Apache Roles Dataset

Dr. Megan Squire
Elon University, NC
FLOSSmole.org
Goal

Collect a giant list of all the Apache participants, what their role(s) are within Apache, and on what project.
Goal

Collect a giant list of all the Apache participants, what their role(s) are within Apache, and on what project.

Timestamp everything so when I collect this data again, I can see changes.
Why Apache?

- Big (200+ subprojects)
- Transparent
- Popular to study
What I'm looking for

Alice is a committer on CouchDB.

- **Person**: Alice
- **Role**: committer
- **Project**: CouchDB
Complete record

On May 19, 2013 I learned from the Minutes of the April Board Meeting that Alice is a committer on CouchDB.
Why is this useful?

Discussions that require knowledge of roles:

- Power relations
- Leadership structure
- Group dynamics
- Promotion and retention
- Decisionmaking
- Workload, responsibilities
Data sources

- List of Committers (CLA signers)
- List of Signers (not yet committed)
- List of Committers (self-submitted)
- List of ASF members
- Minutes of Board Meetings
- Individual Project Team Listings
## List of CLA-signed committers

<http://people.apache.org/committer-index.html>

<table>
<thead>
<tr>
<th>SVN id</th>
<th>Name</th>
<th>SVN Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>a_horuzhenko</td>
<td>Artyom Horuzhenko</td>
<td>openmeetings</td>
</tr>
<tr>
<td>aadachik</td>
<td>Andrus Adamchik</td>
<td>apsite, cayenne, cayenne-pmc, click, incubator, incubator-pmc, member, pmc-chairs, wave</td>
</tr>
<tr>
<td>aadomowski</td>
<td>Aleksander Adamowski</td>
<td>directory</td>
</tr>
<tr>
<td>aadrian</td>
<td>Adrian Antal</td>
<td>click</td>
</tr>
<tr>
<td>aaf</td>
<td>Alexei Fedotov</td>
<td>incubator, openmeetings, openmeetings-pmc</td>
</tr>
<tr>
<td>aanderson</td>
<td>Aaron Anderson</td>
<td>ode</td>
</tr>
<tr>
<td>aaron</td>
<td><strong>Aaron Bannert</strong></td>
<td>apr, apr-pmc, apsite, httpd, httpd-pmc, incubator, incubator-pmc, infrastructure, member</td>
</tr>
<tr>
<td>ab</td>
<td>Andrzej Bialecki</td>
<td>apsite, gora, gora-pmc, incubator, incubator-pmc, lucene, lucene-orp, lucene-pmc, member, nutch, nutch-pmc</td>
</tr>
</tbody>
</table>
List of CLA Signers, not yet committers

### Unlisted CLAs

Persons with signed CLAs but are not committers:

<table>
<thead>
<tr>
<th>Hari A V</th>
<th>Software AG</th>
<th>Abdullah Ali AL-Dokkan</th>
<th>Cristina Lucia Abad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steve Abatangle</td>
<td>Reda Abdi</td>
<td>solid</td>
<td>Walied Amer Abdullah</td>
</tr>
<tr>
<td>Jalud Abdulmenan</td>
<td>Abhilash</td>
<td>Bryce Acer</td>
<td>Jin Adachi</td>
</tr>
<tr>
<td>Henry Adams</td>
<td>David Adcox</td>
<td>Rolf Ade</td>
<td>Igor Afanasyev</td>
</tr>
<tr>
<td>Neeti Agarwal</td>
<td>Vinay Kumar Agarwal</td>
<td>Kapish Aggarwal</td>
<td>Tulika Agrawal</td>
</tr>
<tr>
<td>Hasan Ahmad</td>
<td>Onur Aktas</td>
<td>Daniel Albers</td>
<td>Nick Alcock</td>
</tr>
<tr>
<td>Rail Aliev</td>
<td>Patrick Allaert</td>
<td>Alan Allegret</td>
<td>Jon Allen</td>
</tr>
<tr>
<td>Miguel Almeida</td>
<td>Luis Alves</td>
<td>Guido Amarilla</td>
<td>Louis Amodeo</td>
</tr>
</tbody>
</table>

http://people.apache.org/committer-index.html#unlistedclas
<table>
<thead>
<tr>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shinichiro Abe</td>
</tr>
<tr>
<td>Erik Abele</td>
</tr>
<tr>
<td>Saminda Abeyruwan</td>
</tr>
<tr>
<td>Andrus Adamchik</td>
</tr>
<tr>
<td>Alessandro Adamou</td>
</tr>
<tr>
<td>Glenn Adams</td>
</tr>
<tr>
<td>Rahul Akolkar</td>
</tr>
<tr>
<td>Jim Ancona</td>
</tr>
</tbody>
</table>
Committers by ID (self-submitted)

