A Strategy for Plant Breeding Data Management in International Agricultural Research

Introduction

Exchange of germplasm boosted crop improvement for subsistence agriculture during the 70s and 80s, and now we have the opportunity of integrating molecular techniques into breeding projects to obtain a boost for the new millennium. However these techniques are more demanding of information management practices and are most effective when information is exchanged in addition to germplasm.

We examine desirable information flows for crop improvement in international agricultural research and consider roles of partners and resources required to achieve the desired level of information management and sharing.

In Figure 1 the basic structure of crop improvement projects in international agricultural research is depicted. Breeding projects at the bottom level are implemented on the ground by Advanced Research Institutes (ARIs), National Agricultural Research Systems (NARS), Networks of ARIs and NARS and by Small and Medium Sized Enterprises (SMEs).

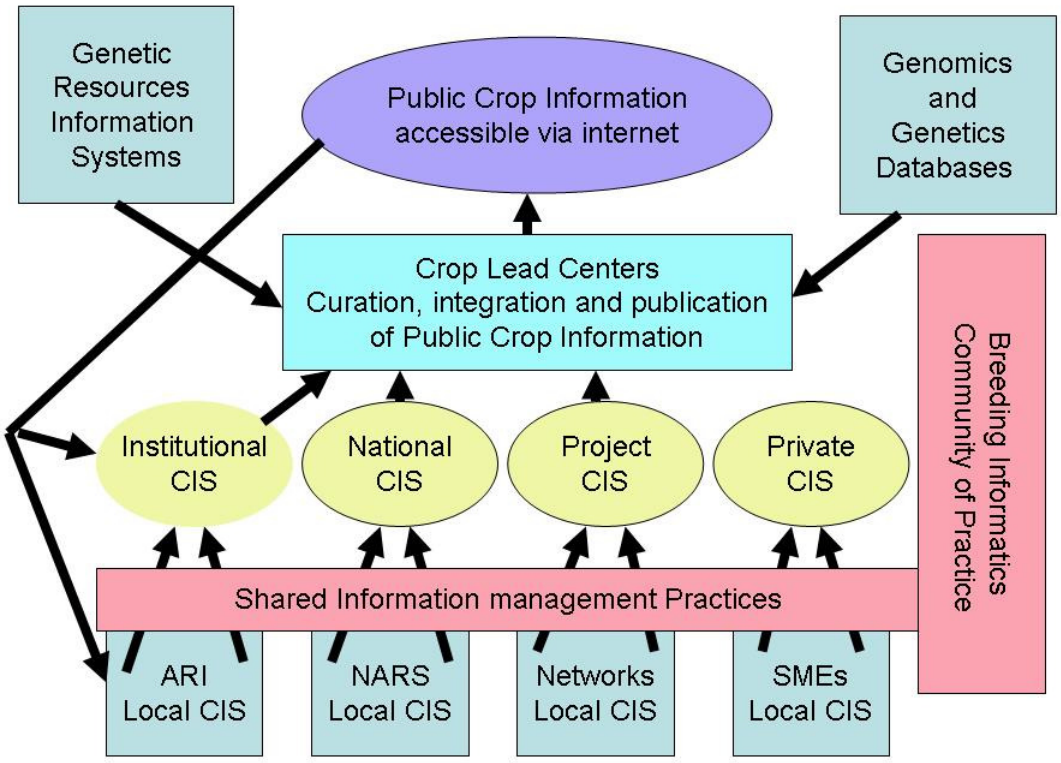

Figure 1. Flows of Information in Plant Breeding in International Agricultural Research

Each breeding project requires a local crop information system (CIS) which are generally not very sophisticated at present and which need to be upgraded to take full advantage of new breeding technologies. There are numerous 'best practices' for breeding information management which should be shared across all breeding projects and partners. These range from generic procedures for managing pedigrees, nurseries, and characterization and evaluation trials to crop specific practices for propagation, trait evaluation and, in some cases, germplasm nomenclature.

