

# Open collaboration in an institutional context

Thomas Metz, Michael Jonathan Mendoza, Rowena Valerio  
International Rice Research Institute, DAPO Box 7777, Metro Manila, Philippines

## Background

In 2003, the Generation Challenge Programme (GCP), a research and capacity-building network of initially 17 centers of the Consultative Group on International Agricultural Research (CGIAR) and national research institutions and universities, started with the aim to produce a public system for accessing and developing new genetic resources. This public system utilises plant genetic diversity, advanced genomic science and comparative biology to develop tools and technologies that enable plant breeders in the developing world to produce better crop varieties for resource-poor farmers.

This system was designed to involve the capture, storage, integration, analysis and dynamic dissemination of substantial volumes of diverse and dispersed genomic, genetic resource and crop improvement information in a generic way that would be usable for all 22 mandate crops in the CGIAR. The linking and integrating of these information components into a coherent information gateway required extensive new software development as well as the use and adaptation of existing software.

With several partners around the globe developing and integrating software into a single coherent system, there was a clear need for a collaborative platform for software engineering, which was commissioned by the GCP in 2004. Analogous to the overall principles of the GCP, this platform had to be inclusive, transparent, participatory, and scalable, supporting the production of global public goods.

None of the partners had such a comprehensive platform in operation but it was also clear that such a platform could not be built from scratch. The basic strategy for implementing a software engineering platform was to adopt systems that had been proven highly successful for open source software as well as open content development. In 2004 there were two outstanding collaboration sites on the web, SourceForge.net, hosting more than 50,000 software development projects (currently 150,000), and Wikipedia, having more than 300,000 articles (currently 2 million) in its English version.

## Choice of tools

Collaboration tools are developed with particular target audiences in mind, be they commercial institutional users on one hand, or voluntary independent users on the other. Depending on the main target audience, the tools emphasize different sets of functionalities and principles. In a corporate environment the emphasis may be on security, monitoring, management and efficiency. In an environment with mainly individual volunteers the emphasis may be on cost, transparency, simplicity and effectiveness. Software used in institutional environments tends to be commercial software, whereas in voluntary environments it is often free, open source software. While commercial software often has a licensing model that charges per user and could potentially be expensive for a large group of participants, it has often more functionality and may generally be considered more user-friendly. Free, open source software may not carry a price tag for the software, but the cost of the hardware, setup, and maintenance may be considerable.

In the context of the collaborative software engineering platform for the GCP with a potentially large group of participants from different countries, continents and types of organizations, using their different ICT infrastructure, the choice fell on two web-based collaboration platforms that had proven to be extremely successful in a voluntary environment, namely SourceForge and

Wikipedia.

SourceForge is a collaborative software development environment that provides revision control and a web front-end to several communication and collaboration tools (mailing list, discussion forum, issue tracker, file release system, etc.) for each hosted project. SourceForge itself was originally open source software but was later commercialized. The SourceForge site provides free project hosting and has remained highly popular, with currently more than 150,000 registered projects and more than 1.5 million registered users. A fork of the open source version was developed and maintained under the name GForge, which is the system that was selected for the GCP software engineering platform. In 2004, there were practically three choices, using a commercial software engineering platform with a high annual license fee per developer, hosting the GCP projects on the SourceForge site, or installing and using the SourceForge derivative GForge. The commercial solution would have been too expensive as it would have severely limited the number of potential contributors, while hosting projects on SourceForge would have provided too little control and motivation and would have left some uncertainty about the continued provision of free services.

Wikipedia has become ubiquitous, and the underlying MediaWiki software is freely available for use. In 2004 there were a number of free Wiki implementations available, some of them with more functionality or easier to install than MediaWiki. The choice was determined by the popularity of Wikipedia which was considered a guarantee for the further development and maintenance of the underlying MediaWiki software.

While the GForge collaborative software development was meant to be used mainly by software developers, the MediaWiki implementation was meant to be used by all collaborators for creating and maintaining documents related to planning, design, implementation, support, etc. At the time of the installation of the two systems, there was very little collective experience with such tools and communities, and the approach was one of learning by doing.

