

Generation Challenge Program Information Systems Platform and Network Design Workshop

31 May – 4 June 2004, CIMMYT headquarters, Mexico

WORKSHOP REPORT

Contents

Section 1: Topics and summaries of discussion

TOPIC 1: Project Organization.....	2
TOPIC 2: User Needs and Design Requirements Analysis.....	3
TOPIC 3: Review of Existing Systems and Initiatives Available as Potential GCP Information Subsystem Components.....	3
TOPIC 4: Developing a set of related use cases and a list of technology options available.....	5
TOPIC 5: In-Depth Discussion of Resources Available and Process for Implementation.....	6
TOPIC 6: Database and Molecular Marker Data Modelling.....	6

Section 2: Meeting outputs

1) Decision to use web services technology supported by BioMOBY.....	7
2) Use case template developed.....	7
3) Consolidated list of use cases identified.....	7
4) Criteria established for choosing software tools.....	8
5) Criteria established for choosing database model for storing molecular marker data.....	8
6) Approach developed for implementing the distributed architecture of the GCP informatics platform..	9
7) Implementation procedures identified for the distributed architecture approach.....	9
8) Tasks and timeline identified to move forward with this design.....	9

Section 3: Annexes (zip file)

- 1) Meeting agenda
- 2) List of participants
- 3) CIAT Crop Information Systems
- 4) INIBAP
- 5) CIMMYT
- 6) IRRI
- 7) CIP
- 8) Germinate
- 9) Maize GDB
- 10) Soaplab, BioMOBY, and Taverna – presented by Martin Singer
- 11) Semantic MOBY
- 12) PLANET
- 13) GDPC
- 14) Embrapa
- 15) Use Case Template
- 16) Breakout Group A: Set of Related Use Cases
- 17) Breakout B: Technology Options
- 18) Breakout Group A: Resources Available for Crops
- 19) Breakout Group B: Resource Integration
- 20) Databases and Molecular Marker Data Modelling

Section 1: Topics and summary of discussion

Introduction

Held at CIMMYT's headquarters in Mexico 31 May through 4 June, the Information Systems Platform and Network Design Workshop brought together experts from within and outside the Generation Challenge Programme to develop a comprehensive architectural blueprint of the GCP platform, network, and registry. This workshop was a follow-up of the SP4 consultation workshop held at IPGRI in February. The goals of the Information Systems Platform and Network Design Workshop were:

- To identify the components needed for a GCP information system using use case methodology
- To discuss the technological options for the interoperability of the various information systems components
- To choose a model for the high level architecture of the GCP platform and network

The above goals were only partially fulfilled in the workshop, but significant progress was accomplished towards them. First, some further discussions of use cases (with CIMMYT biological scientist inputs) was undertaken and the methodology for use case documentation clarified, although the actual documentation of use cases remains very incomplete. Second, significant design decisions were made concerning pioneering technologies for GCP platform and network development were identified, including the general system interoperability. Third, a "Model Driven Architecture" design paradigm was endorsed and some preliminary elements of such an architecture identified, although very incompletely.

This section of the report covers the topics that were discussed during the course of the workshop and the outcomes of those discussions. The outcomes of the workshop discussions often determined the subject and course of subsequent workshop sessions. Section 2 details the decisions and outputs of the workshop. Section 3 is the compiled presentations and other documents that were used or developed during the workshop.

TOPIC 1: Project Organization

Goal: To discuss general issues of project design and implementation process.

Discussion: In the discussion of general issues of project design and implementation, workshop participants agreed that user needs should drive this process. There was also agreement that to accelerate learning, create ownership, promote fast delivery among users, a strong, consistent process for consultation with the users is necessary. This iterative process should be lengthier and more intense than normal to ensure that the designers listen to the needs and reactions of the users and modify the platform accordingly.

Several conceptual approaches for the design of the platform and network were discussed. There was consensus that whatever approach was to be adopted must be both iterative and sequential, agile and robust. Some elements will require sequential development by their nature, but the preferred approach will be an iterative process.

The “Model Driven Architecture” (MDA) approach was suggested as an approach for the GCP information systems platform. This approach models systems/domains in a technology/network independent way first (“platform-independent model”), and then expresses the same entities and their relationships in a platform-specific way (platform = network technology, “platform-specific model”). This is followed by implementation. It was agreed that the success of the project should be measured in terms of prioritized satisfaction of GCP research and organizational (Consortium level) needs. Research needs were furthermore recognized to have near term (“year 1”) and longer term (2-5 year) prioritization timelines.

