The Provenance of a Tweet

We welcome your comments on this working paper. Use Google Doc's Comments feature on this copy, email sfm@gwu.edu, or connect with us via Twitter @SocialFeedMgr.

Dan Kerchner
Justin Littman
Christie Peterson
Vakil Smallen
Rachel Trent
Laura Wrubel

In "A Manifesto for Data Sharing in Social Media Research", Katrin Weller and Katharina E. Kinder-Kurlanda draw the connection between data sharing, validity, and documentation for social media research (Weller and Kinder-Kurlanda 2016):

> [Data sharing supports] validity by advancing reproducibility and comparability: reproducibility of research results is an important requirement for achieving validity in many scientific disciplines. Even in those domains where reproducibility is not required or even desired, it needs to be made transparent how results were generated. ...
>
> In the case of social media data user-friendly documentation may include e.g. explanations about the hashtags used for collection or lists of accounts, which also need to be archived in order for the research to become reproducible. Detailed documentation of the dataset and the provision of code and syntax allow everyone to check how collection, cleaning and analysis were performed.

Providing this research documentation then becomes a crucial component of their "call to action for the broader research community to advance current practices of data sharing in the future."

This notion of research documentation is closely aligned with the archival field's concept of provenance, a concept foundational to the field and referring to information that traces the origin and chain of custody of archived information. In the context of digital curation, there is a particular emphasis on tracing provenance back to the origination of the data, including the particular settings and configurations used to collect and create data ("Reference Model for an Open Archival Information System (OAIS)" 2012; "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information" 1996).

The W3C's PROV working group offers a definition of provenance that provides a common ground for the social media research community and the archival community ("PROV-Overview" 2016):

> Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

Social Feed Manager (http://go.gwu.edu/sfm) is open source software for collecting social media data. It is intended to satisfy the requirements of both social media researchers across a broad array of disciplines to collect datasets of social media and archivists/librarians to build collections of social media. As such, one of the goals of SFM is to record and make available provenance metadata for the social media data. The purpose of this paper is to describe the metadata the SFM team has chosen to record about the source of the social media items SFM collects, with the aim of advancing the conversation around social media data.
The approach described in this paper is influenced heavily by discussions in the social media research community about research methodology (Jürgens and Jungherr 2016; Weller and Kinder-Kurlanda 2015; Mayr and Weller 2016; Hutton and Henderson 2015). Further, it draws from archival approaches to digital curation and digital provenance ("DCC Curation Lifecycle Model | Digital Curation Centre" 2016; Lee and Allard, n.d.). And lastly, it leverages approaches taken by the web archiving community for collecting from the web (International Internet Preservation Consortium 2016).

The approach taken by SFM is informed by and could probably be mapped to several metadata standards, including W3C PROV provenance metadata ("PROV-DM: The PROV Data Model" 2016) and PREMIS preservation metadata ("PREMIS Data Dictionary for Preservation Metadata, Version 3.0" 2015). However, SFM is not attempting to implement any current specification.

Areas of provenance metadata for a harvested tweet

SFM collects social media posts. SFM can collect a variety of types of social media (e.g., Flickr photos, Weibo posts), but for illustrative purposes this discussion will focus on the tweet.

In particular, it will focus on this tweet:

The provenance metadata for this single tweet spans three areas: creation, collection, and selection. Creation is the authoring of the tweet. Collection is how SFM retrieved the tweet from Twitter's API. And selection is what decisions by the user led SFM to harvest the tweet.

In what follows, we will explore the provenance metadata around these three areas. This will explain how and where provenance metadata is recorded, as it is recorded across different parts of the SFM system.

