

GCP ARM 2009
Brainstorming I—Mon 21st Sep (1330–1500)
(Coordinator: G McLaren)

How to Standardize Information Management for the MBP

A BACKGROUND:

One of the key requirements of molecular breeding, and hence one of the key elements of the Platform concerns Data and Information Management. The overall strategy adopted by the project is to have one objective (2.1) making sure that existing informatics tools are available to users for logistical operations and pedigree, genotype and phenotype data management. Objective 2.2 will add analytic and decision support tools to that information system, and Objective 2.3 integrates public breeding data and develops the next generation of logistical and analysis tools in a configurable workflow system.

One key activity in this chain is the interaction with users to ensure that their data management is of appropriate quality, that their data can be shared as required and that they can use the analysis and decision support applications being developed.

The key problem is that many users and use cases have existing partial or complete information systems which may or may not be completely shareable with other partners in the same breeding project, and for various, legitimate reasons these users may be unwilling to change to the platform recommended information system.

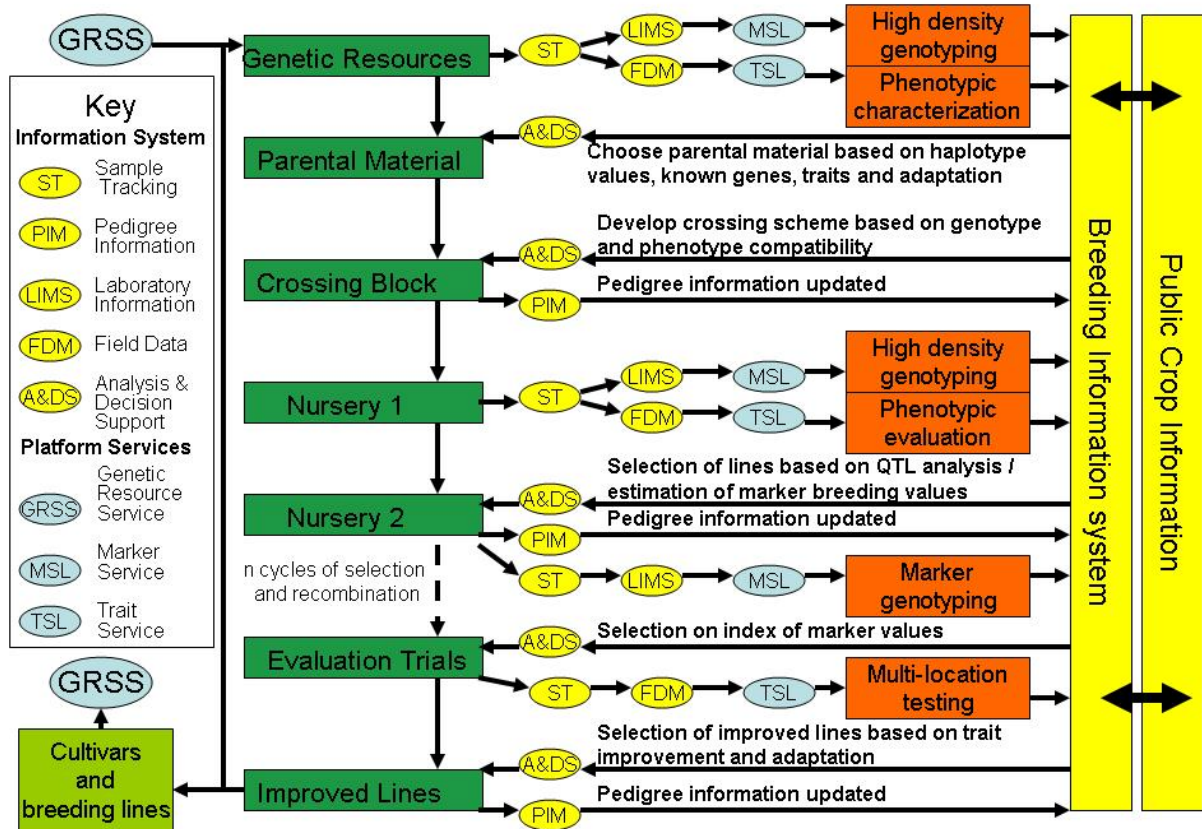
In order to develop a workable modus operandi for this activity we set out some basic principles:

- Users and developers recognise the complexity of data management for molecular breeding which requires high standards of sample and pedigree identification as well as integration of field and lab data in ways that are much more demanding and much less forgiving than in conventional breeding.
- Data quality for molecular breeding is of paramount importance because large investments in lab and field characterization follow on a relatively small number of genotypes identified by analysis of primary data.
- Users of the platform may have existing data management systems which need to be accommodated provided they are of sufficient quality and capacity and need to be changed or replaced if they are not.
- Data management needs to be compatible across all members working on the same breeding project.
- Genotypic and phenotypic data collected during breeding processes has immediate value for breeders to make selections, and also cumulative value over years and populations. Users of the platform will be able to access this public cumulative information from many breeders, and will also be willing to contribute breeding data to these public crop databases.

- Use of platform analysis and decision support tools supported by the platform and sharing of data across platform partners will require data to be formatted and stored in defined ways which may have to be accommodated by users with existing systems.
- The platform provides services to train users on how to manage and use the Platform Information System or on how to format data to be compatible with platform supported tools and sharing. These services are available at cost to users so that they can be a sustainable part of the platform.

Now we have in Figure 1 a diagram indicating points of interaction between users and the MBP. It is the yellow interactions which concern us at this time.

Figure 1. Interaction of breeding workflow and platform elements



For each activity in the information pathway, there are a number of specific requirements to ensure complete data recording and facilitate data integration. The MBP will provide an information system and user applications to meet these requirements, but in cases where users need to use different systems data must also be available in specified template formats so that it is accessible to analytical tools and available for data sharing

and integration with public crop information. We can consider each activity in the information pathway in turn:

1. Sample Tracking (ST).

This is the process of making a list of samples of germplasm (seed, plant structures, leaf samples, DNA etc.) for treatment, analysis, shipment, storage, characterization, evaluation or any other reason. The key features of the process are that a list of germplasm is developed with unique identifiers tightly linked to local identifiers – usually sample or entry numbers and to source locators (where the sample came from). It is essential that a system is used which allows the automatic production of labels which can be attached to physical samples and these labels may have bar codes or not. The important feature of the labels is to uniquely identify the list and the sample number. They may also have the germplasm identifiers and other useful information.

The most frequently used application for sample tracking in plant breeding is the excel spreadsheet, and while this may be convenient and even effective for data collection it is not recommended for the formation of sample lists because this invariably involves cutting and pasting and all the attendant disastrous errors associated with those functions. The applications supported by the Platform will be the SetGen application of ICIS and the maize Workbook application from CIMMYT which, although Excel based, produces Seed Prep lists via programmed algorithms.

Users wishing to use other applications need to be able to produce data files (usually Excel) for the Platform following a Germplasm Template which will contain header information on purpose, person responsible, date and place, and then for each entry, Sample Number, Unique Germplasm ID, Designation (indicating the development history of the germplasm), Pedigree (indicating the recent cross history of the germplasm), Seed Source.

2. Pedigree Information management (PIM)

The management of Pedigree Information is a recursive process through generations. Many Pedigree management systems consist only of a designation encoding recent development history and a source record in breeders' fieldbooks which allow a breeder to chain back from book to book to the founders of a particular line. These systems are often managed by cut, past and update and are prone to mistakes and 'abbreviations'. Even when flawlessly managed, the worst part about these systems is that they are not 'computable' that is, not amenable to computer analysis. Modern breeding decision support tools are now able to mine information from relatives in many ways, but this depends on the pedigrees being managed in a computable format. The first step in this process is to separate pedigree management from naming conventions.

The Platform will support the ICIS Genealogy Management System (GMS) which supports remote allocation of unique germplasm identifiers and fully computable pedigree information as well as considerable freedom in naming and annotating germplasm. It tracks methodology and chronology of germplasm genesis in a fully

computable way. The key application for this process is the SetGen application that is also used for sample tracking.

The Platform will provide a service to train users in the use of the GMS and SetGen as well as in the process of capturing historical pedigrees for those users who wish to update their Pedigree Information management. Users wishing to use other systems need to be able to deliver germplasm lists following the Germplasm Template as described above in order to use the platform analysis and decision support tools and to share breeding information. This template can be parsed into GMS with at least one crossing generation of history.

