

Data release email message sent by J-M Ribaut to PIs on 9th September 2009

Dear GCP partners,

As I am sure you can appreciate, one of the major challenges of a big network such as GCP is the quality control, the release, be it via our GCP Central Registry or through another public database, and the proper documentation of the data produced by the projects you are leading.

Every year we have a portfolio of about 70 projects involving more than 100 institutions.

Data quality and availability is one of the most important indicators of success for the Programme, and our reviewers consider this parameter very carefully. As such, we must ensure that we are managing our data in a proper and efficient way. We need, therefore, to work together to ensure that GCP fulfills its obligation vis-à-vis the scientific and donor communities but without making your life too complicated through unnecessarily complex or unclear procedures.

Currently our Central Registry is not as complete as expected and some projects have not been closed because data has not been released. GCP procedures have not been followed, possibly due to difficulty in using the templates or to lack of time on the part of PIs or both.

The Management Team is ready to facilitate the process of collecting this data and the objective of this message is really to ask for your support so that we can now move forward together. In this context, please let me remind you of some principles and clarify some procedures:

1) Data release

- It is a requirement to indicate in all GCP project proposals the nature and the format of the data that will be produced.
- Those data must be described on the Central Registry and released, in the registry or in a public database before the end of the project.
- Released data must be properly documented – to the level where external users can understand the context, quality and limitations of the data for secondary use. Templates are available to indicate the necessary level of documentation and convenient formats.
- Data released to public databases must be described in the CR and be reachable from the CR (via web services or specific links to those databases). Public Databases are specialist databases which add value to GCP data by integrating them with other public data and adding context, analysis, or interpretation not available in the simple template format. These databases

must have institutional support and commitment for long term conservation; links to datafiles on project web pages are not sufficient.

- You may ask for a grace period of a maximum of six months beyond the official end-date of the project, in the event that there is a need for data polishing or if you are facing confidentiality issues because data have not yet been published.
- During that grace period, data should be stored in a restricted area of our Central Registry or be made accessible to us through your institutional website.
- GCP will not release the 20% of the final year's funds and SP Leaders will not approve project final technical reports without the release of expected data.

2) Quality control

- It is your responsibility to ensure that all data produced in your respective projects, by you and collaborators, are of acceptable quality and are well documented so that they can be accessed and used by scientists not involved in your projects.
- It is the responsibility of the SP Leaders to ensure that expected data sets are available before they approve technical reports. SP Leaders will focus on the nature and the presentation of the data, verifying that data sets match expectations, but you should not expect SP Leaders to conduct a detailed review of your data quality.
- As is common practice in crop DBs, the quality of data posted to the GCP Central Registry is under your responsibility, as project PIs.
- For data related to products, including germplasm that will be distributed through the GCP platform, the Management Team may request independent quality checks on the data.

3) Data templates and trouble-shooting

- For data to be user-friendly, accessible and well-documented, good data templates are a must. GCP templates are available via our bioinformatics portal (<http://www.generationcp.org/bioinformatics.php>).
- The MT realizes that filling these templates might be challenging at the beginning, so we support a half-time position to help with this task.
- **In case you have problems loading your data to the GCP Central registry please, contact: Elisabeth Van Strien at sp4helpdesk@generationcp.org.**
- Elisabeth will not only help you to load your data, but will also collect feedback on the templates and suggestions on how to implement them.

- Good templates are not generated over night and we need your input to help us improve their shape and content!
- If suggestions are sound, Elisabeth will pass the information to the GCP bioinformaticians who will then modify the templates accordingly.
- Contact the SP Leader who supervises your project if you are still facing problems with your data.

In short, we must have a **good** and **complete** repository, and you are the main actors to help us achieve this objective. So please, take the time to upload and document your data sets. Our Central Registry is the showcase for research conducted by GCP scientists, and we should all be proud of the information and data sets contained therein.

Please, let me know if you have any comment or if anything is unclear, and feel free to share this message with your Co-PIs and collaborators.

Kind regards,

Jean-Marcel Ribaut

On behalf of the GCP Management Team