CA Journal of **Cultural Analytics**

Archival Circulation on the Web: The Vine-Tweets Dataset

Ed Summers and Amy Wickner

06.04.19

Peer-Reviewed By: Anon.

Clusters: Data

Article DOI: 10.22148/16.040

Journal ISSN: 2371-4549

Cite: Ed Summers and Amy Wickner, "Archival Circulation on the Web: The Vine-Tweets Dataset," Journal of Cultural Analytics. May 4, 2019.

At Ethics and Archiving the Web, a conference convened in March 2018 at the New Museum in New York City, a group of artists, archivists, activists and researchers met to critically examine the ethical implications of our ability to collect and archive content from the web. In a session focused on the ethics of digital folklore, Frances Corry asked the audience to consider, "What's *the right way* to shut down a social networking site?" (Figure 1).

Corry specifically discussed the closure of Vine, a social media service designed for sharing six-second video clips. She highlighted how, despite Vine's best intentions to sunset their service "the right way," mistakes were made regarding missing content and leaked personal information in users' archives.[2] Despite these difficulties, Vine's effort remains noteworthy because of how it mobilizes the concept of an archive in dismantling its online service. Vine's commitment to preserving content with user consent stands in stark contrast to the way many web platforms have shuttered their windows, barricaded their doors, and pulled vast amounts of content offline in the process.

[1] Christine A George, "Frances Corry provides a 'burning yellow' background for a burning ethical question," Twitter, March 23, 2017.

[2] Vine Support, "Fixing a bug in the Vine Archive," Medium, May 19, 2017.

Figure 1: Figure 1. Frances Corry presenting at Ethics and Archiving the Web [1]

In this essay, we describe the Vine-Tweets dataset generated during the first author's participation in an effort by the Archive Team to preserve and provide access to Vine via the Internet Archive's Wayback Machine.[3] The Vine-Tweets dataset, itself stored at the Internet Archive, is a simple mapping of identifiers—tweet identifiers to Vine URLs—that can be downloaded as a set of text files. We describe the Vine platform and its sunset strategy (Vine Archive), and examine the Vine-Tweets dataset as consisting of "data fumes" created at the juncture of two large media platforms.[4] As data fumes, the dataset reflects the use of attentional signals in the construction of web archives, while continuing to act as an index to material preserved via Vine Archive and the Internet Archive. In addition, we show how the dataset can be used to measure and analyze how content that is designated as archival continues to circulate in social media. Finally, we point to patterns of continuing circulation as evidence of a new paradigm of "convivial decay" for sunsetting web services.[5] We offer this dataset case study to illustrate the multiple, interdependent forms that archives can take on the web,

[3] Internet Archive (vine-tweets; accessed June 1, 2018), https://archive.org/details/vine-tweets.

[4] Jim Thatcher, "Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data," *International Journal of Communication* 8 (2014): 1765-1783.

[5] Marisa Leavitt Cohn, "Convivial Decay: Entangled Lifetimes in a Geriatric Infrastructure," in *CSCW '16 Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ed. Darren Gergle, Meredith Ringel Morris, Pernille Bjørn, and Joseph Konstan (New York: Association for Computing Machinery, 2016), 1511-1523.

and further contribute to an understanding of how archives are built on and of the web.

A Short History of Vine's Short History

Vine was a social media service that allowed its users to create and share three-to-six-second looping videos with their followers. Users recorded video by opening the app on a mobile device, pressing and holding a camera icon on the screen, and releasing to pause recording. This allowed for both continuous takes and montages, as well as creative forms of looping. Features and affordances in the Vine app included a grid, focus tool, a "ghost" tool used to overlay the last frame of a previous take, the ability to save drafts, and the ability to like and "revine" or re-share content by other users. The Vine feed and browser interface displayed likes, revines, and loops (number of times played) alongside each video (Figure 2). Viners could discover and share new Vines by following one another individually or via channels, playlists, and tags.