Creation

Tweets that are retrieved from Twitter's API contain metadata that isn't available from the website. Here's the full response received from the API to a request for this tweet, as of April 20, 2016: https://gist.github.com/justinlittman/462a398d161002a8caff0905bf4e5f7f

The metadata includes information about the Twitter user who authored the tweet. Here's a subset of the user metadata[1]:

```
"user": {
    "id": 216776631,
    "verified": true,
    "description": "Join our campaign for president at
https://t.co/nuBuflGIwb. Tweets by staff.",
    "location": "Vermont",
    "screen_name": "BernieSanders",
```

[1] User metadata is documented at https://dev.twitter.com/overview/api/users

```
    "lang": "en",
    "name": "Bernie Sanders",
    "url": "https://t.co/W6f7Iy1Nho",
    "created_at": "Wed Nov 17 17:53:52 +0000 2010",
    "time_zone": "Eastern Time (US & Canada)",
},
```

The metadata includes how it was posted:

```
"source": "<a href=\"https://about.twitter.com/products/tweetdeck\"
rel=\"nofollow\">TweetDeck</a>",
```

and when it was posted:

```
"created_at": "Fri May 20 16:02:03 +0000 2016",
```

The tweet metadata might even include from where it was posted. @BernieSanders' tweet doesn't, but @justin_littman's tweet (https://twitter.com/justin_littman/status/721034915396063232) does:

```
"place": {
    "full_name": "Iceland",
    "url":
"https://api.twitter.com/1.1/geo/id/c3932d3da7922986.json",
    "country": "\u00cdsland",
    "place_type": "country",
    "bounding_box": {
        "type": "Polygon",
        "coordinates": [
            [
                [-24.7942167, 63.1861245],
                [-13.2194228, 63.1861245],
                [-13.2194228, 66.5999889],
                [-24.7942167, 66.5999889]
            ]
        ]
    },
    "contained_within": [],
    "country_code": "IS",
    "attributes": {},
    "id": "c3932d3da7922986",
    "name": "Iceland"
```

},

Thus, the tweet metadata includes fields describing some of the circumstances of the tweet's posting as reported by Twitter.[2]

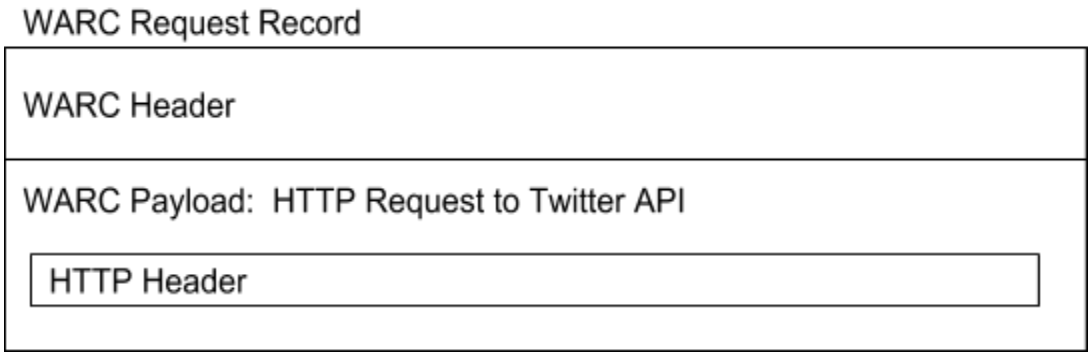
Collection

Twitter allows tweets to be collected from its API. In SFM, calls to the API are made by a component called a harvester; in the case of retrieving the @BernieSanders timeline, the API is called by the Twitter REST harvester.

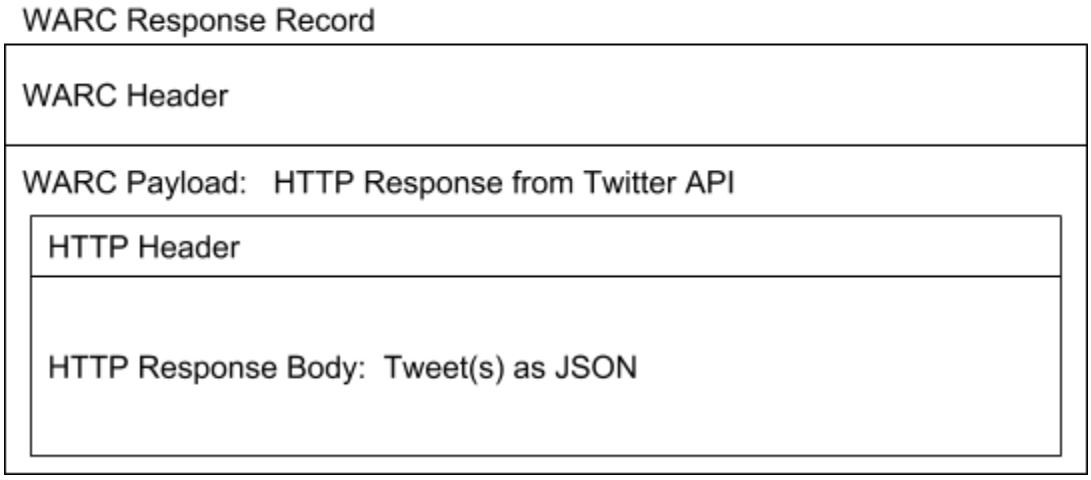
Like the API of other social media platforms, Twitter's API is accessed over HTTP. For each call to the API, the SFM harvester records the *exact HTTP transaction* between itself and the API, providing precise documentation of the collection of the social media data. Let's consider how collection is recorded and what provenance metadata it makes available.

