

# GCP-SP4 Data Quality, Platform Development and LIMS Workshop, Feb 2005, Wageningen

Draft Report on the GCP Workshop on Data Quality, LIMS and Platform Development, Wageningen (WICC), 16-18 February, 2005

## Introduction

This workshop was the first activity of the GCP SP4 task on **Improvement of Quality of Existing GCP Databases**. The objectives of the workshop as specified in the task workplan were:

- To develop a collaborative workplan and assign responsibility for software modules to be developed by each partner.
- To review requirements for Quality Assurance of GCP data and software and commission a base-line survey of current quality status.
- To develop plans to stimulate and support a community of practice to exchange information and best practice on LIMS implementation and operation in labs carrying out GCP activities.

Twenty one colleagues from 10 consortium member sites attended the meeting with apologies from CIAT and EMBRAPA who were unable to attend due to conflicting and unforeseen circumstances. Five experts from SCRI, EBI, PLANET, GRAMENE and WUR (ALTERRA) attended the meeting by invitation.

## Quality Assurance

Junt Halbertsma, Quality Officer from Alterra, joined the workshop on Wednesday afternoon to give a presentation on Quality Assurance principles and practices for databases and models in the natural sciences. This generated extensive discussion on Quality Assurance for the GCP. It was decided that the scope of QA activities for 2005 under this project should be restricted to the SP4 activities of data curation, database infrastructure and software development. Extension of QA activities under the auspices of SP4 to data collection activities in other SPs in 2006 would be discussed with program managers.

Group brainstorming on SP4 quality control and assurance generated the following table of ideas:

Task	Quality Control Measure
Data Management: Verification of the quality of incoming data	Compare data points reported against curated data
	Checking validity of data values
	Ask for ID's of duplicate samples and compare data

	Check what data processing has been done on any record
	Identify outliers
	Verification of meta-data
LIMS	Enumerate desirable QC features of GCP LIMS
Domain modeling	Validate semantics & completeness of our models against other public models
	Assess if the domain model is being used and by whom
	Are we getting effective user feedback on the models: are the models complete and accurate?
Template development	Same concerns as domain modeling, plus...
	Validate input-template-output document transformations
	Build in data parsing failure conditions
	Develop a test data set from each data source
	Validate inventory of templates against inventory of target formats
	Retrieve templated data out of the central repository and run analysis, to compare with independent analysis of original input data
Platform Development	Independent software testing (by software engineers outside the team)
	End user scrutiny of software outputs
	Test data and script suites (e.g. JUnit); integration builds; test often
	Publish source code (open source)
	Work in progress should be visible
	Deliver rapidly to end users (to stimulate user feedback)
Repository	See template retrieval out of the central registry
	Assay usage profile of repository

QA activities agreed for 2005 are:

- A base-line survey of quality of current GCP data in crop information systems. Graham McLaren will develop a set of quality measures for different data types in collaboration with SP4 members and then those responsible for curation of SP4

information systems will evaluate current data in those systems and report on the status with proposals for improvement if required.

- A survey of quality of existing crop information databases according to criteria proposed by Junt Halbertsma which include:
  - Theory – validate simplifications of real world representation in the database design,
  - Documentation – has the structure and metadata of the database been completely documented
  - Tests – test data for completeness and valid range
  - Validation – validate data in the database against independent observations of the real world
  - Management – technical management such as backup status, feedback processing and improvement execution and documentation.
- Software development activities undertaken by SP4 will be carried out using the GCP Use Case Database, CropWiki content development environment and the CropForge software engineering project management environment which supports CVS, bug and feature tracking, and release management. The impact of these facilities on quality and accessibility of software products will be evaluated over the current development cycle.

## Platform Development

Objectives, principles and strategies for SP4 platform development were reviewed and discussed.

