Apache-Affiliated Twitter Screen Names: A Dataset

Megan Squire
Dept. of Computing Sciences
Elon University
Elon, NC, USA
msquire@elon.edu

***Abstract*—This paper describes a new dataset containing Twitter screen names for members of the projects affiliated with the Apache Software Foundation (ASF). The dataset includes the confirmed Twitter screen names, as well as the real name as listed on Twitter, and the user identification as used within the Apache organization. The paper also describes the process used to collect and clean this data, and shows some sample queries for learning how to use the data. The dataset has been donated to the FLOSSmole project and is available for download (https://code.google.com/p/flossmole/downloads/detail?name=apacheTwitter2013-Jan.zip) or direct querying via a database client.**

***Index Terms*—open source software, free software, Apache, Twitter, names, dataset, user name, screen name, identification, follower, committer.**

I. Introduction

The purpose of this paper is to describe a dataset comprised of the Twitter screen names for people and projects affiliated with the Apache Software Foundation. The Apache Software Foundation (ASF) was created in 1999 as a not-for-profit corporation supporting multiple related software projects all released under the Apache open source license. Since that time, the ASF has been home to hundreds of affiliated projects, and has served as a valuable source of software engineering research data (e.g. [1]), as most of its artifacts (email, source code, bug reports) are publicly available for viewing.

One newer form of social media used by some Apache-affiliated projects is Twitter. Twitter is a social media web site where writers, identified by a screen name preceded by an "@" symbol, post short messages (called *tweets*) for others to read. Readers can choose to subscribe to the tweets of others (called "following"), and writers can have their tweets read by ("followed by") others. To assist readers in finding messages that are interesting to them, writers may also tag their tweets with words or phrases preceded by a "#" (hashtag) symbol.

Researchers are studying how software project teams communicate using Twitter (e.g. [2] or [3]). To do this, it is helpful to have a list of all the possible screen names for the projects or people to be studied. With such a list, the researcher can direct the Twitter API to collect the biographical information or entire list of tweets for those users or the follower/followed-by list, which can then be used to build a social network or to mine the textual content of tweets, etc.

This paper therefore assists such research by detailing the construction of a Twitter screen name list for projects and people affiliated with the ASF. Section II describes where the data came from and how the names were collected. Section III is a description of how the data was stored and cleaned, including instructions for accessing and using this dataset. Section IV describes the limitations and challenges of this dataset, as well as some suggestions for future work.

II. Data Collection

The data for this set comes from various pages and text files posted on the ASF web site, as well as from Twitter itself. Some of the semi-structured data was collected through scripts, some of the screen names were discovered using the Twitter API, and a significant number of the screen names were collected or verified using manual searches. This section describes each of these data sources. It is important to know that all of discovered Twitter screen names for this dataset were checked manually to ensure that they referred to the correct Apache-related person. Section III gives more details about this aspect of the curation procedure.

A. Apache project Twitter names: Twitter search

In this dataset, Twitter screen names can refer to Apache people or to Apache projects. To gather screen names for Apache projects, we used the official list of nearly 200 Apache projects [4], then searched Twitter for related screen names. In choosing a screen name on Twitter, most of the projects took the form of @Apache<ProjectName> (for example *Cayenne* becomes @ApacheCayenne) or @<ProjectName> (for example *Hadoop* becomes @hadoop. There were a few projects, such as *Struts* or *Bean Validation* that took non-standard forms such as @TheApacheStruts and @BValTeam.

This process yielded 51 unique Twitter screen names for ASF projects. Not every ASF project uses Twitter to communicate. Even among those projects that did use Twitter, the job of finding screen names was complicated by multiple screen names or project name changes. There are also a number of ASF projects that are not listed on the official web site project listings. (For example @ApacheTomEE and @OpenEJB are related projects. However, neither is listed on the main Apache project page, and these two projects have separate Twitter screen names.) Some projects used Twitter hashtags but did not have a dedicated screen name. For example, Apache Aries states on its apache.org project page that "Our Twitter tag is #apache-aries", but it does not list an actual Twitter screen name for the project. Thus, this project is not listed in our dataset of screen names.

B. Apache people Twitter names: official web site

In addition to screen names for projects, we are also interested in gathering screen names for Apache people. The first source of data for this set was Apache's own web pages, including:

- the committer list [5] and links to the Apache-hosted user FOAF (friend-of-a-friend; an RDF file format) files, and
- the mdtext files on the Apache web site (template data used to automatically generate the Apache web site), and
- the project-by-project lists of contributors with contact details (for example "Who we are" and "team list" pages usually found on each project linked from [4]).

