



Generation CP Domain Modeling Task Meeting

Wageningen, the Netherlands

Feb. 14th-16th, 2005



Introduction

- General welcome with general meeting logistics/announcements (Theo/Richard)
- Round robin group introductions
- Terms of reference for GCP domain modeling task (Theo)
- Review proposed workshop agenda (clarify/amend/accept - Richard)



Day 1 - Agenda

- Task overview (Richard)
 - Review 2004 activities & achievements.
 - Overview of goals, objectives and outputs for the GCP domain modeling task in 2005.
 - Review of proposed technical strategy for the task.
 - Discussion of Generation CP use cases and brief introduction to the GCP use case database.
- Introductory presentation from each assigned domain model editorial teams, for their thematic modules
- Pertinent external initiatives:
 - OMG (SNP initiative) - Martin Senger
 - GDPC/DM - Terry Casstevens



Day 1 - Agenda (cont'd)

- Summary discussion about task work plan & priorities (Theo)
- Look ahead to 2nd day (Richard)
 - Installation of software tools (e.g. Protégé, XSLT, etc.)
 - Think about 2004 high level model (be prepared to comment and propose amendments as required): module scope and “strata”
- Planning of (optional?) *ad hoc* evening activities(?)
 - Dinner plans (finalized)
 - Curate white paper use cases into database
 - Other SP4 task project satellite meetings



Day 2 - Agenda

- Tutorial introduction to Ontology modeling tools and protocols (Richard):
 - Domain modeling principles
 - Introduction to Protégé
 - OWL/RDF XML
- XML model transformation technologies (Guy Davenport)
- Relationship of domain models to template task (Guy)
- Relationship of domain models to GCP platform and network architecture and software (Alex Cosico)



Day 2 - Agenda (cont'd)

- Formal group review of high level 2004 (UML) model architecture to validate high level concepts (entities) and relationships
- Editorial team breakout sessions:
 - Revisit proposed strategy and process
 - Discuss and refine context of editorial theme in overall domain modeling
 - Identify modeling target outputs and milestones
- Planning of (optional?) *ad hoc* evening activities(?)
 - Dinner plans (finalized)
 - Protégé domain model development activity(?)
 - Other SP4 task project meetings



Day 3 - Agenda

- Editorial team reports on breakout sessions
- Discuss 2005 modeling priorities and strategy
 - Revise 2004 domain model architecture to suit
- Summary of domain modeling workshop and reaffirm task work plan and associated milestones for 2005
 - Phase I* of preliminary planning for proposed mid-year intensive software engineering workshop (“hackathon”)

* Phase II in data quality workshop later this week



Questions?



2004 Activities & Achievements I.

- **Nov. 2003-Feb. 2004:** elaboration of SP4 technical white papers on user needs and existing technologies
- **Feb. 2004:** IPGRI SP4 project & white paper review meeting; tasks refined
- **May. 2004:** CIMMYT platform & network design meeting:
 - Several external database and MOBY experts involved
 - SP4 team generally adopts MDA paradigm for system development
 - Elaboration of use cases and preliminary UML modeling of same
 - Web services paradigm (mainly MOBY) adopted for network integration
- **July 2004:** IRRI “hackathon” (actually an extension of CIMMYT meeting):
 - Elaboration of web services & LSID design strategy
 - Elaboration of use case development strategy
 - First generation domain models developed and published
 - Ontology development methodologies discussed



2004 Activities & Achievements II.

- **September 2004:** Brisbane annual review meeting:
 - Demonstration of prototype web services applications
 - Report (and project meeting) on commissioning of HPC cluster/grid
 - Prototype of use case database presented
- **November 2004:** NCGR-hosted MOBY meeting
 - A handful of GCP scientists swamp the meeting to learn more about MOBY
 - GCP scientists discuss GCP domain modeling strategy both before and during meeting
 - MOBY-S and S-MOBY projects agreed to collaborate more closely to identify convergence of technology; Mark Wilkinson announced that MOBY-S is adopting MyGrid registry solution; Martin Senger popularized Taverna once again...



2005 Domain Modeling Task Overview

- Overview of goals, objectives and outputs for the GCP domain modeling task in 2005
- Review of proposed technical strategy for the task
- Discussion of Generation CP use cases for 2005 and brief introduction to the GCP use case database and proposed modeling infrastructure
- First introduction to the 2004 domain model



Task Goal

- Specify domain (“data”) models to be mapped onto scientific use cases at the information platform, network (web services) and user interface level.
 - ***Rationale:*** the development and adoption of common domain model standards underlies any sensible attempt at Generation CP interoperability of information systems, including a web services driven architecture for global integration of diverse data types across diverse platforms



Task Objectives (by end of 2005)

- ***Develop GCP domain models***

- Assist in the compilation of scientific use cases, to elucidate domain models pertinent to the Generation CP.
- Extend domain models initiated in year 1 to meet the requirements for priority scientific informatics use cases, including capture of priority SP* data sets.
- Commission a community editorial process for domain model development.