Stefan Bodewig

Projects:
- Apache Ant
- Apache Commons
- Apache Gump
- Apache Jakarta
- XMLUnit
- Apache Logging
- Apache Lucene.NET

PGP Keys:
- ID: 5F6B8B72 Fingerprint: CE80 75A2 5154 7BEE 249B C151 A211 5AE1 5F6B 8B72

Weblogs
- Stefan Bodewig's Weblog - Mixed Content
- Stefan Bodewig's Weblog - Gemischtwarenladen

Bernd Bohmann

Projects:
- Apache MyFaces

PGP Keys:
- ID: C4FC9A65 Fingerprint: C4D5 1950 987B E71C B9C5 1E35 ABD6 32EC C4FC 9A65

http://people.apache.org/list_B.html
# Members of the ASF

## Members of The Apache Software Foundation

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>erikabele</td>
<td>Erik Abele</td>
<td>HttpComponents, httpd, httpd-docs</td>
</tr>
<tr>
<td>samisa</td>
<td>Samisa Abeysinghe</td>
<td>Axis2/C, Axis C++, Axis2</td>
</tr>
<tr>
<td>aadamchik</td>
<td>Andrei Adamchik</td>
<td></td>
</tr>
<tr>
<td>rahul</td>
<td>Rahul Akolkar</td>
<td></td>
</tr>
<tr>
<td>gregames</td>
<td>Greg Ames</td>
<td>httpd, APR</td>
</tr>
<tr>
<td>jta</td>
<td>Jean T. Anderson</td>
<td>Derby</td>
</tr>
<tr>
<td>cezar</td>
<td>Cezar Cristian Andrei</td>
<td>XMLBeans, VXQuery</td>
</tr>
<tr>
<td>danny</td>
<td>Danny Angus</td>
<td>James, Jakarta</td>
</tr>
<tr>
<td>assaf</td>
<td>Assaf Arkin</td>
<td></td>
</tr>
</tbody>
</table>

Board Meeting Minutes

RESOLVED, that the persons listed immediately below be and hereby are appointed to serve as the initial members of the Apache CloudStack Project:

* Joe Brockmeier <jzb@apache.org>
* Will Chan <willchan@apache.org>
* Chip Childers <chipchilders@apache.org>
* Mohammad Nour El-Din <mnour@apache.org>
* Sebastien Goasguen <sebgoa@apache.org>
* Matt Richard Hogstrom <hogstrom@apache.org>
* Wido den Hollander <widodh@apache.org>
* Alex Huang <ahuang@apache.org>
* Jim Jagielski <jim@apache.org>
* Alex Karasulu <akarasulu@apache.org>
* John Kinsella <jlk@apache.org>
* Kevin Kluge <kluge@apache.org>
* Olivier Lam <olamy@apache.org>
* David Nalley <ke4qqq@apache.org>
* Brett Porter <brett@apache.org>
* Noah Slater <nslater@apache.org>
* Marcus Sorensen <mlsorensen@apache.org>
* Edison Su <edison@apache.org>
* Hugo Trippaers <hugo@apache.org>
* Chiradeep Vittal <chiradeep@apache.org>

NOW, THEREFORE, BE IT FURTHER RESOLVED, that Chip Childers be appointed to the office of Vice President, Apache CloudStack, to serve in accordance with and subject to the direction of the Board of Directors and the Bylaws of the Foundation until death, resignation, retirement, removal or disqualification, or until a successor is appointed; and be it further
Attachment V: Report from the Apache Karaf Project [Jean-Baptiste Onofré]

Apache Karaf provides higher level features and services specifically designed for creating OSGi-based servers.

Community

Indicators during the last 3 months:
- Download: 31,490 downloads (all distributions, all branches) between December, 1 and February, 28.
- User mailing list: 557 messages between December, 1 and February, 28.
- Dev mailing list: 382 messages between December, 1 and February, 28.
- Commits: 621 commits between December, 1 and February, 28.

