PROCEEDINGS OF SCIENCE

ScienceSoft: Open Software for Open Science

Alberto Di Meglio[1]
CERN – European Organization for Nuclear Research
1211 Geneva, Switzerland
E-mail: alberto.di.meglio@cern.ch

Florida Estrella
CERN – European Organization for Nuclear Research
1211 Geneva, Switzerland
E-mail: florida.estrella@cern.ch

Most of the software developed today by research institutes, university, research projects, etc. is typically stored in local source and binary repositories and available for the duration of a project lifetime only. Finding software based on given functional characteristics is almost impossible and binary packages are mostly available from local university or project repositories rather than the open source community repositories like Fedora/EPEL or Debian. Furthermore general information about who develops, contributes to and most importantly uses a given software program is very difficult to find out and yet the widespread availability of such information would give more visibility and credibility to the software products. The creation of links or relationships not only among pieces of software, but equally among the people interacting with the software across and beyond specific project and communities would foster a more active community and create the conditions for sharing ideas and skills, a more rapid improvement of the software quality and the creation of more sustainable open source communities. This paper presents the work performed as part of the EMI project in collaboration with other partners in setting up an open community dedicated to the development of software for scientific research. The community goals and the benefits for developers and users are outlined. A conceptual prototype of the community portal, the services and the collaboration tools are described.

The International Symposium on Grids and Clouds (ISGC) 2012
Academia Sinica, Taipei, Taiwan
February 26 – March 2, 2012

[1] Speaker

1. Introduction

There is a wealth of open source software in use across scientific communities but the value of its contribution to science is under-estimated, under-utilised and often poorly coordinated. Some websites such as ohloh (http://www.ohloh.net/) offer directories that attempt to rate the quality and impact of open source software projects, but currently lack the means of attracting developers and users from academic communities and harvesting a large enough body of essential data to make their results meaningful for the scientific research environments. Being able to use the power of cataloguing services, trends and statistics would provide a sound basis for judging the popularity of specific software, enable social-networking amongst users and developers, create active communities and promote citizen science. Rating software and providing a means by which it can be cited in a similar manner to publications and datasets would enable the authors to gain merit and career advancement for their work and accelerate the open source software movement in scientific communities. Being able to quantify the impact of open source software would allow funding agencies, companies and venture capitalists to better target their investments leading to a more vibrant and sustainable open source market for open science.

During September 2011 the EMI project started a discussion on a general proposal to investigate, design and establish an open source software initiative as part of the EMI's long-term sustainability plans. The main objective of this proposal was initially to create the conditions for the continuing development, support and use of the EMI software products after the end of the EMI project by establishing a broad open source community of developers and users.

However, it became rapidly clear from discussions with users and developers from scientific communities and research projects that similar concerns about long-term sustainability were shared by many other projects developing or using middleware, applications, tools and related services of critical interest for the European scientific research communities.

The discussion on establishing a broad open source community around software for scientific research was therefore extended to a number of interested parties. The goal of the initiative, called ScienceSoft, is currently to define and setup a truly open community of software developers, system engineers and users in the context of global scientific research by including scientists, developers, service providers, research institutes and commercial companies in the process since its inception.

This paper outlines the state-of-the-art of the ScienceSoft open source activities in the scientific research environments and describes a number of problems and potential solutions identified by discussing with user and developers of existing projects and communities. The potential benefits of the proposed solutions are described and how scientists, developer, administrators, managers and funding bodies could exploit them to take decisions and plan their activities.

2.Market Analysis

As a starting point for the discussion, a brief market analysis was performed. Five main types of computing and data environments were considered:

- Grid (High-Throughput) computing
- High-Performance Computing
- Cloud computing
- Desktop or volunteer computing
- Stand-alone computing

Combination of these basic types can be considered as well, like for example the provision of grid services using on-demand IaaS clouds or the provision of cloud-like access to grid resources.

The software layers involved in the use of the selected types of environments include:

- Operating system software
- Middleware software
- Service management software
- Applications and community specific services
- User interfaces and portals
- Tests, tools, benchmarks and other supporting software

The roles involved in the use of the software can further be classified as:

- Developers
- Users
- Service providers
- System engineers, platform integrators
- Institutes and Companies
- Collaborations (projects, initiatives, communities, virtual organizations, etc.)
- Funding bodies

The roles are not mutually exclusive. On the contrary the complex relationship among people in various roles is indeed one of the fundamental issues that ScienceSoft proposes to investigate. For example the developer of a software product can be at the same time the user of another product (a dependency) or a service (a testing service or a cloud IaaS service) and a provider of applications for a group of scientific researchers.

