DATA
CURATION
NETWORK

Twitter Data Curation Primer

Topic	Description
File Extensions	.csv, .txt, .json
Structure	Tabular, Javascript Object Notation
Primary fields or areas of use	A variety of fields including social sciences, education, and public health.
Source and affiliation	Twitter, a social media platform that makes data available in a variety of formats
Key questions for curation review	1. Will the data be restricted or open access? 2. Was the data deposited in compliance with the Terms of Service? 3. Are the Tweet IDs separate from the tweet or user content? 4. Is the content accessible, or has much of the content been lost since data collection?
Tools for curation review	DocNow Hydrator, Microsoft Excel, R, Python, Social Feed Manager, twarc
Date Created	6/26/2020
Created by	Marley Kalt, Johns Hopkins University Dorris Scott, Washington University in St. Louis
Date updated and summary of changes made	

Suggested Citation: Kalt, Marley and Scott, Dorris. (2020). Twitter Data Curation Primer. http://hdl.handle.net/11299/216584

This work was created as part of the "Specialized Data Curation" Workshop #3 held at Washington University in St. Louis, MO on November 5-6, 2019. These workshops have been generously funded by the Institute of Museum and Library Services # RE-85-18-0040-18.

Table of Contents:

1. Description of Format

Twitter is a social media site founded in 2006. The Twitter platform, started as a "microblogging" site where individuals could post statements, or "tweets," comprised of 140 characters or less (raised to 280 characters in 2017), has grown into one of the largest social networking sites worldwide with over 320 million active users (Molina, 2017). The site's popularity and ability to share information quickly and with a wide audience – users can "retweet" a tweet to share it with their followers in addition to the followers of the original poster, and can use "hashtags" to connect their post to all other tweets about an event – has made Twitter a go-to platform for sharing breaking news and current events. Many individuals use Twitter to respond to current events, as well, making Twitter a hotbed for political and social commentary and sense-making in real time. Twitter data can serve as a rich resource for researchers, due to its ability to capture the flow of information and public sentiment surrounding a particular event or point in time.

Twitter data can be obtained through four main methods; retrieving tweets using one of Twitter's APIs, re-using an open Twitter dataset that is archived online, directly purchasing a dataset from Twitter, and accessing or purchasing tweets from a Twitter service provider (Littman, 2017).

This primer will focus on Twitter data obtained from an Application Programming Interface (API). An API is an interface that allows two applications to talk to each other (MuleSoft, 2020). Twitter has several APIs that allow users to interact programmatically with the platform through posting or retrieving tweets and their associated metadata.

Three APIs of interest are the Standard, Premium, and Enterprise API. With the Standard API, users can post, interact, and retrieve tweets and timelines, post and receive direct messages (DMs), search, follow, and get user IDs, get trends, manage and retrieve account information, and create and manage lists (Twitter, 2020a). The Premium and Enterprise APIs come with more advanced functions such as advanced filtering and access to historical tweets from a longer time range. One important distinction between these three APIs is the access to historical tweets. While the Standard API allows users to search for tweets that were created within the last seven days, the Premium API provides access to the last 30 days and the Enterprise API allows users to search for Tweets created from 2006. While the Premium API has both free and paid access to these tweets, users must pay to use the Enterprise API.

Another common method to get Twitter data is through a website that archives Twitter data, such as TweetSets or the Documenting the Now Catalog (DocNow). Twitter's Terms of Service do not allow for the actual tweets to be published online, but the content of a dataset can still be retrieved through tweet IDs, which are allowed to be publicly shared. Once a dataset of tweet IDs is obtained, users can use an API or download software with a graphical user interface, such as DocNow's Hydrator, to retrieve the tweet content associated with each unique tweet ID.

Users can also buy Twitter data directly from Twitter through the Historical PowerTrack API. The Historical PowerTrack API gives users access to the entire archive of Twitter data and has an advanced set of filtering options (Twitter, 2020b). Twitter data can also be accessed from a number of commercial and academic Twitter service providers such as DiscoverText and Brandwatch for a fee.

To preserve tweet content or design, Twitter datasets can also be created through web archiving services such as Webrecorder and Archive-It.

Twitter data can be used in a variety of formats, but the most common are JSON, text files and spreadsheet formats including .xlsx and .csv. Twitter researchers and curators should be comfortable with JSON as Twitter's APIs return tweet data and metadata in the JSON format. However, for preservation and sharing, Twitter data is often stored in text files or spreadsheets.

