



Improvement of quality of existing GCP databases

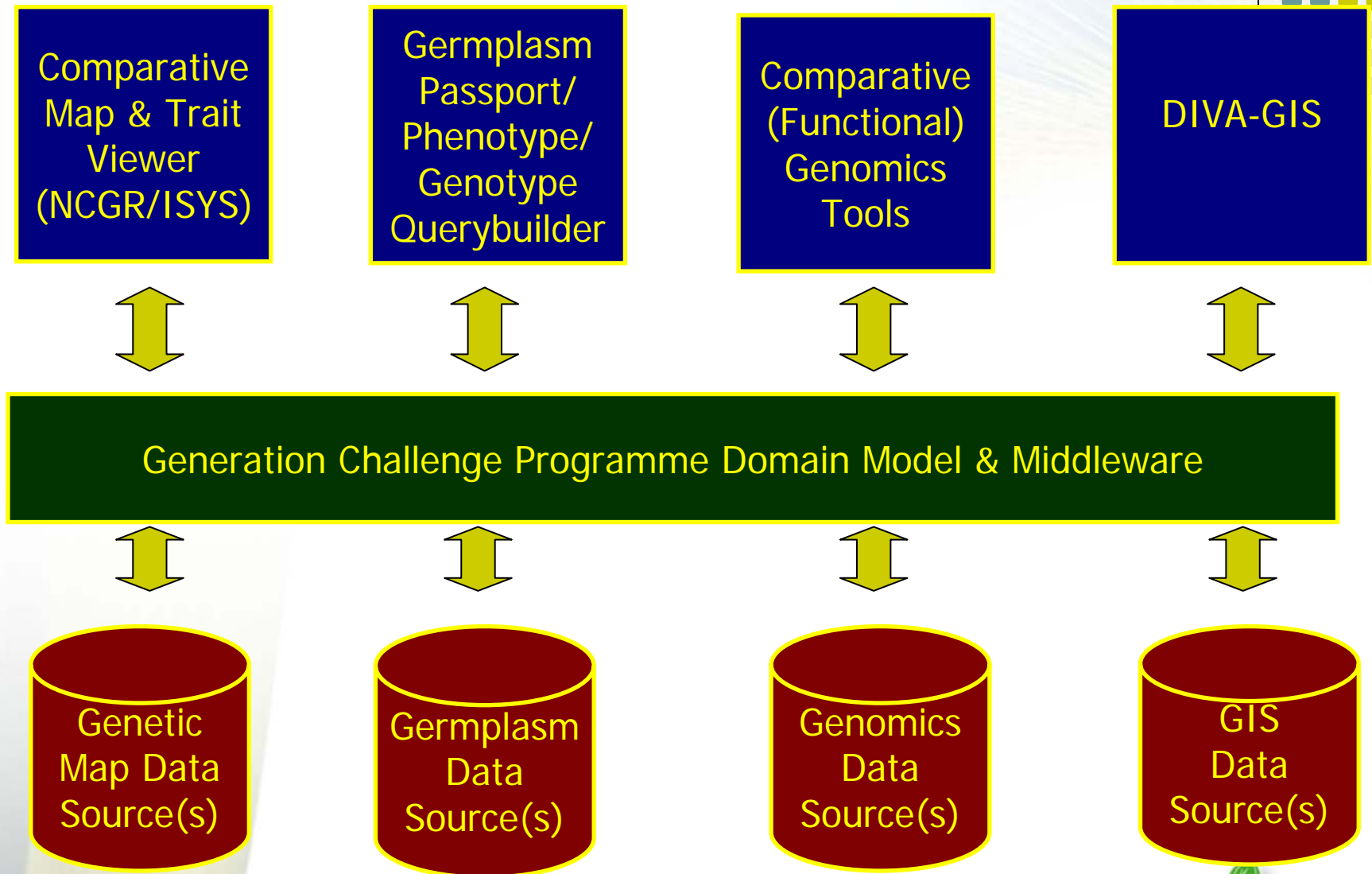
SP4 Task 28

Graham McLaren

Vision

- High quality GCP data collected and stored in local, curated databases
- Integration with existing data of known quality
- Specialist applications able to access GCP and related data
- Applications interoperate with each other via a semantically rich domain model

An Example Crop Research Scenario



Task 28 Team



- Guy Davenport - CIMMYT,
- Reinhardt Simon – CIP,
- Edwin Rojas - CIP,
- Manuel Ruiz - CIRAD,
- Jayshree Balaji. - ICRISAT,
- Akinola Akintunde - ICARDA,
- Visvanathan Mahalakshmi - IITA,
- Mathieu Rouard - INIBAP,
- Graham McLaren - IRRI
- Richard Bruskwiech - IRRI

Objectives

- Implementation of agreed data models
- Connection of local databases to the model layer
- Development/adoption of data curation and analysis tools
- Mounting tools on the domain model layer
- Development and comparison of LIMS
- Definition of quality standards and quality assurance protocols



Principles

The emphasis of the GCP contribution is on generic solutions and collaborative development so that outputs have wide application across the CP partners. Generic solutions are those that can readily be employed for different databases and crops.