An HTTP transaction is composed of an HTTP request from the client (the Twitter REST harvester) to the server (Twitter API) and an HTTP response from the server to the client. The HTTP transaction is recorded as a pair of WARC request and response records in a WARC file. A WARC file, short for Web ARChive file, is a file format designed by the web archiving community specifically for recording HTTP transactions, typically that are collected during web crawling (International Internet Preservation Consortium 2016).

A WARC record is composed of a WARC record header and a WARC record payload. The WARC record payload is the HTTP message, which is composed of the HTTP message header and the HTTP message body ("Hypertext Transfer Protocol -- HTTP/1.1" 2016). In this case, the HTTP message body from the WARC response records contains tweets encoded as JSON.



[2] Further information on tweet metadata is available from https://dev.twitter.com/overview/api/tweets.



Given this overall structure for recording the harvesting of social media data from APIs, let's unpack the various records to consider what provenance metadata is available, starting with the HTTP request message. The HTTP request message contains the request made to the API. Here is an example HTTP request message for retrieving a Twitter user timeline: https://gist.github.com/justinlittman/ccd731df697cdfd42b7605c3a61dea0f#file-warc-request-record-L11. The following is the request line, which shows the URL from the API that was retrieved and the query parameters that were passed:

```
GET
/1.1/statuses/user_timeline.json?count=200&max_id=682336123457155073&
user_id=216776631 HTTP/1.1
```

In this case, it was a request for 200 records (`count=200`) starting after a particular tweet (`max_id=682336123457155073`) from the timeline (`user_timeline.json`) of @BernieSanders (`user_id=216776631`). Following this line are message headers that provide additional parameters for the request that might affect the social media data that is returned. In particular, note that the last line contains the API keys (i.e., the credentials) that were used for the request:

```
authorization: OAuth oauth_nonce="158709461152068544681463772498",
oauth_timestamp="1463772498", oauth_version="1.0",
oauth_signature_method="HMAC-SHA1",
oauth_consumer_key="EHdoTe7ksBgflP5nUalEfhaeo",
oauth_token="481186914-c2yZjbk1np0Z5MWEFYYQKSQNFBXd8T9r4k90YkJl",
oauth_signature="z8sHkIQ2WsN6fqPFF83rpoVSN7U%3D"
```

In addition to the tweets themselves (i.e., the HTTP message body), the HTTP response contains additional metadata about the response from the API. Here's the example HTTP response:

https://gist.github.com/justinlittman/518a73c92606039af739ae4f7afe7419#file-warc-response-record-L12. The first line is the status line, and it contains the status code indicating how the request was handled by the server. In this case, the status code indicates that the server successfully handled the request. Twitter provides guidance on the meaning of different status codes: https://dev.twitter.com/overview/api/response-codes.