- ***Feed domain models into GCP platform development***

- Collaborate with "Creation and maintenance of templates for Generation CP data storage in repositories" to translate domain models into data templates.
- Commission software tools to translate the domain models into data type and service type ontology specifications for web services implementation.
- Commission software tools to translate the domain models into components of the reference platform for "Improvement of quality of existing databases".



Task Outcomes (for 2005)

- Prioritized scientific use cases compiled from year 1 user needs white papers and other consultations into a GenerationCP use case database and applied towards elaboration of domain models.
- Domain modeling design and application methodology fully designed, implemented and deployed by task participants, for domain model design and application.
- Public version-managing model repository commissioned and populated with domain models meeting year (1 &) 2 Generation CP requirements, in particular, completed core passport, germplasm, phenotype and genotype models.
- Domain models applied to design of data templates, web services and platform software components deployed for GenerationCP production use.



Task Implementation

- Five domain model themes elaborated by editorial teams:
 - Generic core models, germplasm (genealogy), phenotype, genotype data (IRRI, Richard/Graham)
 - Passport data (IPGRI, Tom Hazekamp)
 - Mapping data incl. sequence maps (CIRAD, Manuel Ruiz)
 - Locational and Environment Data (CIP, Reinhard Simon/Edwin Rojas)
 - Functional genomics data (NIAS, Masaru Takeya/Shoshi Kikuchi)
- Two phases of domain model development:
 - Phase I (Jan-Apr, 2005) - delivery of priority domain models for key entity classes, for templates and basic web services
 - Phase II (May-Dec, 2005) - elaboration of basic models and full application to platform and network implementation, including scientific publication of modeling outputs



Editorial Team Responsibilities

- Assist with refined use case curation pertinent to theme
- Apply agreed modeling methodology to their assigned model theme
- Consult with other pertinent experts both within and outside the GCP, including pertinent international initiatives (e.g. non-GCP modelling and ontology consortia)
- Coordinate development and publication of version-managed domain models in the GCP model/template repository
- Track and assist development of GCP template, web service and platform products relating to their models
- Represent task at SP* meetings relating to their assigned model theme



What do we mean by “ontology” and “domain model”?

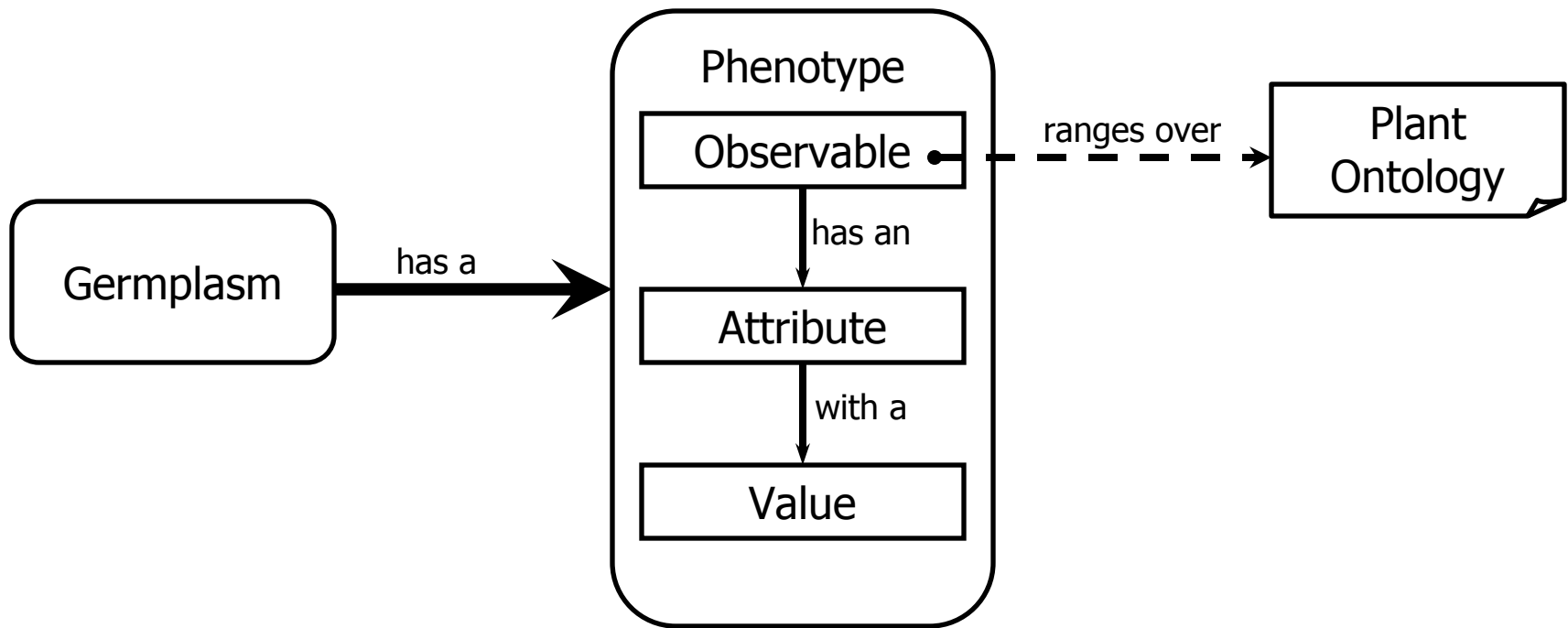
- ***Ontology:*** concepts and relationships, named (CV) and defined
- ***Domain model:*** an ontological description of all pertinent entities and their relationships in a given domain of discourse.