Sharing Twitter data can be limited, as Twitter's Developer Agreement and Policy (Twitter Developers, 2020) governs the amount of content and metadata that can be shared. This policy only allows content and metadata for public tweets to be shared in publicly-available datasets. Twitter's APIs only return tweets that are publicly available, but users are able to make their profile settings private or delete tweets at any time. Over time, tweets that were once public can be made private, which means datasets with public content and metadata shared in compliance with Twitter's developer policy may become non-compliant over time. It would be possible for researchers to share their entire dataset and periodically check Twitter to verify the tweets are still public.

However, an easier way to publicly share Twitter datasets in compliance with Twitter's policies is to only share the unique identifier (tweet ID) for each tweet. For researchers to access the full content and metadata of a dataset, they would "hydrate" the dataset of tweet IDs, by making requests to one of Twitter's APIs (or using a third-party service) using each tweet ID as a search query. If the corresponding tweet is still publicly available, researchers will be able to retrieve the full content and metadata of that tweet; if not, researchers will receive a response indicating the individual tweet is no longer available.

It is important for curators to remain up-to-date on Twitter's latest developer policies, as Twitter regularly updates its terms of use.

2. Examples

Documenting the Now, a community and tool that is focused on responsible social media data archiving has an online catalog of publicly available Twitter datasets

TweetSets is an online archive of twitter datasets based out of the George Washington University Library. These tweets were collected with their Social Feed Manager tool

Data.world is a data-sharing platform that hosts Twitter datasets (registration required).

3. Sample Citations

Dataset Citations:

Littman, Justin, 2017, "Hurricanes Harvey and Irma Tweet ids", https://doi.org/10.7910/DVN/QRKIBW, Harvard Dataverse, V1.

Phillips, Mark Edward. Notre Dame Cathedral Fire Dataset, dataset, 2019-04-08/2019-04-29; (https://digital.library.unt.edu/ark:/67531/metadc1477117/: accessed January 3,2020), University of North Texas Libraries, UNT DigitalLibrary, https://digital.library.unt.edu.

Summers, E. (2019). "ACH2019 Tweets." Internet Archive. https://archive.org/details/ach2019-tweets.

Individual File Citations (for a dataset with multiple files):

Littman, Justin, 2017, "harvey_filter_tweet_ids.txt", Hurricanes Harvey and Irma Tweet ids", https://doi.org/10.7910/DVN/QRKIBW/REEPPK, Harvard Dataverse, V1.

4. Key Questions to Ask Yourself as a Curator

Do you have enough contextual and technical information to reproduce the dataset?

- Are the parameters of data collection (e.g., time frame of tweets, specific hashtags, users or keywords followed) clear and complete?
- Do you have the tweet ID for each tweet, to be able to retrieve full JSON records for each tweet using a Twitter API or third-party service?
- Do you know which API or service the depositor used to collect the data?

Is there potentially sensitive or identifiable information in the dataset?

- Consider the balance between archiving tweets that were publicly available at the time of data collection with protecting Twitter users' right to privacy.
    - For example, you may consider stricter anonymity protections for tweets by private citizens than for tweets by public figures and business or institutional accounts.
- Always refer to Twitter's Developer Terms and Tweet Compliance guide for the most up-to-date information on how Twitter API users can share the data they collect.

Is there a high risk of loss for this dataset?

- For large Twitter datasets (e.g., millions of tweets), there is a lower risk of loss.

    - It is likely that, over time, a high proportion of these tweets will still be publicly available and findings from the dataset can be reproduced.
    - It is best practice in these cases to archive only the tweet IDs and collection parameters for the dataset.
- For small Twitter datasets, there is a higher risk of loss.
    - Over time, a smaller proportion of these tweets may still be publicly available; it may be more difficult to reproduce the results of the original dataset.
    - When depositing in a public archive, it is best practice to only archive tweet IDs and data collection parameters to ensure compliance with Twitter's terms of service. However, curators and depositors may consider alternate archiving options if the risk of loss to the data and to reproducibility of findings from the data is deemed too great.

Can this dataset be publicly shared?

- See Twitter's Developer Agreement and Policy, Twitter's terms on the Redistribution of Twitter Content, and Twitter's Tweet Compliance guide for more information on what can be publicly shared.
- Are there any data use agreements in place?

Is data collection ongoing?

- Could the researcher deposit additional tweets collected using the same collection parameters later on?