## **Implementation and use**

In early 2005, the GForge collaborative software development environment as well as the MediaWiki software were installed for testing on desktop computers that had been reconfigured as Linux-based servers. By April 2005, both systems had been configured and tested and were launched at a workshop under the names CropForge and CropWiki, with initial training given on Wiki editing. During the year, training was provided in groups and individually on several occasions. Dedicated servers were purchased by the end of 2005, and the existing systems were transferred and upgraded to the dedicated servers. While the transfer and upgrading of the GForge system was done with the assistance of a support company, all other server installation, configuration, application installation, testing, and trouble-shooting were performed by an in-house systems administrator with extensive Unix/Linux experience. Funding for the server purchase, the GForge transfer and upgrading, as well as the staff time was provided under a project of the GCP, and this support is still ongoing. Both systems were not restricted for use by the GCP, but were open to provide hosting services for other projects and activities. Although all software, including the operating system, is free/open source, the entire deployment represents a considerable investment, as it includes dedicated server hardware, systems administrator support, face-to-face meetings, workshops, training materials development, and training. The full project documents for the deployment and support of these collaboration platforms are available at the GCP website.

The CropForge system currently hosts 71 software development and support projects and has 180 registered users. Each project is independently managed by one or more administrators who can configure the functionalities, access control, as well as membership of their respective project. The projects differ in activity level, with some of them being very active, while others are dormant. Some projects primarily use the source code version control functionalities and address

the needs of the software developers, while other projects use the communication and support functionalities that address the interaction between software developers and users. Generally, all project content, e.g. source code, mailing lists, discussion forums, trackers, file download, etc. are publicly accessible, but contribution and modification are generally restricted to project team members.

The Generation Challenge Programme pantheon project is a very active software development project with 4 administrators and 47 developers. Since the establishment of the project 17 months ago, 29 developers located in 6 different organizations worldwide have made more than 27,000 changes or additions to the source code and the documentation files of that project. The source code repository is the only functionality used in this project and it contains more than 4,500 Java source code files and 5,800 html documentation files.

The purpose of the ICIS communication project is the distribution and support of the International Crop Information System (ICIS) software. The project was established 31 months ago, has 21 team members from 6 different institutions worldwide, and doesn't contain any source code. However, the project maintains 2 mailing lists, a file download area, 3 issue trackers and 3 discussion forums to provide distribution, communication and support between the ICIS developers and users. The issue trackers for software problems (bugs) and feature requests are probably the most important feedback mechanisms from the user community. The content of the issue trackers and the discussion forums is publicly accessible, creating a transparent development and support environment and a record of the project history, which is useful for new users of the software.

The CropWiki system currently hosts more than 20 independent Wiki sites with very different configurations for access control and member roles. Some Wiki sites are used as part of the institutional Intranet, accessible only from the local area network, while other Wiki sites are globally visible. None of the Wiki sites has been configured for anonymous editing. The two most developed Wiki sites are the GCPWiki and the ICISWiki which started in February 2005 and March 2005 respectively. The ICISWiki initially allowed self-registration of users, but this had to be restricted due to the appearance of automatic account creation and content vandalism software. Following is a comparison of the content amount, access and participation for the above two Wiki sites.

<b>Wiki site</b>	<b>GCPWiki</b>	<b>ICISWiki</b>
Number of content pages	580	200
Number of uploaded files (jpg, pdf, ppt)	620	1,100
Number of page views	66,000	718,000
Number of page edits	15,000	7,000
Number of registered users	220	100
Number of users with one or more edits	82	94
Number of users with 10 or more edits	42	26
Read access	closed	open
Months of operation	31	30

The differences in content and participation reflect the different communities and purposes of the two Wiki sites, while the differences in page views is most likely caused by search engines indexing the pages of the open ICISWiki.

## **Discussion and conclusions**

The choice of software should neither be driven only by cost nor only by functionality. Commercial

software has often more functionality and has more user-friendly interfaces than free/open source software and would usually win a comparison on total functionality. Therefore, comparisons should be made based on the required functionality, rather than the total functionality. However, it may sometimes be difficult to determine which functionality is required, and needs may also change over time. Less functionality and fewer options may actually improve adoption and use. If the software choice is made only based on cost, free/open source software is hard to beat. However, the major cost of deployment and support is not the software, but hardware, connectivity, configuration, user training and support. Commercial software that is charged on a per-user basis may become prohibitively expensive, if the goal is to reach a wide community of contributors. Assessing the long-term development and support for a particular open source software may be difficult. In this respect we made a conservative choice with the MediaWiki software, as it was the basis for the very popular Wikipedia site. When choosing free/open source software, more technical expertise for installation, configuration, upgrading, etc. may be required than with commercial software.

