Project Roles in the Apache Software Foundation: A Dataset

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

Abstract—**This paper outlines the steps in the creation and maintenance of a new dataset listing leaders of the various projects of the Apache Software Foundation (ASF). Included in this dataset are different levels of committers to the various ASF project code bases, as well as regular and emeritus members of the ASF, and directors and officers of the ASF. The dataset has been donated to the FLOSSmole project under an open source license, and is available for download (https://code.google.com /p/flossmole/downloads/detail?name=apachePeople2013-Jan.zip), or for direct querying via a database client.**

***Index Terms*—open source, free software, Apache, roles, dataset, email, leadership, hierarchy, project organization.**

I. Introduction

The Apache Software Foundation (ASF) is a not-for-profit corporation created as the umbrella organization to support multiple software projects released under the Apache open source license. The original Apache HTTP Server was released in 1995 and the ASF was founded four years later in 1999. In the years since, the ASF has been home to hundreds of affiliated projects.

Because of the success of the ASF and its relatively long standing in the free, libre, and open source software (FLOSS) community, artifacts from the Apache projects are very interesting to researchers who work to understand how FLOSS is made. For example, email is a frequently-studied artifact of the FLOSS development process (see for example [1] and [2]), and ASF projects are some of the most-frequently studied in this category [3].

Email archives from other (non-FLOSS, non-software) domains have also been used with some success to understand leadership structure or power relationships. Specifically, two strong motivators of this project were the work done by Gilbert [4] and by Prabhakaran [5]. These papers use the Enron email corpus [6] to study and predict usage of power language within a corporate structure. A critical addition to this email dataset was the addition of the Enron roles dataset by Shetty and Adibi [7]. So, for a FLOSS researcher to study leadership structure in this way with all or some of the ASF emails would also require knowledge of the roles played by each participant in those email conversations.

Thus, the purpose of the new dataset described in this paper is to assist researchers who are using artifacts of the development process for ASF projects (artifacts could include email, source code, bug reports and the like) to know the roles of each participant mentioned in those artifacts. The dataset created and described in this paper is a timestamped collection of people, projects, and roles on those projects.

The next section of this paper describes where the data came from and how it was collected. Following that is a description of how the data was stored, cleaned, and curated. Next are instructions for interested users of this dataset, as well as some simple usage examples. The final section of the paper 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. Some of the semi-structured data was collected through scripts, while other data needed to be collected and parsed manually.

A. Committers by ID

The first source of data for this set was the list of "ASF committers by ID" [8]. A committer is someone who has signed a Contributor License Agreement (CLA) with Apache and has committed code to one or more Apache projects. ASF maintains a web page listing the user ID used to commit the code into the source code management system (this ID is sometimes called "SVN id" on ASF pages), each user's real name, and the user's affiliated projects. There is also a list of people (real names only) who have signed a CLA but who have not yet committed code to a project. Table I shows some sample data from a "Committers by ID" data source. This process yields about 9000 committer records each time it is run.

TABLE I. "Committers by ID" Sample Data

Item	***Sample Data***
Svn_id	husted
Real_name	Ted Nathan Husted
Project_name	struts
Role_on_project	Committer

B. Committers List

The next data source for this set was the list of committers listed in order by last name [9], with information provided by the committers themselves. This page includes links to additional pages (generated with alphabetical listings, for example list_A.html, list_B.html, etc.) showing the user ID of the person, the real name, any projects the person is affiliated with, associated web pages, and a few additional fields, such as a map and sometimes a PGP key. Table II shows some sample data. This process yields about 1200 records each time it is run.

TABLE II. "COMMITTERS BY NAME" SAMPLE DATA

Item	***Sample Data***
Svn_id	husted
Real_name	Ted Husted
Web_site	http://jroller.com/page/TedHusted
Project_name	Apache Struts
Role_on_project	Committer

Note that in Table I, the project name was not capitalized, and does not include the word "Apache", and the user is identified by first, middle, and last name. However, in the data for Table II, the user is listed by first and last name only, the project name is slightly different, and the web page is listed.