5. Key Clarifications to Get From Researcher

Curators should ask for clarifications regarding understanding and documenting the dataset along with sharing the data and reproducing results from the researcher. It is important to get these clarifications in order to ensure compliance with the policies of the repository or publishing guidelines of academic journals, and making sure that the methods that were applied was accurately captured in the documentation.

Key Clarifications Regarding Understanding and Documenting the Dataset

- Is there a codebook with column-level descriptions of metadata? It is important to provide column-level metadata descriptions (e.g., does the "id" column refer to tweet IDs or user IDs?)
- Is there a data dictionary or README file with a brief description of the purpose of the collection, a list of the hashtags used for the Twitter collection, date of collection, and file names?

Key Clarifications Regarding Sharing the Data and Reproducing Results

- Are the method(s) used to collect the tweets included in a README file?
    - Did you collect the tweets via a programming language such as R or Python or did you use a program such as NodeXL or the Hydrator tool?

- Has the dataset been de-identified to remove any sensitive personal information? An example of this are phone numbers on personal accounts.
- If tweets were collected using a programming language, can the code used to create the collection be shared?
- Is the data collection and creation in compliance of Twitter's term of service? The curator should not feel like they need to enforce the terms of service but rather remind the researcher of the importance of complying with this policy.

6. Metadata Standards and README Requirements

At the time of publication of this primer, the authors were unaware of any metadata standards designed specifically for social media data.

For explanations of the metadata contained within each JSON record retrieved from Twitter API services, the Twitter Developer documentation and the Social Feed Manager documentation have comprehensive data dictionaries for tweet objects.

A README file should be created for every archived Twitter dataset, for users to understand how the data was collected and used, and how the dataset could be reproduced. In addition to standard README elements such as dataset title, author and related publication(s), useful information the README, specific to Twitter datasets, includes:

- Number of tweets in the dataset
- Time period covered by the archived tweets
- Search terms (e.g., hashtags, keywords, usernames) used to collect the archived tweets
- Tools or methods used to collect the archived tweets
- Explanation of file structure, if there are multiple files containing Twitter data
- Contextual information surrounding the subject matter of the dataset

7. Resources and Software for Reviewing Data

Twitter datasets can be reviewed using text editors such as Notepad++, Sublime Text, Atom, and TextEdit, or spreadsheet software such as Google Sheets and Microsoft Excel. In addition, there are a number of libraries and packages that can be used in Python or R to review these data. See Appendix B for a list of Python and R packages used to collect, hydrate, and analyze Twitter data.

Doc Now Twitter Hydrator tool: If you want to hydrate the tweet IDs to get the content of the tweets, you can use this tool. Hydrating tweets involve using tweet or user IDs to retrieve the content of the tweets. The results of the hydration will be exported in a .txt format. The Hydrator tool uses a graphical interface, so you do not need to use the command line or a programming language to use this tool.

Social Feed Manager: This tool, created by George Washington University Libraries, allows users to build a collection of tweets based on a tailored search. The Social Feed Manager allows for building and managing various collections of Twitter data. Data can be exported to a spreadsheet or through to the command line.

R and Twitter packages: See Appendix B for a list of Python and R packages used to collect, hydrate, and analyze Twitter data.

8. Preservation Actions

- Data files should be converted to open, non-proprietary file formats (e.g., .txt, .csv, .json)
  - Note: Microsoft Excel file formats (e.g., .xlsx, .xls) are not recommended for the preservation of Twitter data. Excel software has a tendency to convert tweet IDs, which are long strings of numeric values, into scientific notation. Converting data out of this format can result in irreversible corruption or data loss

- Data files should be accompanied by a README file to provide context about data collection parameters and how to rehydrate the data, if necessary

- Any code used to process the data or reproduce results should be archived along with the raw data:
  - If the data is associated with a published work, it is best practice to archive a current version of any code, even if the code is also available on a public repository, such as GitHub
  - Make sure there are no API keys, or other information related to the researcher's Twitter credentials, in the archived code

- Make sure the preserved dataset complies with Twitter's terms of service:
  - This may mean preserving only the tweet IDs from a dataset
  - See Twitter's Developer Agreement and Policy, Twitter's terms on the Redistribution of Twitter Content, and Twitter's Tweet Compliance guide for more information on how Twitter data can be archived and shared

9. What to Look for to Make Sure Files Meet FAIR Principles

Findable

- The dataset should be publicly available in a data repository
- The dataset should have a persistent identifier (e.g., DOI)

Accessible

- When possible, data should be made public through the removal of personally identifiable information (PII). In many cases, this may mean archiving files containing tweet IDs, rather than the full-text or metadata of the tweets

