
Development and Application of Generation Challenge Programme Domain Models

**Richard Bruskiewich¹, Guy Davenport², Tom Hazekamp³,
Thomas Metz¹, Manuel Ruiz⁴, Reinhard Simon⁵, Masaru
Takeya⁶, Jennifer Lee (U. Dundee/SCRI)⁷, Martin Senger¹,
and Graham McLaren¹**

¹IRRI, ²CIMMYT, ³IPGRI, ⁴CIRAD, ⁵CIP, ⁶NIAS, ⁷University of
Dundee/Scottish Crop Research Institute

Presenter: Mr. Victor Jun Ulat, IRRI

Overview

Task Motivation, Scope & Definition

- The GCP scientific and organizational integration challenge
- What are domain models? How will they be used?

Domain Model Development & Application

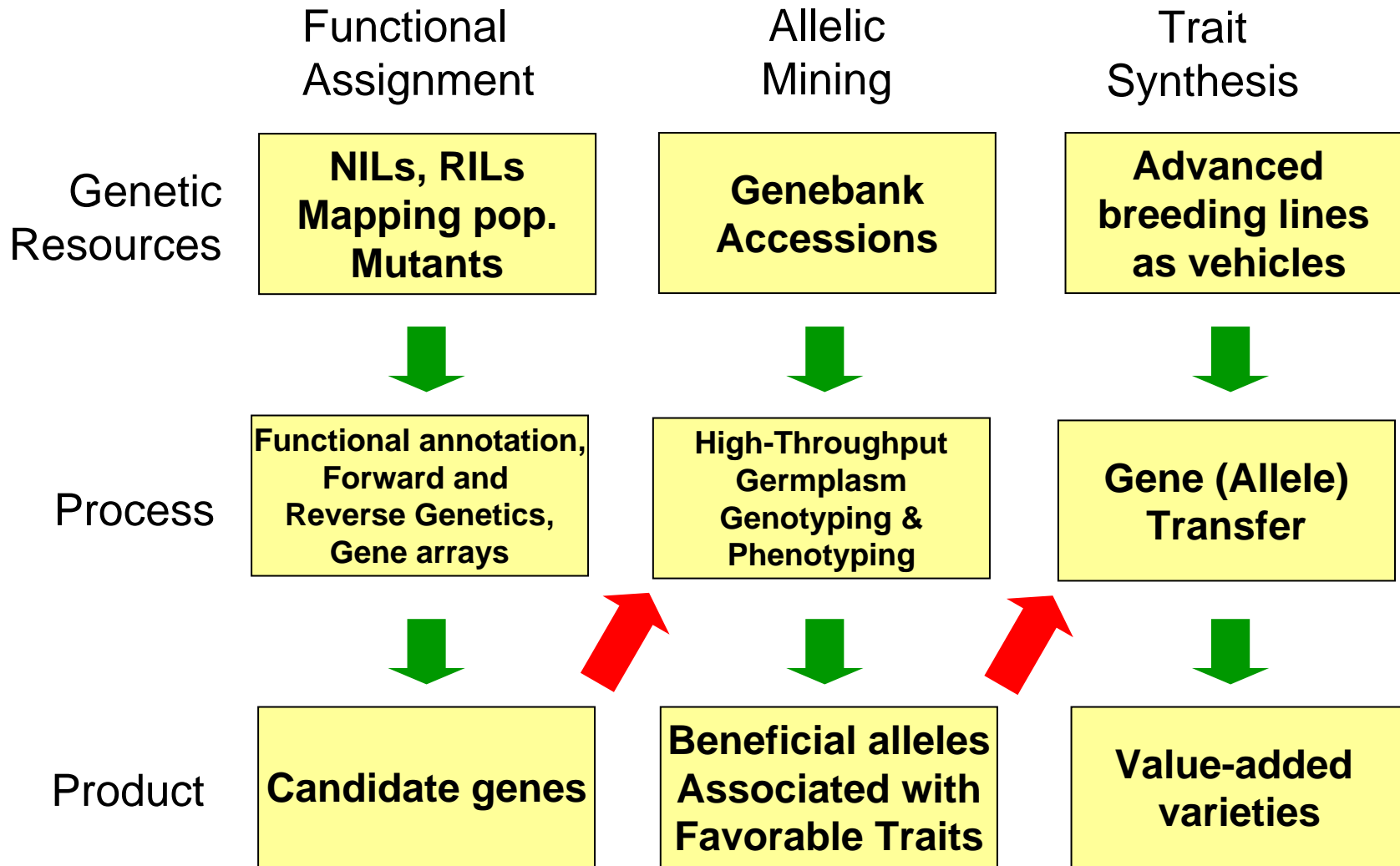
- Task Strategy and Execution
 - Domain model scope
 - Editorial teams
 - Progress report
- Sample GCP platform scenario to be built on the domain models
- Future Directions