Outcomes:

- 1) The participants agreed to adopt the Model Driven Architecture approach.
- 2) Agile methodology was also adopted to ensure that the system is developed in an iterative and incremental way and provides continuous delivery.
- 3) UML (Unified Modelling Language) will be used to diagram the use cases that will drive the development of the system.

TOPIC 2: User Needs and Design Requirements Analysis

Goal: To formalize project use cases and high level system design requirements for the GCP informatics system.

Discussion: Jens Riis presented his white paper on Germplasm Information Systems Use Cases. Participants requested more specificity in the details of the use cases presented. There was a suggestion to establish an online contribution system for users to comment on use cases and to provide feedback on how well the new architecture is working, possibly using WIKI. It was decided that a template is needed to better describe use cases. See Annex 15: Use Case Template.

CIMMYT scientists Maartin Van Ginkel, Marilyn Warburton, Jiankang Wang, and Mark Sawkins presented on their informatics needs. Maarten Van Ginkel, wheat breeder and head of CIMMYT’s wheat gene bank and the international testing unit, identified a number of informatics areas he would like to see developed, including how to link data gathered in Mexico to data gathered around the world with GIS data. Dr. Van Ginkel acknowledged that the International Crop Information System (ICIS) was now being considered as a core Germplasm platform for CIMMYT. Marilyn Warburton, biotechnologist in the Applied Biotechnology Center at CIMMYT, discussed the need for a platform that could cut down on her data processing time for her association genetics work and make the resulting data more accessible to the public. Jiankang Wang, CIMMYT wheat scientist and designer of a wheat breeding simulation program, talked about system functions that could augment breeding programs. Mark Sawkins, CIMMYT biotechnologist who studies drought tolerance in maize, listed a number of functions he would like to see improved in the genomics capabilities for public domain research, specifically the generation and management of pathway data. See Outputs #3 (page 7): List of Consolidated User Needs Identified from CIMMYT scientist inputs at the GCP Platform and Network Design Meeting for a full list.

Outcomes:

- 1) The user requirements discussion helped familiarize the outside experts participating in the meeting with the types of data requirements GCP scientists and partners have that the information platform must be able to handle.
- 2) A fresh enumeration of user needs based on CIMMYT scientist discussions was captured on flip charts and later consolidated into a list as mentioned above. This list could help confirm and/or refine GCP SP4 white paper discussions of user needs.

TOPIC 3: Review of Existing Systems and Initiatives Available as Potential GCP Information Subsystem Components

Goal: To elaborate what databases, software, and other tools GCP partners are currently using and to learn about other initiatives underway around the world that the GCP may be able to use.

Discussion: Each participant presented the systems and initiatives with which they are currently involved.

GCP partners:

- CIAT
 - Web searches for germplasm
 - Barcode technology
 - Genetic resources information system for beans, cassava, forages on Solaris
 - SINGER
 - ICIS
 - ACeDB

- INIBAP
 - MGIS
 - TropGENE DB

- CIMMYT
 - IWIS
 - Maize Fieldbook
 - Maize Genebank System
 - MaizeFinder

- IRRI
 - ICIS
 - IRIS, including functional genomics extensions and the Java ICIS platform

- CIP
 - Barcode technology
 - Pocket-PC
 - LIS
 - DIVA-GIS
 - SAS

○ R

Other initiatives:

- Germinate: Plant Data Management System
- Maize GDB
- BioMoby
- Soaplab: web services for accessing analysis tools
- Taverna: acts as an intermediate layer between user level applications and workflow enactors
- Semantic MOBY
- PLANET: A Network of European Plant Databases
- Genomic Diversity and Phenotype Connection (GDPC)
- Embrapa Genetic Resource and Biotechnology (BIOFOCO)

Outcomes:

- 1) The presentations and discussions following allowed for identification of missing elements in some existing systems and some of the problems which the new platform should help overcome.
- 2) The outside initiatives helped illuminate some of the options for interoperability and integration, and also presented some solutions to problems that had been mentioned, such as the storage of molecular marker data.
- 3) A consensus was achieved that the International Crop Information System (ICIS) and the Germinate project, with some cross-fertilization of the GDPC “object model” system, form the initial starting points for the development of a reference implementation of the germplasm core of the GCP platform. The endorsement of these technologies was recognized as NOT precluding alternative implementations based on extant systems (e.g. IWIS, MGIS/TropGene).