Levels of Ontology in Domain Models

- Ontological principles enter into the descriptions of domain entities at three levels:
 - ***System architectural level:*** e.g. black box entities (superclasses) and their high level relationships (i.e. the 2004 UML data model)
 - ***Entity class level:*** “internal” entity attributes and behaviors (i.e. Java object classes)
 - ***Attribute value level:*** attribute values that range over an ontology (e.g. Gene Ontology (GO) term values for a gene product entity)

Example of Ontology Levels





Key Paradigm for this SP4 Task

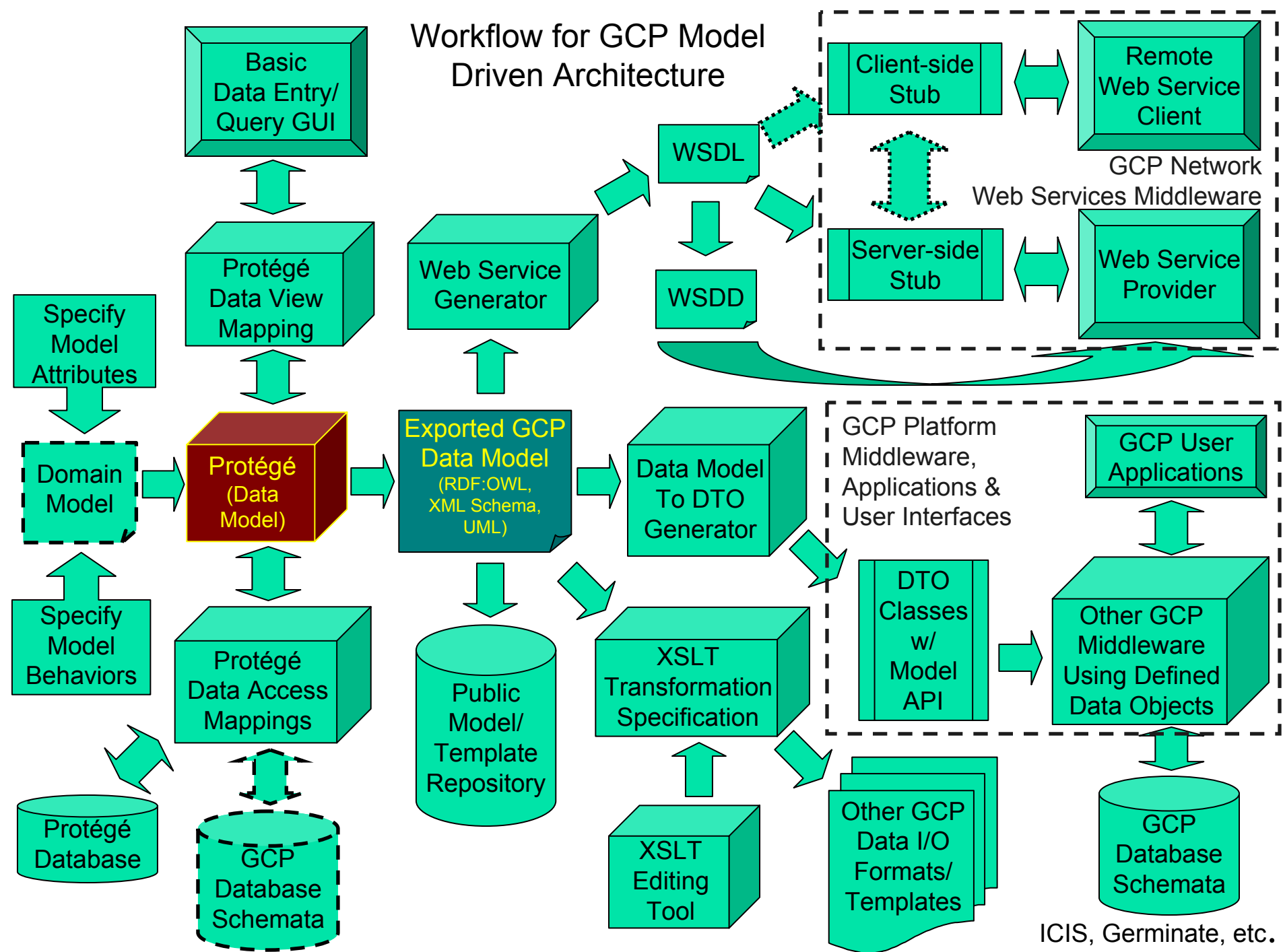
- To represent all levels of the domain model as an ontology description of entity attributes and behaviors, encoded in a XML-based ontology language (OWL/RDF)
- Use the resulting XML domain model as a blueprint for GCP template, platform and network (web services) software design