We are pleased to count two new committers elected in February:
- Filippo Balicchia
- Christoph Grtischenberger

location within the data source

VP/PMC chair

"committers"
Individual Project "Team" Pages

### APACHE ACCUMULO CONTRIBUTORS

<table>
<thead>
<tr>
<th>username</th>
<th>name</th>
<th>organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>acordova</td>
<td>Aaron Cordova</td>
<td></td>
</tr>
<tr>
<td>afuchs</td>
<td>Adam Fuchs</td>
<td></td>
</tr>
<tr>
<td>bimargulies</td>
<td>Benson Margulies</td>
<td>Hortonworks</td>
</tr>
<tr>
<td>billie</td>
<td>Billie Rinaldi</td>
<td>Praxis Engineering</td>
</tr>
<tr>
<td>brianloss</td>
<td>Brian Loss</td>
<td></td>
</tr>
</tbody>
</table>

**PROJECT**
- Home
- Downloads
- Features
- License

**COMMUNITY**
- Get Involved
- Mailing Lists
- **People**

**DEVELOPMENT**
- Source & Guide
- Issues
- Builds

**DOCUMENTATION**
- Manual 1.4
- Javadoc 1.4
- Examples 1.4
- Screenshots

**company and URL**

---

*Accumulo*
Team Pages, another version

Committers

- **Damien Katz, damien@apache.org**
  Original developer and remains the project lead.

- **Jan Lehnardt, jan@apache.org**
  Worked on the original UNIX port and now works on all consultant and gives presentations around the world.

- **Noah Slater, nslater@apache.org**
  Developed and maintains the Autotools build system and maintains a number of related packages.
Team Pages, yet another version

## Members

The following is a list of developers with commit privileges that have directly contributed to the project in one way or another.

<table>
<thead>
<tr>
<th>Image</th>
<th>Id</th>
<th>Name</th>
<th>Email</th>
<th>Organization</th>
<th>Roles</th>
<th>Time Zone</th>
<th>Actual Time (GMT)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="olamy.png" alt="Image" /></td>
<td>olamy</td>
<td>Olivier Lamy</td>
<td><a href="mailto:olamy@apache.org">olamy@apache.org</a></td>
<td></td>
<td>PMC Chair</td>
<td>Australia/Melbourne</td>
<td>Wed May 15 2013 05:14:40 GMT-0400 (EDT)</td>
</tr>
<tr>
<td><img src="aheritage.png" alt="Image" /></td>
<td>aheritier</td>
<td>Arnaud Héritier</td>
<td><a href="mailto:aheritier@apache.org">aheritier@apache.org</a></td>
<td></td>
<td>PMC Member</td>
<td>+1</td>
<td>Tue May 14 2013 20:14:40 GMT-0400 (EDT)</td>
</tr>
<tr>
<td><img src="baerrach.png" alt="Image" /></td>
<td>baerrach</td>
<td>Barrie Treloar</td>
<td><a href="mailto:baerrach@apache.org">baerrach@apache.org</a></td>
<td></td>
<td>PMC Member</td>
<td>Australia/Adelaide</td>
<td>Wed May 15 2013 04:14:40 GMT-0400 (EDT)</td>
</tr>
<tr>
<td><img src="bimargulies.png" alt="Image" /></td>
<td>bimargulies</td>
<td>Benson Margulies</td>
<td><a href="mailto:bimargulies@apache.org">bimargulies@apache.org</a></td>
<td></td>
<td>PMC Member</td>
<td>America/New_York</td>
<td>Tue May 14 2013 14:14:40 GMT-0400 (EDT)</td>
</tr>
<tr>
<td><img src="brett.png" alt="Image" /></td>
<td>brett</td>
<td>Brett Porter</td>
<td><a href="mailto:brett@apache.org">brett@apache.org</a></td>
<td>ASF</td>
<td>PMC Member</td>
<td>+10</td>
<td>Wed May 15 2013 05:14:40 GMT-0400 (EDT)</td>
</tr>
</tbody>
</table>
End result

everything on FLOSSmole:
-- in MySQL, and
-- at our Google Code site
## Sample data

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Key</th>
<th>Nullable?</th>
<th>Sample Row</th>
</tr>
</thead>
<tbody>
<tr>
<td>svn_id</td>
<td>PK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>real_name</td>
<td>PK</td>
<td></td>
<td>Luis Bernardo</td>
</tr>
<tr>
<td>web_site</td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>datasource_id</td>
<td>PK, FK</td>
<td></td>
<td>375</td>
</tr>
<tr>
<td>project_name</td>
<td>PK</td>
<td></td>
<td>Apache XML Graphics</td>
</tr>
<tr>
<td>role_on_project</td>
<td>PK</td>
<td></td>
<td>Committer</td>
</tr>
<tr>
<td>details</td>
<td></td>
<td>Yes</td>
<td>Attachment BC</td>
</tr>
<tr>
<td>email</td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>organization</td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>last_updated</td>
<td></td>
<td></td>
<td>2013-04-14 20:25:08</td>
</tr>
</tbody>
</table>
Dig in...

SELECT *
FROM `apache_people_projects`
WHERE real_name LIKE "Alice%"
AND role_on_project = "PMC Chair" OR "Vice President";

more inspiring examples in the paper...
Got a better idea?