<http://www.generationcp.org>

- 10 year international research programme formally initiated in 2003 and involving global partners from CGIAR, advanced research institutes and developing country national research & extension partners (core NARES: China (CAAS), India (ICAR), Brazil (EMBRAPA), African Centre for Gene Technology (South Africa)).
- Research directed toward molecular-driven crop improvement:
 - Genomics and comparative biology across species
 - Characterization of genetic diversity for allele mining
- Five subprogrammes:
 - SP1 – Characterization of Genetic Resources
 - SP2 – Comparative Crop Biology
 - SP3 – Gene Transfer into Breeding
 - ***SP4 – Crop Information Systems and Bioinformatics***
 - SP5 – Training & Capacity Building

GCP in a nutshell: Gene Discovery to Trait Synthesis



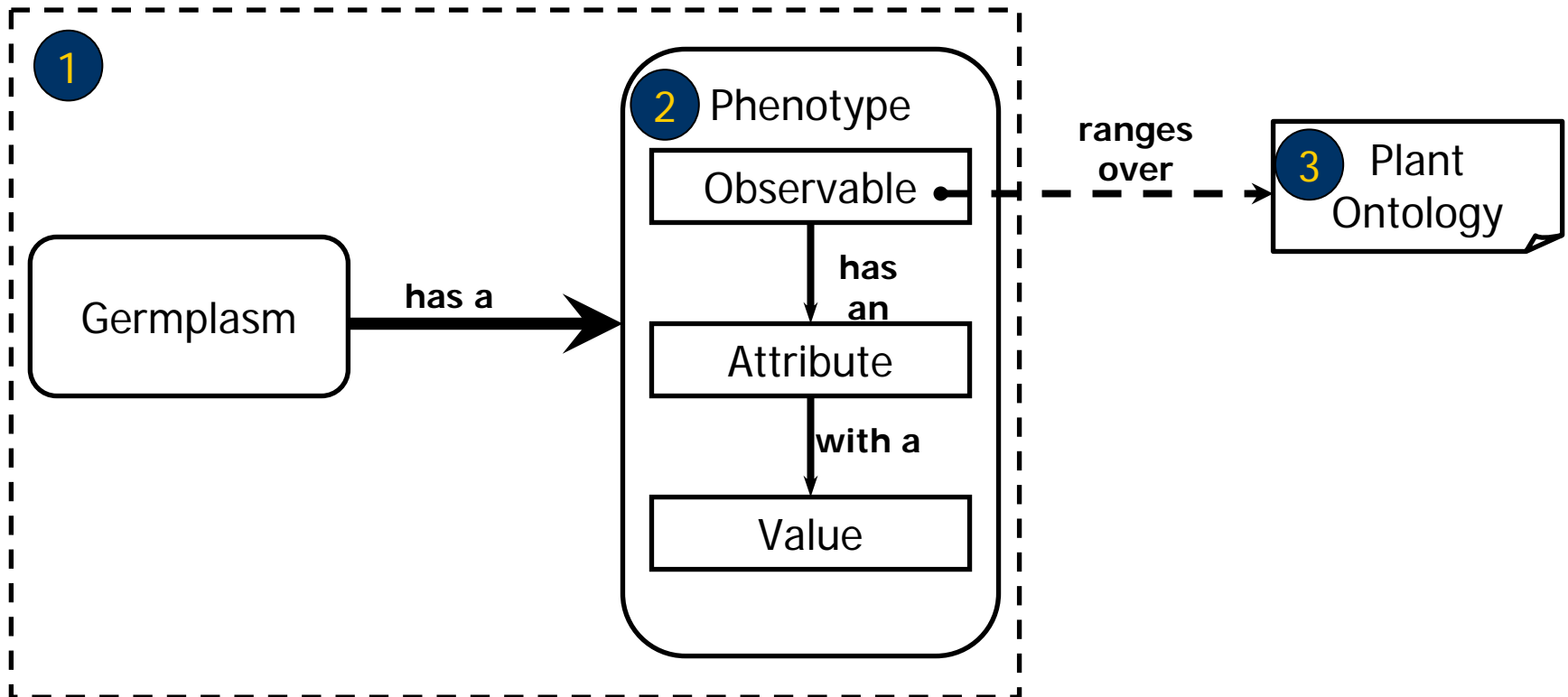
GCP scientific and organizational integration challenge