Interoperable

- Data should be archived in non-proprietary file formats (e.g., .csv, .txt, .json rather than Excel)
    - Many of the existing tools for hydrating and analyzing Twitter data, such as Twarc, are built for plain text, .csv or .json files. Archiving data in these open formats allows users to make use of these tools without needing additional software to read or reformat the data

Reusable

- Datasets should include complete information about how the data was collected
- Datasets that archive tweet IDs should have guidance for rehydrating the data
- Datasets that archive tweet IDs should have enough tweets in the dataset to remain meaningful if some content were to be lost (e.g., made private or deleted from Twitter's platform) over time

10. Ways in Which Fields May Use This Format

Twitter data is of interest to a variety of research fields and academic disciplines ranging from the humanities to public health. One interesting facet of Twitter data is that it is used in both qualitative and quantitative analysis, through methods such as content analysis and machine learning.

Examples:

Digital Humanities

#silentsam Social Media Analysis Project: The project was conducted by the UNC Chapel Hill Digital Innovation Lab which involved collecting and analyzing tweets regarding the confederate monument Silent Sam. Another aim of this project was to establish best practices regarding the use of social media data in digital humanities.

Geography

Mapping the EU Referendum on Twitter: Tweets related to the EU referendum were collected to better understand which parts of Great Britain supported staying or leaving the EU.

Geology

USGS Twitter Earthquake Dispatch (USGSted): The United States Geological Survey (USGS) created a software application which collects tweets that contain the word "earthquake" and related words and stores them in a database. The National Earthquake Information Center (NEIC) uses this information to create maps of the tweet locations and to analyze the severity of the earthquake.

Public Health

HealthMap: Healthmap was created by a team of researchers to mine Twitter for foodborne illnesses and other kinds of health conditions. Data was collected with an R package and NodeXL. The data was stored as csv files and as Excel files respectively.

11. Documentation of Curation Process: What to Capture

- Steps taken to understand or view the data
  - Document any scripts or tools used if a dataset was hydrated during curation

- Transformations to the data
  - Document any actions taken to make the data both FAIR and sharable (e.g., file format migrations, limiting a dataset to adhere to Twitter's terms of use)

- Decisions about what data to share
  - Document how you came to any decisions about which parts of the dataset will or will not be shared publicly

12. Bibliography

HealthMap (n.d.) About. Retrieved from http://www.diseasedaily.org/about.

Littman, J. (2017). Where to get Twitter data for academic research. Social Fee Manager. Retrieved from https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data.

Molina, B. (2017, October 26). Twitter overcounted active users since 2014, shares surge on profit hopes. USA Today. Retrieved from https://www.usatoday.com/story/tech/news/2017/10/26/twitter-overcounted-active-users-since-2014-shares-surge/801968001/

MuleSoft (2020). What is an API (Application Programming Interface). Retrieved from https://www.mulesoft.com/resources/api/what-is-an-api

Twitter (2020a). Getting Started. Retrieved from https://developer.twitter.com/en/docs/basics/getting-started

Twitter (2020b). Get Batch Historical Tweets. Retrieved from https://developer.twitter.com/en/docs/tweets/batch-historical/overview

Twitter Developers. (2020). Developer Agreement and Policy. Retrieved from https://developer.twitter.com/en/developer-terms/agreement-and-policy

U.S. Geological Survey (2011). Twitter Earthquake Dispatch (@USGSted). Retrieved from https://www.usgs.gov/software/usgs-twitter-earthquake-dispatch-usgsted

Appendix A – file type CURATED checklist

CHECK Step

CURATE Action	Curator Checklist
Check data files and read documentation Review the content of the data files (e.g., open and run the files or code) Verify all metadata provided by the author and review the available documentation	❑ Files open as expected o Issues __________ ❑ Metadata quality is rich, accurate, and complete o There is a README / Codebook / Data Dictionary / Other documentation o Documentation fully describes the parameters of data collection o Documentation fully describes the technology used to collect data o All variables are defined ❑ Documentation/Metadata missing or no documentation/metadata ❑ Documentation/metadata needs work ❑ There is no sensitive information present o The dataset does not include potentially sensitive content tweeted by individuals o If code was deposited with the data, personal API keys have been removed

