citations will capture only scholarly impact, representing the use of that small minority of researchers with the professional and political wherewithal to cite these objects in their work (Priem, Piwowar, & Hemminger, 2012). Citations cannot be used to measure impacts among the general public, nor are they likely to capture the use of most academics who, for a variety of reasons, might not cite such objects in their research. The former group is an especially significant audience for sound recordings, given that most users, including researchers of sound, listen to recordings for pleasure. Measuring the recreational use of this kind of digital object requires an entirely different metric.

Approaches drawn from altmetric and social media research may offer new ways of measuring the reach of historical sound recordings. Altmetric
research tends to focus more specifically on the online sharing of scholarly resources, although data gathered from social media messages may help to understand the impact of other types of resources, including audiovisual materials (Bornmann, 2014; Fenner, 2014). There is strong evidence that altmetrics provide a new source of data for evaluating arts and humanities outputs as contrasted with traditional bibliometric measures (Hammarfelt, 2016; McLaughlin et al., 2015). Researchers have demonstrated that humanities outputs have a high density of social media mentions per work, and that humanities research often garners more references on social media than in the scholarly record (Costas, Zahedi, & Wouters, 2014). Of social media sources, Twitter provides the greatest number of references in all disciplines (Costas et al., 2014; Haustein, Costas, & Larivière, 2015). Social media research, which shifts the focus from objects to interactions, models ethnographic approaches to the study of online communities. The study of these interactions allows the researcher to consider user motivations and sociocultural factors, which may help to contextualize references to digital heritage objects.

Empirical studies relying on social media data are often vulnerable to critiques pertaining to data quality and representativeness. Most researchers agree that tweets and other social media metrics point to impact that is not captured in the citation record (Bornmann, 2014; Costas et al., 2014; Priem et al., 2012; Sud & Thelwall, 2013). Data quality, however, is an issue of legitimate concern. Previous research has shown significant levels of commercial and automated tweeting (Hammarfelt, 2014; Robinson-Garcia, Costas, Isett, Melkers, & Hicks, 2017). Researchers have observed high mentions of research outputs with short or funny titles, and have characterized Twitter and Facebook in particular as “content-poor” sources of data (Bornmann, 2014; Haustein et al., 2015). Others caution that altmetric data, which includes social media data, point more towards attention than impact (Sugimoto, Work, Larivière, & Haustein, 2017). Furthermore, the sampling frames implemented by researchers often introduce bias by focusing on a narrow set of features, to the exclusion of other potentially relevant data (Tufekci, 2014). Any social media approach to analyzing the impact of digital heritage resources should therefore consider additional sources of data before drawing conclusions.
RELATED RESEARCH

Relatively few empirical studies focus on the impact of digital resources in the humanities, and of these, fewer still discuss digital sound recordings. The following studies use both quantitative and qualitative measures. The LAIRAH and the SPHERE studies combined webometric and ethnographic data to assess digital libraries, information resources, and, in the latter case, a digital collection of *The Stormont Papers* (Hughes et al., 2015; Warwick, Terras, Huntington, & Pappa, 2008). The LAIRAH study combines an analysis of deep log data and user interviews to identify criteria of use, which include simple descriptive titles, and a preference among users for general over specialist information resources (Warwick et al., 2008). Hughes et al. drew upon the Toolkit for the Impact of Digitised Scholarly Resources (TIDSR) for the selection of quantitative and qualitative methodologies, emphasizing that these measures are best used in combination in order to capture a deeper view of impact (Hughes et al., 2015). Their methods pointed to potential enhancements to be achieved by the integration of new content to the original resource (Hughes et al., 2015). Sinn used quantitative data—ten years of citations—to study historians and their use of digital archives. Her findings show that digital materials appeared more often in figures than in citations, and that digital audio and multimedia resources are cited with far less frequency than text and visual materials (Sinn, 2012).

Green and Courtney use survey and interview data to focus more particularly on the issue of the usability of digital collections. Although they identified greater use of audio among humanists than Sinn, audio recordings were still the least used resources (H. E. Green & Courtney, 2014). The authors’ findings suggest that better metadata and improved searchability would improve the discovery and reuse of these objects, while copyright restrictions remain potent barriers.

Many altmetric studies that include the arts and humanities compare the effectiveness of various metrics across all disciplines, including the social sciences and the sciences. Of the research to focus more specifically on arts and humanities outputs, there is much discussion of the potential for altmetrics, but few applications. Konkiel et al. provide a comprehensive overview of webometrics, including altmetrics, to be leveraged when measuring the value of digital special collections and institutional repositories (Konkiel, Dalmau, & Scherer, 2015). They argue for the need of standards and best practices for implementing such metrics. In one of few studies to show an application
of altmetrics in the humanities, Hammarfelt analyzes a sample of humanities publications on various social media platforms and reference managers. His findings suggest that Twitter is better adapted to analyzing monographs, while Mendeley is more promising for analyzing articles (Hammarfelt, 2014). Hammarfelt’s study relies on the manual assignment of document object identifiers (DOIs), where available, to retrieve data. DOIs are not widely used as document identifiers in the humanities, potentially complicating a researcher’s ability to cross-reference data from services like Altmetric and Plum Analytics. For this reason, a core concern of altmetrics research will remain how best to implement a study of humanities outputs.