- Many diverse data sets generated for many purposes and in many formats
- Many diverse available and novel tools and data sources required for this analysis
- Complex scientific integration protocols
- Data being generated by 18 core partners + many external collaborators, globally distributed

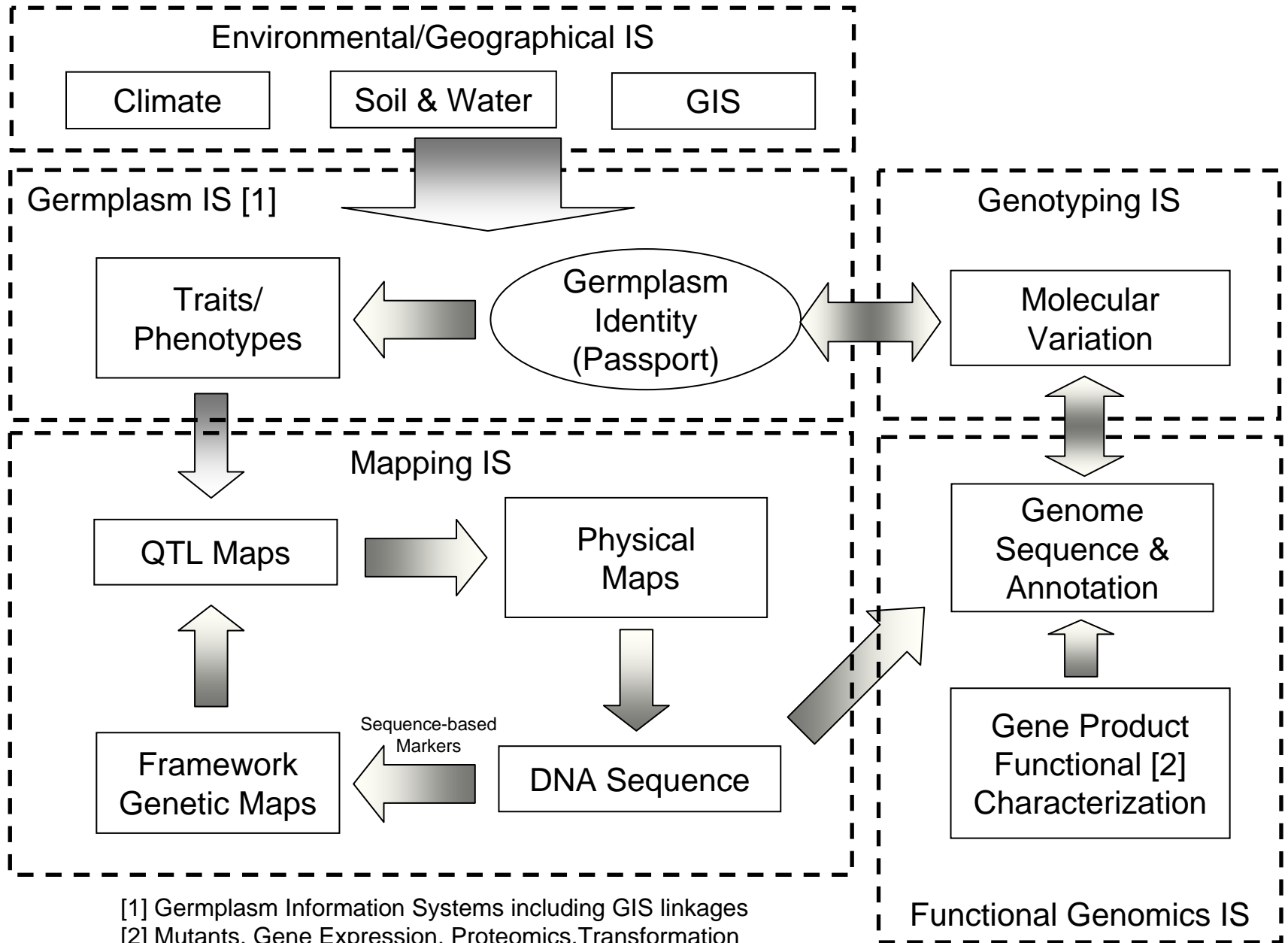
Domain Modeling Task Scope and Definition