Creators and consumers often cross-posted the videos to other social media services like Twitter and Facebook, where they saw further dissemination. Founded in June 2012, Vine was sold to Twitter in October of that year. The Vine app for iOS launched in January 2013, followed by Android and Windows versions in June and November 2013, respectively. Like the original 140-character constraint of Twitter, the app's brief, looping format provided a creative constraint that gave rise to the micro video genre, which was eventually taken up by platforms such as Snapchat, Instagram and YouTube.[6] A year before it closed in January 2017, Vine had over 200 million active users.[7]

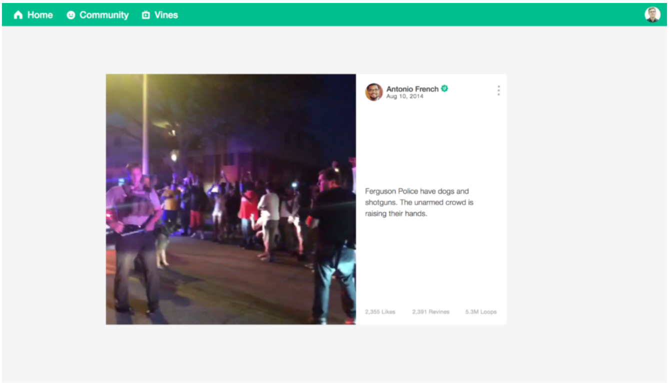
Unlike many social media companies that shut down platforms or services, Vine intentionally deployed the concept of "archive" to talk about how Vine videos would continue to live on the web after the service was turned off. When it first announced in October 2016 that it would discontinue service, Vine suggested that the content would remain online in some form:

> Nothing is happening to the apps, website or your Vines today. We value you, your Vines, and are going to do this the right way. You'll be able to access and download your Vines. We'll be keeping

[6] David Pogue, "12 Micro Video Apps Let You Shoot for Social Stardom," *Scientific American*, May 1, 2013.

[7] Craig Smith, "27 Amazing Vine Statistics and Facts | By the Numbers," *DMR*, August 12, 2018.

[9] Antonio French, "Ferguson Police have dogs and shotguns. The unarmed crowd is raising their hands," Vine, August 10, 2014. Also available at: "Ferguson Police have dogs and shotguns. The unarmed crowd is raising their hands," Internet Archive, Antonio French, archived March 10, 2016.



Antonio French
Aug 10, 2014

Ferguson Police have dogs and shotguns. The unarmed crowd is raising their hands.

2,355 Likes 2,391 Revines 5.3M Loops

Figure 2: Figure 2. A Vine posted by then St. Louis Alderman Antonio French that, as of this writing, has been looped 5.3 million times.[9]

> the website online because we think it's important to still be able to watch all the incredible Vines that have been made. You will be notified before we make any changes to the app or website.[10]

Two months later, Vine again announced that it planned to keep all video content online:

> All of your Vines will continue to live on the vine.co website so you can browse all of the amazing videos you created over the years. See our FAQ for more details and stay tuned on our website, app and @Vine on Twitter for updates.[11]

And on January 20, 2017, Vine disabled the creation of new content and introduced its new archive, using language that emphasizes persistence and the ability to relive experiences of the live app:

> Today, we made an update to the website: the Vine Archive is a time capsule of all posts made to vine.co from 2013-2017. Jump into a classic meme, have a laugh, or look up a profile.[12]

Rather than turning off its website completely Vine gave its users an opportunity to download and/or delete their data. After a certain point, their videos would be collected into a read-only public archive accessible at https://vine.co. As of

[10] Team Vine & Twitter, "Important News about Vine," Medium, October 27, 2016.
[11] Team Vine & Twitter, "Vine Update," Medium, December 16, 2016.
[12] Team Vine & Twitter, "The Vine Camera & Archive," Medium, January 20, 2017.