Most breeding projects do not, and should not, operate in isolation, they are conducted by a group of partners from the same or different types of institution. These partners need to share information with each other, but will often not be willing to share information openly, at least their most recent information. Hence there is a need for project level crop information systems – institutional, national, network or private company systems. These are easiest to assemble and maintain if all the partners use common standards and best practices and they should also integrate with global public information for each crop to be most effective.

Just as exchange of germplasm between projects, institutes and countries was the key to sustained crop improvement earlier, the same broad exchange of information is required now. Just as the international centers of the CGAIR provided the focus, technology, and honest brokerage for the exchange of germplasm they have the mandate and responsibility to do the same for information now. This can be achieved by Crop Lead Centers for each crop taking the responsibility to develop and promote effective breeding technology and best practices for breeding logistics and information management for each crop and for curating, updating and publishing public crop information from all sources. These are part of the international mandate they hold for each crop.

Modern breeding technology must be derived from genomics and genetics research conducted in laboratories around the world with knowledge disseminated through publications and genomics and genetics databases. Modern technologies are also facilitating the use of genetic resources and the information which enables this use also needs to be integrated and disseminated as public crop information. The CLCs are again in the key position to facilitate this.

In order to develop, maintain and disseminate these crop informatics best practices, some organizational group needs to facilitate communication, training and knowledge management for each crop – some sort of breeding informatics community of practice (COP).

Information management and flow at the project level requires the participation of several key actors. Figure 2 illustrates a typical multi-partner breeding project in international agricultural research.