Following the status line are message headers. Those not beginning without an x- are generally described by a W3C specification. While the provenance value of most of the headers remains to be evaluated, at the very least the date is significant for establishing when the data was collected:

```
date: Fri, 20 May 2016 19:28:18 GMT
```

Message headers prefixed with x- are considered non-standard. In this case, the x- fields are Twitter-specific. Among other uses, Twitter uses x- message headers to provide information on rate limits which might be significant for understanding gaps in collecting.[3]

The WARC headers document the HTTP messages, including metadata about how each message was sent or received. Some of this is a duplicate of metadata from the HTTP message. For both record types, the unique metadata includes the date that the record was created:

```
WARC-Date: 2016-05-20T19:28:18Z
```

and fixities for the HTTP message:

```
WARC-Block-Digest: sha1:d4f5ddcfbe1c814fdee445ff145abebf22411bf8
```

The WARC response record also includes the IP address of the server:

```
WARC-IP-Address: 199.16.156.199
```

All of the provenance metadata described to this point in the paper is recorded in WARC files stored on the filesystem. The provenance metadata described in what follows is recorded as database records[4] in SFM's database. This begins with a database record for each WARC file that is created. The WARC database record includes the location, size, fixity, and creation date of the WARC.

---

[3] We are not going to address Twitter-specific issues with rate limiting. For further discussion of the challenges of collecting Twitter see (Jürgens and Jungherr 2016; Nejdl et al. 2016).

[4] For clarity, we are using the term "database record" to distinguish from the use of the term "record" in archives, meaning an individual historical document.

Each WARC database record is associated with a harvest database record. The harvest database record contains information about a single collection activity, including information about the request that was made to the SFM harvester to perform the harvest and information provided by the harvester on the outcome of the collection activity. Information about the harvest request include the request date and an association with the collection database record. The collection database record specifies the harvest type (e.g., Twitter user timeline), the credentials (API keys) to be used, seeds, and harvest options.

Seeds are the specific targets of collection. What a seed is will vary depending on the harvest type. For example, for a Twitter user timeline harvest type, each seed will be a Twitter account; for a Twitter search harvest type, the seed will be the search query. The number of seeds will also vary depending on the harvest type. For example, a Twitter user timeline harvest type may have multiple seeds, but a Twitter sample stream will have no seeds.

Seeds may have a token and/or a uid. Like seeds, the semantics of these fields vary by harvest type. For example, for a Twitter user timeline, the token is the screen name and the uid is the user id.

Harvest options are additional specifications provided to the harvester. These also vary by harvest type. For a Twitter user timeline harvest type, they include whether to perform an incremental harvest (i.e., retrieve all tweets in a user timeline or only new tweets), and whether to extract media and/or web resource links from the tweets. (If extracted, these links are then passed on to a web harvester to collect.)

It is important to note that the collection database record documents the settings as they existed when the harvest request was made, not as they currently exists. Users may change collections to reflect changes in selection, e.g., adding new seeds or changing frequency of harvests, and these changes will not be reflected in the database records already created.

Each SFM harvester provides metadata on the outcome of the collection activity, which is recorded in the harvest database record. This includes the date started and ended; some basic statistics; any informational, warning, or error messages; token updates; and uid updates.

Notice that not all of the metadata that is recorded is exposed in the UI.

The basic statistics are a count, by day, of the type and number of social media items collected. (In the UI, only a rollup of the stats are displayed, not the counts by day.)

Token updates are changes in tokens detected by the harvester. So, for example, if when retrieving a user timeline, the Twitter REST harvester detected that the screen name had changed, it would report back the new screen name.

The uid updates are the reverse of token updates. If given only a token and the harvester found a uid, the harvester would report back the uid. So, for example, if the Twitter REST harvester was only given a screen name and it found the user id for that screen name it would report back the user id. (The motivation for this is to perform collection using uids, which don't change, instead of tokens, which do. However, tokens are often more convenient for users to provide.)