Step one was to write several simple scripts to collect these pages and parse out Twitter screen names using variations of *@Name* and *http://twitter.com/name*. Then, each of these discovered Twitter screen names was checked manually via the Twitter web site. When validated, we added the screen name to the database, along with the "real name" listed on Twitter and the crosslinked Apache "svn ID" data (when available). For example, the Apache Rave and Mahout projects listed the Twitter screen names for all their contributors on their project web pages, as well as the svn ID for each member. This process only yielded only about 25 valid screen names.

C. Apache people Twitter names: participant web sites

The next data source was the list of external FOAF files and web sites linked to by Apache committers [5] who have a profile on the Apache web site. Not every committer has a profile, and of those that do, only about 400 list a web site or blog or externally-hosted FOAF file. We wrote some simple scripts to visit each external site, retrieve the HTML, and parse those files for Twitter screen names. We then checked each screen name manually to ensure that the script had pulled the one for the correct user (since many of the sites listed Twitter screen names for friends or other non-ASF people). Most of the web sites did not list any Twitter names at all. This process yielded approximately 100 valid screen names.

D. Apache people Twitter names: Twitter API

After completing the preceding semi-automated search strategies, there were only about 200 confirmed Twitter screen names in the dataset, including screen names for projects. Yet, from collecting a related dataset of ASF-affiliated people [6], we knew that there were at least 4500 people involved with Apache projects at the contributor level and higher. So, the next step was to use the Twitter API [7] to try to guess screen names for these remaining people. Knowing that we would have to manually confirm each discovered screen name, we decided to focus on finding screen names for the known Apache committers that we *did not* already have in our list, guessing the screen names were as follows:

- The Twitter screen name might be the same as the Apache ID: use the Apache svn IDs discovered and documented in another dataset we had made of Apache people [6];
- The Twitter screen name might be the same as the Apache email: use the left-hand side of the email address listed for that person, also taken from [6];
- The Twitter screen name might be a combination of first and last name: use the list of names (also from [6]), and remove white space.

There were a few significant difficulties with using the Twitter API to yield a list of possible Twitter screen names. First, the Apache ID and email prefixes yielded an enormous number of "positive" hits, however most of these Twitter profiles, upon inspection, were revealed to NOT be the correct person. (Or, more specifically, we were unable to confirm them as Apache affiliates from information given in the profile.)

The next difficulty we ran into was with our patterns for guessing screen names. First, automatic identification of the correct person was complicated by the use of abbreviations, middle initials, and underscores in Twitter screen names. The Apache ID might be *jwhitlock* but the Twitter screen name would be *whitlockjc*. It was impractical from a time standpoint to guess and manually confirm all of these variants.

Finally, there were different character sets in the first and last names, the presence of which leads to flattening of characters (especially vowels) in usernames, and thus a difficulty in finding Twitter screen names. Here is an example of how that happens: one confirmed Apache contributor is listed as "Anne Katherine Petterøe" on Apache, and has the Apache ID 'akpetteroe' (note the absence of the 'ø' character). On the Twitter site there is both a screen name of akpetteroe (without the 'ø') and completely unrelated screen name with a similar real name. Without manual inspection of these two profiles by a domain expert, it would be difficult to discern that the first one is actually an old account for this same person, and that both accounts are actually maintained by the same person as the Apache contributor. (This situation is uncovered easily via a manual search of Twitter using the first name and last name with 'ø' included.)

Our experience in making this dataset was that multiple accounts and shadow accounts and similar names and spelling variants in multiple alphabets are very common. This makes using the 'automated' discovery of screen names via the Twitter API somewhat less-than-ideal.

E. Apache people Twitter names: Twitter search

As a final attempt to collect the Twitter screen names for the remaining users, we decided to use a hybrid approach. Using our Apache people dataset [6], we collected a list of the svn_id, email, web site, and first and last name of the known Apache people. We then removed any users for whom we already had a Twitter account listed (from the procedures described in II.A-II.D above). We then sorted these remaining Apache people in order of how many times they were mentioned as contributors in the Apache people dataset. In other words, we ranked them by how many projects they were on and how many roles they had on those projects. We then performed a manual Twitter search using first and last names in order to find the most active people first.

This process was somewhat time-consuming, but it had the advantage that it was not necessary to try to guess the screen names, and it excluded the users we already had names for.

Using this method, it was possible to uncover Twitter screen names for approximately 200 of the remaining Apache people.

In all, the steps in section II yielded approximately 500 confirmed Twitter screen names for Apache affiliates.

III. Data Storage, Cleaning, and Curation

The data from the sources described above are stored in the publicly-accessible FLOSSmole [8] MySQL database, and delimited text files of the data have been released on the FLOSSmole data downloads page on Google Code [9].

As the data was collected using each of the methods described in part II, each data row was given a timestamp and a datasource_id number. These two numbers help indicate which data source it came from originally and when it was added to the database. The datasource_id allows the data rows to be compared over time, and allows an analyst to include or remove rows that are unwanted or untrusted. (Don't trust the script? Don't use that dataset.)