- The GCP needs a common understanding of crop research concepts and their interrelationships (“ontology”) to be shared across (internet-integrated) platform tools and data sources.
- This is the underlying rationale for the SP4 task of formally specifying a public, shared GCP domain model.
- The underlying meaning (“semantics”) of the model needs to be specified at three distinct but interconnected structural levels:
 - **System architectural level:** e.g. the general semantics of high level entities and their inter-relationships
 - **Entity class level:** attributes and behaviors internal to the high level entities (i.e. data entity class attributes and behaviours)
 - **Attribute value level:** attribute values that range over an ontology (e.g. Gene Ontology (GO) term dictionary for gene product function)

Illustration of the Three Levels of Domain Modeling Semantics (“Ontology”)



GCP Domain Model Scope



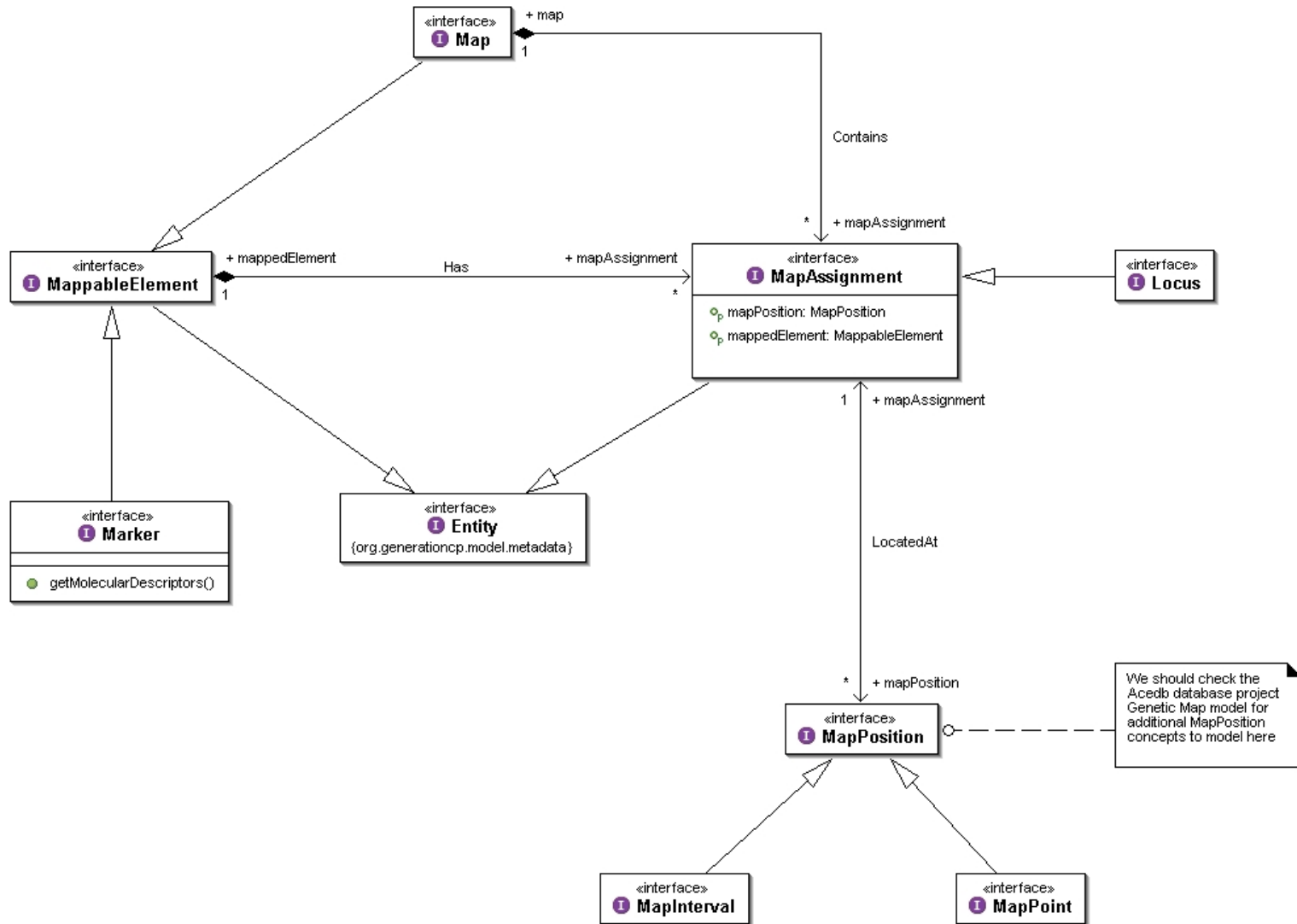
Domain Model Editorial Teams

Scope of Models	Leading Institute	Principal Contributors*
Core domain model and generic models	IRRI	Richard Bruskiewich, Guy Davenport
Germplasm Genealogy Genotype and Phenotype	IRRI	Graham McLaren, Thomas Metz, Jennifer Lee (SCRI)
Germplasm Passport	IPGRI	Tom Hazekamp
Location (GIS) and Environmental Data	CIP	Reinhard Simon, Edwin Rojas, Isaiah Mukema (IRRI)
Genetic and Genomic Maps	CIRAD	Manuel Ruiz, Guy Davenport
Functional genomics	NIAS	Masaru Takeya

Progress Report

- SP4 meetings with discussion relating to domain models held in Wageningen (February), IRRI (May) and Vancouver, BC (August).
- In Wageningen, editorial teams presented their work plans and strategies; a team discussion reaffirmed* use of a technical standard for general model specification called the Unified Modeling Language or “UML” for short. Where possible, existing public standards to be reused.
- Some initial (“Phase I”) models developed by May were fed into data template task and initial platform prototyping.
- Problems experienced getting different UML software tools to share models was partially resolved at the May meeting by deciding to integrate models under one tool, EclipseUML*
- Each team developed a set of EclipseUML models by the August meeting, at which time an attempt was made to cross-link all the modules of the model into one cohesive set of models.