TOPIC 4: Developing a set of related use and a list of technology options available

Goal: To discuss in two parallel groups the use cases that were presented in the meeting and the technology options that are available in the existing informatics universe.

It was decided that the group should use a bottom-up approach to determine what the high level architecture for the platform should be to ensure that the user needs to drive the choices of tools used as well as the design of the overall package. So the big group broke up into two small groups working in parallel to discuss the use cases that had been presented in the meeting and the technology options.

Breakout Group A: Set of related use cases – The groups developed a set of related use cases based on one basic scenario (breeder wants to identify a range of markers that can be used to manipulate a quantitative trait such as drought resistance). They broke down this scenario to identify the resources needed in terms of databases and analysis in each use case (3 use cases identified: Germplasm Selection, Marker Trait Association, Germplasm Diversity Analysis and Archival of Results). They also tentatively identified

some of the data types that were involved in the use cases. In the follow-up discussion on resources available by crop (see below), they then identified by data type and crop what resources are currently available to complete these jobs. See Annex 16: Set of Related Use Cases for full presentation of their discussion.

Breakout Group B: Technology group – This group fleshed out the alternatives for integration of the various information systems based on interoperability technologies. They discussed the use of uniform vs. mixed information systems components (software, databases, applications). Should the institutes adopt standard components across the GCP and then link through the connectivity layer, or should they keep the mixed components that are in place now and connect those via connectivity layer?

Outcomes:

- 1) Use case group:
 - a. Identified explicitly the issues involved in trying to meet the needs of users as identified in use cases.
 - b. Some basic UML modelling of the sample use cases was attempted.
- 2) Technology group:
 - a. Recommended web services (more specifically, BioMoby) for interoperability and listed the pros and cons of the technology.
 - b. Identified candidate software for implementation of the interoperability technology and listed a number of different softwares available and appropriate.
 - c. Gave recommendations on the integration process to guide SP4 as to how most efficiently implement the technology and the requisite software.

TOPIC 5: In-Depth Discussion of Resources Available and Process for Implementation

Goal: To discuss in parallel breakout groups the informatics resources that are currently available by crop and the process for integrating resources.

Breakout group A: Resources available by crop – The group listed the data needs by subprogram and the resources available by crop for each type of data. This exercise was to determine criteria for choosing tools to be used.

Breakout group B: Resource Integration – This group presented the steps necessary to implement a resource in the platform and to integrate the existing resources within the GCP.

Outcomes:

- 1) Resources by crop group:
 - a. Created a table to help evaluate various options for tools.
 - b. Gave recommendations as to how to interpret the criteria assessment and how to make choices based on the evaluations. See Annex 1: Resources Available by Crop for the full presentation.
- 2) Resource integration group:

- a. Gave recommendations on what must be in place to successfully integrate GCP resources. See Annex 19: Resource Integration for the full presentation of their discussion.

TOPIC 6: Database and Molecular Marker Data Modelling

Goal: To discuss options for molecular marker data modelling. It was identified in earlier discussion that there is a great need in the GCP for databases that handle molecular marker data.

Discussion: A task force was set up to explore Germinate, Maize GDB, GDPC/GDPDM, and ICIS as possible options for storing molecular marker data. The group listed attributes of each option and came up with some recommendations for criteria to choose a particular molecular database model.

Outcome:

- 1) Created list of criteria for choosing molecular marker database. See Annex 20 for full presentation.

Section 2: Meeting Outputs

1) Decision to use web services technology supported by BioMOBY.

2) Use case template developed.

ID number:	
Title:	
Summary:	
Data required:	
End result:	
User interface components:	
Data sources:	
Exceptional cases:	
Priority:	
Actors:	
Assumption	
Detailed Description/Steps:	
User contact:	
Existing tools addressing this UC	

3) Consolidated list of user needs identified from CIMMYT scientist discussions.