January 2018, https://vine.co highlighted curated lists of Vines by year; by channel (including Animals, Art, Music & Dance, Sports, and Weird); editors' picks from throughout Vine's short history; and playlists highlighting seminal Vines and memes like *on fleek*, *yeet*, and *FRESHAVOCADO*.[13] As of this writing, however, the Vine Archive no longer provides multiple access pathways to collections of Vines; direct links are now the only way to view individual videos. The current Vine FAQ page reads:

> In April of 2018, we moved the Vine Archive into a more static archived state to allow us to better preserve the public, creative expression of the Vine community. This new version of the Vine Archive allows you to watch Vines on vine.co (or anywhere that URL might be embedded) and share Vines via their unique URLs.[14]

These recent changes deactivate dynamic, linked aspects of Vine Archive, leaving the Vine-Tweets dataset to now act as an open, archival index for locating individual posts. Vine's transition from active use into a more static, maintenance mode offers an example of what Marisa Leavitt Cohn calls "convivial decay," in which an organization actively manages, plans and attends to a sunsetting process:

> Decay has a temporal quality that disrupts the assumed progressive temporality of technological change. It asks: were prior alignments as livable as we thought? Livability thus deals in rhythms but also in durations. It requires a consideration of which rhythms of work are sustainable or livable, and an awareness that the lifetimes of systems are themselves finite. The negotiation of decay is a negotiation of multiple lifetimes that are entwined - how to carefully cut these away from each other, or allow them to be companionable. It is this process of alignment that I call convivial decay.[15]

In engineering and provisioning a static archive to replace an actively used service, Vine proposed an alternative archival form for the web, challenging the assumption that retired web platforms and superseded content disappear from view - even as material may grow gradually more difficult to access.

Appraising Vine

Between October and December of 2016, when little was known about what would happen to Vine's content, a group of volunteer archivist activists known as

[13] "Vine," Internet Archive, archived January 9, 2018.

[14] "Vine FAQs," Internet Archive, archived October 26, 2018.

[15] Cohn, 1521.

Archive Team worked to preserve what they could of the platform at the Internet Archive.[16] Archive Team describe themselves as:

> … a loose collective of rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritage. Since 2009 this variant force of nature has caught wind of shutdowns, shutoffs, mergers, and plain old deletions—and done our best to save the history before it's lost forever.[17]

Archive Team volunteers developed several ad hoc channels through which to preserve Vine content. People could submit specific Vine videos URLs via a Google Form, or by sending a message to the @archiveteam Twitter account, both of which would add the URL to a tracking database. The database was then queried by instances of ArchiveTeam Warrior, a customized web client for crawling Vine content at a specified URL.[18] Warrior saved the results of this crawling activity in the Web ARChive (WARC) file format and uploaded WARCs to the Internet Archive, where they could be viewed in the Wayback Machine. To access crawled Vines, users must enter https://vine.co/ or a known Vine URL - such as https://vine.co/v/MVTjXW5tXwa, the unique URL for the video shown in Figure 2 - as a search term in the Wayback Machine (https://archive.org/web/). The search returns a calendar view of all crawls performed on that URL or, if the site has been crawled just once, redirects to that capture. If the calendar view appears, users next choose a date and timestamp to see the site as crawled on that day. Although Archive Team was able to capture a large number of Vine URLs, discovery works differently than it did on the live site or in the app. Exploring a November 5, 2016, capture of the Vine homepage—crawled during Archive Team's Vine project—shows that embedded videos perform with fidelity to the original site while search, tags, profiles, channels, and other features do not.[19] Direct video URLs remain the most effective way to access the Internet Archive collection of Vines.

An Archive Team volunteer extended the Warrior software to discover user profiles and lists of favorites, which expanded the population of Vine URLs to archive.[20] In addition, another volunteer identified a hidden Vine application programming interface (API)—referenced in the site's JavaScript but not mentioned in public API documentation—which made it possible to retrieve millions of user profile identifiers. The total number of Vine videos and their

[16] "Vine," Archive Team, updated January 21, 2017.

[17] "Main Page," Archive Team, updated June 28, 2015.

[18] "ArchiveTeam Warrior," Archive Team, updated March 6, 2018.

[19] "Vine," Internet Archive, archived November 5, 2016.

[20] "vine-grab," GitHub, Archive Team, updated January 15, 2017.

size was not known at the time, but was estimated to be somewhere between 40 and 100 million Vines, or about 200 to 300 terabytes of data.

