Cultural Analysis Paper Using Kibble

The complete paper is included as an attachment Cultural Analysis Paper Using Kibble

Below is a short summary

Introduction

This paper will focus on culture and how it has evolved at the Apache Software Foundation (ASF). It will review existing literature and research to find the main cultural elements that comprise ASF culture and using tools and indicators, show whether these cultural elements can be successfully transmitted.

The ASF was founded in 1999 on a single open source project, called the Apache HTTP Server project. The values, behaviour, knowledge and governance model that developed as part of the creation of the Apache HTTP Server project is the source of the ASF culture that is called “The Apache Way”.

The Apache Way is dependent on one central tenet – meritocracy, and this is embedded into all layers of the ASF from the formal governance model and the election of directors and members all the way through to the projects and the recognition of individual contributions.

Today the ASF is made up of over 350 projects and software initiatives and it is claimed that each of these projects demonstrate and accept the Apache Way as their cultural model. This paper will attempt to define and create a cultural baseline for the ASF, and then test the baseline against other ASF projects to see if they have the same or similar profile.

To understand the ASF culture, we need to look closely at the origins of the first ever Apache project - Apache HTTP Server.

Methodology

This paper will analyse the main elements of the ASF culture (“the Apache Way”) and attempt to provide evidence to show the extent that this culture can be successfully transmitted to new projects. It will test and explore the following hypothesis:

- As the original source of ASF culture, data from the Apache HTTP Server project can be extracted and used to create a cultural model
- This cultural model can be used as a baseline to compare new ASF projects to see if they exhibit a similar cultural profile or not

Based on the previous assumption, this paper will focus on the following:

- Define the main concepts and values of the ASF cultural model using publicly available information
- Mine the public data available from the ASF code repositories and mailing lists for the Apache HTTP Server project to create a cultural baseline
- Mine the public data available from the ASF code repositories and mailing lists for a series of Apache projects that were created after the Apache HTTP Server project to look for data indicators that validate or disprove the demonstration of elements found in the Apache HTTP cultural baseline

Data Source

All ASF projects have publicly archived mailing lists. The Apache mailing list are a permanent searchable archive that are publicly available. They are a legacy communication medium from the creation of the ASF that now forms an integral part of any ASF project. It is the heart of a project and a place where people interact, communicate, collaborate, argue, agree and disagree. This means that it is an appropriate place to mine data to look for cultural indicators.

Research Tools

The research will be carried out using the following tools, formulas and indicators.

Apache Kibble

Apache Kibble is a suite of tools for collecting, aggregating and visualising data and activity in software projects.

The following Kibble indicators will be used:

Pony Factor

The Pony Factor (PF) measures the diversity of a project based on the contributions from individual contributors. It can be defined as:

- “The lowest number of contributors whose total contribution makes up the majority”

of whatever is being measured (e.g. lines of code written, number of messages sent etc.).

A higher Pony Factor means that a project has a good tolerance for continuing to survive if one or more of the core contributors leaves.
NOTE: Pony Factor includes all contributions from contributors whether they are still active or not.

**Augmented Pony Factor**

The Augmented Pony Factor (APF) is an adjustment to the standard Pony Factor calculation where contributions from contributors that are no longer active are not included.

**NOTE:** The Augmented Pony Factor will not be used as part of the assessment but a description of it has been included here for completeness.

**Meta Pony Factor**

The Meta Pony Factor calculation is a work in progress. It attempts measure the affiliation of a contributor based on the email address linked to the contribution. If developed further then this could help identify distinct organisations that are contributing.

**Sentient Analysis**

Pang and Lee (2008) defined “Sentient analysis” is a tool that is

> "a kind of text mining, which is used to predict human mind, specifically the emotional state of a person by extracting specific emotional expressions from the text”

This means that it can be used as an indicator to gauge people’s opinions and reactions to certain ideas. Data is collected in the form of text and an algorithm is used to identify keywords associated with an emotion. Any communication can be linked to several emotions so weightings are used to highlight the strength of the sentiment.

**Key Phrase Extraction**

Key Phrase Extraction (KPE) is a method where key phrase or words are extracted that can summarise the main ideas or themes of a document. It has been successfully used for indexing journals and online content but in this paper it will be used to extract any text that could indicate cultural ideas or language.

**Contributor Retention**

Contributor retention is related to how successful a project is at attracting and retaining contributor and can be broken down into the following areas:

- **Active Contributors:**
  - How many contributors are active within the project
  - Are contributors regular and remain active over a longer timespan

- **Retained Contributors:**
  - How long a contributor has been contributing
  - The longer a contributor has been retained the more successful a project is at retaining them

- **Contributors that have Left**
  - How many contributors are leaving

- **Past Contributors that have Returned**
  - How many contributors have contributed in the past and have returned to rejoin the community

**Data Selection**

Data to create the cultural baseline will be extracted from the Apache HTTP Server project and the following data will be used:

- Apache HTTP Server project Mailing List Archives 1996 – Current
- Apache HTTP Server project Code Repositories 1996 – Current

The following 15 projects have been selected for comparison because:
They have different varying years of being an ASF project ranging from 1 to 11 years
They have data available in Apache Kibble
None of them existed at the time that HTTP Server project was created
The culture they exhibit would have been created after ASF was established

List of 15 Apache Top Level Projects (TLPs) for Cultural Comparison

The projects can be broken down into two sub groups as follows:

<table>
<thead>
<tr>
<th>Projects 5 Years and Less as TLP</th>
<th>Projects Over 5 Years as TLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache Beam</td>
<td>Apache Jena</td>
</tr>
<tr>
<td>Apache Fineract</td>
<td>Apache OFBiz</td>
</tr>
<tr>
<td>Apache Ignite</td>
<td>Apache Pivot</td>
</tr>
<tr>
<td>Apache Kudu</td>
<td>Apache Sling</td>
</tr>
<tr>
<td>Apache Netbeans</td>
<td>Apache Stanbol</td>
</tr>
<tr>
<td>Apache Phoenix</td>
<td>Apache Subversion</td>
</tr>
<tr>
<td>Apache Clerezza</td>
<td>Apache Traffic Server</td>
</tr>
<tr>
<td>Apache Cloudstack</td>
<td></td>
</tr>
</tbody>
</table>

- Source data from project mailing lists, source repositories and issue trackers for these projects was be loaded into two Kibble views
- Data for each Apache Kibble indicator was reviewed against the Apache HTTP Server project baseline.

Conclusions

This paper focussed on examining the culture that has evolved at the Apache Software Foundation (ASF) and investigating the extent to which their values and culture can be effectively transmitted to new projects. It used a set of tools and indicators to create a cultural baseline based on the values and behaviours shown by the first ever ASF project, Httpd. A set of 8 indicators was used to create the baseline by mining the data publicly available from the ASF project archives.

For the Apache HTTP Server baseline, the following indicators were used to capture, highlight and measure a potential range of cultural elements:

- Pony Factors: Diversity of the community, confirmation that merit is being rewarded, indication of community growth, retention of contributors
- Sentient Analysis: Dominant emotions being displayed in community interactions, communication style, overall mood of the communication (negative, positive or neutral) over time
- Key Phrase Analysis: Identifying the most common important phrases and words being used, indication of collaboration, identifies the use of unique cultural language

Fifteen ASF projects were selected and divided into two groups for comparison against the Apache HTTP Server cultural baseline model. The groups were as follows:

- Projects that have been ASF Top Level Projects (TLPs) for Over 5 Years
- Project that have been ASF Top Level Projects (TLPs) for 5 Years or Less

The results of the comparison against the Apache HTTP Server baseline is shown below:

Pony Factors

The older projects appeared to follow the Apache HTTP Server baseline, showing that merit is being rewarded (i.e. no difference between codebase authorship and committership). These projects also showed a significant increase in Pony Factor meaning increased diversity of contributors and also community growth.

The younger projects did not appear to follow the Apache HTTP Server baseline and had a difference between codebase authorship and committership showing that they are perhaps not as frequent in recognising merit based on contributor activity. The increase in Pony Factor shows that they are also increasing the diversity of their contributors but not at the same rate as the older projects.
**Contributor Retention**

For the older projects, over 45% of their total contributors are new (i.e. have been contributing for less than a year) which means these projects are very successful at attracting new people. Nearly 40% of their contributors have 2 – 5 years experience meaning they have a good flow of people coming into the community and staying. Contributor retention for these older projects is good and very different to, and a lot higher than the Apache HTTP Server baseline.

For the younger projects, over 50% of their contributors are new and have been contributing for less than a year which means these projects are also very successful at attracting people. Over 45% of contributors have 2 - 5 years experience meaning they have a good flow of people coming into the community. Contributor retention is extremely high (almost triple the number of the older projects) which is significantly different to the Apache HTTP Server baseline.

**Sentient Analysis**

Both older and younger projects have similar profiles to the Apache HTTP Server baseline. The communication style was consistently positive and although negativity was listed in the top 5 for all groups, over time it did not affect the overall communication mood.

**Key Phrase Analysis**

Once again both older and younger projects have similar profiles to the Apache HTTP Server baseline. They all showed elements of standard everyday communication and technical interactions. It was interesting to see that they also showed indications of cultural expression of ASF values, such as openness, collaboration and community.

The most significant cultural element which appeared in the baseline as well as both older and younger projects was the "+1" indicator. This is unique to the ASF as an indicator of consensus.

**Summary**

The results show that although Apache HTTP Server was a very successful project during the past, it is currently going though a period of potential decline in attracting new contributors. Even so, the culture that created the Apache HTTP Server project is still very strong and is being transmitted to other ASF projects. The values, communication and expected behaviours are reinforced in the community via their mailing lists.

It appears that as ASF projects age they become more confident and proficient in demonstrating key aspects of the culture and recognising merit. Older projects appear to be experienced at retaining a good mix and flow of contributors, while the younger projects are the most successful at attracting new contributors.

This paper highlights that values and culture can be transmitted and the extent to which the transmission occurs is dependant on how long a project is exposed to the culture. The longer the exposure, the more naturally the culture appears to manifest itself. Perhaps the focus of certain cultural elements can be interpreted and adopted more easily than others, and this could be the areas where the younger projects are more successful than the older projects. Ongoing exchange between older and younger projects could provide good ways to create cultural balance in both.

Culture is always something that is evolving so it would be useful to be able to build on the results of this paper with further research in the following areas:

- Analysis of non ASF related open source projects and comparison with an ASF projects to confirm if ASF values and behaviour are cultural specific / unique
- Analysis of an ASF incubating project that has a completely different cultural profile to see if cultural changes occur as part of the project evolution to graduation to Top Level Project
- Investigation of ASF projects that have gone into decline to see if the decline could have been predicted using any of the indicators used in this paper
- Interviewing or surveying ASF projects about their interpretation of the Apache Way culture
- Interviewing people (e.g. Directors, Board Members, Officers) involved with managing open source governance to collect their feedback about the values and culture that they expected or want to create and do a comparison of the results.