

# Improving community awareness in software forges by semantical aggregation of tools feeds

Quang Vu DANG  
Christian BAC  
Olivier BERGER

Institut TELECOM, SudParis

9, rue Charles Fourier, 91011 Evry Cedex, France

{quang\_vu.dang; christian.bac; olivier.berger}  
@it-sudparis.eu

Xuan Sang DAO

Institut de la Francophonie pour l'Informatique

42, Ta Quang Buu, Hai Ba Trung, Hanoi, Vietnam

dxsang@ifi.edu.vn

## ABSTRACT

It is rather difficult to visualize what can be the contribution of a member in a project, especially when the project uses multiple tools to produce its results. This is the case for collaborative development of FLOSS software, that use Wiki, bug tracker, mailing lists and source code management tools. This paper presents an approach to data collection by using aggregation of feeds published by the different tools in software forges. To allow the aggregation, collected data is semantically reformatted into semantic Web standards: RDF, DC, DOAP, and FOAF. Resulting data is processed and displayed in a supervision module that has been integrated into the PicoForge platform. This module is able to draw a live graph of the social community out of the different sources of data, and export semantic feeds for other uses.

## Keywords

free and open source software development, public data, semantic Web, social network analysis, community of practice, social filtering, RDF.

## 1.INTRODUCTION

Free libre and open source software projects (FLOSS) often use development platforms called software forges (such as SourceForge, Savannah, Gforge, Trac, PicoForge...). A forge helps them organize their community, provides collaborative tools to the members (such as Source versioning, Mailing list, Wiki, Bug tracker, forum ...). In order to help analysis by researchers on FLOSS there are many tools that retrieve information about FLOSS development. These tools analyze the data stored by the collaborative tools such as CVS/SVN log, database of bug, mail archive... To facilitate the mining, data is collected from forges, anonymised and then processed. This allows only differed studies on the projects and on independent data sources. It also doesn't give a real time vision to the project members. Moreover there are FLOSS projects which are developed on multiple forges so one needs to integrate project data from multiple sources [5].

In this paper we propose an approach to data collection in FLOSS development using aggregation of feeds provided by tools in forges to better monitor activities. This approach also allows to collect data of multiple projects across multiple forges and across multiple communities. The fresh information in feeds can help members to have an accurate vision of their project's current state. We plan to apply metrics on the resulting data to help understand the quality of the community. In our longer term plan we also want to find relationships is any, between, the quality of the produced software and the liveliness of the community.

In section 2 we recap some research initiatives and their tools focusing on public data about FLOSS, as well as the use of semantic Web standards for representation of metadata. Section 3 describes our approach and methodology. Section 4 presents a case study using our approach to implement a supervision tool in the PicoForge forge.

## 2.BACKGROUND

### 2.1 Existing research initiative and tools

In order to provide data to researchers interested in FLOSS projects, there are many research tools that retrieve and analyse information about FLOSS development.

The FLOSSmole<sup>1</sup> project provides public data about FLOSS development for academic research. It includes data and analysis from SourceForge, Freshmeat, RubyForge, ObjectWeb... [5]. The FLOSSMetrics<sup>2</sup> project aims at constructing, publishing and analyzing a large scale database with information and metrics about libre software development coming from several thousands of software projects, using existing methodologies, and tools already developed [4]. The SQO-OSS<sup>3</sup> project aims at providing a platform with a pluggable architecture for software development organizations to observe the OSS quality by using novel techniques and algorithms in data mining and metric analysis of source code[2].

There are many tools which were used in the above projects to retrieve the information from libre software projects posted in forges: CVSanaly, MailingListStars, pyTernity... Each tool has a proper data schema, and it seems difficult to integrate information across tools or to share information between communities.

### 2.2 Useful Semantic Web standards

Semantic Web standards can help to annotate, organize or integrate the information on projects, actors and their production. This can help finding and sharing public data on FLOSS project as was proposed in[7].

RSS generally refers to an XML dialect used in the syndication of Web content. This standard is usually used to get updates information whose nature changes frequently, typically in FLOSS this can be lists of news, new or changed items of wiki, e-mails, bugs or "commits" to source code. RSS 1.0<sup>4</sup> (aka RSS/RDF) is formulated using the RDF (Resource Description Framework)

<sup>1</sup> <http://ossmole.sourceforge.net/>

<sup>2</sup> <http://www.flossmetrics.org/>

<sup>3</sup> <http://www.sqo-oss.eu/>

<sup>4</sup> <http://Web.resource.org/rss/1.0/>

standard, and may be consumed either as a naive XML format or interpreted as a labelled graph model. A channel has a basic set of properties (link, title, description) and is associated with an RDF Sequence of items. Each item itself has a link, title, description and optional attributes such as Dublin Core<sup>5</sup> elements(**dc:creator**, **dc:contributor** ...).