The problem of deciding what to select for archiving from within a large body of material is what archivists know as as an *appraisal* problem. While there are many methods and theories for archival appraisal, the Society of American Archivists' *Glossary of Archival and Records Terminology* offers a relatively uncontroversial definition: ”the process of identifying materials offered to an archives that have sufficient value to be accessioned.”[21] One technique that Archive Team has used in the past when confronted with large amounts of web content to collect was to measure usage or popularity as a way to prioritize.[22] For example, Archive Team undertook to preserve the video streaming service Justin.tv after the company announced it would eliminate archiving and video-on-demand features given low view counts— over half of on-demand videos had been played 0 or 1 times. Archive Team's appraisal criteria acknowledge their use of video views as a proxy for archival value, given constraints of scale:

> Due to the nature of the Justin.tv archives, not all videos could be stored in the Internet Archive due to the immense size. The Justin.tv video archives are about 1 PB. Unfortunately, there is no practical way to determine which videos are ‘important.’ Videos with 10 or more views were selected (scraped through their search function or by a CSV manifest provided by Justin.tv staff) for archiving by the Warrior.[23]

Without labeling it as such, Archive Team volunteers used popularity as an appraisal measure for content, an approach firmly grounded in traditional archival appraisal practice:

> The documentary heritage should be formed according to an archival conception, historically assessed, which reflects the consciousness of the particular period for which the archives is responsible and from which the source material to be appraised is taken.[24]

Measuring “consciousness” is not a tractable problem by any stretch of the imag-

[21] ”Appraisal,” Glossary of Archival and Records Terminology, Society of American Archivists, accessed June 1, 2018.

[22] Ed Summers and Ricardo Punzalan, ”Bots, Seeds and People: Web Archives as Infrastructure,” *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2017): 821-834.

[23] ”Justin.tv,” Archive Team, updated January 17, 2017.

[24] Hans Booms, ”Society and the Formation of a Documentary Heritage: Issues in the Appraisal of Archival Sources,” *Archivaria* 24 (Summer 1987): 69-107.

ination, but social media platforms nevertheless deeply engage in collecting and measuring attitudinal or attentional data for a variety of economic and political purposes.[25] We see Archive Team's use of attentional data for appraisal as similar to the use of proxy measures in data science. Rather than measure consciousness directly, data scientists identify proxies for the desired information in order to indirectly observe the desired behavior or property. For example, per-capita gross domestic product (GDP) might be used as a proxy for quality of life. As Sheila Jasanoff, Cathy O'Neil, James C. Scott, and others have demonstrated, the choice of proxy is of great significance because of the biases a proxy can introduce, as well as the ways in which measurements co-produce the very behavior and systems that they are designed to make legible.[26]

Confronting the problem of identifying Vine videos to archive, Archive Team expanded its data collection to consider what Vines people were talking about on Twitter. In collaboration with other Archive Team volunteers, the first author developed a bot or piece of software named vine_urls that uses calls to the Twitter API to identify any URLs from the host *vine.co* that had been mentioned in tweets, and to log the corresponding tweet identifier and Vine URL in hourly, time-stamped files in the current working directory.[27] These data could then be aggregated and loaded into the Archive Team tracking database.

Jim Thatcher has described phenomena in which the output data from an existing computational process provide the input for a new and separate computational process as the application of "data fumes."[28] In our case study, output data from the computational process of sharing content on Twitter provides input data for the identification of Vine videos in need of archiving. Thatcher coined the term to describe downstream reuse of spatial data shared during social media check-ins, but it also provides a critical lens for examining the efficacy of appraisal in data archives. Indeed, using social media signals as a measure of appraisal value are part of a sea change in contemporary archival practices:

> Next generation mobile network infrastructure, including more powerful mobile devices that support ubiquitous computing, gives rise to this imaginary that mutually influences and is influenced by contemporary networked recordkeeping practices. This imagina-

[25] Tim Wu, *The Attention Merchants* (New York, Knopf, 2016).

[26] Sheila Jasanoff, *State of Knowledge: The Co-Production of Science and Social Order* (London: Routledge, 2006); Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Broadway Books, 2017); James C. Scott, *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (New Haven: Yale University Press, 1998).

[27] "vine_urls.py," GitHub, edsu, updated December 31, 2016.