In this classification, the different roles have different problems and are looking for different specific solutions. The ideal open source community must be able to take into account and express the diversity across the market (vertical dimension), while at the same time recognising the commonalities across software layers (horizontal dimension). The distinction

between vertical and horizontal dimensions is at the core of the ScienceSoft initiative will be further explored later on in this paper.

Figure 1 illustrates the previously defined categories and their relationships.

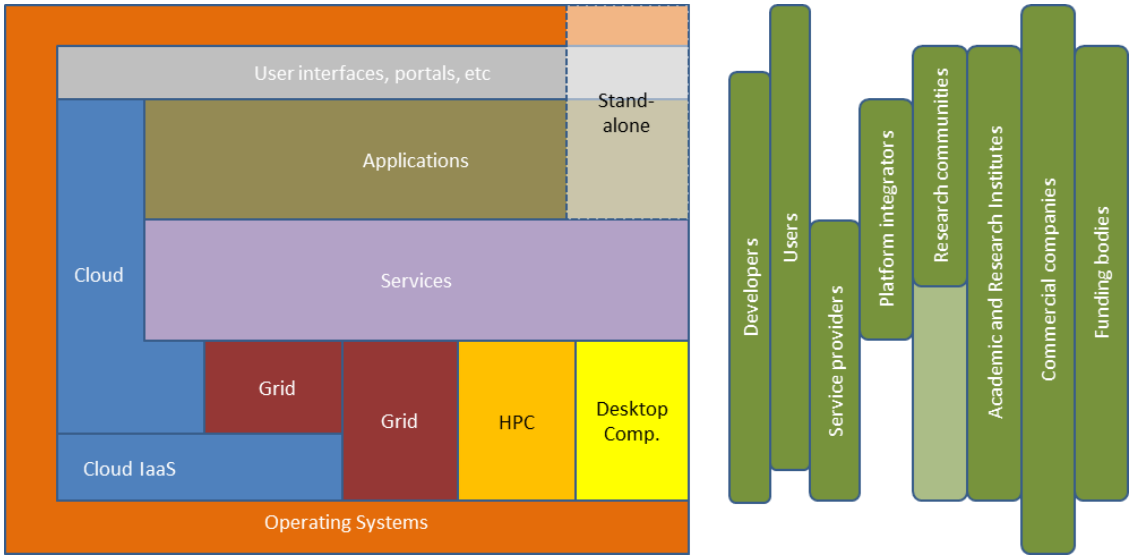

Figure 1: The software and services market classification

3.Identified Problems

Users and developers of software for scientific research projects have been contacted during conferences and presentations and asked about what problems they are confronted with in their normal working activities. Among the identified problems the following are considered the most critical:

- Lack of continuity in support, development, coordination of software
- Non-optimal communication between users and developers
- Lack of consistent real usage information
- Limited access to other users' experience
- Limited or complex ways of finding what exists already
- No way of influencing the production of software
- Lack of visibility of the software activities
- No way of assessing the user "market"

Most of the software developed today by research institutes, university, research projects, etc. is typically stored in local source and binary repositories and readily available for the duration of the project lifetime only. Finding software based on given functional characteristics or field of application is very difficult especially for new projects or young researchers. Binaries to be run on the most used operating systems are available from many different places ranging

from local university repositories to mainstream community repositories like Fedora, EPEL or Debian. Cases of conflicts are often found between different versions distributed by different people from different places. Source code is even more difficult to locate and access and contributing with comments, patches and fixes, which is a very common activity in the open source world, is traditionally very difficult to do in the research communities. This has been for years a primary complaint from users.

Most of the reported problems can be categorized as a lack of consistent and transparent information about the software being used in scientific research. The problem is not necessarily a lack of technical information (such as documentation or user guides, although this has also been described as a problem in many cases), but rather a lack of metadata. Information about who develops, contributes and uses a given program is very difficult to find out and yet the widespread availability of such information would give more visibility and credibility to the software products. In addition, the EC invests considerable amounts of money into funding projects that directly or indirectly need to develop software. A single repository of metadata information about software products would allow projects to avoid re-developing existing solutions and would provide valuable statistics about software usage. Such information could be used also by the EC to monitor the outcome and impact of funded projects, the extension of adoption of open source software and the compliance with OSI licenses and possible as input to future EC calls objectives and framework programmes. In the same way, the information could give more strength and credibility to project proposals, which could be backed by realistic information about usage, impact and exploitation of the software.