List of Use Cases identified at the GCP platform & network design workshop		
Number	Title	UC Template
1.0	Germplasm Selection to define a Diverse Set	Yes
2.0	Comparative Genomics	Yes
3.0	Access to/integration with multiple/distributed databases	
4.0	High-Throughput Data Capture (LIMS)	Yes
5.0	Data Analysis	
6.0	Deeper understanding of genetic basis of wide & specific adaptation of germplasm	Yes
7.a	What alleles confer what values?	Yes
7.b	Where (GIS) are these alleles found?	Yes
8.a	References to existing systems	
8.b	Wrappers to existing systems	
9.0	Developing standards for managing & disseminating germplasm lines	
10.0		
11.0	Controlled Vocabularies and Ontologies (CVO)	
12.0	Marker standards and nomenclature	Yes
13.0	Pre-filtering germplasm for crosses based on genetic simulation	Yes
14.0	Gene Transfer	
15.0	Data mining of large numbers of (allele) data points from GCP and public databases	
16.0	Association Genetics	Yes
17.0	Functionally complete genotyping	
18.0	Manage the overload of batch processing of public data sets	
19.0	User Friendly Processing	
20.0	Manage the mix of public domain and commercial tools	
21.0	Consensus QTL maps within crops (comparative across crops)	
22.0	Gene Expression	Yes
23.0	Access to Functional Genomics Tools	Yes
24.0	Comparative Profiling	
25.0	Generation/management, alignment to public data of pathway data	Yes

4) Criteria established for choosing software tools.

Criteria	Software			
	Structure	GeneFlow	Joinmap	Arlequin
Quality of documentation (either existing or can be generated)	Good, but domain knowledge needed	Good, big user manual, domain knowledge needed	Good	Good, extensive manual
Access to source code.	On application?	Proprietary	No	No
Licensing issues	None			
o Cost		\$17,000 per license	\$2-3 ,000	Free
o Potential for site /consortium licenses		Minimal potential	Yes	
o Public domain?			No	
o License management	None	Hardware dongle	None	None
Extent of use and user perception	Very widely used	Being used, basically ok	Widely used and liked	Widely used and favorably viewed
Programmability (can be driven rather than point and click)	Scriptable and very nice user interface	Not in any way	Unknown	Perhaps
Compatibility / Deployment	Java	Uses own DB, can not link to your own.	Windows specific	Java (Windows at least)
Data format issues (input/output)	Conversion cumbersome	Some issues with input	Okay but...	Input difficult, output good
Degree of customizability	Flexible parameters	At a price	Lots of parameters	Lots of choices
Ease of use.	Moderate	Good, once you get data in	Easy for a domain expert	Easy for a domain expert, dangerous for novice
Platform/database dependence	Independent	Windows only, depends on commercial DB backend		Requires conversion script for database
Quality of technical support	Single author, very responsive	Single author, very responsive	Okay	No major development effort (currently)
Suitability for deployment across platforms and centers	Yes	No	Platforms no, Centers yes	Yes

5) Criteria established for choosing database model for storing molecular marker data.

- ❖ Benchmark several database/data models to see how they perform.
- ❖ Use a data object model instantiated in middleware as (Java? Perl?) data objects to interact with the database and is able to link to multiple databases using XML (BioMoby data type?) encodings of the data objects.
- ❖ Set up a plant genotype data working group
 - Find a stupid acronym
 - Benchmark existing data schemas and work towards developing a common schema (may not be possible, but should at least be discussed)
 - Discuss incorporation of new molecular data models into existing systems which lack good molecular data models.
 - Adopt and develop data models for integration of analysis tools
 - Attempt to define standard data objects that the data models used conform to.
- ❖ Databases/Data models adopted need to be complete enough to handle all aspects the CP would require.
- ❖ Must be available to the CP program
- ❖ Data model designed in an open source platform but can be ported to other platforms.

6) Approach developed for implementing the distributed architecture of the GCP informatics platform.

- ❖ Whole system will not be integrated at one time, but in a stepwise fashion.
- ❖ Define realistic timelines for early implementation (and stick to it!!)
- ❖ Training on integration tools (developer's workshops)
- ❖ Early as possible start modeling system (as soon as use case solutions are defined). Address issues with experiences with modelling
- ❖ It is imperative to the success of the CP that those who can share data do. Attitude is everything.
- ❖ Use of *Model-View-Control* design pattern

7) Implementation procedures identified for the distributed architecture approach.

- ❖ Build system in short repeating cycles.
- ❖ Make services BioMOBY compliant to the extent possible
 - Be prepared for non MOBY compliant web services.
- ❖ Integration of data format will be at the Challenge Program level.
- ❖ Integration of web services into workflow will also be addressed at the Challenge Program level.
- ❖ The Challenge Program level will be responsible for the empowerment of the centers to develop and share their specific model and workflow requirements.
- ❖ Consider other components of the integration process, e.g.
 - Security
 - Different levels of access to data

8) Tasks and timeline identified to move forward with this design.