[28] Thatcher, "Living on Fumes."

> tion and the practices of selection and destruction of mobile records are subject to device settings, terms of service across platforms and service providers, hardware and software constraints, as well as law enforcement and surveillance programs, among other influences.[29]

As we discuss below, the Vine-Tweets dataset reflects one way in which mobile-generated data already shape archival practices and will continue to do so.

Archival Circulation on the Web

The vine_urls bot ran between November 2016 and November 2017, during which time it collected 127,655,208 tweets and 2,115,360 unique Vine URLs. The data may be accessed via the Internet Archive by downloading a set of 13 gzip-compressed tar archives ranging in size from 37 to 293 megabytes and named by month of data collection: *201611.tar.gz*, *201612.tar.gz*, and so on. Users must uncompress each gzip file, then extract the contents of the resulting tar file using a file compression or archiver utility. Uncompressed folders are sized between 135 megabytes and 1 gigabyte. Each folder contains between 450 and 750 text files listing tweet ID-Vine URL pairs separated by spaces. Text files can be manipulated as is or imported into spreadsheet software; one pair appears on each row.

It is important to note that the bot continued to run for eight months after Vine had turned off the ability to create new videos. As well as providing a stream of Vine URLs for Archive Team to collect and deposit at the Internet Archive, vine_urls captured a unique picture of how social media content circulates not only through a dynamic platform but also via static, archival representations. Vine's decision to place video content in a read-only archive allowed content to continue to circulate even after the service was officially closed.

We can observe this circulation by using the Vine-Tweets dataset to visualize the number of tweets containing Vines in the months leading up to and after the sunsetting of the Vine service (Figure 3). The code for generating this and following visualizations from the Vine-Tweets dataset is available separately as a Jupyter Notebook.[30]

The rate at which Vines circulate on Twitter diminishes greatly after sunset, particularly after the ability to create vines is turned off. However, archival Vines continue to circulate for the remaining period of data collection, sometimes in

[29] Amelia Acker, "Radical Appraisal Practices and the Mobile Forensic Imaginary," *Archive Journal* 5 (November 2015).

[30] "Vine-Tweets Dataset," GitHub, edsu, updated May 31, 2018.

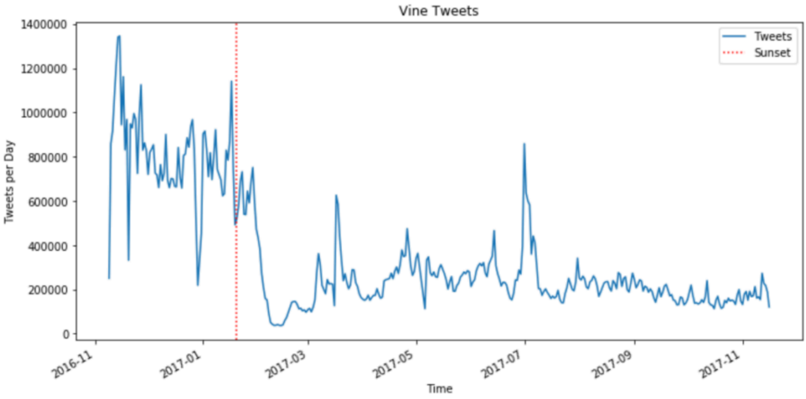


Figure 3: Tweets containing Vines per day.

great numbers - for example, a spike of 858,765 tweets on July 1, 2017. Ascertaining the cause of this and other anomalies is outside the scope of the current research. We can also examine the number of unique Vine videos being shared over the same time period (Figure 4).

Figure 4 is a more conventional graph in that it shows a crescendo of new-video sharing leading up to the sunset, followed by gradual tapering off over time. And yet, 20,000 unique archival videos shared per day on Twitter is hardly insignificant. Such sharing would have been impossible had Vine decided to take the content completely offline when sunsetting the service.

We can furthermore use the ratio of unique Vines to Vine tweets to generate a measurement of variety. Variety in this case operates as an index of the diversity of Vine videos that are shared on a given day. The more a specific set of videos are shared, such as through retweets, the lower the variety.