Text analysis methodologies in both altmetric and social media research permit researchers to explore the content of social media messages, in addition to counts. Altmetric researchers have used sentiment analysis to score tweets about scholarly articles with the aim of improving our understanding of tweets as measures of scholarly impact (Friedrich, Bowman, Stock, & Haustein, 2015). Social media research in sociological and humanistic disciplines has focused variously on hashtag co-occurrence analysis and Latent Semantic Analysis (LSA) topic modeling, in addition to spatial and network approaches, to investigate and frame arguments on topics as diverse as racism denial, online advisory message boards, hyperlocal media, and political activism (Bingham-Hall & Law, 2015; Gonzalez-Bailon, Wang, Rivero, Borge-Holthoefer, & Moreno, 2012; Hashimoto, Shirota, & Chakraborty, 2016; Sharma & Brooker, 2017). The methods of data collection implemented in social media research are of particular relevance to the study of arts and humanities outputs, given that they do not rely on document identifiers.

This study proposes to contribute to the understanding of a text analysis approach to the study of references to sound recordings, and by extension, to other arts and humanities outputs. Although this paper analyzes Twitter mentions of digital sound collections, the methodology may be applied in other research contexts. More specifically, this paper aims to answer the following research questions:

- What motivates users to interact with digital sound archives on Twitter, and to tweet about digital sound?
- What can be inferred about user values with regard to digital sound recordings through the analysis of this record of interaction?
METHODOLOGY

APPROACH

This study uses a combination of quantitative and qualitative approaches to analyze a social media dataset collected using the Twitter Search API. Tweets, retweets and mentions of five sound archives were collected over a period of three years. I use two exploratory techniques – word frequency analysis and topic modeling – to identify high frequency thematic and topical trends in the text of the tweets. Using the Balanced Value Impact (BVI) Model as a framework for labeling algorithmically generated topics, I characterize and evaluate topics that reference digital sound collections for evidence of user values. In what follows, I outline the steps taken at the stages of data collection, analysis, and interpretation.

DATA COLLECTION

Five sound archives – British Library Sounds, Europeana Sounds, the Internet Archive Audio Archive, PennSound, and UbuWeb – were selected chiefly for their prominence, but also in light of the references they generate on Twitter. Other than the commonality of making digital sound recordings of cultural and historical significance freely available online, these archives share relatively little in common, rendering the empirical comparison of tweet counts and topical content of dubious value (Table 1 provides a description of the archives’ collections). With the exception of creative uses of sound (described under Innovation Impacts), the analysis focuses on the aggregated data. Two of the archives are government funded digital library projects: British Library Sounds and Europeana Sounds. One, PennSound, is affiliated with an academic institution. And two, the Internet Archive and UbuWeb, are independent non-profits.

<table>
<thead>
<tr>
<th>Sound archive</th>
<th>Description of collections</th>
</tr>
</thead>
<tbody>
<tr>
<td>British Library Sounds</td>
<td>90,000 digital recordings of speech, music, wildlife and the environment</td>
</tr>
<tr>
<td>Europeana Sounds</td>
<td>600,000 recordings of music, spoken word, radio broadcasts, and environmental sounds; in a project phase through January 2017</td>
</tr>
</tbody>
</table>
Sound archive | Description of collections
---|---
Internet Archive Audio Archive | 3.6 million recordings of mostly user contributed content, including radio broadcasts, podcasts, music, sound art, soundscapes, and speech
PennSound | 45,000 recordings of poetry, performance, art, and lectures about poetry
UbuWeb | Specializes in avant-garde sound and film, including poetry readings, conceptual art, ethnopoetic sound, lectures, documentaries, and sound ephemera

Over a thirty-five month period (August 15, 2014 to July 15, 2017), a total of 25,649 tweets referencing these sound archives were collected. Using the Twitter Archiving Google Spreadsheet (TAGS) tool (Hawksey, 2014), separate data collection was conducted for each sound archive using queries that contained the institutions’ Twitter handle and in some cases additional identifying keywords (see Table 2 for queries). For instance, to capture tweets relevant to the Audio Archive of the Internet Archive, the keywords “audio” and “music” were added to the search query.

Table 2: Sound archive and search query. Note that the query for Europeana Sounds was modified over the course of data collection to accommodate a change in their Twitter account name.

| Sound archive | Search query |
---|---|
British Library Sounds | soundarchive |
Europeana Sounds | EuropeanaMusic |
Internet Archive Audio Archive | internetarchive AND audio OR music |
PennSound | pennsound |
UbuWeb | ubuweb |

Data collected using the TAGS tool were processed and analyzed in the R language (R Core Team, 2017) using the following libraries: dplyr (Wickham, Francois, Henry, & Müller, 2017), ggplot2 (Wickham, 2016), litdata (Goldstone, 2015), lubridate (Grolemund & Wickham, 2011), mallet (Mimno, 2013), stringr (Wickham, 2017), and tm (Feinerer & Hornik, 2017).
Tweets about the five sound archives that were collected separately were merged and sorted chronologically. The text of the tweets was normalized to remove punctuation (including # and @ signs), URLs, and stop words in preparation for word frequency analysis and topic modeling. The data, scripts, and custom stop word list used in this study are available at https://github.com/giannetti/digital-sound.

ANALYSIS

This study combines the quantitative techniques of topic modeling and word frequency analysis to identify events of particularly high relief. In a first step, to compose a basic picture of the text of the tweets, I generated lists of words and the frequencies with which they appeared to identify events and patterns of particular interest. In the next step, the text of the tweets was used as the input of a topic modeling algorithm to explore these events and patterns in further detail.

Topic modeling serves as a method of content analysis in this study. Topic modeling is a probabilistic technique that sorts words with a tendency to co-occur into “topics.” In digital humanities as well as in other computational fields, topic modeling is used as an exploratory method to find thematic or topical groupings within large collections of texts. Its comparative advantage over word frequency analysis is a greater sensitivity to the context of use. Given a known set of texts or documents, topic modeling offers a way to infer the hidden structure that might have generated them (Blei, 2012; Underwood, 2012). Topic modeling assumes two things: each text or document exhibits multiple topics, and each topic contains distinctive terms (Riddell, 2012). When the documents in the analysis can be arranged chronologically, the variation in topic proportions may uncover connections between topics and thematic trends over time.