Excerpt of the Current GCP Domain Model

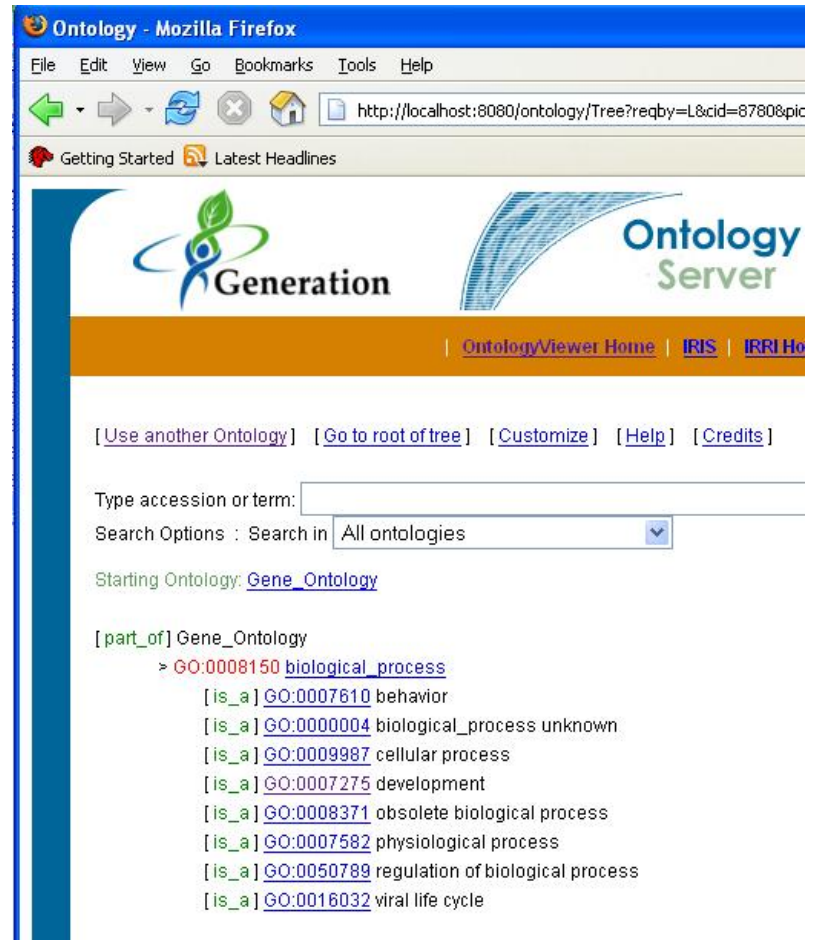


Model Publication

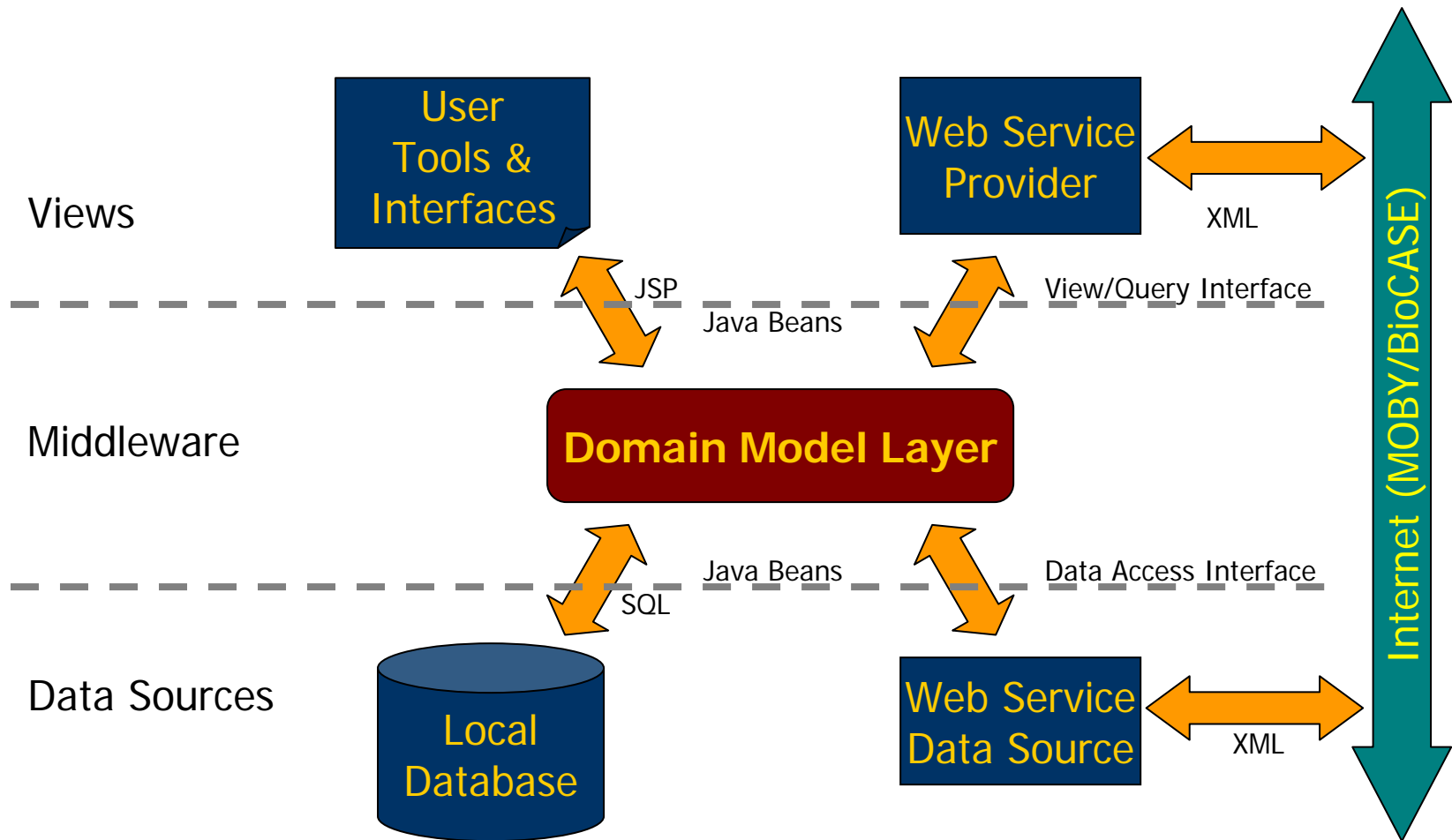
- Model internet “name space” established as <http://www.generationcp.org/model>
- The current set of EclipseUML models posted to CropForge (<http://cropforge.org>) under the “Demeter-DMS” subproject of the GCP Middleware.
- Further technical details available on the Task 22 poster and posted on the GCP CropWiki (<http://cropwiki.irri.org/gcp>)

Ontology Viewer

- Establishing a catalog to manage GCP ontology, to complement the UML domain model specification.
- Based on GMOD Chado database schema
- Web-based ontology management tool developed and deployed @ <http://www.iris.irri.org/ontology>
- Other tools being adapted (e.g. Protégé)