FOAF<sup>6</sup> (Friend Of A Friend) is an XML/RDF schema for describing people and the relationships between them and the things they create and do. FOAF is also used to draw the social network of community of practice by using **foaf:know** attribute.

Each software development project may be described by using the DOAP<sup>7</sup> (Description Of A Project) schema that is an XML/RDF vocabulary to describe open source projects. It is able to internationalize description of a software project and its associated resources, including participants and Web resources.

Through the use of RDF, a single RSS feed can contain semantic information combining different vocabularies (for instance, FOAF + DOAP). For example, SourceKibitzer generates DOAP/FOAF metadata from hosted projects and members profiles.

In [8], Simmons and Dillon propose an ontology based approach to address knowledge management in open source software development. This ontology covers the following concepts: Participant, Role, Activity, Procedure, Artefact, Tool. Other ontologies describe data in community. The SIOC project defines an ontology that contains concepts necessary to express information contained in online community sites (Ex: boards, blogs, etc)[15]. Baetle [16] is an ontology to describe software bugs and trouble tickets that aims at becoming the standard used by bugzilla and other repositories to enable people to query for bugs across repositories.

### 3. APPROACH AND METHODOLOGY

#### 3.1 Improving community awareness

Our general questioning is whether one can improve community awareness for project members. We'll address this goal in trying to answer, based on the recent member activity, to two questions:

1. What is happening in my projects?
2. How does the community of my project work?

The answer will be provided by trying to measure and visualize several related factors:

- Who is working
- How are they doing
- What are they doing
- Development degree of project
- Resource available of community
- Communal services of community
- Where/When do members work
- Capacity development of community
- Sharing knowledge, cooperation, communication process

<sup>5</sup> <http://dublincore.org/documents/1999/07/02/dces/>

<sup>6</sup> <http://xmlns.com/foaf/spec/>

<sup>7</sup> <http://usefulinc.com/doap/>

#### 3.2 Semantic metadata extracted from projects activity

In general, in a Forge, an item in a RSS feed is the result of a member action, so aggregating RSS feeds allows to enlighten one person's activity. The aggregated RSS helps members to easily review all recent actions in their project.

Moreover it can help measure activity of the whole community through some parameters such as: the number of actions, the number of active members, the number of used tools with number of actions in each tool.

Analyzing the aggregation of different feeds also helps constructing the social network for that community of practice by combining relations in different activities conducted in heterogeneous tools. Two members are related when they discuss the same subject in mailing lists, edit the same topic in a Wiki or commit the same module in Subversion. These combined relations give a more comprehensive vision about the collaboration in the community than the results analyzed in single source, such as [9].

Historical data can be used to analyze activity trends of members of a project: monthly activity, daily activity, hourly activity.

#### 3.3 Semantic aggregator and processor in a forge

Our approach consists in integrating, in the very forge used by developers of the open source projects, a semantic aggregator using RSS/RDF 1.0 to qualify and aggregate the initially non-semantic RSS feeds of the different tools.

To enrich information of RSS items, we integrate FOAF schema which describes the authors of actions in the RSS description.

Each project is in addition described with DOAP in the forge's portal. It then publishes a public feed that provides the information integrated from feeds of tools used this project. To include a pointer to the public feeds in DOAP we use the **rdfs:seeAlso** attribute and add an object **rss:channel**. By contrast the **doap:project/doap:name** attribute is used to point to its project. The items of the converted RSS/RDF feeds is made from items generated by the forge's tools in the original non-semantic RSS feeds. In the RSS 1.0 RDF schema there will then be FOAF attributes such as **foaf:Person/foaf:nick**, **foaf:Person/foaf:mbox** which describes the contributor of an item. The figure 1 shows our integrated RDF schema.