Because some of the columns from this dataset are related to (but not dependent on) [6], we show the related table *apache_people_projects* in the data model in Figure 1.

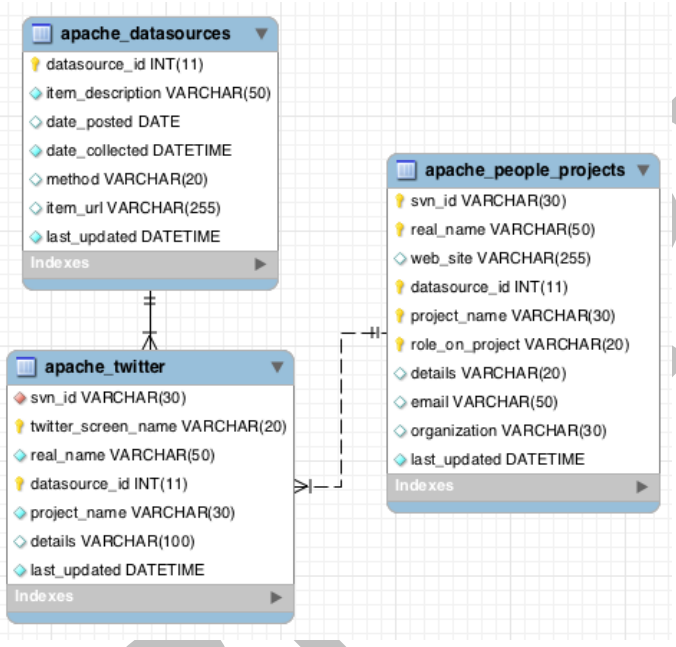

Fig. 1. Apache Twitter Data Model in FLOSSmole Database

The data model for the dataset is straightforward. Figure 1 shows the tables involved:

- The *apache_datasources* table holds a numbered list of all the sources of data in the Apache portion of the FLOSSmole database. The Apache Twitter dataset is made up of approximately four (as of this writing) data sources, each described in section II.
- The *apache_twitter* table lists each Twitter screen name that was found inside each different source of data.
- We connect these to a known Apache svn_id and/or real name when possible via *apache_people_projects*. When the Twitter real name conflicts with the Apache real name (John Smith vs. John K. Smith for example), we use the Twitter version of the name to populate the *real_name* column.

It is possible to answer the following sorts of questions with this dataset:

1. List all Twitter screen names having to do with Apache for people with a last name like 'Smith':

```
SELECT DISTINCT twitter_screen_name
FROM apache_twitter
WHERE real_name LIKE '%Smith%';
```

2. List Twitter screen names for Apache projects (not people):

```
SELECT * FROM apache_twitter
WHERE svn_id IS NULL
AND real_name IS NULL;
```

3. List the Apache details and Twitter screen names for everyone involved on the Apache Cayenne project.

```
SELECT DISTINCT p.svn_id, p.real_name,
    p.role_on_project,
    t.twitter_screen_name
FROM apache_people_projects p
INNER JOIN apache_twitter t
ON p.svn_id = t.svn_id
WHERE p.svn_id IS NOT NULL
AND p.svn_id !=""
AND p.project_name LIKE '%cayenne%';
```

(One caveat: This query assumes that for the people in the apache_people_projects table, the svn_id is known. However, this is not always the case [6], and for some datasource_ids in that table, this field will be empty because the svn_id was not given. The svn_id is also not given for many contributors.)

4. List Twitter accounts and other details for everyone who has ever held the role of Vice President / PMC Chair of any Apache project.

```
SELECT DISTINCT p.svn_id, p.real_name,
p.project_name, p.role_on_project,
t.twitter_screen_name
FROM apache_people_projects p
INNER JOIN apache_twitter t
ON p.svn_id = t.svn_id
WHERE p.role_on_project LIKE 'V%P%'
OR p.role_on_project = 'PMC Chair';
```

IV. Limitations and Challenges of the Dataset

This dataset is unique, and will be helpful for people looking for information about Apache's contributors and their Twitter IDs, but it has some challenges too. This section lists some of the things that users should know about this dataset.

A. Domain Expertise

The effectiveness of the data gathering steps described in II.A-II.D is ultimately dependent upon the expertise of a human being. This domain expert visually inspects each discovered Twitter account (or in the case of step II.D, each search result) and attempts to verify whether it belongs to an Apache-related contributor, committer, or not. Some of the Twitter profiles are very easy to identify, with 'Apache Committer' written in the bio section, for example.

Other Twitter profiles were trickier to confirm. Some of the matches required looking at email domains, web site domains, tags used, the presence of apache-related words or tags, and even the presence of @screen_names of other known Apache people. In some cases, there was a candidate Twitter screen name and real name that perfectly matched a known Apache user, but because there was no evidence that the Twitter user was the same person as the Apache person, the two entities were left unconnected. In cases where it was not clear that the person was the same (based on multiple pieces of evidence) we took a conservative approach: we did NOT include the record.