Variety draws a quite different picture from the previous graphs (Figure 5). It is easy to see that, despite some unremarkable peaks and troughs, the general shape of Vine variety remains unchanged before and after the sunset. One exception is a huge spike in variety that appears roughly one month after the closure of the service. However, this measure demonstrates that even aggregate numbers of tweets and vines can provide insight about how Vine content circulates as archival material in social media.

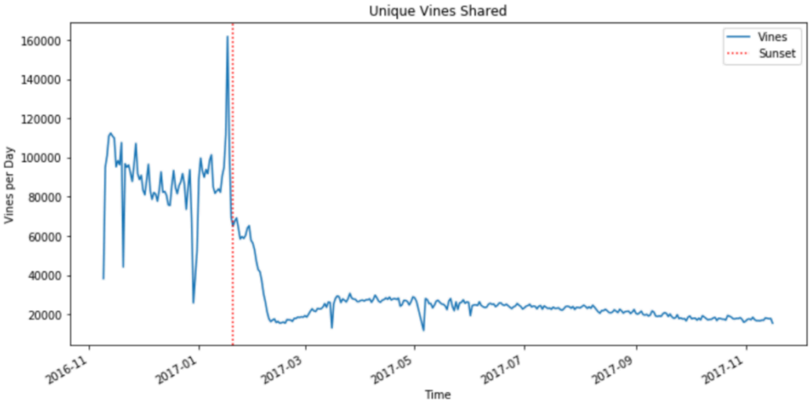


Figure 4: Unique Vines shared in tweets per day.

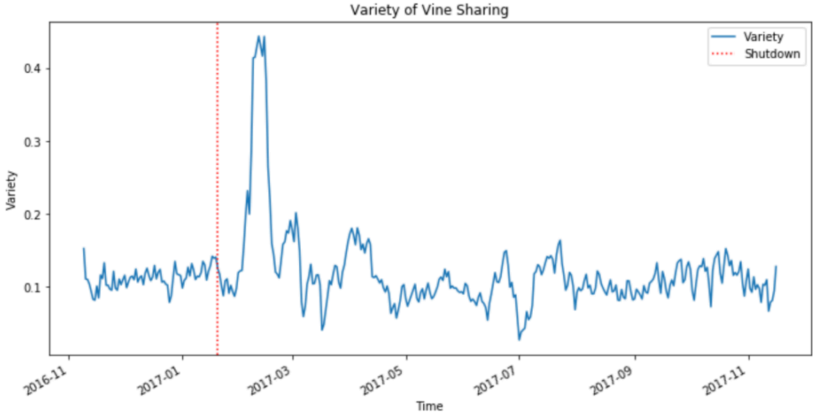


Figure 5: Variety (unique Vines over Vine tweets) per day.

Discussion

What does the activity above signify about the nature of social media archives and the web as an archival form? For one, the Vine-Tweets dataset provides a unique window into the attentional dimensions of web archives. The web archives in this case encompass both the read-only archive that Vine deployed as part of its sunsetting process, as well as the archive assembled by Archive Team volunteers for storage and access at the Internet Archive.

The activity we observe both before and after the sunset of Vine suggest possible futures for social media and web platforms that, through "convivial decay," diverge from the inevitable infrastructural shutdown. It is commonplace for organizations to selectively decommission or completely turn off websites when they dissolve or have deemed content obsolete. As the World Wide Web enters its third decade of existence, however, methods of care for its unsustainable infrastructures require focused attention. The management of the Vine archive suggests that web publishers can find useful places along the axis of online and offline in the lifecycle of their organizations. Administrative changes to the Vine Archive's design - such as the elimination of curated playlists and other points of access - may influence the social media circulation of archival Vines going forward from the period of data collection.