3. Management of Genotyping Data (LIMS)

The problem of managing genotyping data starts with a germplasm list and continues with a sample list identifying the biological material for each sample (leaf, seed etc.) and a DNA list. All these processes depend on lists as defined in 1. The DNA list must then be mapped onto the laboratory protocols (plates, tubes, wells etc.) in a Laboratory Workbook. Workbooks have the feature of containing a field indicating the destination of the sample. One complexity of this last step is that repeats (replicates) and standards may be introduced. The data collection fields must also be fully annotated with a description of the property being measured (eg marker) the scale or units (eg bp or band) and the method or protocol being used.

Once in the laboratory, the LIMS takes over and ensures that data collected in the lab emerges connected to the correct sample number, and that the process of collecting the data is traceable and verifiable for quality assurance purposes.

The list processing for entry into the laboratory can all be managed with the ICIS list processing tool SetGen. The ICIS workbook supports the process of transforming the sample lists into Laboratory Workbooks. Most analyses supported by the Platform will be conducted through contracted laboratories, but if users wish to conduct their own analyses, the Platform will support the ICRISAT LIMS with help in customizing, installing and managing it.

Users wishing to manage genotyping data with external systems must be able to deliver sample lists in the Germplasm Template, and genotyping data relating to those lists in a Genotyping Template derived from the Laboratory Workbook which will allow automatic loading of the genotyping data into the Breeding Information System with computable links to germplasm identifiers.

4. Management of Phenotyping Data or Field Data Management (FDM)

Phenotyping Data is probably the most complex breeding data to manage because of its diversity, the complexity of measurement protocols and the difficulty of managing field conditions. Once again the data collecting process starts with a Germplasm List. In this case there is usually no sample list derived from it, but there is a Fieldbook which assigns Germplasm samples to field plots where they will be grown for characterization or evaluation. Replications and checks are usually introduced and there is often an experimental design and randomization to be accommodated. In

addition the fieldbook contains columns for data to be collected and this data must be fully documented with Trait, Scale (units) and method.

The ICIS Workbook and the Maize Fieldbook both manage this process of forming fieldbooks from lists and these will be supported by the Platform. Users wishing to use a different system must be able to deliver phenotyping data in a Phenotyping Template that allows automatic loading of the data into the Breeding Information System with computable links to germplasm identifiers so that genotype and phenotype data can be linked and analysed together.

5. Analysis and Decision Support (A&DS)

Analysis and decision support tools will be developed to assist breeders in selecting parents by extracting and displaying trait and genotype information about candidate parents from public and private plant breeding databases. This information will be combined with mapping information and knowledge on adaptation. Other tools will assist breeders in selecting lines from genotyped breeding populations according to genotype and phenotype information for line development or recombination. Tactical simulation tools and identity by descent algorithms will be included in these decision support tools.

A key feature of the tools is the need to access to diverse but integrated breeding data and this will be accomplished by accessing the local and public breeding databases. For this reason primary data must be available in a format suitable for loading into these databases. The Platform will provide training in the use of these formats, the loading of the data and the use of the decision support tools. The tools themselves will be developed using the GCP data consumer interface and hence will be able to access data in any GCP compliant database, however the Platform will concentrate on access to data stored in ICIS databases.