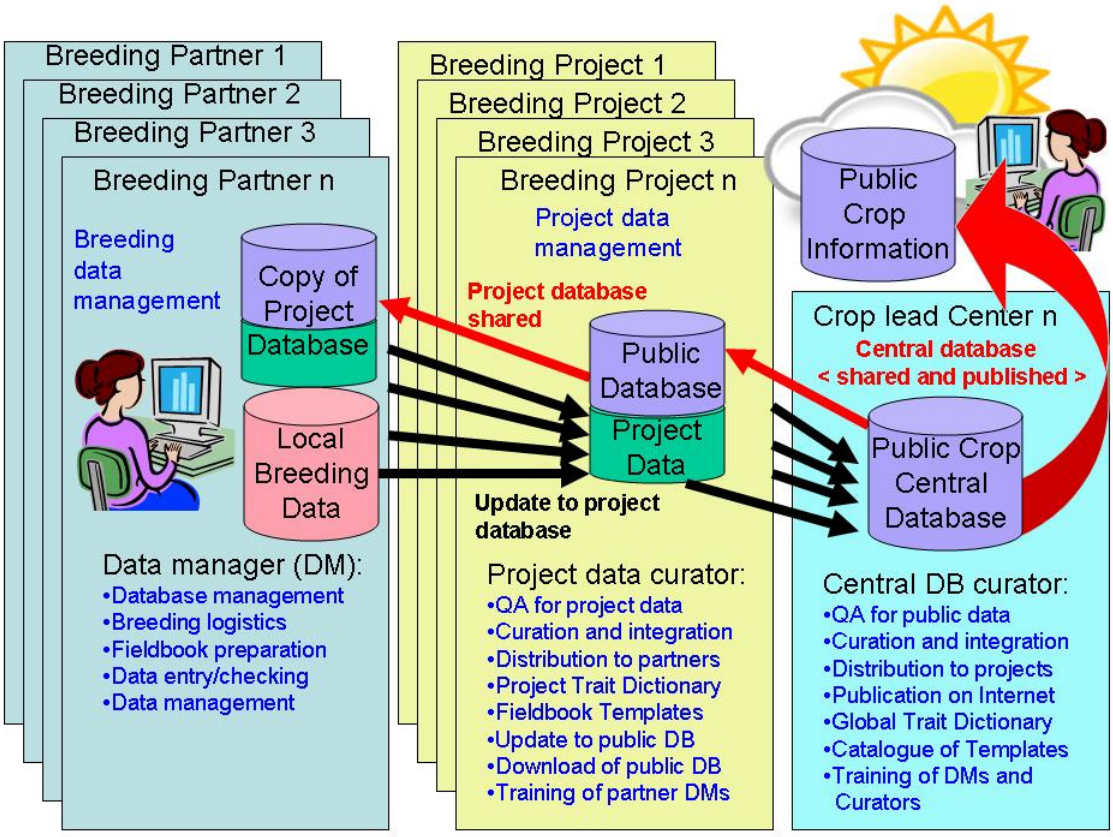

Figure 2. Informatics activities and information flow in a typical multi-partner breeding project in international agricultural research.

It is clear that having the middle layer of project data management is a great complexity, and where possible it should be avoided so that data flows directly from the local breeding databases to the Crop Central Database with the project curation done at the same time as the central curation. Most International Centers should work like this. However when external partners join the breeding effort, sharing data outside the partnership is very often a major issue so the ability to handle the intermediate layer is essential for collaborative breeding projects.

Examples

An example which fit this model well can be found in international Wheat breeding. The India – Australia Molecular Wheat Breeding Project sponsored by ACIAR, has as a strategic objective the deployment of molecular technologies in wheat breeding across several institutes in India and as practical objectives the exchange of wheat germplasm and technology between the breeding programs in India and the wheat improvement program of Sydney University in Australia. Partners are The Directorate of Wheat Research in Karnal, Pubjab Agricultural University in Ludhiana in India and the Plant Breeding Institute at Cobbitty in New South Whales, Australia. The project data curator

is at the National Research Center on Plant Biotechnology (NRCPB) in Delhi, and the central curator for public wheat information is at CIMMYT in Mexico.

Another example is the set-up of the information management of the Rice Challenge Initiative of GCP. The rice information is centrally managed in IRRI, Philippines while the project data of the Rice CI is managed at AfricaRice, Benin. Nomenclature and standards for the germplasm and traits of the project are established in AfricaRice. Germplasm lists submitted by the partners are curated and fieldbooks for the projects are created there. These are transmitted to the partners namely, INERA (Burkina Faso), IER (Mali) and NCRI (Nigeria) for data collection. Then, the data are integrated by the project data curator at AfricaRice. As partners becomes more acquainted with the system through orientation and training, the curation of the germplasm lists and field books can then be transferred to them.

Breeding Partners

Each partner in a modern breeding project requires manpower and resources dedicated to breeding data management. In a small unit this may be the breeder, but in larger ones and in research institutes it may be an informatics specialist dedicated part or full time to the project. In some cases a project data manager may take on the data management tasks for each partner as well as the project data curation. In any case the key responsibilities of the data manager, at the partner level, are:

- Database management – putting together the project database and the local CIS so that the informatics applications which support breeding operations operate on the full database.
- Breeding logistics – preparation of germplasm lists, labels, seed management and inventory management.
- Fieldbook preparation – integration of germplasm lists, trait templates and field designs into fieldbooks which can be printed or loaded on data collection devices.
- Data entry and checking – entering and checking pedigree, characterization and evaluation data into the local CIS.
- Data management – loading, querying, reporting data for breeding operations and transferring data to the project data curator for integration and sharing with partners.

Project data curators

At the project level, informatics tasks are more involved with setting standards and coordinating information flows. Each project needs to have dedicated resources for project data curation and a project data curator (who may also take the role of partner data manager in some cases) has the following responsibilities:

- Quality Assurance for project data. This involves verifying germplasm sample identity for lab and field measurements, ensuring that collected data are with valid ranges, accurately and completely recording environmental and experimental conditions and running statistical procedures to identify outliers and suspicious values.