4.Possible services and benefits

From the discussions with users and developers and the outcome of the CERN Workshop in February [1], a number of desirable functionality has been defined:

- **Software, services and people catalogues**: the first and foremost desired functionality is the provision of catalogues of information about software products, software-related services and people. The catalogues should allow to group together sets of related products, services and people based on flexible search criteria.
- **Generation of statistics**: the information collected and processed should not only be used to search about software, but also to general relevant and useful statistics.
- **Honour system**: community users should be able to rate the registered software and services based on their experience. Ratings can be provided based on predefined categories such as reliability, support quality, documentation, ease-of-use, standards support, etc.
- **Citation system to allow software to be referenced in papers**: registered software should receive some sort of unique identifier like the DOIs used for papers, so they can be reliably cited in scientific publications [2].
- **Marketplace for products, services, and people**: this is one of the most interesting features and one that may well define a community. Matching demand and offer of

software products, service and people skills should be enabled based on the catalogues maintained by the community tools.

- **Links to technical services**: one of the marketplace-related activities is the provision of technical services. A range of services could be designed and provided by community members and provided to other members for free or for a fee depending on conditions.
- **Platform integration support**: using the collected information and the product catalogues, it should be possible to define community-specific software stacks to be supported by platform integrators. These community-specific profiles or stacks can then be pre-packaged and easily deployed using the more and more standard virtualization and cloud technologies
- **Support for creation of ad-hoc communities and groups**: this requirement is what may actually define ScienceSoft as a community-enabler. ScienceSoft per se should act a super-community provide a framework and tools to enable more specific communities to interact without being isolated from other communities.
- **Coordination, collaboration and discussion tools**: collaboration tools typical of distributed online infrastructures should be available through the ScienceSoft community portal.
- **Support for organization of technical events**: it should be possible to advertise and manage technical events related to the hosted communities or projects. Support can range from dedicated event pages, to agendas, advertising, collection of material, etc.

5.State-of-the-Art of Open Source Communities

The first approach being considered to address the described issues and provide the desired functionality is to exploit lessons learned from successful open source software communities.

Most of the software used in scientific research and developed by academic institutes is generically "open source" in the sense that it uses some type of OSI license. However, it takes more that source code and a license to have a "community". The general definition of an open source software community is a group of developers and users interacting to produce free-software. The interaction among users and developers, the sharing of resources and common objectives and the benefits deriving from sharing are of course fundamental to have a community and not just software in a repository.

We can distinguish primarily between four different types of open source communities:

- **Technology-specific (or horizontal) projects**: this type of communities includes projects focused around a specific technology or framework which all members contribute to. Usually the membership rules are quite stringent both in technological and legal terms. For example it's not uncommon to have to adopt a mandated IP model and a license for all contributing products. Notable examples of this category are the Apache Foundation [3], the Eclipse Foundation [4] or the Drupal Association [5].
- **Operating system distributions**: communities focused around different flavours of Linux operating systems have been among the first to emerge and have in many cases enabled

most profitable open source business models. Although in general there is no formal membership into these communities, the engagement rules to contribute are quite strict and require a peer-review level of competence and quality for both contributors and products. Most notable examples of this category are Fedora [6], EPEL [7], Debian [8], CentOS [9], etc.

- **Services and tools**: these open source communities usually provide a software application and often services based on that application. They have dual usage models, whereby access to the service is free for personal or non-commercial use, while professional use is charged a fee. Most notable examples include SourceForge [10], GitHub [11], Zarafa [12], and many others.
- **End-to-end (or vertical) open source communities**: at the end of 2011 Andrew Aitken, president of the Olliance Group, a leading open source consulting firms, wrote an article about the appearance of "super-communities" or communities of communities [13]. The super-communities instead of focusing on a particular piece of open source technology are built around the entire end-to-end supply chain of an industrial sector, like the aerospace industry (Polarsys launched by Airbus [14]), stock exchange management (OpenMama launched by the New York Stock Exchange [15]) or electronic healthcare records management (OSEHRA launched by the US Department of Veteran Affairs [16]). The Olliance Group predits that this kind of communities will rapidly increase in number. Microsoft launched in 2011 the OuterCurve [17] community to host open source software and communities and provide general-purpose IP management services with a focus on Windows applications. In the scientific communities, portal like NanoHub [18] or CyberSKA [19] focus on the general needs of the nano technologies and radio astronomy communities respectively.