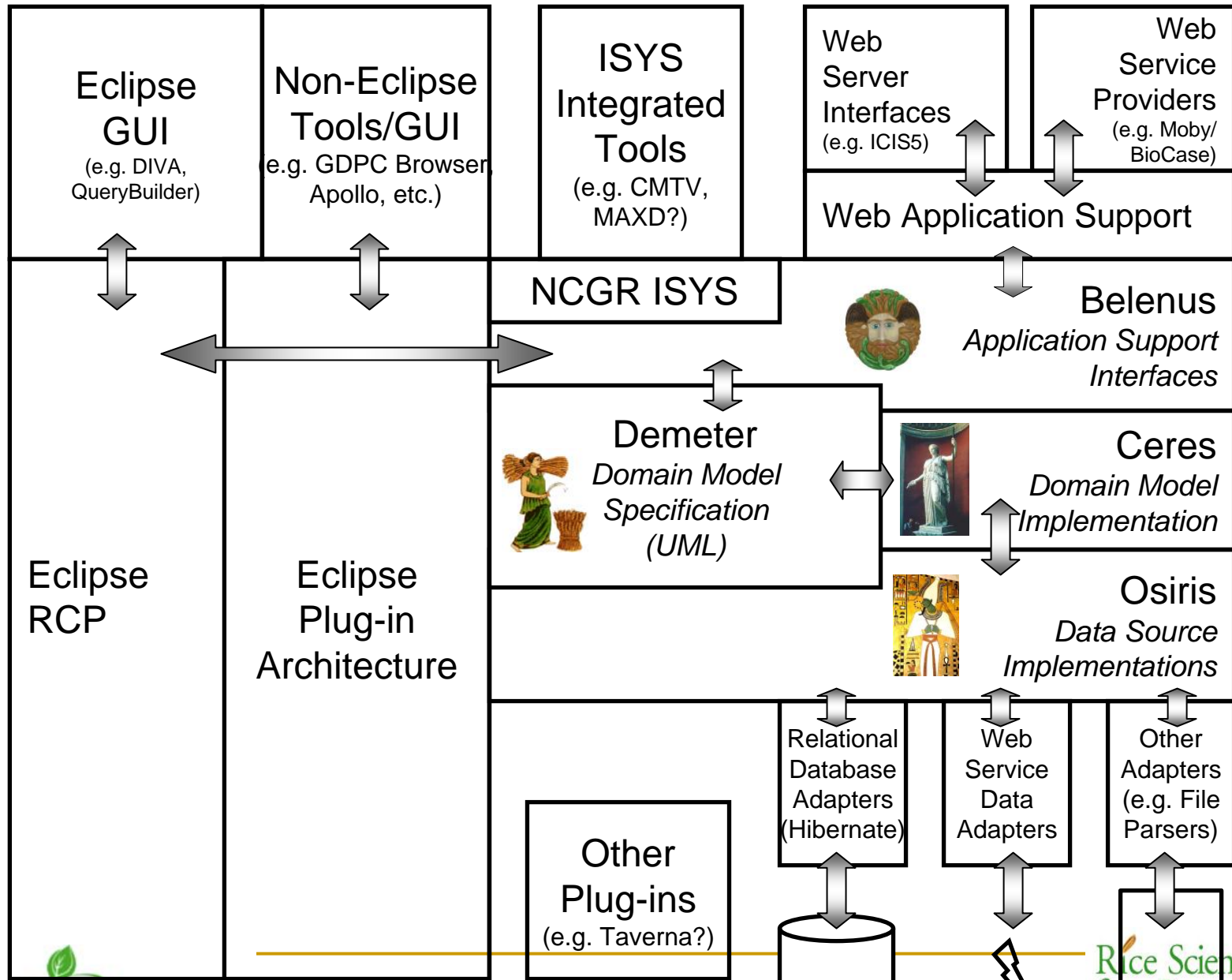


Relationship of GCP Domain Model to Platform

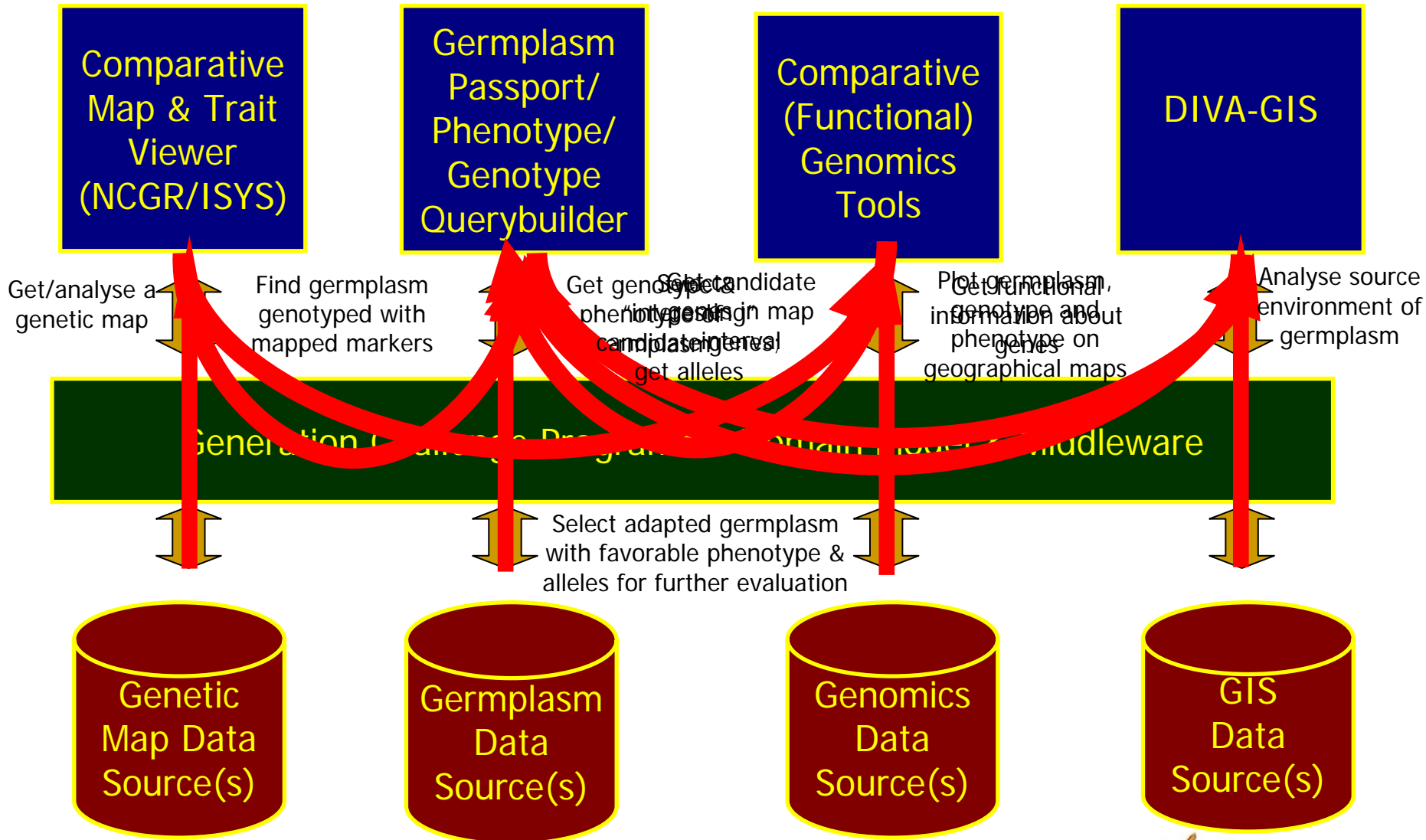


Domain Model Implementation in the GCP Platform

- EclipseUML models structured into a logical hierarchy of “packages”, for example, the **org.generationcp.model.germplasm** package contains germplasm related models.
- EclipseUML tool directly generates and cross-links models with Java language software files that are now being directly used in the (Java) GCP platform implementation efforts of several prototypes



An Example Crop Research Scenario



Future Directions

- Formal release of version 1.0 models coupled with a scientific paper submission about the project in December 2005
- So far, the modeling activity has been mostly an armchair analysis of domain entities and attributes... In the near future, more attention will be paid to the iterative validation of the domain models in GCP platform implementation of high priority use cases (such as the scenario just presented).
- Wish to develop a strategy to automatically generate data templates (essentially, XML schemata) directly from the domain model.
- Wish to (automatically?) map the domain models onto web service protocols (i.e. onto BioMOBY data types).
- Scope of domain models will be extended to meet evolving needs of the GCP