Word frequency analysis and topic modeling form the first phase of analysis; in a subsequent phase, the outputs of these processes were investigated and manually labeled, drawing in part upon a framework called the Balanced Value Impact (BVI) Model. The BVI Model is a framework to measure the impact of digital heritage resources (Tanner, 2012). In particular, attention has been paid to what Tanner refers to as social and innovation impacts, balanced with the value propositions of community, inheritance and education. Tanner and Deegan explain social and audience impacts to mean that “the audience, the beneficial stakeholders and wider
society have been affected and changed in a beneficial fashion” (Tanner & Deegan, 2013, p. 18). Innovation impacts indicate that the digital resource enables innovation, and therefore social and/or economic benefits accrue to the community. Tanner and Deegan define the aforementioned value propositions as follows:

Education Value. People are aware that digital resources contribute to their own or to other people’s sense of culture, education, knowledge and heritage and therefore value it.

Community Value. People benefit from the experience of being part of a community that is afforded by the digital resource.

Inheritance/Bequest Value. People derive benefit from the inheritance passed down to them and satisfaction from the fact that their descendants and other members of the community will in the future be able to enjoy a digital resource if they so choose (Tanner & Deegan, 2013, p. 17).

The BVI Model is intended to be a flexible heuristic that aids the reflection and analysis involved in an impact assessment. Weighing these impacts with cultural modes of value provides an evidence-based approach that accounts for tangible and intangible benefits, and allows the researcher to consider data points of his or her choosing.

As this study will demonstrate, the BVI model provides a useful rubric for the analysis of a topic model of data about digital sound collections. Topic modeling algorithms sort words into groups, but it falls to the human interpreter to study and assess the results for coherence. During this assessment process, the researcher may alter one or more parameters to adjust the picture of evidence. The number of topics is one such parameter; for this study, the number has been set to twenty-five to reduce the quantity of less meaningful topics, even if this did bring about the occasional merging of topics that were only distantly related. Another parameter relates to the way in which the tweets were segmented into documents. A topic modeling algorithm models the topic as “an intersection of vocabulary and context” (Goldstone & Underwood, 2014, p. 361). In a practical sense, this means the algorithm requires the context, or document, to be defined. While it may seem logical to view the tweet as the document, in practice it is usually too small a unit to yield satisfactory results. In this study, the context...
is primarily influenced by chronology. I have segmented the merged and sorted tweets into “documents” of nine tweets each, which results in bags of words that are small enough to produce many themes, while not being so numerous as to obscure patterns. Regarding the top words of each topic presented in Table 5, I have not attempted to label every topic, and have chosen instead to focus on topics that, upon investigation, clearly referenced sound collections. In most cases, the topic is composed primarily of isolated events (i.e. highly retweeted tweets), and trends emerge primarily at the stage of topic labeling and interpretation.

Topic modeling was performed on the aggregated corpus of tweets, inclusive of all retweets, and separately on the Internet Archive corpus with retweets removed. As discussed under Results, tweets collected for British Library Sounds, Europeana Sounds, PennSound, and UbuWeb demonstrated varying but strong broadcast characteristics in which the sound archive’s Twitter account was the most dominant actor in the collected data, which included high levels of repeated text (mostly retweets). Due to this structural feature of the data, the output of topic modeling revealed events, more so than trends. This characteristic is largely absent in data collected for the Internet Archive, where only 22 tweets were authored by @internetarchive, of which 9 were original tweets. The sensitivity of Latent Dirichlet Allocation (LDA) topic modeling to repeated text has been turned to advantage in the topic model of the aggregated corpus to reveal the most retweeted tweets, many of which originated from the archives’ accounts (Schofield, Thompson, & Mimno, 2017). Repeated text in this study has been interpreted as a measure of user interest in the subject matter. Creative, user-driven uses of sound and platforms were more prevalent in the Internet Archive tweets. For this subset of the data (1,270 tweets), a more canonical use of LDA topic modeling, showing trends over time, revealed topics relating to community archiving and the development of tools that interact programmatically with the archive’s data and metadata.

RESULTS

DATASET CHARACTERISTICS

The dataset exhibits a few broad characteristics, the most significant of which is a promotional character driven by the activity of social media managers at the sound archives themselves. The most extreme case is
Europeana Sounds: of the 5,228 tweets collected for this archive, 4,649 (89%) were from the archive’s own account, as shown in Tables 3 and 4. This may be owed to the fact that Europeana Sounds was in a project phase for most of the period under examination, during which time tweets did not reference URLs to existing or new collections. This lower level of user engagement is also reflected in the relatively small number of unique accounts present in the Europeana Sounds tweets (see Table 3). Secondarily, one observes a high number of retweets overall (at 12,143, nearly half of the total collected), which are particularly concentrated in the tweets collected for British Library Sounds and Europeana Sounds. For these two sound archives in particular, much of the observed activity originated from the archive itself. Tweets by the archive about its collections, or about digital sound in general, were retweeted, or replied to by the archive’s followers. British Library Sounds, the Internet Archive, PennSound, and UbuWeb showed varying but high levels of user engagement, whether measured by the number of unique accounts, original and quoted tweets, retweets, or replies.

Table 3: The composition of tweets and unique accounts collected for each archive. Note: the quoted tweet values are approximative, given that the quoted_tweet metadata has been included in Twitter API payloads only as of May 2015. Prior to this, quoted tweets counted towards the 140 character limit of an original tweet. See https://twittercommunity.com/t/api-payloads-to-include-original-quoted-tweet-objects/38184.