C. Members List

The next source of data for this set was the list of ASF members and emeritus members [10]. Membership in the ASF is by invitation only. The web page implies that members will have contributed to one or more Apache projects, although there are numerous members listed on the page who do not have an affiliated project listed. (And the converse is true as well: there are many more committers who are not actual members of the ASF.) The members page lists the user ID for each member (just called "id" on the page), the name of the member, a web site associated with that member, and any list of projects with which the user is affiliated. For emeritus members, only the id and name are listed. A few emeritus members are missing user IDs, and most do not have a web site listed. Table III shows some sample data. This process yields about 450 records each time it is run.

TABLE III. "MEMBERS OF THE ASF" SAMPLE DATA

Item	***Sample Data***
Svn_id	clr
Real_name	Craig L Russell
Project_name	Apache Software Foundation
Role_on_project	Member

D. Minutes of board meetings

Another source of data for this set is the Apache Board meeting minutes [11]. The ASF board meets monthly online to perform tasks like approving new projects, discontinuing projects, changing project leaders (called Vice Presidents of a project), and recording other reportable administrative tasks such as new committers added or project management committee (PMC) members added. After several failed attempts to use automated entity identification techniques through NLTK [12], we finally just gathered this data through a manual process. The procedure consists of reading the minutes, finding names of people and their associated role and project, and noting what section of the minutes that information came from. This dataset now includes the relevant "people-project-role" information present in all the ASF Board Meeting minutes from 2012. (Prior years Board meeting minutes are available, but were not parsed for this study.) Table IV shows some sample data from the minutes. This process yields between 50-200 records for each set of minutes.

TABLE IV. "BOARD MEETING MINUTES" SAMPLE DATA

Item	***Sample Data***
Svn_id	clr
Real_name	Craig L. Russell
Project_name	Apache Software Foundation
Role_on_project	Secretary
Details	Item 2

Note that details are included for where in the minutes the information was found ("Item 2" or "Attachment M" for instance). Note also the slight spelling difference in the middle initial compared to Table III (here, it is spelled with a period).

E. Individual Project Team Listings

The final source of data for this project is the individual ASF project web pages that list team members. There were 130 ASF projects that had listings of their team members. These details can be located in one of two ways: either by finding a people.mdtext file located in the project SVN tree, or by manually looking at each project page, locating the team pages (called various non-standardized names) and parsing these pages manually for names, emails, user ID, organization, and role on the project. Table V shows some sample data.

TABLE V. "INDIVIDUAL PROJECT TEAM LISTINGS" SAMPLE DATA

Item	***Sample Data***
Svn_id	ajith
Real_name	Ajith Ranabahu
Project_name	Apache Axiom
Role_on_project	Committer
Organization	WSO2
Email	ajith@wso2.com
Web_site	http://www.apache.org/~ajith

This work gathering the contributor information from individual project team pages was fairly tedious since each

ASF project had their web page laid out differently, and there were at least 10 different "formats" for the mdtext pages. This process yields approximately 4400 records each time it is run.

III. Data Storage, Cleaning, and Curation

The data from the sources described above are stored in the publicly-accessible FLOSSmole [13] MySQL database, and delimited text files of the data have been released on the FLOSSmole data downloads page on Google Code [14].

As the data was parsed from the ASF web pages and text files, 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. The most data cleaning steps were performed on the Board Meeting minutes data, as this was the most unstructured and non-uniform of the sources. In the Board Meeting minutes data, there were many rows in which either the username or real name was missing. When it was easy to find the missing data from a similar row, the data was added to fill out the incomplete row. The Board Meeting minutes also had some misspellings which were corrected upon reading, and before data was added to the database (for example, the minutes showed a user called 'wavd' instead of 'wave'). This type of error in the original data was corrected when caught.