B. Maintainability

Gathering this data using the methods described above was not particularly fast or easy. A recommended improvement for maintaining this dataset into the future is to attempt only to identify Twitter screen names of new Apache contributors. However, only adding new users is ultimately not really a sufficient or satisfactory condition. Only adding new Apache contributors does not take into account the possibility that Twitter screen names for existing Apache contributors could change or disappear. Yet, we know that updating the whole dataset will be an ongoing maintenance challenge since its validity currently depends on domain expertise.

We hope that someone who finds our dataset useful will develop a better method for collecting these screen names and matching them to Apache contributors. And as this maintenance occurs, we encourage researchers to donate the improved data back to the community. For example, within the FLOSSmole repository, each dataset is given a number ("datasource_id") so researchers can pick and choose which ones to use or discard from their analyses. As new methods are developed to gather Twitter screen names, the new dataset can easily be added alongside this original dataset.

C. Privacy

None of the information gathered in this dataset was from any non-public source. Everything used to construct the Apache Twitter dataset (and the Apache contributors dataset that we got the names from for II.C and II.D) was found on public web sites. The automated collectors did not violate and robots.txt files or terms of service. However, there may be a concern in that two existing sources of public data were joined together in a new way. As explained in III.A ("Domain Expertise") above, we did not include any Twitter account for which there was no evidence that the person was actually an Apache contributor *even if they were named identically* (in the form of Apache project self-identification, Apache-related hashtags, Apache-related links, tweets to or from other confirmed Apache contributors, or the like). The production of "you should follow" lists has some precedent on the popular web ([10] and [11]) and in the literature [12].

D. Missing Data

As with any data collection project, the data can always be bigger, better, and more complete. Some of the missing data from this project includes:

- Private Twitter accounts: these made it difficult to confirm whether the person was actually an Apache contributor, unless it stated such in the public Twitter "bio";
- Missing real names or Apache IDs: for example not all Apache contributors in [6] have an Apache ID. In this case, we only have the name value to match on, so this makes doing SQL joins as shown in Section III example 3 and 4 somewhat more difficult.
- Missing rich data about Twitter users: because the focus was Twitter screen names, we did not use the Twitter API to collect other data about the Twitter users (e.g. tweets, bio, followers, following). This Apache Twitter dataset will probably enable that sort of collection by someone else in the future, but that was not the focus of this work.

V. CONCLUSION

This paper describes a new dataset of Twitter screen names for Apache Software Foundation projects and known committers to those projects. The data was collected through scripts and manual processing and is publicly available for download and for querying. This dataset should be useful to any researchers who are interested in Twitter mining or in following particular people or projects. For example, if the researcher wishes to compare the structure of Twitter networks across projects or contributors, this list can serve as a seed to that activity. The dataset currently consists of Twitter screen names gathered in January 2013, but it can be expanded and tagged to include new names discovered in the future.

REFERENCES

[1] A. Mockus, R. Fielding, J.D. Herbsleb (2002). "Two case studies of open source software development: Apache and Mozilla". *ACM Trans Sw Egr. Meth, 11(3).* pp. 309-346.

[2] A. Java, X. Song, T. Finin, and B. Tseng. "Why we twitter: understanding microblogging usage and communities." In *Proc 9th WebKDD and 1st SNA-KDD 2007 Wkshp. Web mining and social network analysis* (WebKDD/SNA-KDD '07). pp. 56-65.

[3] G. Bougie, J. Starke, M. Storey, and D.M. German. "Towards understanding twitter use in software engineering: preliminary findings, ongoing challenges and future questions." In *Proc. 2nd Int. Wshp. Web 2.0 Sw. Egr.* (Web2SE '11). 2011. pp. 31-36.

[4] Apache Projects. http://projects.apache.org/ indexes/alpha.html

[5] ASF Committers. http://people.apache.org/ committers.html

[6] M. Squire, "Roles on Apache Software Foundation projects". In *10th Working Conf. Mining Software Repositories (MSR2013).* San Francisco, CA, USA. May 18-19 2013. 4 pages.

[7] Twitter API. http://dev.twitter.com

[8] FLOSSmole, http://flossmole.org

[9] Apache Twitter Names 2013-Jan, https://code.google.com/p/flossmole/downloads/detail?name=apacheTwitter2013-Jan.zip

[10] Listorious. http://listorious.com

[11] Twibes. http://twibes.com

[12] J. Hannon, M. Bennett, and B. Smyth. "Recommending twitter users to follow using content and collaborative filtering approaches." *Proc. 4th ACM Conf Recommender Sys.* (RecSys '10), pp.199-206.