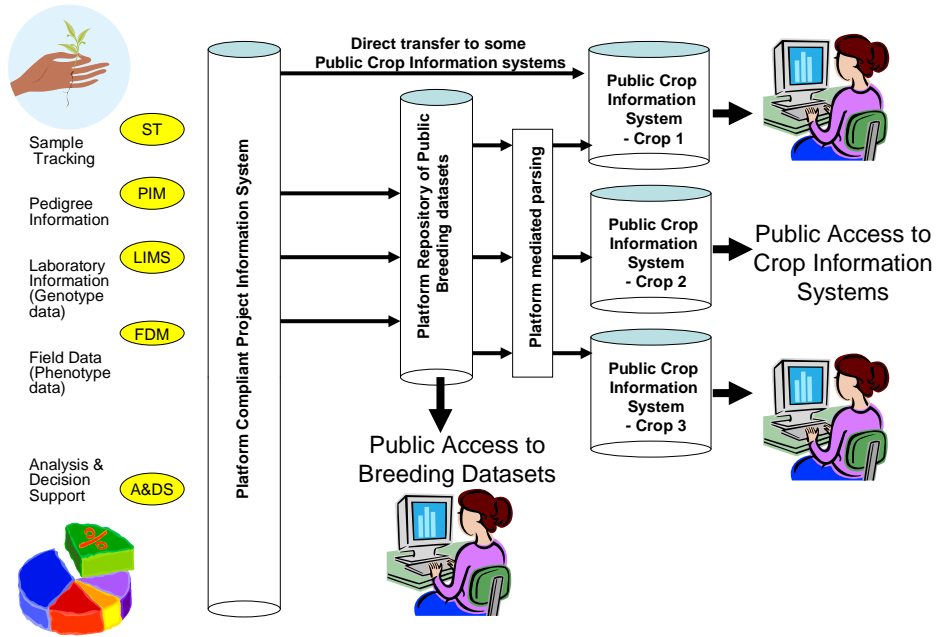
6. Public Crop Information

Users of the platform will deposit primary breeding data – pedigree information, genotypic data and phenotypic data as individual datasets in appropriate template formats to a platform repository. These datasets will be publically available (after a suitable embargo period if necessary).

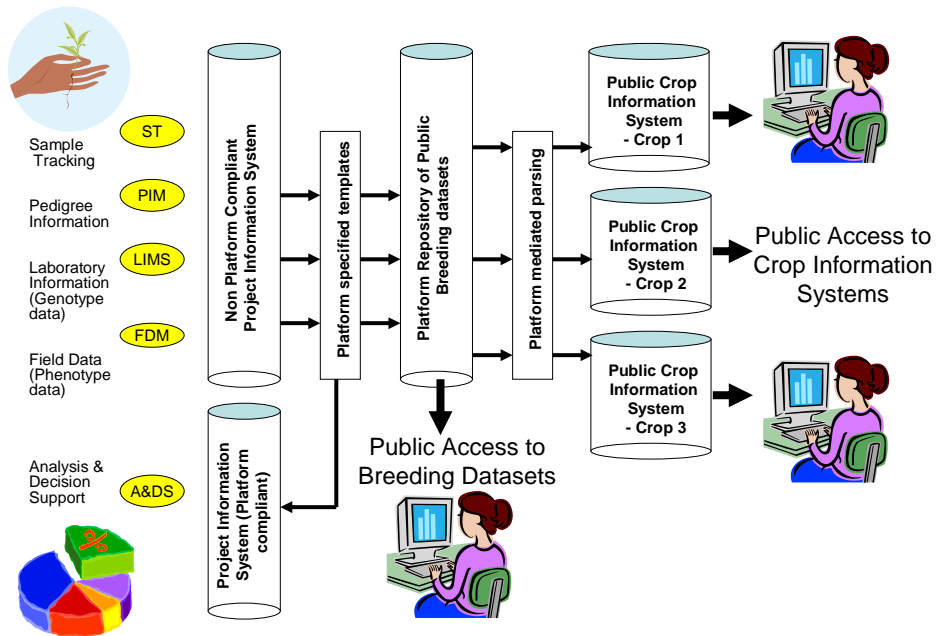
However the main power of cumulative historical data is to have it integrated through pedigrees in a Crop Information System which can be queried by germplasm, genes or traits or mined for associations which only become apparent when the quantity of data available becomes significant. In order to achieve this integration and access the Platform will partner with Institutions willing to take on the long-term curation of public information for particular crops. Curators will then be assisted in the loading of public platform data into these information systems with recognition and attribution for the breeders contributing this data and with quality assurance provided by Platform procedures.

The procedures outlined in these six cases are summarized in Figure 2 which indicates the essential differences in data management which will enable use on non-platform information systems while maintaining compatibility with platform data standards and analytical tools.

Figure 2. Molecular Breeding Platform Data Flows:
a) For users with Platform Compliant Breeding Information Systems



b) For users with non-compliant Breeding Information Systems



B DISCUSSION POINTS:

1) The problem is that there are decision support tools in MBP that we want users to adopt but they are be tied up with a particular information system (ICIS) but some users have their own favourite system.