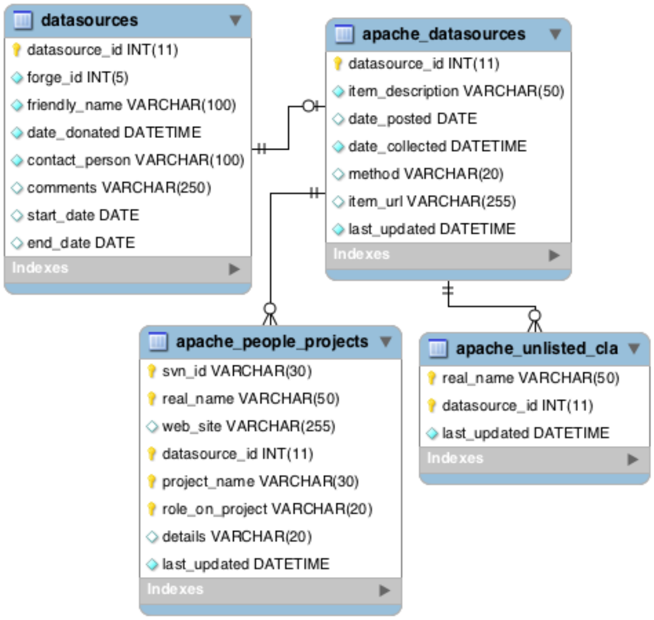

Fig. 1. Apache Roles Data Model in FLOSSmole Database

The data model is straightforward, but is somewhat de-normalized (flat). Figure 1 shows the tables involved in this dataset. The *datasources* table holds a numbered list of all the sources of data in the FLOSSmole database. The Apache dataset is made up of approximately 20 (as of this writing) data sources. The *apache_datasources* table adds a few more columns of information that are specific to the Apache data, for example the name of the script that gathered the data, and the URL for where the data came from. The table called *apache_people_projects* lists each person-role-project combination that was found inside each different source of data. The *details* column lists the part of the document where the information was found, if any. If the person has an associated web site, that was listed. The *apache_unlisted_cla* table is just a simple listing of anyone who was listed as a signed committer but who has no associated project listing (and therefore cannot be listed in the *apache_people_projects* table.)

As shown in the figure, the main entities in this dataset are data sources, people, projects, and roles. Right now the entities are not entirely normalized. (As data continues to be collected for this project, lookup tables of Apache Projects or Users or Roles - for example - will be more correct and complete, at which point normalization can be considered.) The primary key for *apache_people_projects* is a composite key of data source id, user id, project name, and role on that project.

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

1. List everything about the person called Craig L. Russell, including a list of projects and roles. Put these in order by when they happened:
```
SELECT * FROM apache_people_projects app
INNER JOIN apache_datasources ad
ON ad.datasource_id = app.datasource_id
WHERE real_name like 'Craig%Russell'
OR svn_id='clr'
ORDER BY ad.date_posted, ad.date_collected;
```

2. List everyone who has served in the role of "Vice President" for Apache Struts since 2012 and put them in order by date. (Note that we use a wildcard on both ends of the project name since some data sources prefix 'struts' with the word 'Apache' and some other data sources list the PMC for struts like 'struts-pmc'.)
```
SELECT * FROM apache_people_projects app
INNER JOIN apache_datasources ad
ON ad.datasource_id = app.datasource_id
WHERE role_on_project = 'Vice President'
AND project_name LIKE '%struts%';
```

3. List once the names, roles, and contact information for everyone who has worked on Struts at any point (with a named role such as 'Vice President' or 'Committer'), including the PMC.
```
SELECT DISTINCT svn_id,
       real_name, role_on_project
FROM apache_people_projects
WHERE project_name LIKE '%struts%';
```

4. List information for everyone who has ever held the role of Vice President / PMC Chair of any project.
```
    SELECT DISTINCT svn_id, real_name,
project_name, role_on_project
    FROM apache_people_projects
    WHERE role_on_project LIKE 'Vice President'
   or role_on_project = 'PMC Chair';
```

IV. Limitations and Challenges of the Dataset

This dataset is unique and relatively large, and will be helpful for people looking for information about Apache's contributors and roles, but it has some challenges too. This section lists some known issues with this dataset.

A. Maintainability

First, owing to the size and unstructured quality of the data, especially the portion of the dataset that is based on Board Meeting minutes, it will not be terribly easy to keep updated and maintained in the future. Updating the dataset will be an ongoing challenge since it is currently an entirely manual process to enter the Board Meeting minutes data. (The committer and member data sources are scripted, so these are trivial to update.)

The upside is that since FLOSSmole takes data donations from contributors, anyone can assist with reading and entering the data from the minutes. Each donation is given a number ("datasource_id") so researchers can pick and choose whether to trust a particular subset of the data or not.