Another aspect of the Vine-Tweets dataset that we explore here is its transactional quality, in which conversations on Twitter about Vine videos perform as proxies informing archival assessment, recorded through just two pieces of information: the identifier for a tweet and the URL for a Vine video. The application of data fumes from one social media platform to another is not without its epistemological concerns. To use the popularity of a piece of platform content as a measure of archival value privileges users of that platform over users of another site or service. Social media companies claim to create flattened, democratic spaces for giving voice and agency, but these platforms are contingently uneven, affording particular types of interactions while privileging specific types of users and communities.[31]

In collecting tweets that mention Vine videos, to what extent did Archive Team simply measure Twitter activity (like shares, retweets, and mentions) rather than the cultural value of particular Vine videos? For example, it is easy to imagine that Vines were shared on Facebook in addition to Twitter. Neither the authors nor users of the Vine-Tweets dataset can be familiar with that sharing activity - in part because, to our knowledge, it hasn't been documented. That Vine and Twitter

[31] José van Dijck, *The Culture of Connectivity* (London: Oxford University Press, 2013).

share an administrative link has shaped the data collection process. What is lost by looking only at conversations about Vine on Twitter rather than in conjunction with other social media platforms?

Collecting data from Twitter entails actually operating its internal search indexes and processual data flows that are effectively hidden from view behind the corporate sheen of APIs and their documentation. Applications of machine learning and other computational algorithms to classify and rank - for example, to determine the relevance of data to the parameters of API calls - are subject to corporate secrecy, lack of widespread technical skill, "mismatch" between algorithmic and human reasoning, and other forms of opacity.[32] Without knowing more about how an API works or its fault tolerances, how can we know that we are accessing *all* of the Vine-sharing behavior on Twitter? When measuring the value of six-second micro videos, the stakes are relatively low; but imagine applying data fumes to appraising and archiving content by creators such as elected representatives or transnational corporations. How much more should we know about the infrastructural details of platforms in order to use their output data as proxies for value?

Conclusion

The seeming simplicity of the Vine-Tweets dataset presents challenges for researchers who would like to perform additional data processing that would be required for a more detailed study of the dataset as a situated, cultural artifact. As a simple mapping between tweet identifier and Vine URL, the Vine-Tweets dataset exists largely in a state of potential. For example, a researcher who wants to analyze the context in which a Vine was mentioned may need to examine the text of the tweet itself, what hashtags it used, the profile of the user who sent the tweet, and how many times it was retweeted. This information is not available in the tweet identifier itself, which is simply an integer. Researchers must "rehydrate" identifiers, using a tool such as the Hydrator to call Twitter's API, retrieve the metadata associated with each tweet identifier, and write it to an analyzable format like JavaScript Object Notation (JSON) or Comma-Separated Values (CSV).[33]

Similarly a Vine URL is itself only an identifier for a Vine post. In order to analyze the content of videos, video comments, and users, researchers must look up these URLs in either the Vine Archive or the Internet Archive. Information of interest

[32] Jenna Burrell. "How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3, no. 1 (June 2016).

[33] "Hydrator," GitHub, Documenting the Now, updated March 21, 2018.

could then be extracted from the page HTML. This represents significant additional work that a researcher would need to perform in order to do more than measure the rates at which tweets and Vine videos circulated before and after the closure of the Vine platform, as we have done here.

Finally, the Vine-Tweets dataset offers a new vantage point for understanding the precarity of cultural production on the web. Traditionally, the brittleness of the web has been understood in terms of link or reference rot, otherwise known as broken links.[34] Yet the Vine-Tweets dataset provides evidence of what can happen when an archive is deployed as a strategy in the dismantling of an active service, rather than simply abandoning the content completely. Twitter's investment in sustaining Vine as an archive has allowed the videos to continue to circulate on the web even after the contemporaneity of the content was past. Returning to Corry's question, with which we began, Vine Archive provides a significant example of how to shut down a website. In addition, Cohn's concept of convivial decay provides a generative model for thoughtfully managing web content along a continuum from fully online to completely offline. The Vine-Tweets dataset is a lens for observing tangible effects of using the web as an archival form.

[34] Robert Sanderson, Mark Phillips, and Herbert Van de Sompel, "Analyzing the Persistence of Referenced Web Resources with Memento" (paper presentation, Open Repositories 2011, Austin, TX, June 6-11, 2011); Jonathan Zittrain, Kendra Albert, and Lawrence Lessig,"Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations," *Legal Information Management* 14, no. 2 (2015): 88-99.