Consortium members from CIP, CIRAD, IPGRI/INIBAP, ICRISAT, ICARDA, CIMMYT, IITA, and IRRI presented reports and gave demonstrations of 2004 activities and products. It is clear that a great deal of platform development was achieved in 2004 with tremendous opportunities for adoption across different institutes. It was agreed that a more collaborative mode of development was required to avoid duplication and increase speed and efficiency of development. Collaboration tools were described, demonstrated and discussed. These included a Wiki for open-ended collaborative document development, and a Gforge site, with a CVS for software product versioning and release, bug and proposed feature tracking, and project discussion fora. It was decided that they provide a suitable environment for collaborative software development and therefore, should be adopted by all GCP partners to coordinate and manage 2005 GCP funded software engineering activities. The system could also be available for non GCP projects related to crop information or related analysis. A detailed review of platform components available, as well as gaps and future development plans/priorities of each partner, was undertaken. The following table summarizes some of the items listed by each partner.

Institute	Already Have, Can/Will Share	Would like to Obtain/Work On
ICARDA	Completed LIMS, Gene integration	Web interface (to ICIS), linking location to DIVA

CIP	DIVA GIS, Data Mart, Mobile Operations, Integration with barcode & wireless	Registry of experiments, Field books & data, extend data mart application
IITA	Complete LIMS, mapping of workflows	Gene bank seed inventory management for gene banks, location data QC management, genotype module
CIM MYT	GIS, (maize & wheat) fieldbook systems, modeling genebank, comparative map viewer	LIMS, genomic (genotyping, phenotyping, QTL, gene expression) systems, consolidated crop information systems
INIB AP/IP GRI	MCPD management, characterization data (Musa), web services experience (MOBY, GBIF)	Sample tracking, OS independent standalone query interfaces
ICRI SAT	LIMS including equipment logging, ICRIS: phenotyping, genotyping, pedigree; mapping workflows	Completion of ICRIS modules, Location data management
CIRA D	Mapping & genomics information systems (ortholog pipeline), map viewers, Perl objects, generic interfaces, web services	More systems integration of mapping to genomic data, gene expression data, phenotype data
CIAT		
IRRI	ICIS as LIMS (prototype), MEDAL, genealogy, field data and seed inventory management, ICIS5 Java architecture + WWW + web services prototype + workbench	Genebank management IS, location data, GEMS & genotyping, more ICIS 5 web interfaces, completion of the Java genomics workbench and MOBY web services prototype, mobile data loggers & bar coding
NIAS	Rice genome information systems: INE, RiceGAAS, RAD, Tos17 Phenotyping + flanking sequence database, rice EST database, KOMA (full length cDNA database), microarray analysis (RMOS), RED (rice expression database), rice annotation pipeline system, rice proteome database, genebank database, rice genome resource center delivery system	Extend RED towards oligo array database, QTL mapping data
NASC	Plant ontology, PlaNET moby services, phenotype database, MIAMEPlant, Affy gene expression to MAGE-OM	
Corn ell (Gra mene )	GDPC, diversity analysis tools	Diversity analysis module & tools

U. Dund ee/ SCRI	Germinate	
JIC	Java display software, database access	
WUR		
CAA S		
EMB RAP A	Genoma	

In order to facilitate collaboration, a partition of platform architecture on the basis of information domain (germplasm, phenotype, marker, genotype, map etc) and on the basis of software layers - data sources, domain model middleware and applications (user interfaces) - was discussed. Analysis of available components and platform needs indicated that a primary partition on the basis of software layers, with a secondary subdivision (of the middleware layer activity) into generic and domain model driven activities, would support collaboration best.

Accordingly, four categories were proposed for 2005 platform development activities:

- Domain Model Middleware Architecture
  - Combine the best concepts and practices of the available middleware options: ISYS, GDPC, ICIS5 ...
  - Provide for flexible and extensible framework for integration of diverse domain models, including provisions for filter driven queries automatically adapting to domain model changes
  - Develop sample application and data source adaptors for the range of implementation component types (e.g. interfaces to relational databases, web services, web browser and standalone applications)
- Domain Model Layer
  - Develop the middleware modules to support the specific GCP domain model semantics and logic for:
    - Passport data
    - Genealogy data
    - Phenotype data
    - Location and Environmental data
    - Genotype data
    - Map data
    - Functional genomics data
- Web Interfaces