- Curation and integration of project data. Collecting pedigree, characterization and evaluation data from all partners and integrating it into a single data resource which can be shared with all partners in the project.
- Distribution of integrated project data to partners. Once data has been integrated it must be shared with partners via databases and reports.
- Preparation and maintenance of a project trait dictionary. Ensure that all traits, protocols and scales used in the project are agreed, understood and followed by all partners.
- Preparation and maintenance of Fieldbook templates. Agree on the nurseries and trials which will be conducted by each partner and develop the design and trait templates for the fieldbooks which will be used to collect the data.
- Update of released project data to the public crop database. Agree with partners which data can be published and when and make this data available with quality assurance and intellectual property status to the central crop curator.
- Download of public information and integration with project data. Obtain updated central databases from the central crop curator and make these available to project partners, integrated with project data where possible.
- Training of partner data managers in use of tools and standards. Prepare training materials and run training courses for partner data managers to ensure efficient use of informatics tools and accurate and efficient collection of data.

Crop Lead Centers

Each crop needs some coordinating body to develop and promote breeding best practices, to focus the community on maintaining an up-to-date trait dictionary and integrating and publishing public crop information from all relevant sources. This should be coordinated by a crop information curator at a Crop Lead Center with the following responsibilities:

- Quality Assurance for public crop information. The key issue here is attribution of all public data to the owners and producers of that data who must be responsible for its quality. At the curator level, the responsibility is to develop, maintain and publish best practices for quality control and quality assurance.
- Curation and integration of data from different sources. The curator is responsible for identifying connections and conflicts in information from different projects and resolving these with partners before integrating data from different sources into a single data resource.
- Distribution of public information to projects once data has been curated and integrated.
- Publication on Internet of public crop information to ensure access to the widest possible audience of researchers interested in improving staple crops,
- Maintain catalogues and supporting information on best breeding practices for the mandate crop, including effective markers, and breeding strategies coming out of genomics and genetics research and validation.
- Maintain a central trait dictionary with all the traits, scales and protocols gathered and integrated from projects and communities. Integrate trait ontologies with the ontologies of the broader biological research community.

- Maintain a catalogue of public trait templates which can be used by any researchers and breeders needing to implement field trials for breeding or physiology
- Coordinate and participate in training of project data curators and partner data managers so that they have the capacity to implement best practices, use the facilitating applications and contribute quality data and information to the crop improvement community.

Some of these functions are specific to particular crops, but many are generic across any crop, and some central platform designed to support these generic activities and technologies would be the most efficient way of supporting crop improvement for food security and development.

Technical considerations

Crop specific breeding technology such as marker systems and validated marker-trait associations, or methodology for DNA sampling or rapid generation advance are important for efficient collaboration and scaling up of crop improvement. CLCs are conducting research leading to this breeding methodology and they need to make the results of their research as well as results of relevant research for other laboratories available to the breeding community in ways that are accessible to breeders. This includes integrating information from genomic and genetic information resources in the crop information they publish.

Providing access to genetic resources is a major responsibility of CLCs, who generally also manage the largest public collections of resources for their mandate crops. Recent initiatives to develop crop registries across all germplasm collections is a major new task for CLCs, and it requires an information system which can manage complex relationships between seed samples in different genebanks. Integrating the information about these genetic resources is a critical element is providing facilitated access. This integration is greatly facilitated by using compatible information systems for managing genetic resources and breeding information.

The key technical constraint to the efficient management of crop information across the layers of implementation is standardization and consistency. At the crop level, the most important key to data integration is a community accepted trait dictionary – an ontology of traits of interest for each crop together with a set of effective protocols for their evaluation including scales or units of measurements and data quality standards.

At a more generic level is the need for efficient information systems which support sample tracking, pedigree management, breeding logistics, characterization and field evaluation are critical. The more standardized these systems, the easier is the process of data integration, firstly from partners to projects, then from projects to CLCs and finally for the publication of crop information from many crops. Systems must support the unique identification of germplasm (unique across all projects if possible), a computable pedigree management system or a rigorously implemented germplasm nomenclature system which can be parsed into pedigree relationships, a nursery and fieldbook system

which uses the trait dictionary to unambiguously annotate data records. If a common system is not used by most partners the role of the project data curator and the central database curator become infinitely more complicated and expensive.