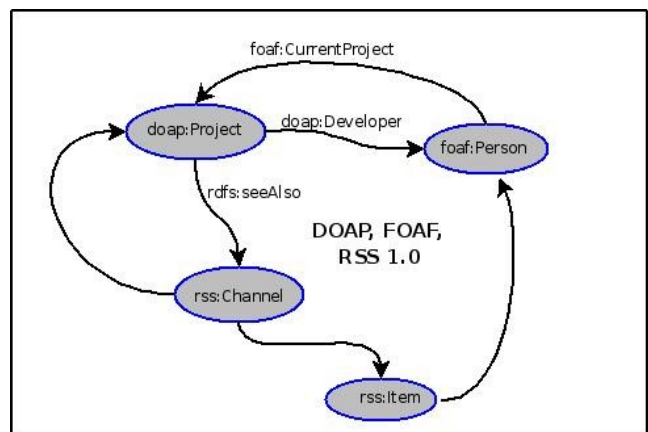


Figure 1: Integrated RDF Schema

An example of RSS channel description with a pointer to a DOAP document of project.

```
<rss:channel>
<rss:title>projeta </rss:title>
<rss:description>WebSVN RSS feed - projeta
</rss:description>
<rss:link>http://forgedemo.org/projeta </rss:link>
<doap:Project>
  <doap:name>projeta </doap:name>
</doap:Project>
</rss:channel>
```

An example of RSS item description with pointer to the FOAF document of contributor.

```
<rss:item>
<rss:title>minor fixes in version 2 branch
</rss:title>
<rss:link>http://forgedemo.org/projeta </rss:link>
<rss:description>
Rev 2463 - toto (3 file(s) modified)
minor fixes in version 2 branch
</rss:description>
<foaf:person>
  <foaf:nick>toto </foaf:nick>
</foaf:person>
<dc:date>Mon, 02 Jun 2008 22:18:02 +0100</dc:date>
</item>
```

## 4. CASE STUDY ON PICOFORGE

### 4.1 Adding a supervision tool in the forge

Started in order for use as a pedagogical platform, PicoForge is a libre-software system released under the GNU GPL license, which eventually evolved as a general-purpose forge. It provides a Web-based collaborative work platform built on top of phpGroupware<sup>8</sup> and other libre software tools. PicoForge provides project hosting facilities for small teams of software developers. It is mainly oriented to teaching and research environments. The forge integrates several libre software Web applications: TWiki, Sympa, CVS, Subversion, WebSVN, Mantis, much of which include RSS feeds to track activity. Our Supervision tool is a module developed in order to be integrated to a future release of PicoForge.

It will fetch, mix and process the non-semantic RSS feeds already published in the collaborative tools such as TWiki, Sympa or WebSVN for each project. It is also able to add other RSS feeds from outside PicoForge which are related to the project. Figure 2 shows the architecture of the Supervision tool.

It allows “querying” public projects of the platform to export a list of projects in RSS format or in RDF as FOAF + DOAP. Here, DOAP describes project, FOAF describes their members, and public RSS/RDF feeds will publish semantic notifications of project activity.

In order to have a vision of community of practice of projects, the Supervision tool also constructs a social network of project members based on their actions.

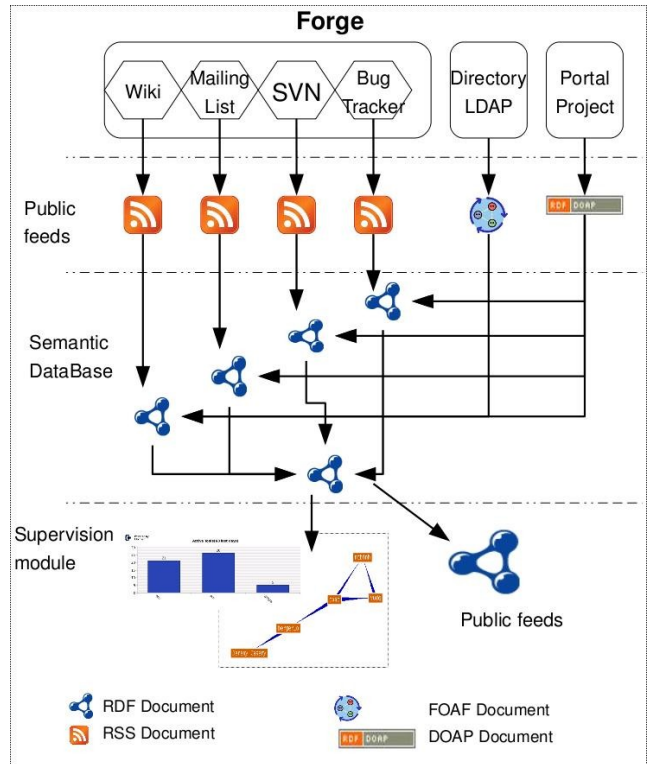


Figure 2: Supervision on PicoForge

Several third libraries is used to implement Supervision tool in PicoForge:

- RDF API for PHP<sup>9</sup>(aka RAP) is a Semantic Web toolkit written by PHP language. It allow to parse, query, manipulate, serialize and serve RDF model(for example MemModel, DbModel, InfModel, ResModel, OntModel). It support for the RDQL query language. In Supervision tool we use RAP library to manipulate and query RSS documents in Mysql<sup>10</sup> Database.
- RSS\_PHP<sup>11</sup> is a RSS parser and XML parser for PHP using DOMDocument. This library is used to parse, reformat RSS documents before store in RDF database model via through RAP library.
- Libchart<sup>12</sup> is a PHP library to create charts such as Bar charts (horizontal or vertical), Line charts, Pie charts. It is used to generate statistical charts in Supervision tool.
- TouchGraph<sup>13</sup> allow for the visualization of graph such as social network. It is a Java application. In Supervision tools we use TGLinkBrowser library to show members network.

### 4.2 First results on a collaboration project

In our case study we used this Supervision tool to observe a public project for software development hosted in the PicoForge installation at Institut TELECOM, SudParis<sup>14</sup>. In the following,

<sup>9</sup> <http://www4.wiwiss.fu-berlin.de/bizer/rdfapi/>

<sup>10</sup> <http://www.mysql.com/>

<sup>11</sup> <http://rssphp.net/>

<sup>12</sup> <http://naku.dohcrew.com/libchart/pages/introduction/>

<sup>13</sup> <http://www.touchgraph.com/>

<sup>14</sup> <http://picoforge.int-evry.fr/>

<sup>8</sup> <http://www.phpgroupware.org/>

we provide screenshots of the graphs produced by the developed tool, available to project members, which try and answer the questions proposed in section 3.

Figure 3 shows the total activity in project with different used tools in the 60 last days. It helps project manager know what is the kind of common activities in recent time. The figure 4 shows the active users with number of their activities.

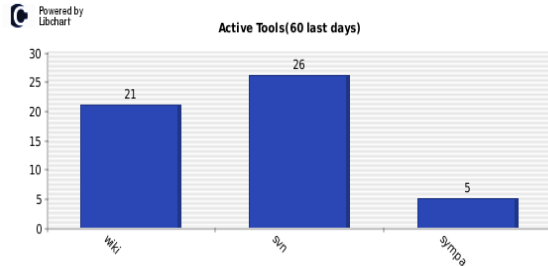


Figure 3: Total activities in project in 60 last days

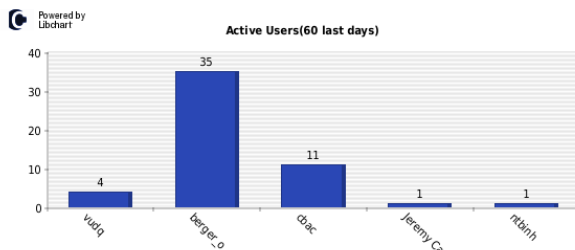


Figure 4: Active users in 60 last days

The active degree of project can estimate by using total number of activities in window time. Moreover based on interaction of members Supervision tool can draw live network of the active members and relations between them in a window time (for example, in figures 5, a network of the current month). In this case we have combined three kind of relations: co-commit on Subversion, co-drafting on Twiki or discussion on Sympa mailing list. So it give a vision more complete about the collaboration in the community of project than the results analyzed on single source. The vision is fresh, dynamic and updated in real time.

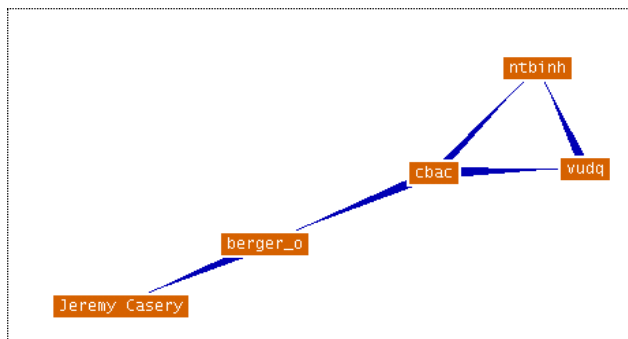


Figure 5: Members network

## 5.CONCLUSION

This paper presents an approach to help collect, transport, and correlate live data about multiple projects across multiple software forges to enlighten group awareness. We extended RSS 1.0 RDF schema by integrating FOAF, DOAP and Dublin Core

schema which help to enrich information in RSS feeds. By aggregating data from different sources, using standard semantic Web formats, we seek better interoperability, which may allow the use of generic high-level supervision tools which may be plugged in various forges systems. The produced graphs help members of teams to get a better real time visualization of the inter-person cooperation in their project.