B. Consistency

There are some consistency issues with the data: Some data sources included middle names or middle initials and there were a number of name misspellings. Some of these were caught and corrected, but many more are left uncorrected. There are also instances of projects changing names (for example the project called RAT changed its name to Apache Creadur upon "graduating" from the project incubator to being an official project of the ASF).

Also, different data sources used different words for different roles (VP versus PMC Chair) and different character sets for people's names (for example sometimes the VP of Apache Camel is spelled 'Christian Mueller' and sometimes it is spelled 'Christian Müller'). Sometimes middle names are included, and sometimes they are not. There are different rules about punctuation for the different datasets on the ASF web site. This results in slightly different spellings in the database.

C. Missing Data

Some of the projects reporting in the Board Meeting minutes referred to roles, but used first names only (no last name and no username, e.g. "We welcome Mary as a committer"), making it hard to determine anything more about Mary. In addition, some of the committer listings on Apache's own web site were giving 404 ("file not found") errors: for example, any committer listing page where the last names started with lowercase letters were missing. This issue was reported to Apache 'infra' team but unfixed as of this writing. The scripts to collect this data will automatically pick up the files once it is corrected, but in the meantime, data is missing.

Another important piece of missing data happens when a person at the Committer level leaves a project. There is no official role for "former member" or the like. Some of the individual project pages reported this, but many other data sources did not. A few projects reported in the minutes on something they called an Emeritus status for (former) project management committee (PMC) members, but the minutes also stated at one point that the Board does not actually recognize Emeritus PMC members, so the projects should quit reporting them.

V. Conclusion

This paper describes a new dataset of people, projects, and roles in the Apache Software Foundation and its family of projects. The data was collected through scripts and through manual processing and is publicly available for download and for querying. This dataset should be useful to any researchers who are interested in knowing about the roles in ASF projects, who has held those roles, and for how long. For example, if the researcher is examining software artifacts such as email or source code, knowing what official role is held by a person may be helpful. It is also possible to look up the person's role on other, related ASF projects, and to find out additional contact details about that person that may not be apparent from the email or source code artifact itself. The dataset currently consists of 2012 data, but can be expanded to include historical data from 1999 and onward, as reflected in the Apache Board Meeting minutes.

References

[1] A.C. MacLean, L.J. Pratt, C.D. Knutson, B. Noble, and E.K. Ringger, "Knowledge homogeneity and specialization in the Apache HTTP Server project," *Open Source Systems: Grounding Research (OSS 2011),* pp. 106-122.

[2] P.C. Rigby and M. Storey, "Understanding broadcast based peer review on open source software projects," *Proc. 33rd Int. Conf. Sw. Eng. (ICSE 2011)*, pp. 541-550. May 2011.

[3] M. Squire, "How the FLOSS Research Community Uses Email Archives," *Int. J. Open Source Sys. Proc*: 4(1). Jan-Mar 2013. *In press*.

[4] E. Gilbert. "Phrases that signal workplace hierarchy," *Proc. ACM 2012 Comp. Supp. Cooperative Work (CSCW '12),* pp. 1037-1046.

[5] V. Prabhakaran, O. Rambow, and M. Diab. "Predicting overt display of power in written dialogs," *Proc. 2012 Conf. N. Amer. Ch. Assoc. Comp. Linguistics: Human Language Technologies* (NAACL HLT '12). pp 518-522.

[6] B. Klimt and Y. Yiming. "Introducing the Enron corpus." *First conference on email and anti-spam (CEAS)*. 2004.

[7] J. Shetty and J. Adibi, "The Enron email dataset database schema and brief statistical report", *Information Sciences Institute Technical Report, U. So. California,* 4 pages.

[8] Apache Committers by ID, http://people.apache.org/committer-index.html

[9] Apache Committers by Name, http://people.apache.org/committers.html

[10] Apache Members List, http://www.apache.org/foundation/members.html

[11] Apache Board Meeting Minutes Index, http://www.apache.org/foundation/records/minutes/

[12] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*. O'Reilly. 2009.

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

[14] Apache People-Project-Roles 2013-Jan, https://code.google.com/p/flossmole/downloads/detail?name=apachePeople2013-Jan.zip