Activities - I

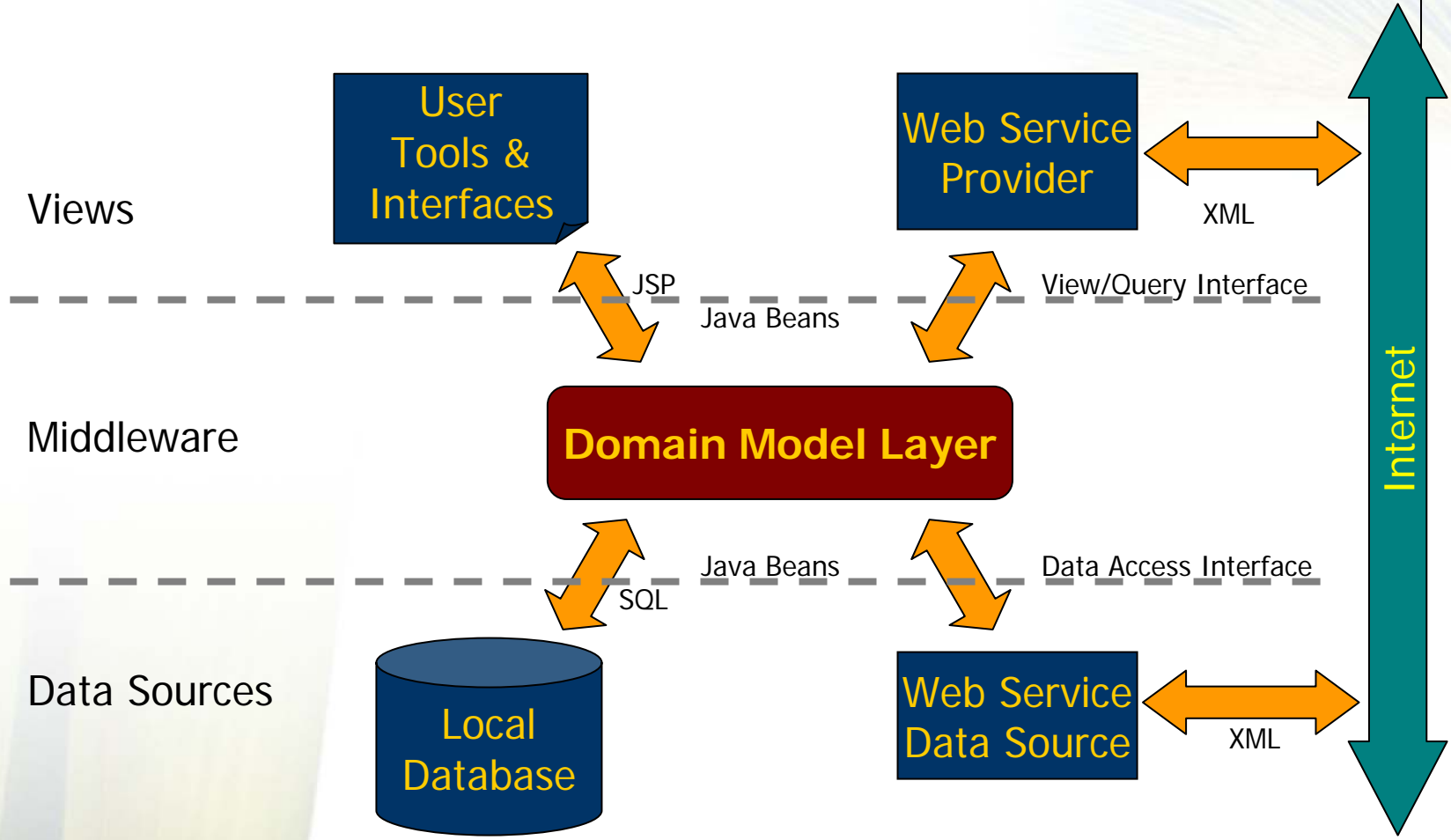
- General Platform Architecture and integration of the CMTV into the GCP platform – G. Davenport
- Implementation of middleware for germplasm, genotype and phenotype domain models – R. Bruskiwich
- Implementation of middleware for passport domain model – M. Rouard
- Implementation of middleware for location and environment domain models – R. Simon
- Implementation of middleware and interface for mapping domain model – M. Ruiz