Martin: Isn't this integration the GCP platform intention?

Graham: Yes, but it is still a problem now since most systems are not wrapped as ICIS is wrapped through a layer.

Martin: This layer exists

Anthony: Templates are also constructed so we have 1) templates 2) ICIS 3) data submitted to central registry. Are the data submitted in the data registry stored in templates?

Graham: When the GCP started, all partners are collecting various data that would be analyzed in a central way. But we have 15 institutes that were involved with only part of their time for GCP and they are not willing to change their system. We decided to develop a platform that has a domain model and has data sources that will use the domain model and a transformer to use these data and render to their own visualization or viewing tool.

But we still need to publish the data accumulated through the GCP projects and hence the central registry. But there was also need to document the data in a standard form. Hence, we decided to have templates. However there were people who didn't submit because their data do not comply with the templates.

The central registry is meant to have a home for data that are not managed by a system. Wheat genotyping datasets are examples that were initially submitted to the crop registry but where taken out because they are stored in IWIS now.

Anthony: Where is the configurable workflow now? Can we insist that users use this system?

Graham: Yes, we can use the approach that if you want to use our decision tool, then you have to use the database that we are using but we will exclude a lot of users. Hence we need a more accommodating approach and I suggest an approach based on templates.

The question today is users wishing to use other applications need to be able to produce data files following a template.

Analysis decision support tools assume that the data come in a particular format because if they are not, it will require efforts from us to make it compatible.

Martin: The template is good thing because it solved some of the problems we faced. People can put data in the template and just create a data source using that template.

Shawn: Services are cost-recovery. So, it is possible to hire a person who can wrap a system of the users.

2) Are breeders willing to share and publish their breeding data.

Graham specified the basic principles for information management. Set out in the document above.

Some breeders don't agree that cumulative information has value. But they are willing to share and publish their own data.

How should this data be published?

Graham: GCP will approach each institution which has a mandate for a certain crop and ask them to include the MBP data, with attribution to the researchers, and manage them together with the rest of their central data. For example:

ICIS-based: Rice, wheat, maize

Non-ICIS based: beans, chickpea, cowpea, sorghum.

3) How should the Platform Interact with different types of users?

There are two main groups of users:

- Users with Platform-compliant system
- Users with non-platform compliant system.

For the second class of users, data can be taken in the form of templates. And we can create a wrapper to this template to integrate with the platform.

Anthony: I think we need to step back. One issue is if generating a simple template does not encourage users to use it, is it because of advertizing. A second issue is the wrapper to the template. The third issue is if users are not using the template, then what is the point of developing them?

Graham: I think the issue about the template is not because of the template per se. In standard GCP projects, the data are a key product but in a breeding project, the data will be used for decision support and if this is only available through the template people will use it. The agreement is that the user will not write the wrapper but if we can agree about the format, we can write a wrapper for it.

Martin: The reason why users are not using the database or template is maybe because they have small, simple datasets but as the data become enormous in molecular breeding, they will need a system to handle it. The users didn't find the need because they don't deal with complicated data yet.

Guy: If you want to use the platform, then provide money to wrap your own system.

Graham: The solution I am proposing with low entry is asking users to fit their data in a set of templates.

Anthony: How many templates do we have now?

Graham: For breeding tasks, maybe it is based on the breeding system they are using: Type of marker, SSR, AFLP, SNP and then pedigree data, phenotype data and evaluation data.

Mike: I think the complication is not from the users but from the end part which is the institution who will adopt the data.

Martin: I think ICRISAT will develop a data source because it is tasked to do it.

Guy: I think one issue is how many tools the users want to use from the platform. If they are using most of the tools, then maybe it is better to wrap their system to the platform.

Jean-Marcel: I think the institution should be asked to enforce the use of electronic way of entering data.

Guy: I don't want to waste time formulating a generic template for everyone. But maybe a template for each crop is possible.

Jean-Marcel: The analytic pipeline is not crop-specific

Shawn: But if the MBP offers a service to wrap users existing systems then users can pay for the conversion , they just need it for the initial start and then they can use the system.

Guy: Who should be responsible to interact with those use cases?

Graham: The Platform manager will maintain the portal and he/she and the informatics coordinator will deal with users and the developers.