<table>
<thead>
<tr>
<th>Archive</th>
<th>Original Tweets</th>
<th>Retweets</th>
<th>Quoted</th>
<th>Replies</th>
<th>Accounts</th>
<th>Total Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>British Library Sounds</td>
<td>2238</td>
<td>6079</td>
<td>295</td>
<td>709</td>
<td>4085</td>
<td>9321</td>
</tr>
<tr>
<td>Europeana Sounds</td>
<td>1839</td>
<td>2450</td>
<td>177</td>
<td>762</td>
<td>363</td>
<td>5228</td>
</tr>
<tr>
<td>Internet Archive</td>
<td>1076</td>
<td>911</td>
<td>34</td>
<td>160</td>
<td>1523</td>
<td>2181</td>
</tr>
<tr>
<td>PennSound</td>
<td>3273</td>
<td>2190</td>
<td>77</td>
<td>308</td>
<td>2531</td>
<td>5848</td>
</tr>
<tr>
<td>UbuWeb</td>
<td>1044</td>
<td>513</td>
<td>26</td>
<td>1488</td>
<td>1445</td>
<td>3071</td>
</tr>
</tbody>
</table>
Table 4: The top ten tweeters overall include the Twitter account names of four of the five archives. Furthermore, Al Filreis (@Afilreis), @Bagolina, Cheryl Tipp (@CherylTipp), and the British Library World and Traditional Music Collection (@BL_WorldTrad) are affiliated with PennSound, Europeana Sounds, and British Library Sounds.

<table>
<thead>
<tr>
<th>User Account</th>
<th>Total Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>eu_sounds</td>
<td>4649</td>
</tr>
<tr>
<td>soundarchive</td>
<td>1855</td>
</tr>
<tr>
<td>ubuweb</td>
<td>1217</td>
</tr>
<tr>
<td>PennSound</td>
<td>904</td>
</tr>
<tr>
<td>Afilreis</td>
<td>150</td>
</tr>
<tr>
<td>ANBOLIVIA</td>
<td>134</td>
</tr>
<tr>
<td>Bagolina</td>
<td>110</td>
</tr>
<tr>
<td>CherylTipp</td>
<td>108</td>
</tr>
<tr>
<td>CelizMurray</td>
<td>100</td>
</tr>
<tr>
<td>BL_WorldTrad</td>
<td>97</td>
</tr>
</tbody>
</table>

Retweets in this study are regarded as a measure of user interest in the subject matter. Researchers who work with Twitter data regard the retweet as, at minimum, a form of “rebroadcasting” (Ediger, Jiang, Riedy, Bader, & Corley, 2010) or “information exposure” (Tufekci, 2014) that in turn points to a wide variety of potential user motivations. In the context of this study, retweets may be summarized as acts of appraisal, curation and (re)circulation. The retweet is an opportunity for users to deepen involvement with the sound archive; it allows them to express something about themselves, call attention to content, and cultivate followers. Occasionally, users will retweet to bookmark content for later consultation (Holmberg & Thelwall, 2014). The concept of “spreadable” media may provide a useful framework for analyzing retweets (J. Green & Jenkins, 2011). More so than virality or “stickiness,” the word *spreadability* accounts for the ways that users decide which messages spread. Similar to the findings of Ediger et al. (2010), the retweets in this dataset refer to a small subset of original tweets, the text of which, though scrambled, is visible in nearly all of the topics presented in Table 5. With the caution that the topics of Table 5 present only the top ten most prominent words in each topic, the keywords of “viral” tweets, and
consequently the tweets that users have established as the most “spreadable” in this dataset, are traceable to a total of 49 original tweets, or 0.2% of the total. Inasmuch as there are patterns to be detected when evaluating the messages that do and do not spread, a social media dataset such as this one provides useful material for reflection on user values as they align with the access and discovery goals of the sound archive.

Table 5: Top ten most prominent words in each topic.

<table>
<thead>
<tr>
<th>topic</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>heritagelottery funding 95m abbotsbury pavlova swans dreamily creely worldballetday anna</td>
</tr>
<tr>
<td>2</td>
<td>gertrude stein valentine sherwood anderson 1934 rec mp3 lamborn gifs</td>
</tr>
<tr>
<td>3</td>
<td>sings helen computer eula joylandfiction sonybmgs womens entire egypt chorus</td>
</tr>
<tr>
<td>4</td>
<td>eusounds15 sonores bagolina bnfmonde projet actubnf r merci phaune machines</td>
</tr>
<tr>
<td>5</td>
<td>auld syne lang ago happynewyear 117 remarkable years land matter recordists</td>
</tr>
<tr>
<td>6</td>
<td>eusounds16 vilnius johnny surabaya dunaway judy weill hanson balloons conference</td>
</tr>
<tr>
<td>7</td>
<td>uk directory help collections create shoresounds save coastal uks soundmap</td>
</tr>
<tr>
<td>8</td>
<td>tape monoskop mp3 film beautiful short art recorders issue berlin</td>
</tr>
<tr>
<td>9</td>
<td>loy mina 1960 blackburn monroe harriet interview weldon johnson among</td>
</tr>
<tr>
<td>10</td>
<td>christmas mention frost mentions followers advent howd reach twitter collection</td>
</tr>
<tr>
<td>11</td>
<td>blog thanks today stories daily look conference latest radio europeanamusic</td>
</tr>
<tr>
<td>12</td>
<td>free streaming nature download quality internet 200 dead soundcloud top</td>
</tr>
<tr>
<td>13</td>
<td>like can one just wikipedia editathon find get ubu good</td>
</tr>
<tr>
<td>14</td>
<td>v internetarchivee groove sitting x dangermindsblog stylus magnified anbolivia stunning</td>
</tr>
<tr>
<td>Topic</td>
<td>Label</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>15</td>
<td>segmented william rts ginsberg thanksall newly daily stories came burroughs mydekel469</td>
</tr>
<tr>
<td>16</td>
<td>instruments radiolulu tag latherevival meanstsotrpod tagdaythursday radio oldtimesounds otrnewgen booksr4life</td>
</tr>
<tr>
<td>17</td>
<td>womad womadcharltonpk womad2015 bbcarts exploring 30 travel form years blworldtrad</td>
</tr>
<tr>
<td>18</td>
<td>womad2016 family tom archived memoriam raworth blworldtrad 100 spoken recorded</td>
</tr>
<tr>
<td>19</td>
<td>baraka amiri browse everybody map digs warning addictive stations appropriation</td>
</tr>
<tr>
<td>20</td>
<td>poem poetry podcast poems read performs creeley ashbery john poemtalk</td>
</tr>
<tr>
<td>21</td>
<td>soundandmusic embedded residency composers uknatarchives btarchives guitarist clash teenaged ditondo</td>
</tr>
<tr>
<td>22</td>
<td>syliphone label using odnb bleap 1000 stored now firenzesoundmap tracks</td>
</tr>
<tr>
<td>23</td>
<td>mirror fine petabyte peoplert textfiles heads soundcloud people 60000 sfx</td>
</tr>
<tr>
<td>24</td>
<td>cd blacklivesmatter releases japan keyword theyll texts upload player ago</td>
</tr>
<tr>
<td>25</td>
<td>rain tommilesz lacquer digitising cheryltipp 100000 disc 40000 preservation 1990</td>
</tr>
</tbody>
</table>