The curation of public crop information is a task requiring significant skill and time. Curators need to know something about the biology and breeding methods of the crop, and they need to be skilled in data management. The introduction of a new source of data is the most difficult, and the difficulty depends on the quality of data management which has been practiced in the past. Once a source is integrated, keeping it up to date is a much easier task. Historical data tends to be held in a variety of formats ranging from old field notebooks to computer files held on magnetic tapes. Decisions need to be made about the value of converting this historical data into modern formats. The range of options is from complete conversion to just starting from scratch with current material. The appropriate decision depends on the quality and consistency of the historical records as well as their relevance to modern breeding approaches.

The question of which breeding data should be published is often raised, and it may depend of particular crops to some extent, but in general, any data that has been collected, that is in good enough shape to publish cheaply should probably be published. If data is well collected and curated for its primary use, it costs very little to publish. The most important data to publish is pedigree data going back to genetic resources – particularly public genetic resources. Secondly, good characterization data anywhere along the pedigrees – morphological, agronomic and pathological (anything that researchers spent time and money collecting – volumes are not large and getting it is expensive). Molecular characterization should also be published, particularly where trait associations are known. Volumes and negligible at the moment, but if ultra high throughput genotyping is used then volumes will be an issue, but presumably the point of collecting it is to make it available. Finally, and possibly most important, is evaluation data on promising lines, with associated environmental context so that adaptation can be determined. This should include breeders' opinions where possible.

Data quality is a critical element and one way to address the issue is to ensure that all public data is unambiguously attributed to the owners and collectors of that data. The reputation of these owners and collectors is then the first level of quality assurance. In addition to this there may be analytical processes which can be used to evaluate data quality and there could even be an element of peer review by recording the degree to which public datasets are used by external users.

Hardware and Software Requirements

Computer hardware for managing breeding information is not specialized, at least not at the current level of throughput that public crop improvement programs are working. At the partner level ordinary personal computers are sufficient although attention needs to be paid to label printing, bar-code reading, hand-held data capture devices and automatic data capture form scales and other measurement instruments. The support and management of this array of instruments is not a trivial task and must be adequately resourced.

Current crop databases run between 50MB and 2GB so it is still possible to host them on a personal computer although more high powered relational database management systems (RDBMS) than the conventional standard of Microsoft Access are becoming necessary at the top end. However this is still quite manageable with free tools like MySQL, these tools do take more technical knowledge to install and maintain.

At the project and center level ordinary servers are currently adequate for managing breeding data which tends to be complex but not voluminous. The thing that will change this in the not too distant future is the deployment of ultra high throughput genotyping and phenotyping which is beginning to make an impact in large commercial breeding operations. This includes genome-wide characterization of breeding lines by next generation sequencing, and deployment of automatic imaging systems for high throughput phenotyping. It is not clear what the computing infrastructure implications of this will be, but probably large data files will be stored in the cloud on remote mass storage devices and analysis tools will be moved to the data rather than the reverse. This infrastructure will be available for rent on a commodity basis like electricity, gas or other utilities.

For the publication of crop information, facilities with good internet access are required although, as above, massive storage is not yet necessary. The complexity of crop information – highly structured in terms of germplasm, environment and management contexts requires efficient data warehousing and search algorithms to provide responsive query interfaces. These are difficult and expensive to build and maintain, and shared technology across different crops has obvious efficiencies.

Software required to support breeding informatics covers the full range from RDBMS, breeding applications, data capture software, analysis software and decision support tools. These are available from commercial vendors or they can be obtained freely, and sometimes with open source code, for the cost of maintenance and development. It is difficult to make a cost comparison of the two approaches, but it is certain that if independent systems are used for each crop, institute or breeding project the cost will be substantial. The use of common strategies and tools offers a clear opportunity for economies of scale for development, maintenance, support and training.