Task 1 – build platform

1. Describe/validate use cases

TASK LEADER: Carlos Lopez

TIMELINE: posted to virtual workspace by end of June

- a. Must use 'use case template.'
 - i. Transfer use cases from flip charts and white papers to use case template and get feedback from relevant task leaders.
 - ii. Record them in a Annex 2dos format on the virtual workspace for maximum recording of feedback and input.
 - iii. Partner with biologists to evaluate use cases.
 - iv. One final document to be drafted by three task leaders of Germplasm, Central Registry and Interoperability.
- b. UML documents created to integrate template documents.
- c. VWS must be clearly set up to display this and encourage interaction.
- d. Validation will be incorporated in this process.

UPDATE 28/7/04: IRRI follow-up meeting in July has initiated design of a user requirements database to formally capture and organize use cases out of the GCP white papers and CIMMYT meeting documents.

2. Create/revise data structures for datasets

TASK LEADERS: Raj Sood, Graham McLaren, Richard Bruskiewich

TIMELINE: end of July (following hackathon) and by Brisbane meeting.

- a. Germplasm: Raj will do schema for passport data for implementation for EURISCO and SINGER. By Brisbane meeting, we have draft for review. Eukarpia meeting in Europe to discuss as well. Raj to draft outline of process.
- b. Comparative Gene Catalog and universal marker list: Genomics task list. Richard to draw in more external experts. To draft outline for process.
- c. Repository for Comparative Microarray data: Richard also drafting (in Functional Genomics task). Structure will be designed, physical location will be discussed elsewhere.
- d. Documenting experiments: related to above. Also issue of ontologies for germplasm and phenotypic data. Phenotype data: question is the ability to be able to have a language to describe in-the-near-future experiments. [LIMS in short term.] Note that there is a request to inform scientists that this process is underway. (RMB note 27/7/04: issue discussed at Montpellier phenotyping meeting resulting in that Graham McLaren has initiated a formal process of capturing phenotype documentation).

UPDATE 28/7/04: IRRI follow-up meeting in July has initiated design of a high level GCP data models for the “model driven architecture”.

3. Create Archive for Conversion Tools

TASK LEADERS: Raj Sood

TIMELINE: beginning of August.

- a. Part of Central Registry task. A lot of tools are already there. Raj to follow up on this (Dave Marshall and Reinhard have tools available for this). Should be made available to public.

Task 2: Network First Cycle

4. Define XML schemas for:

- a. Germplasm data
- b. Passport data
- c. Phenotypes including environments
- d. Genotypes including marker types

All covered above.

5. Contact BioMOBY on adopting technology: part of Interoperability task.

TASK LEADER: Richard

TIMELINE: No urgent action needed, but Richard will create active prototypes. Ongoing and going well. "Prime the pump."

6. Make first web services based on BioMOBY

TASK LEADERS: Raj, Richard

TIMELINES: Prototypes running by Brisbane meeting.

- a. SINGER + EURISCO + web interface: Central Registry/Raj will add this task. Which budget will this be covered by? Raj will contact Theo about this.
- b. ICIS + Gramene (or similar) web services integration: Interoperability Group will handle this.

7. Prototyping Tool Integration.

TASK LEADER: Richard will lead the development of an initial GCP platform prototype integrating some tools. Martin Senger's "Taverna" tool (and the underlying workflow engine) to be investigated further for use in the GCP for network integration of web services.

TIMELINES: By Brisbane, we will discuss what the next steps on this should be.

8. Model entire system

TASK LEADERS: Richard (Interoperability), with help from Raj and Carlos.

TIMELINE: July meeting will result in document. Public iterative process.

Task 3: Capacity Building

9. Organize training

TASK LEADERS: ?

TIMELINES: training to be conducted at IRRI during hackathon.

- a. System modelling: will be done at IRRI in July. Martin Singer will train?
- b. Web service technology: Need proof-of-concept first. Tentatively around the PAG meeting in January. Need a face-to-face training, need experts involved from BioMOBY. Discuss with Carmen. 3-day training for all centers plus optional people from non-CG centers. Maybe adopt one of our use cases for the training. One or two need to be ready. That means ICIS, SINGER, and EURISCO interfaces need to be running.

Task 4: Other things

10. Policy commitment: data availability. Action: discuss in Rome. Talk with Victoria in late June.