The software produced and used in scientific applications is by its nature very diverse. It uses different programming models, technologies, IP and licensing models. In addition, the end users, the scientists using the software to perform their research, are not overly interest in what technologies are used under the hood, but are very concerned with having a working set of tools. The first three types of open source communities described above could therefore be used for parts of the software produced in the academic world, but wouldn't bring the level of communication and organization needed to provide the functionality described earlier. A suitable model for ScienceSoft could therefore be the fourth one, where the overall end-to-end software needs of specific scientific communities could be modelled and addressed with a more global approach than just individual pieces of software.

6.ScienceSoft Organization, Structure, Operations

Based on the preceding analysis the ScienceSoft super-community could therefore be configured as a community of users and developers of software targeted at scientific applications. The community members contribute software and information and provide services to other members. The members within ScienceSoft are organized in focused

communities or collaborations around a specific scientific or research topic. People, software, services, etc. are tagged within the super-communities based on their particular focus in order to build community-specific sets of resources. Resource tagging is not exclusive. The same resource can be tagged with more than one community focus, so that it becomes possible to understand the overall usage of that resource within a more general scientific context.

A community portal gives access to the community services like member and product registration and management and the different functionality described earlier in the document.

Most of the common base functionality required to operate the ScienceSoft portal already exists in some form. As a first implementation, the ScienceSoft portal can be configured as an aggregation of such functionality within a coherent container. Functionality like software inventories, source-code repositories, social networking, forums, etc., can be provided in this way using existing open source services like ohloh.net or other inventories, Drupal modules for the web based collaboration tools and existing social networks like FaceBook, Linkedin or Google+ for user management.

The community specific services would instead be provided by the members through links or applications running in the portal. Community-specific micro-sites can be easily established from common templates to create well-defined identities and focused sets of people, programs, services, tools, information, etc.

Although the initial organization and governance structure of ScienceSoft should be as lean and lightweight as possible, it is foreseen that depending on the success and evolution of its activities it could become in the future a Not-for-Profit Foundation on the model of existing successful open source foundation.

7.Current Activities and Next Steps

The requirements and possible implementation strategies for Science Soft are currently being investigated by the ScienceSoft Steering Committee. Initial requirements and desired functionality have been discussed in occasion of the ScienceSoft Workshop held at CERN in February 2012. The workshop participants have agreed to keep providing feedback and collaborate in the definition of implementation priorities. The proposed timeline is as follows:

- March to June 2012: definition of priorities and identification of volunteers ScienceSoft maintainers
- July to December 2012: progressive implementation of a prototype community portal, dissemination, engagement of scientific communities in trying the functionality and providing feedback
- January to April 2013: start of the regular activities, further requirements and implementation cycles. Until this date the community is incubated with the EMI project, which provides overall coordination
- May 2013 onward: regular operations, fund raising for continuing activities based on the success of the initiative. Phase down and discontinuation if no interest has emerged.

Users interested in taking part of just being informed about the ScienceSoft activities can refer to its web site at http://sciencesoft.org

8.Acknowledgements

This work is partially based on the discussions among the ScienceSoft Steering Commitee members and the participants of the CERN ScienceSoft Workshop in February 2012. Our thanks go to all who have contributed to the discussions.

This work has been partially funded by the European Commission as part of the EMI project (Grant Agreement INFSO-RI-261611).

References

[1] ScienceSoft Workshop, CERN, 8 February 2012, https://indico.cern.ch/conferenceDisplay.py?ovw=True&confId=160503

[2] The Digital Object Identifiers (DOI), http://www.doi.org

[3] The Apache Software Foundation, http://apache.org

[4] The ECLIPSE Foundation, http://www.eclipse.org/org/foundation

[5] The Drupal Association, http://association.drupal.org

[6] The Fedora Community, http://fedoraproject.org/en/join-fedora

[7] EPEL – Extra Packages for Enterprise Linux, http://fedoraproject.org/wiki/EPEL

[8] Debian, http://www.debian.org

[9] CentOS – The Community Enterprise Operating System, http://www.centos.org

[10] SourceForge, http://sourceforge.net

[11] GitHub – Social Codiing, https://github.com

[12] Zarafa Community Hub, http://community.zarafa.com

[13] Aiken, Andrew, *The Advent of Super Communities*, OpenSource Delivers, 20 December 2011, http://opensourcedelivers.com/2011/12/20/the-advent-of-super-communities

[14] Polarsys, http://www.polarsys.org

[15] OpenMama, http://www.openmama.org

[16] OSEHRA, http://osehra.org

[17] The OuterCurve Foundation, http://www.outercurve.org

[18] NanoHub, http://nanohub.org

[19] CyberSKA, http://www.cyberska.org