**Recency**

In addition to the broadcast or promotional nature of the dataset, another characteristic concerns newness or recency. Archives with growing collections of digital audio, as well as those that blog or highlight new research using objects in their collections, attracted the attention of Twitter users. This preference is evinced by the prominence of the word ‘new’, as well as associated terms like ‘now’ and ‘today’, in word frequency analyses. After stemming to remove suffixes, ‘new’ was the seventh most frequent term overall (relative frequency: 0.7%), appearing after the more expected, and less distinctive terms ‘sound’, ‘record’, ‘music’, ‘archive’, ‘audio’, and ‘read’. Figure 1 shows the relative frequency of the word ‘new,’ together with the
next most frequent terms, ‘collect’ and ‘listen,’ over time.

Figure 1: A plot of the frequency of the word ‘new’, and its nearest neighbors ‘collect’ and ‘listen’, across all tweets. All three words are lemmas, and thus include references to ‘new’ but also ‘newly’, ‘news’, etc. The spike observed for ‘collect’ corresponds to the January 2015 launch of British Library Sounds’ “Save Our Sounds” campaign to preserve the nation’s sound collections.

This characteristic is closely intertwined with the earlier discussion of promotion. In drilling down to the level of individual tweets, one sees that the word was used in connection with the archives’ announcements of new digital collections, but also of new blog posts featuring the archives’ recordings, new interviews and podcasts, or in certain cases a new website or interface. Elsewhere, researchers have found that Twitter metrics demonstrate a ‘recency bias’ and are found in higher quantities only for the most recent publications (Costas et al., 2014). An analogous bias can be found in this dataset, inasmuch as users demonstrated a strong interest in
tweets about recently added collections, and secondarily in tweets about new formal or informal research that used the sound archive’s collections as source material.

**EPHEMERA**

Several altmetric studies mention user interest in topics of an ephemeral, prurient, humorous or curious nature. The topic model includes several vivid examples of tweets of ephemeral interest. For example, topic 6 in Table 5 references a PennSound recording of Kurt Weill’s “Surabaya Johnny” performed on balloons (UbuWeb, 2016). In relation to richer social media sources like blogs, tweets may be ill suited to discussions of intellectual or cultural impact, and yet careful attention is occasionally rewarded by finding resources that may not be highly cited but are clearly of broad public interest (Bornmann, 2014; Neylon, 2014). In the topic model, a tweet that at first glance appears ephemeral in nature, but that may point to deeper engagement appears in topic 2 of Table 5; Figure 2 shows its word frequency distribution. This topic is primarily composed of a much shared tweet from UbuWeb on Valentine’s Day referencing a recording of Gertrude Stein reading *A Valentine to Sherwood Anderson* (UbuWeb, 2015).

ubuweb: Gertrude Stein reads “A #Valentine to Sherwood Anderson.” Rec. 1934 [MP3 link]: https://media.sas.upenn.edu/pennsound/authors/Stein/1935/Stein-Gertrude_A-Valentine-to-Sherwood-Anderson.mp3 (@PennSound)

While undoubtedly an object of transitory interest, this tweet may generate alternative readings. The poem, read with Stein’s characteristically round timbre, may be for or about two people, as Alice B. Toklas is plainly the unnamed subject of many sections. A first-level reader might share this tweet because of its thematic appropriateness, and perhaps to indicate an appreciation of poetry or sound. A second-level reader might share it as a form of irony. From the perspective of this reader, the poem’s invitation to celebrate gratitude, in addition to romantic love, relaxes the stakes for would-be celebrants and enlarges the compass of the event. The Stein tweet arguably points to a digital resource of wide public interest, and may, in this light, be interpreted as evidence of “community value.”
Tanner describes community value as a benefit that is derived “from the experience of being part of a community that is afforded by the digital resource” (Tanner, 2012, p. 37). It seems reasonable to ask, in the case of Twitter, whether or not a tweet can create a community of users on the fly, or if a ready-made community must already exist for a tweet to become spreadable. The aforementioned tweet about Gertrude Stein might be an example of the former: a poetry recording that finds a wider audience thanks to the crosscutting power of a particular holiday. In either case, the top words of the topic model in Table 5 provide evidence that large potential communities do exist on Twitter for various genres of recordings.