Activities II



- Mondrian data warehousing for GCP data – E. Rojas
- LIMS: web applications for sample tracking – B. Jayashree
- LIMS: further development of GMLIMS and IGGEMS – A. Akintunde
- Adaptation of functional genomics tools for GCP platform – R. Bruskiwich
- Develop a strategy for Quality Assurance for GCP data – G. McLaren

Relationship of GCP Domain Model to Platform



A Pantheon of Information Tasks



Demeter – The Greek Goddess of Agriculture: Specification of the Domain model



Belenus – The Celtic Sun God: Application Support Interface



Ceres – Roman Goddess of Plants: Domain Model Implementation



Osiris – The Egyptian God of the underworld: Data Source Interface



Task 28 Posters:

- Task 28 – Quality Improvement
 - Improvement of quality of existing data: databasing of molecular scoring data at CIP (including LiCor data)
- Task 28.1 - General Platform Architecture
 - BioClipse - an open source bioinformatics component architecture based on ISYS(TM) and Eclipse RCP
- Task 28.4 - Location and Environment Data
 - Integrating DIVA-GIS into the GCP platform
- Task 28.6 - Mondrian data Warehouse
 - Data warehouse technology in GCP platform - automated reports and tools for cross-validation of data



Task 28 Posters - Continued

- Task 28.7 – LIMS
 - LIMS web applications - sample tracking
- Task 28 – Quality Improvement
 - Integrating genotyping workflow and databases: an update on Bioinformatics at ICRISAT
- Task 28.8 - LIMS
 - Integrating Phenotypic & Molecular data for allele mining of the ICARDA mandated germplasm collection



Next Steps in Platform Development

- Further refinement of the domain model implementation in the middleware
- Integration of applications into the platform via Eclipse RCP, the Eclipse Plug-in Architecture or an ISYS interface.
- Development of Web Service Applications either as Web Service Providers like Moby and Biocase or as Web Server Interfaces to different model implementations
- Integration more of data sources into the platform via Hibernate adapters or Web Service clients or other specialized adapters as needed.
- Development and merger of LIMS systems and integration with templates

Data Quality

The Conventional Wisdom:



- Data quality control and quality assurance are very important
- You should just see how bad my colleagues' data are!
- Something should really be done about it!

A GCP Data Quality Strategy



- Five GCP data curation centers conducted base-line surveys
- Results of these surveys were discussed at the Vancouver meeting with the following main conclusions:
 - Data quality should be everyone's concern
 - The problem is with your own data
 - The problem starts as soon as you collect it

GCP Data Quality Vision



Achieve recognition as a repository of public research data of certified quality

- **Quality** means ‘fit for the intended purpose’
- **Certified** means that the quality control is documented and assured



GCP Data Quality Goals

- To document data resources to a sufficient standard to allow integration and support use for germplasm improvement and comparative biology
- To promote quality assurance for all data sets and establish guidelines for quality control for GCP funded data resources
- To train scientists in the application of data quality procedures

GCP Data Quality Strategy



- Mandatory application of QA/QC best practices for GCP funded research
- Promote the application of such practices to existing data compiled into the project
- Develop a blueprint and policy supporting the generation and maintenance of high quality data

Components of Quality Assurance



- **Accuracy:** whether or not the datum is close to its “real” value (statistically?)
- **Precision:** the level of uncertainty in the “real” value of the datum
- **Consistency:** logical consistency (temporal sequence, causality, etc.) between data elements
- **Lineage:** source of the information
- **Completeness:** completeness in availability
- **Fitness for purpose:** are the assumptions congruent with purpose

QC Depends on Data Type



- **Identification and ownership** of materials and data
- **Passport data:** secondary data (mostly, from existing genebank databases)
- **Characterization** (environment and phenotype): secondary data (from existing genebank databases)
- **Genotyping data:** more local control (i.e. LIMS) with more opportunity for QC as well as QA
- **Evaluation** (environment dependent phenotype) data: field data that is environment specific
Derived genetic data: e.g. genetic (QTL) maps
- **Sequence and molecular expression data:** gene expression, proteomics, metabolomics, etc.



Next Steps

- Get a clear understanding that QC and QA are everyone's problem
- Develop data-type specific QC protocols
- Train data collectors and curators in applying the QC protocols
- Implement QA components in GCP data repositories