In the future, we hope that the availability of semantic data collected by such tools could be used to plug forges with advanced tools in order to detect activity patterns of members and to a trace co-evolution of software and community. Based on social network detected, we hope that it will be possible also to apply social filtering technique to improve peer review process. The content of data collected and the interaction of members could be used to analyze the communication quality, the cognition process of teams in projects hosted on software development forges.

Availability of such advanced tools for projects direct benefit would certainly help improve FLOSS development environments, provided that good use is made of the informations. Issues like access rights to such tools or correlated data, for privacy reasons, were not considered in this initial work, and remain to be discussed with communities using the forges.

## 6.REFERENCES

- [1] Howison, J., Conklin, M., Crowston, K. (2006). *FLOSSmole: A collaborative repository for FLOSS research data and analyses*, International Journal of Information Technology and Web Engineering, 1(3). July-September, 2006. pp 17-26.
- [2] de Groot, A., Kugler, S., Adams, P.J. and Gousios, G., *Call for Quality: Open Source Quality Observation*, in IFIP International Federation for Information Processing, Volume 203, Open Source Systems, eds. Damiani, E., Fitzgerald, B., Scacchi, W., Scotto, G., (Boston: Springer), pp. 57-62.
- [3] SQO-OSS: Software Quality Observatory for Open Source Software (<http://www.sqo-oss.eu/>)
- [4] FLOSSMetrics project(<http://www.flossmetrics.org/>)
- [5] Conklin, M. (2007). *Project entity matching across FLOSS repositories*. In Proceedings of the 3rd International Conference on Open Source Systems. Limerick, Ireland. June 11-14, 2007. pp. 45-57.
- [6] SourceKibitzer - *Recognizing Contributions to Open Source Software* (<http://www.sourcekibitzer.org/>).
- [7] Anupriya Ankolekar, James D. Herbsleb, Katia Sycara, *Addressing Challenges to Open Source Collaboration With the Semantic Web*, in Taking Stock of the Bazaar: The 3rd Workshop on Open Source Software Engineering, the 25th International Conference on Software Engineering (ICSE), 2003. Portland OR, USA.
- [8] Gregory L. Simmons, Tharam S. Dillon, *Towards an Ontology for Open Source Software Development*, In IFIP International Federation for Information Processing, Volume 203, Open Source Systems, eds. Damiani, E., Fitzgerald, B., Scacchi, W., Scotto, M., Succi, G., (Boston:Springer), pp 65-75.
- [9] Luis López-Fernández, Gregorio Robles, Jesús M. González-Barahona and Israel Herraiz, *Applying Social Network Analysis Techniques to Community-Driven Libre Software Projects*, Proceedings: International Journal of

Information Technology and Web Engineering, Vol. 1, Issue 3, September 1st – 2006.

- [10] Jin Xu, Yongqin Gao, Christley S, Madey G, *A Topological Analysis of the Open Source Software Development Community*, System Sciences, 2005. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference, 03-06 Jan. 2005.
- [11] Luis Lopez-Fernandez, Gregorio Robles, Jesus M. Gonzalez-Barahona, *Applying Social Network Analysis to the Information in CVS Repositories*, Proceedings of the Mining Software Repositories Workshop. 26th International Conference on Software Engineering (Edinburgh, Scotland), May 25th – 2004.
- [12] DOAP: Description of a Project (<http://usefulinc.com/doap/>)
- [13] FOAF Vocabulary Specification 0.91 (<http://xmlns.com/foaf/spec/>)
- [14] RDF Site Summary (RSS) 1.0 (<http://Web.resource.org/rss/1.0/>)
- [15] Cioc: Semantically-Interlinked Online Communities (<http://sioc-project.org/>)
- [16] Baetle: Bug And Enhancement Tracking Language ([http://blogs.sun.com/bblfish/entry/baetle\\_bug\\_and\\_enhancement\\_tracking](http://blogs.sun.com/bblfish/entry/baetle_bug_and_enhancement_tracking))