Human and Financial Resources

Every breeding partner and every breeding project needs to have human and financial resources dedicated to information management. The exact level of resources required is not easy to calculate, and it depends on the technical skill of people available. However in a project with four partners, each managing 2000 field plots a cycle (at all stages of breeding) it would not be unreasonable to deploy three data technicians, one as the project data coordinator and one half-time technician with each partner. These technicians would carry out the functions outlined above for Breeding Partners and Project Data Curators. While breeders themselves may do some of these tasks, they are still expected to undertake the data analysis, interpretation and decision making.

At the crop lead center level, the situation is more complicated. Breeding projects carried out by the CLCs need resources as described above, but the further tasks of curation, integration and publication of public crop information as well as coordinating the community development of trait dictionaries and information on breeding technology and training project data curators and partner data managers require significant time of crop information specialists. For important staple crops, with large breeding communities a full time crop information specialist would be required with some level of technical support from programmers, data managers and analysts. Such a team of three full time equivalent staff could also manage more than one of the smaller crops, particularly if common technology is used for all the crops. The size of this curation team also dictated the speed at which historical data can be integrated and published and the number of data sources which can be handled at a particular time.

Apart from the human resources and computing infrastructure discussed above, financial resources are needed for community and training activities. Whether these should be deployed in a top down fashion or from partners up is debatable, but there is a need for training courses for breeders, data managers and curators who each need different skills for their roles. Getting agreement on standards and best practices also requires communication and collaboration which needs resourcing.

Examples

The India-Australia molecular wheat breeding project mentioned above has resources for half time data managers at each partner site – DWS, PAU, and Cobbitty as well as a full time project data curator at NRCPB, and technical assistance from the CLC amounting to one full time technician covering skills in informatics and data management.

The Rice CI project has allotted resourcse for data management. Part of this is for the part-time data curaor at AfricaRice who supports partners and manages the project data. He receives training and technical assistance from rice CLC (IRRI) which has a full time data management staff.

The Role of the Integrated Breeding Platform

It is clear that there are a lot of crop specific tasks to carried out by the different participants in the breeding informatics pathway discussed above – the partner data managers, the project data curator and the CLC crop information curator. It is also quite apparent that the key opportunities for efficiency and scalability will come from using common technologies to collect, manage and publish crop information for any crop. This could best be facilitated through the IBP, which could undertake the following tasks:

- Develop, maintain and support integrated breeding informatics applications. This would include the design of databases to mange crop information from any crop and the development of user applications to facilitate breeding processes. These would need to be configured to the best practices for each crop, as defined by the CLCs, but in general will provide common functionality.
- Provide an environment and tools for crop communities to develop and maintain information on the crop specific practices of breeding and propagation and on traits and protocols.

- Develop training materials and run training courses on the use of the common informatics applications for breeders for all partner classes in the informatics strategy.
- Develop and maintain a common architecture and interface for the publication of public crop information for any crop.
- Train CLC Crop information curators in strategies, tools and methods for curating and integrating crop information from diverse sources such as multiple genebanks or multiple breeding projects.
- Host the public crop information databases as curated by the CLCs on highly accessible computer infrastructure, for example on cloud-based servers.

Conclusions

Information management at all levels of integrated breeding is as critical to the success of breeding projects as is the management of genetic resources and germplasm. The key to efficient and effective breeding information management is standardization and consistency. The task of developing and agreeing on standards for breeding technology and best practices in international agricultural research needs to be coordinated from an institute with a mandate to undertake broadly applicable crop improvement for each crop. These Crop Lead Centers should also take the lead in managing and publishing public crop information from their own and from partner breeding projects.

The task of managing and curating breeding information across all levels of the breeding process is critical and expensive. Massive efficiencies can be achieved if common data management and publications strategies and technologies are used across many crops and collective ownership of a cross-cutting platform to provide informatics technology, training and support would be the most effective way to realize these efficiencies.