UNDERSTAND Step

CURATE Action	Curator Checklist
Understand the data (or try to) Check for quality assurance and usability issues such as missing data, ambiguous headings, code execution failures, and data presentation concerns Try to detect and extract any "hidden documentation" inherent to the data files that may facilitate reuse Determine if the documentation of the data is sufficient for a user with similar qualifications to the authors to understand and reuse the data. If not, recommend or create additional documentation (e.g., a readme.txt template)	❑ Organization of data well-structured ❑ Headers/codes clearly defined ❑ Define headers ❑ Clarify codes used ❑ Clarify use of "blanks" ❑ Quality control clearly defined ❑ Unclear quality control ❑ Update or add Methodology ❑ Associated code runs (note: you may need a Twitter API key to run any deposited code) ❑ Deposit appears to comply with Twitter's developer policies

REQUEST Step

CURATE Action	Curator Checklist
Request missing information or changes Generate a list of questions for the data author to fix any errors or issues	Narrative describing the concerns, issues, and needed improvements to the data submission.

AUGMENT Step

CURATE Action	Curator Checklist
Augment the submission Enhance metadata to best facilitate discoverability Create and apply metadata for the data record, including descriptive keywords When appropriate, structure and present metadata in domain-specific schemas to facilitate interoperability with other systems	❑ Discoverability sufficient ❑ Recommend (circle one) full-text index / file rename / file reorder / file descriptions / zip files into one archive Other _______________ ❑ Keywords Sufficient ○ Suggestions_____________ ❑ Linkages Sufficient ○ Link to report/paper ○ Link to related data sets ○ Link to source data ○ Link to other _____________

TRANSFORM Step

CURATE Action	Curator Checklist
Transform file formats Identify specialized file formats and their restrictions (e.g., Is the software freely available? Link to it or archive it alongside the data) Transform files into open, non-proprietary file formats that broaden the potential audience for reuse and ensure that preservation actions might be taken by the repository in later steps Retain original files if data transfer is not perfect	❑ Preferred file formats in use ❑ Recommend conversion (e.g., Excel to CSV) ○ Retain original formats ❑ Software needed to hydrate tweets is readily available ❑ Unclear version of software ❑ Unclear software used ❑ Visualization of data easily accessible, if visualizations were included in deposit

EVALUATE Step

CURATE Action	Curator Checklist
Evaluate and rate the overall data record for FAIRness Score the dataset and recommend ways to increase the FAIRness of the data	❑ Findable - o Metadata exceeds author/ title/ date, o Unique PID (DOI, Handle, PURL, etc.) o Discoverable via web search engines ❑ Accessible - o Retrievable via a standard protocol (e.g., HTTP) o Free, open (e.g., download link) ❑ Interoperable - o Metadata formatted in a standard schema (e.g., Dublin Core) o Metadata provided in machine-readable format (OAI feed) ❑ Reusable - o Data include sufficient metadata about the data characteristics to reuse o Contact info displayed if the direct assistance of the author needed o Clear indicators of who created, owns, and stewards the data o Data are released with clear data usage terms (e.g., a CC License)

Appendix B – Resources

SOFTWARE APPLICATIONS

Social Feed Manager: This tool, created by George Washington University Libraries, allows users to build a collection of tweets based on a tailored search. The Social Feed Manager allows for building and managing various collections of Twitter data. Data can be exported to a spreadsheet or through to the command line.

Documenting the Now Hydrator: This tool allows users to “hydrate” tweet IDs to receive the full content and metadata of each tweet, as long as the tweets are publicly available on Twitter. The Hydrator tool uses a graphical interface, so you do not need to use the command line or a programming language to use this tool. However, users must have a registered Twitter developer account and Twitter API credentials to use this tool.

PYTHON LIBRARIES

twarc: Twarc is a library that allows you to use the command line to collect Twitter data as a JSON object.

tweepy: Tweepy is a library that allows you to have access to the Twitter API.

pandas: Pandas is a library that allows you to read in rectangular data such as csv, SQL tables, or Excel sheets into data frames.

csv: This built-in python library allows you to read and write csv files in Python.

R PACKAGES

rtweet: This package allows one to access the Twitter API to collect tweets.

readr: This package provides functions to read rectangular data such as csv, tsv, and delimited files into data frames.

readxl: This package allows you to read Excel files into RStudio into data frames. This package supports both the .xls and .xlsx format. This package can read individual Excel sheets as well.

streamR: This package is used to access Twitter’s Streaming APIs, which allow you to collect tweets happening in real-time.

twitteR: Like rtweet, this package is used to collect tweets. Compared to rtweet, this package is older and isn’t frequently maintained.