Selection

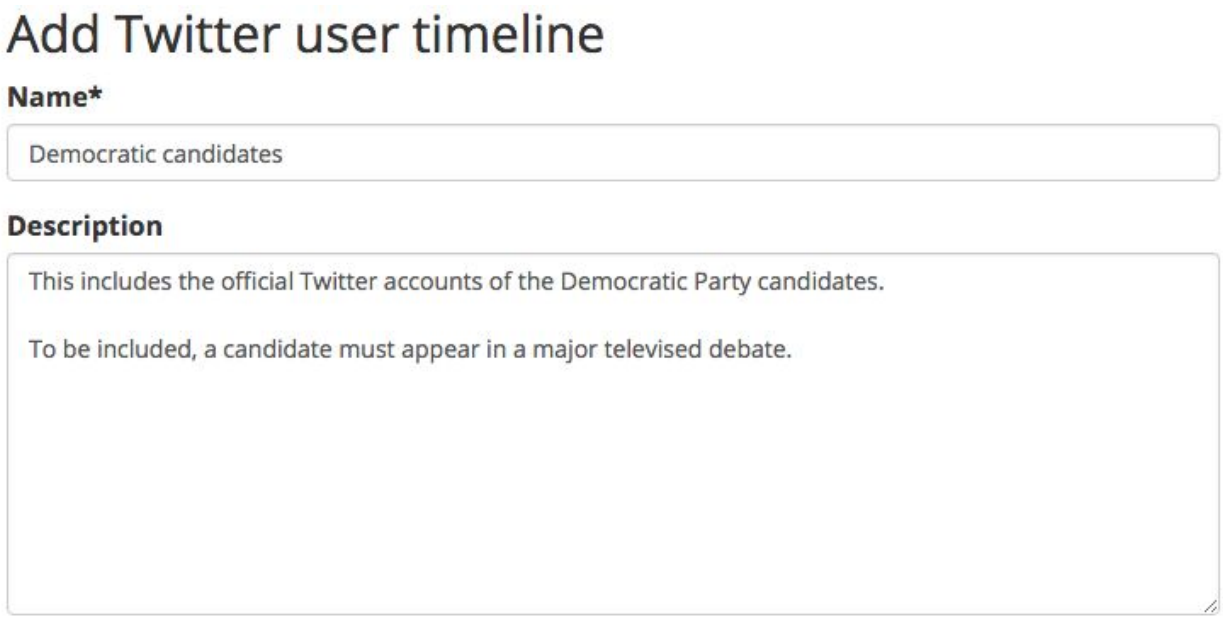
Now that we understand the metadata that is available about how social media items were created and how they were collected, the final area of provenance metadata to consider is selection. Selection is the human process of deciding which social media to collect and which not to collect. SFM provides opportunities for the users performing that selection to document their selection decisions.

The collection (partially described above) is the basic unit of collection. It is for a particular harvest type and is composed of the harvest options, a collection schedule (e.g., every hour,

every day, every week, etc.; only for some harvest types), seeds (again, only for some harvest types), and credentials. Each collection is part of a collection set.

The collection / collection set arrangement provide the user some flexibility in how to organize collected social media. In the example, the collection set is the “2016 election” and the collection is the “Democratic user timelines.”

SFM provides users with means for documenting their organization and selection criteria. In addition to a name, the user can provide a description for the collection and collection set as shown below:



Further, a log is kept in the database of each change to a collection set, collection, seed, or credential. The log automatically includes the time of the change, the user that performed the change, the fields that were changed, and optionally may include a note from the user describing the change. The following shows a change made to the schedule of the “Democratic candidates” collection:

While the 140 characters of a tweet may be parsimonious, as should now be evident, the metadata that documents its creation, collection, and selection is quite expansive. It should be noted that while all of the metadata described above is collected by SFM, not all of it is made readily available to users through mechanisms such as the user interface or data exports. The SFM team has made a “first cut” at exposing the metadata that we think is most significant for users, but stand ready to adjust our strategy.

Perhaps at the expense of providing an excess of details, we have attempted to lay bare the metadata behind the creation, collection, and selection of a single tweet. This metadata is summarized below:

Source	**Unit**	**Subunit**	**Example content**
Twitter	Tweet	User metadata	● Screen name ● Date account created ● Location
		Tweet metadata	● Date ● Tweet text ● Mentions ● Hashtags ● URLs ● Source (how posted)
WARC record (created by SFM harvester)	WARC request record	Record header	● Date WARC record created ● Server information ● Fixities
		Record payload	HTTP request to API: ● URL shows params such as user account id or keywords HTTP message headers: ● API keys used
	WARC response record	Record header	● Date WARC record created ● Server information ● Fixities
		Record payload	HTTP response message body: ● Tweets as JSON HTTP message headers: ● Date ● Response status (e.g. 200 OK) ● Content type