Nature and wildlife recordings (topic 12):

soundarchive: 200 top quality British nature recordings from @britishlibrary free on SoundCloud http://britishlibrary.typepad.co.uk/sound-and-vision/2015/04/british-wildlife-recordings-on-the-move.html

Poetry recordings (identifiable in at least 8 topics, for example topic 9):

jenterysayers: Interview with Mina Loy and Paul Blackburn, 1960, http://writing.upenn.edu/pennsound/x/Loy.php ...

Environmental and industrial soundscapes (topics 22, 25):
citiesandmemory: The best ten sound maps, feat. @London-Sounds @firenzesoundmap @soundarchive @britishlibrary http://citiesandmemory.com/2015/07/top-sound-maps

soundarchive: Enjoy rain, rain and more rain - by listening to it! http://britishlibrary.typepad.co.uk/sound-and-vision/2016/03/here-comes-the-rain-again.html

World and traditional musics (topics 17-18, 22):

soundarchive: Archiving WOMAD: 30 years of @BritishLibrary world music recordings https://shar.es/1s4pCG @WOMAD-CHARLTONPK


Tweets describing crowdsourcing campaigns are evidence of yet more improvised communities. These included Sounds of our Shores (hashtag #shoresounds in topic 7), a three-month campaign in which users contributed their own recordings of the U.K. coast, an effort to tag musical instruments in Europeana Sounds recordings (topic 16), and various Wikipedia editathons to enrich articles on specific themes with references to sound recordings (topic 13).

soundarchive: you can help us create the first ever coastal soundmap of the UK: http://www.bl.uk/sounds-of-our-shores . . . #shoresounds

soundarchive: Calling musical instruments experts! Help Europeana to crowdsource tags on our music collections #eu_sounds http://with.image.ntua.gr/custom/soundspace/index.html . . .

CherylTipp: Busy adding sounds to Wikipedia articles at our @eu_sounds edit-a-thon!
A last example of community value includes professional communities, as evidenced by the presence of conference hashtags #eusounds15 and #eusounds16 (topics 4, 6) and the popularity of a call for applications to British Library Sound’s Embedded residency for composers (topic 21) (soundarchive, 2015).

**INHERITANCE VALUE**

Several highly retweeted tweets point to another user value having to do with real or perceived rarity and uniqueness, which relates to inheritance value. Tanner and Deegan describe inheritance value as “satisfaction from the fact that [users’] descendants and other members of the community will in the future be able to enjoy a digital resource if they so choose” (Tanner & Deegan, 2013, p. 17). Keywords pertaining to inheritance include “save” (relative frequency: 0.2%), “lost” (0.08%), “old” (0.08%), “rare” (0.07%), “past” (0.07%), “treasures” (0.04%), and “unique” (0.02%). Examining these keywords in their context shows that many users are retweeting the appeals of British Library Sounds to support the Save Our Sounds and the related Directory of UK Sound Collections projects (both in topic 7).

soundarchive: New #SaveOurSounds programme by @britishlibrary to save UK’s audio collections http://britishlibrary.typepad.co.uk/sound-and-vision/2015/01/save-our-sounds-15-years-to-save-the-uk-s-sound-collections.html

soundarchive: Do you own a sound collection? Help @britishlibrary create the UK sound directory http://www.bl.uk/projects/uk-sound-directory

Genres of recordings associated with these keywords include poetry and other spoken word recordings, especially early ones featuring well known literary or historic figures, ethnographic field recordings, and avant-garde recordings from UbuWeb.

With few exceptions, the references in tweets are to unique, archival, and not commercial, sound recordings. Commercial recordings, even when out of print and several decades old, are often a lower preservation priority for sound archives because their carriers are considered to be more stable. Some user motivations for engaging with these archival resources might
include sentiment, like nostalgia, and civic or national pride, as evinced by the British Library Sounds topics. In other cases, there is an interest in the history of sound technologies visible in the popularity of tweets about legacy playback equipment (topics 4 and 8) (EuropeanaMusic, 2016; soundarchive, 2016b). In still others, there could be a thrill of recognition among amateur collectors and researchers in finding specialized resources that are exactly suited to their fields of interest. This is perhaps a factor for the popularity of a tweet about the digitized tapes of the Radio Télévision Guinée archives (topic 22) (soundarchive, 2016a). In her study of citations and figures in historical research, Sinn observes a correlation between the usage frequency of digital materials with their uniqueness and relevance to specialized topics (Sinn, 2012, p. 1531). This Twitter dataset appears to validate those findings, inasmuch as user engagement with specialized archival resources is high. Although a complicating factor could be that sound archives tweet less often about commercial recordings, giving users less of this content to discover.

EDUCATION VALUE

As discussed earlier, there is a sizeable community of Twitter users with an interest in poetry recordings, like those of Gertrude Stein (topic 2) and John Ashbery (topic 20) reading their own works. These recordings often include prefatory remarks by the author or reader to situate the work for the listener. In addition to the interest in the primary resource, there is significant evidence that users value secondary and supplementary resources, such as lectures and interviews featuring important literary or historic figures. These users consult sound archives to increase their knowledge of cultural works, their cultural and historical context, and critical reception. Recordings of authors reading their own works were often just as often retweeted as the recordings of authors discussing poetry and culture, as in the case of these tweets referencing Allen Ginsberg and William Burroughs (topic 15), and Amiri Baraka (topic 19).

penmsound: *Ginsberg & Burroughs talking with Studs Terkel* — a PennSound recording now segmented into topics.

shujaxhaider: *Amiri Baraka on cultural appropriation: “everybody digs each other.”* https://media.sas.upenn.edu/
Keywords associated with education value include “interview” (relative frequency 0.17%), “workshop” (0.06%) and “lecture” (0.05%), as in the following examples.