Make a dataset and send it to FLOSSmole.
Apache Twitter Dataset

Dr. Megan Squire
Elon University, NC
FLOSSmole.org
Goal

Collect a giant list of Twitter handles for Apache-affiliated people and projects.
Goal

Collect a giant list of Twitter handles for Apache-affiliated people and projects.

Timestamp this so I can see new entries.
Why is Twitter useful?

- Just another interesting dev artifact
- Useful for social mining
- Everybody's doing it
Data sources

Directly from Apache
--project/team web pages
--listed personal blogs
Data sources

Directly from Apache
--project/team web pages
--listed personal blogs

Indirectly
--Search on Twitter using items from my own Apache Roles dataset
   ○ names
   ○ emails
   ○ svn
   ○ project names
From team pages

<table>
<thead>
<tr>
<th>Name</th>
<th>Mail</th>
<th>Webpage</th>
<th>PMC</th>
<th>A short summary of yourself</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isabel Drost</td>
<td>isabel@...</td>
<td><a href="#">Homepage</a></td>
<td>Yes</td>
<td>Passion for free software (development, but to some extend also the political and economic implications), interested in agile development and project management, lives in Germany. Follow me on Twitter</td>
</tr>
<tr>
<td>Ted Dunning</td>
<td>tdunning@...</td>
<td><a href="#">MapR Technologies</a></td>
<td>Yes</td>
<td>@MaineC</td>
</tr>
</tbody>
</table>

no standard way of writing this on any team page; lots of unstructured data
From individual blogs

<table>
<thead>
<tr>
<th>Name</th>
<th>Mail</th>
<th>Webpage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isabel Drost</td>
<td>isabel@...</td>
<td>Homepage, Blog</td>
</tr>
<tr>
<td>Ted Dunning</td>
<td>tdunning@...</td>
<td>MapR Technologies</td>
</tr>
<tr>
<td>Jeff Eastman</td>
<td>jeastman@...</td>
<td>Windward Solutions</td>
</tr>
<tr>
<td>Drew Farris</td>
<td>drew@...</td>
<td></td>
</tr>
<tr>
<td>Grant Ingersoll</td>
<td>gsingers@...</td>
<td>Homepage</td>
</tr>
<tr>
<td>Sean Owen</td>
<td>srozen@...</td>
<td>Myrrix Ltd</td>
</tr>
<tr>
<td>Robin Anil</td>
<td>robinanil@...</td>
<td>Homepage</td>
</tr>
</tbody>
</table>

write programs to follow the link, try to discern the Twitter name, if any
Using Twitter itself: API

```python
def get_info_by_id(t, ids):
    id_to_info = {}
    while len(ids) > 0:
        # Process 100 ids at a time...
        ids_str = ','.join([str(_id) for _id in ids[:100]])
        ids = ids[100:]
        response = make_twitter_request(t,
                                         getattr(getattr(t, "users"), "lookup"),
                                         user_id=ids_str)
        if response is None:
            break
        if type(response) is dict:  # Handle Twitter API quirk
            response = [response]
        for user_info in response:
            id_to_info[user_info["id"]] = user_info
    return id_to_info
```

Let program try to guess Twitter handle using known Apache data
Using Twitter itself: manual

Curation!

Always confirm that this is really an Apache person.
End Result

about 500 of these

everything on FLOSSmole:
-- in MySQL, and
-- at our Google Code site
## Sample data

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Key</th>
<th>Nullable?</th>
<th>Sample Row 1</th>
<th>Sample Row 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>svn_id</td>
<td>FK</td>
<td>Yes</td>
<td>claudio</td>
<td></td>
</tr>
<tr>
<td>twitter_screen_name</td>
<td>PK</td>
<td></td>
<td>claudiomartella</td>
<td>ApacheShiro</td>
</tr>
<tr>
<td>real_name</td>
<td></td>
<td>Yes</td>
<td>Claudio Martella</td>
<td></td>
</tr>
<tr>
<td>datasource_id</td>
<td>PK, FK</td>
<td></td>
<td>372</td>
<td>370</td>
</tr>
<tr>
<td>project_name</td>
<td></td>
<td>Yes</td>
<td></td>
<td>Apache Shiro</td>
</tr>
<tr>
<td>details</td>
<td></td>
<td>Yes</td>
<td></td>
<td><a href="http://shiro.apache.org/">http://shiro.apache.org/</a></td>
</tr>
<tr>
<td>last_updated</td>
<td></td>
<td></td>
<td>2013-03-27 00:00:00</td>
<td>2013-01-24 14:01:56</td>
</tr>
</tbody>
</table>
How to help

- Anyone can contribute data to FLOSSmole
- We'll help you store it (even if it's very large)
- We'll get you a DOI for it