- Take an inventory and review existing web interface applications and technologies at GCP partner (and external, public) sites, to develop a documented tool kit of web interface design guidelines and reusable software components, compatible to the GCP platform middleware and to partner site customization needs, for future development of use case driven design, implementation and deployment of web interfaces to GCP-hosted information systems
- Develop/adapt core use-case driven web interface applications for:
  - Germplasm by passport and trait queries
  - Genotype data mining
  - Data mart applications for passport, taxonomy, genotype and phenotype data
  - LIMS web applications
- Stand-alone Applications
  - Coordinate with middleware activity to specify the design requirements and develop sample software adaptors to plug in heterogeneous standalone applications into the GCP platform (i.e. domain model layer)
  - Adapt/adopt a selected set of existing diversity analysis, mapping and functional genomics tools for integration into the GCP platform for example:
    - GMOD and GDPC compliant public open source tools
    - NCGR CGMT and/or CMAP software
    - ICIS (genomics) workbench
    - DIVA
    - LIMS standalone applications

Principles for collaborative platform development were discussed:

- Platform development is to be a collaborative project to develop a comprehensive freely available tool box of complementary and compatible technologies.
- Projects should facilitate rapid deployment of platform products and interfaces to GCP end users
- Models and designs will be driven by documented and prioritized use cases (collective responsibility of all partners)
- GCP domain models will be the core logic layer of the platform and must be integrated in a flexible way to allow continuing evolution of the models

The emphasis in 2005 should be on rapid deployment of useful interfaces and applications. However, this software should be compatible with the design of the middleware and domain model layer, which will be essential for general progress in platform development in the future. Work should start immediately on the middleware design pattern specification and the development of first generation domain models, scheduled for publication at the end of April, and incorporation into the platform by mid-May. It is expected that the domain model editorial teams will carry out this integration or

collaborate closely with interested developers, to allow them to meet this timeline. The design of the application and data source adaptors is a high priority. These should be made available as soon as possible through CropForge and CropWiki. The compilation of an inventory and technical compare/contrast of web interfaces and standalone applications can start immediately, mindful of the requirements for the anticipated evolution of the platform. A workshop will be held at IRRI in the middle of May for developers to finalize adaptor protocols and start integration of targeted data sources and applications. This will constitute the work for the remainder of 2005. The table of partners' available technology and priority requirements was useful for illustration, but its further development into a workplan was not possible in the meeting. Hence partners agreed to submit proposals for activities based on the four layers (Middleware, Domain Model, Interface and Applications). All consortium members are free to propose activities alone or in partnership with one another. Proposals will also be considered from non-consortium experts. Partners may submit more than one activity proposal in any number of layers, and these may be logically connected across layers. Activity proposals should be submitted by Thursday, March 10th 2005.

## **Criteria for Activity Proposals**

Activity proposals should precisely identify deliverable products. It should be clear how this task contributes to the overall platform, namely, how the product will be made available and useful to other partners. This includes specifying how design, source code and other components will be available, in case partners want to implement products with different technologies. The proposal should indicate a timeline of activity and specify resources required from GCP as well as resources that will be used from other sources if any. It should clearly state that intellectual property is freely available, or precisely define any restrictions which are required. As an indicator of detail required, it is not expected that many proposals will require more than a page of text. Requested budgets will be assessed on the basis of value for money and if prioritization between good proposals is required whole activities will be declined rather than budgets trimmed. The SP4 leader will make final decisions.

## **LIMS Community of Practice**

Requirements and modalities for a LIMS Community of Practice were discussed. Four action points were decided upon:

- Stimulate communication between LIMS groups in the GCP by reactivating the LIMS email list server with archive facilities. This action item should consider use of other communication options – e.g. discussion forum and/or Wiki.
- ICARDA and CIMMYT groups to evaluate and report on the open source CaLIMS package
- Post solutions to the problem of automatic capture or transfer of data from most common equipment types.
- Propose and discuss strategies for quality control and assurance for data captured and managed by LIMS