BL_Learning: Celebrate @RefugeeWeek with #free #family #poetry workshops (10.30-12.00) inspired by our #FoodStories @soundarchive

ubuweb: Adorno MP3 Archive: radio interviews, lectures, & his own musical works: http://www.ubu.com/sound/adorno.html . . . (@NeinQuarterly)

PennSound and UbuWeb tend to dominate in tweets relating to education value. It is likely not coincidental that the founders of these sound archives are academics, and consciously sought to create digital archives that reflect their pedagogical goals and vision.¹

**INNOVATION IMPACTS AND COMMUNITY**

An early question of this study pertained to the creative reuse and remixing of sound, and the influence on this activity of differing copyright terms for sound recordings, i.e. fifty to seventy years from the date of creation in the United Kingdom and continental Europe versus a tangle of conditions ranging from the life of the author plus seventy years to perpetuity in the United States.² With the understanding of innovation as remix culture, the dataset is more distinctive for what it lacks than for what is present. Relatively little evidence of the creative reuse of sound was identified, and indeed the methodology implemented in this study may be ill-adapted to uncover such evidence. Depending on one’s definition of “creators,” they

---

1. Al Filreis and Charles Bernstein, the directors of PennSound, and Kenneth Goldsmith, founding editor of UbuWeb, are professors of English at the University of Pennsylvania.
2. Cornell University’s “Copyright Term and the Public Domain in the United States” at http://copyright.cornell.edu/resources/publicdomain.cfm provides further explanation of the status of sound recordings in U.S. copyright law.
are not thought to constitute a majority of users.\textsuperscript{3} A more interesting and actionable question in this context is how often users are taking advantage of the affordances of these sound archives to upload new content or extend the platform. When this question is explored, the Internet Archive emerges as distinctive because it archives large quantities of community contributed content, and provides scalable access to metadata and data.\textsuperscript{4}

Some examples of the way in which individuals and nonprofits are using the Internet Archive Audio Archive include the Agencia de Noticias de Bolivia, which regularly archives its news radio broadcasts, and tweets from the handle @ANBOLIVIA (a top ten tweeter in Table 4). In another example, a highly retweeted tweet from the Internet Archive invited users to contribute to a Black Lives Matter collection (topic 24). The hashtag #blacklivesmatter appeared with a relative frequency of 0.06%.

\begin{itemize}
  \item \texttt{internetarchive: Upload video, audio, texts or images with keyword #blacklivesmatter and they'll be added to https://archive.org/details/black-lives-matter} . . .
\end{itemize}

A small developer-related topic emerges when modeling the Internet Archive tweets in isolation. An open source Mopidy extension, written in Python, extends the functionality of the Internet Archive by allowing users to search for and stream its sound recordings, rather than clicking through individual tracks. The term “mopidy” appears in the aggregated corpus with a relative frequency of 0.009%, which gives an idea of its comparatively small scope (by contrast, the Internet Archive’s Grateful Dead collection received more sustained attention).

\begin{itemize}
  \item \texttt{pypi_updates2: Mopidy-InternetArchive 2.0.3: Mopidy extension for playing music from . . . http://pypi.python.org/pypi/Mopidy-InternetArchive/2.0.3}
\end{itemize}

\textsuperscript{3} Based on a 2010 survey, Li and Bernoff report that 23 percent of adult Americans who are online are “creators,” which is to say they have authored a blog post, maintained a website, or uploaded a video or audio recording within the past month. See Li, C., & Bernoff, J. (2008). \textit{Groundswell: Winning in a world transformed by social technologies}. Boston, MA: Harvard Business Press.

\textsuperscript{4} See \url{https://blog.archive.org/developers/} for documentation on the Internet Archive’s tools for developers.
It is important to remark that the Internet Archive and other non-profit digital archives are able to adopt a more liberal attitude towards copyright and fair use than their state peers. The Internet Archive and UbuWeb have a pragmatic approach that only rarely involves identifying copyright holders and securing permissions before publishing sound recordings, although both will remove content on the basis of credible claims. Their example cannot be readily followed by a state-funded organizations, which must adhere strictly to laws governing intellectual property. The ability of the user to interact with the sound archive’s data and metadata freely and at scale is hindered in most cases by copyright law and donor agreements that predate the Internet (Brylawski, 2002; Cleary, 2010; Harkins, 2012). In a related note, the subject of copyright (relative frequency 0.04%) and its contractual trappings did appear in the tweets of this dataset. Some tweets cited legal complexities as a barrier to preservation and dissemination, while others demonstrate a more combative or flippant tone. Topic 3 includes references to a recording at PennSound of a women’s chorus singing Sony BMG’s end user licensing agreement, or EULA.
ubuweb: A women’s chorus sings the entire Sony/BMG’s EULA, composed @joylandfiction [MP3 link]: https://media.sas.upenn.edu/pennsound/authors/Davis-Brian/Davis_Brian_Joseph_Eula_Eula_2007.mp3

Although it is a public institution, British Library Sounds has found ways of tapping user ingenuity through crowdsourcing and other ventures, as discussed earlier. Changes to U.S. copyright law and harmonization of European laws, though elusive and thorny projects, would remove many impediments to innovation and research, and allow sound archives to meet more of contemporary users’ expectations.

CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

DISCUSSION AND CONCLUSIONS

This study demonstrates that primary motivating factors for tweeting about digital sound include the presence of social media outreach from the archive, which gives users material to respond to on the platform, and the promotion of new digital resources. Users consult digital sound archives, and tweet about digital sound, to benefit from the experience of being part of a community, for the pleasure of knowing that a resource exists and is preserved for future generations, to increase their knowledge of culture and history, and, to a smaller degree, for the purpose of innovating and sharing their own creative outputs. The approach to social media research demonstrated in this study provides helpful, even if incomplete, contextual information on the use and impact of digital archival sound recordings, which may have value for the archives’ communication, fundraising, and digital preservation plans.

Spoken word and unique archival recordings tend to attract users’ attention the most. Mentions of commercial recordings were rare, and curiously music was only occasionally referenced, despite the fact that most of these archives have large quantities of music in their collections. The Grateful Dead collection of the Internet Archive’s Live Music Archive, in Figure 3, as well as the world music topics mentioned earlier, are among these exceptions. Although there were few patterns to be detected in the sound archives’ original tweets that users deemed of little interest, as measured by a total lack of retweets or favorites, mentions of music accounted for

24
Why should this be? A possible reason may be that it is easier to verbalize thoughts about spoken word recordings and oral histories, which are manifestations of textual content, than about music. Music is a language composed of pitch, rhythm and timbre. Writing about it is challenging, even when the work includes a textual component, as is the case in song and opera. Another reason may be that the Twitter platform, with its 140 character limit, is not conducive to thoughtful expression about music. This difficulty may be experienced by users of the sound archives as well as by social media managers tweeting on the archives’ behalf, as the relatively low frequency of tweets about music recordings in the context of this study can be traced to a reticence about music resources originating from the archives’ Twitter accounts.

LIMITATIONS AND FUTURE RESEARCH

The results of this study are subject to several important limitations. The Twitter Search API, which the TAGS tool queries, provides access to “a sampling of recent Tweets” where the focus is on relevance, not completeness. Researchers investigating the Search API, and indeed other public Twitter APIs, have documented a bias towards central users who use the service frequently, have following/follower relationships, and interact with other accounts (Driscoll & Walker, 2014; Gonzalez-Bailon et al., 2012). Peripheral activity, including tweets from users with incomplete account information or who tweet infrequently, may be absent from samples collected using the Twitter APIs. As noted above, the queries used to capture relevant tweets for this study assume the presence of the Twitter account name of the archive. This assumption was made in part to ensure the integrity of the data collection and reduce matches of tweets about collections other than sound recordings. This strategy does, however, result in some data loss, notably when a user tweets a link to a sound recording and does not reference the archive from which it comes. Tweets about PennSound recordings were captured whether or not the user referenced its Twitter account name, since the character string “pennsound” is present in the archive’s URLs as well as in its Twitter account name. However, tweets captured for the other

---

5. Other patterns in zero-retweet, zero-favorite tweets from the sound archives: spoken word recordings of lesser known authors and automated tweets.

archives may omit references to URLs alone. This decision was influenced by pragmatic concerns, since queries inclusive of all relevant URLs were more error prone. With this said, the data analyzed in this study favor the communication and promotional activities of the archives themselves, in that tweets to and from the archives’ account names were systematically captured, in addition to user mentions, retweets, and quoted tweets.

The methods used in this study are exploratory in nature, intended to provide some insight into user values as they relate to digital sound recordings and archives. Triangulation with a second, qualitative or quantitative metric, will provide an important counterpoint to the findings presented. For example, connecting the results of a Twitter data analysis with data from a more academically oriented source such as Mendeley will provide a deeper analysis of the scholarly impact of these digital resources, especially as Mendeley has been found to correlate well with citations (Haustein et al., 2015; Mohammadi & Thelwall, 2014; Priem et al., 2012). The Zotero altmetrics API, when ready, may serve as a crucial source of quantitative data on sound recordings, as well as on other arts and humanities outputs, as the adoption of Zotero as a reference manager among humanists may be higher than Mendeley (Heller, 2013). Finally, qualitative ethnographic research will refine and extend the results of social media analysis, for both scholarly and recreational uses of sound, and may help address questions raised in this study regarding the use of commercial versus archival recordings, and the disparity in the uptake of music as compared to spoken word recordings.

ACKNOWLEDGEMENTS

The author is grateful to David Hunter, Kevin Mulcahy, and Pritika Pradhan for their insightful comments and suggestions on a draft of this paper. The anonymous peer reviewers provided detailed and generous commentary that considerably strengthened the observations set forth in this study. This work draws together and expands on remarks made at the annual meetings of the Association for Recorded Sound Collections (ARSC), held in Pittsburgh in

---

7. “Queries can be limited due to complexity”: https://dev.twitter.com/rest/public/search.

8. For more information on Zotero’s public API for bibliometric data that is in development, see Fahringer, A. (2015, September 3). Studying the Altmetrics of Zotero Data [blog post]. Retrieved from https://www.zotero.org/blog/studying-the-altmetrics-of-zotero-data/
2015, the Atlantic Chapter of the Music Library Association in Philadelphia in 2014, and Digital Frontiers, held in Denton in 2014.

REFERENCES


2010 39th International Conference on Parallel Processing (pp. 583–593). https://doi.org/10.1109/ICPP.2010.66


Hammarfelt, B. (2016). Beyond Coverage: Toward a Bibliometrics for the Humanities. In M. Ochsner, S. E. Hug, & H.-D. Daniel (Eds.), Re-
search Assessment in the Humanities (pp. 115–131). Springer International Publishing.


tool MALLET. Retrieved from https://CRAN.R-project.org/package=mallet


tweet. Retrieved from https://twitter.com/soundarchive/statuses/611907298227847168