SFM database	WARC record		Info about the WARC file: • File location • File size • Fixity • Creation date
SFM database	Harvest record		Info about the harvest: ● Date ● Collection ● Date harvest started ● Date harvest ended ● Messages (informational, warning, or error) ● Token/seed updates ● Basic stats on number of items collected
SFM database	Collection record		● Harvest type ● Harvest options (e.g., incremental, harvest web resources) ● Credentials (API keys) ● Seeds belonging to collection ● Description of collection
SFM database	Seed record		Info varies by platform, may include: ● Screen name ● UID ● Keywords to filter on
SFM database	Change Log		List of changes made to a Collection, Seed, or Collection Set. Includes: ● Change note ● Fields changed ● User who made change ● Date of change

For Consideration

In “A Manifesto for Data Sharing in Social Media Research”, Weller and Kinder-Kurlanda, entreat social media researchers to establish standards for provenance metadata for social media data (Weller and Kinder-Kurlanda 2016):

> Ideally, researchers would need to agree on documentation standards for social media datasets: which basic information needs to be documented in order to understand how a dataset was collected or in order to reproduce another dataset with the same parameters?

As toolmakers, our contribution has been to detail what metadata is available to inform this discussion. In that spirit, we conclude with some questions for consideration:

- Which of this provenance metadata do you (researcher, archivist, librarian, etc.) want access to?
- How do you want access to this metadata? In SFM's UI? In reports when exports are created? Exposed via SFM's software libraries? A REST API? Machine-readable? Human-readable?
- What metadata have we missed?
- Do the answers to the previous questions vary by discipline (e.g., humanities, social science, etc.)?
- Are there other relevant specifications or standards that we should consider? Is there value in a mapping to or providing output in accordance with metadata standards such as PREMIS or PROV?

Bibliography

"DCC Curation Lifecycle Model | Digital Curation Centre." 2016. Accessed June 13. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Hutton, L., and T. Henderson. 2015. "Towards Reproducibility in Online Social Network Research." *IEEE Transactions on Emerging Topics in Computing* PP (99): 1–1.

"Hypertext Transfer Protocol -- HTTP/1.1." 2016. Accessed May 24. https://www.w3.org/Protocols/rfc2616/rfc2616.

International Internet Preservation Consortium. 2016. "The WARC Format." Accessed February 10. http://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/.

Jürgens, Pascal, and Andreas Jungherr. 2016. "A Tutorial for Using Twitter Data in the Social Sciences: Data Collection, Preparation, and Analysis," January. doi:10.2139/ssrn.2710146.

Lee, Christopher A., and Suzie Allard. n.d. "Open Data Meets Digital Curation: An Investigation of Practices and Needs." http://ils.unc.edu/callee/idcc16-open-data-lee.pdf.

Mayr, Philipp, and Katrin Weller. 2016. "Think before You Collect: Setting up a Data Collection Approach for Social Media Studies." *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1601.06296.

Nejdl, Wolfgang, Wendy Hall, Paolo Parigi, and Steffen Staab. 2016. "Twitter as a First Draft of the Present – and the Challenges of Preserving It for the Future." In *Proceedings of the 8th ACM Conference on Web Science*, edited by Wolfgang Nejdl, Wendy Hall, Paolo Parigi, and Steffen Staab, 183–89. ACM.

"PREMIS Data Dictionary for Preservation Metadata, Version 3.0." 2015, June. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

"Preserving Digital Information: Report of the Task Force on Archiving of Digital Information." 1996, May. http://www.clir.org/pubs/reports/pub63watersgarrett.pdf.

"PROV-DM: The PROV Data Model." 2016. Accessed June 6. https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

"PROV-Overview." 2016. Accessed June 6. https://www.w3.org/TR/prov-overview/.

"Reference Model for an Open Archival Information System (OAIS)." 2012, June. http://public.ccsds.org/publications/archive/650x0m2.pdf.

Weller, Katrin, and Katharina Kinder-Kurlanda. 2015. "Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research." In *Standards*

and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Work. Proceedings Ninth International AAAI Conference on Web and Social Media, May 26, 28–37.

Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2016. “A Manifesto for Data Sharing in Social Media Research.” In *Proceedings of the 8th ACM Conference on Web Science*, 166–72. ACM.