Including the deployment and support of new types of collaboration systems into a project with defined deliverables and a dedicated budget may help in integrating these new systems into the IT infrastructure of the hosting institute. Hosting with a commercial provider should be considered if too many integration and institutional support problems are anticipated. Virtual server hosting for low-bandwidth collaboration systems or free services may be worthwhile considering, but this will probably not reduce the major support cost. Starting a collaborative work environment without a budget and a clear mandate may not be very sustainable in an institutional context.

In an institutional environment, contributors to collaboratively developed content are expected to provide their contributions as part of their normal work. This work is governed by the contractual agreement between the employer and the employee, and often restricts employees in terms of ownership of their intellectual output, or freedom to release information concerning their work without review. While web-based platforms significantly lower the technical transaction cost for large-scale collaborative content development, they don't address the institutional constraints that employees may face when contributing their work. Explicit permissions as well as a general institutional policy on intellectual property release governing collaborative online work may be required before institutional contributors can freely participate in collaborative content creation.

The ownership of collaboratively created content may be a contentious issue. In the extreme case, such as a feature article on Wikipedia with hundreds of edits from many individual contributors, it may be practically impossible to establish who owns what. However, if the content is created and released under an open content license, the focus shifts from content ownership to the freedom of use, as defined by the particular license. If everyone is permitted to copy the content, distribute it further and create derivative works, ownership becomes largely irrelevant. The main issues are then with the particular license compliance, which may allow or prohibit commercial use, require attribution or share alike distribution of derivative works. At IRRI, a change in the intellectual property policy that requires that information products are generally to be released under open content licenses has helped in establishing a favorable intellectual property framework for collaborative content creation.

The collaboration platforms reported here have generally been configured to prevent anonymous content contribution. The requirement to have an account and to login before contributions can be made was a deliberate choice meant to stimulate communication among contributors, which requires that the source of contribution is known. We believe that such a setup helps to promote high quality contributions and professional interactions. This seems to be very acceptable, as it usually reflects the institutional work environments of the contributors.

Online collaboration systems often create a very transparent environment by preserving the history of every contribution and allowing every user to review the content creation process over time. This is a very important resource for new members to familiarise themselves with the history of a particular content object and the main contributors. A transparent environment also facilitates

attribution and recognition, which may be important in institutional environments where authorship may be part of the recognition and reward system.

Even in an institutional environment, content in online collaborative workspaces will develop unevenly in terms of quality, coverage and maintenance. Depending on the particular point of view, orphaned or patchy content may be considered an asset or a liability. The content of collaborative workspaces should be considered as work in progress and a disclaimer on the site should clarify the nature of the content and possibly point to sites where the finished, quality-reviewed products may be found. We found that exposing such work in progress on the Internet seems to cause the most discomfort in a quality-conscious institutional environment.

## Links and acronyms

- IRRI - International Rice Research Institute <<http://www.irri.org>>
- GCP - Generation Challenge Programme <<http://generationcp.org>>
- CGIAR - Consultative Group on International Agricultural Research <<http://cgiar.org>>
- SourceForge <<http://sourceforge.net>>
- Wikipedia <<http://en.wikipedia.org>>
- GForge <<http://gforge.org>>
- MediaWiki <<http://www.mediawiki.org>>
- CropForge <<http://cropforge.org>>
- CropWiki <<http://cropwiki.irri.org>>
- ICIS - International Crop Information System <<http://www.icis.cgiar.org>>
- ICIS communication project <<http://cropforge.org/projects/iciscomm>>
- Generation Challenge Programme pantheon project <<http://cropforge.org/projects/pantheon>>
- GCPWiki <<http://cropwiki.irri.org/gcp>>
- ICISWiki <<http://cropwiki.irri.org/icis>>

© IRRI, 2007

This work is published under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License. <http://creativecommons.org/licenses/by-nc-nd/3.0/>