Proposed Domain Modeling Methodology

- Develop GCP domain model, including model behaviors, using a powerful, freely available ontology-driven knowledge domain modeling tool (Protégé)
- Use automatically generated data interfaces & database mapping facilities (in Protégé) as scaffold for basic data entry/query interface and persistence (e.g. user interfaces for GCP data templates?).
- Export domain model in a suitable XML format for publication and community review (in a model repository).
- Use XSLT scripts to convert domain model into other data formats (e.g. templates/scripts for input/output of data from other tools)
- Use domain model to generate data transfer objects (e.g. Java beans, Perl objects) directly usable by GCP platform implementations.
- Use domain model for (semi-)automatic generation of web services provider and client software

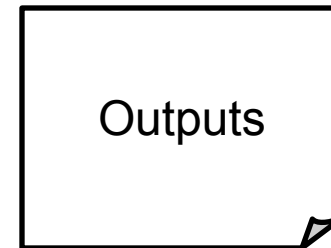
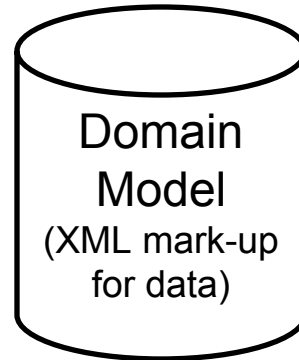
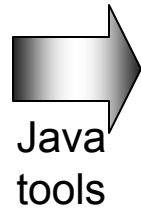
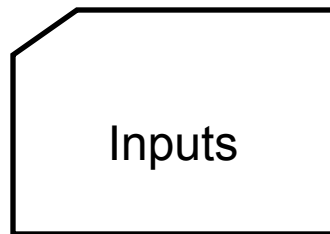
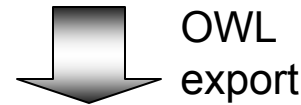
Workflow for GCP Model Driven Architecture



Domain Modeling ↔ Template Tasks

Priority Data Types:

- Germplasm
 - Passport
 - Genealogy
 - Phenotype
- Markers
- Genotypes



Templating task

Modeling task

MOBY tasks

Joint and other tasks

1. Web template input form (Tomcat/Struts/JSP)
2. Spreadsheet (Apache POI)
3. LIMS exported data (ICIS/CIPPEX)
4. Web service (clients)
5. Flat file (text)

1. Web (template) data display
2. Spreadsheet (Apache POI)
3. Database repositories (ICIS/Germenate)
4. Web services (providers)
5. Flat file (XML, registry)



January-February Activities/Milestones

- Use case database commissioned and populated with year 1 white paper use cases
- Inventories compiled/refined for:
 - Priority GCP data types (entities)...
 - ... with pertinent public standards
 - Existing data sets...
 - ...and priority data formats
- 2004 domain model architecture reviewed and refined, for prioritized implementation of model (attribute/behaviour) details providing quick delivery of data templates and web services for year 1 and priority year 2 data, using consensus domain modeling protocols & tools
- Elements of domain modeling workflow prototyped

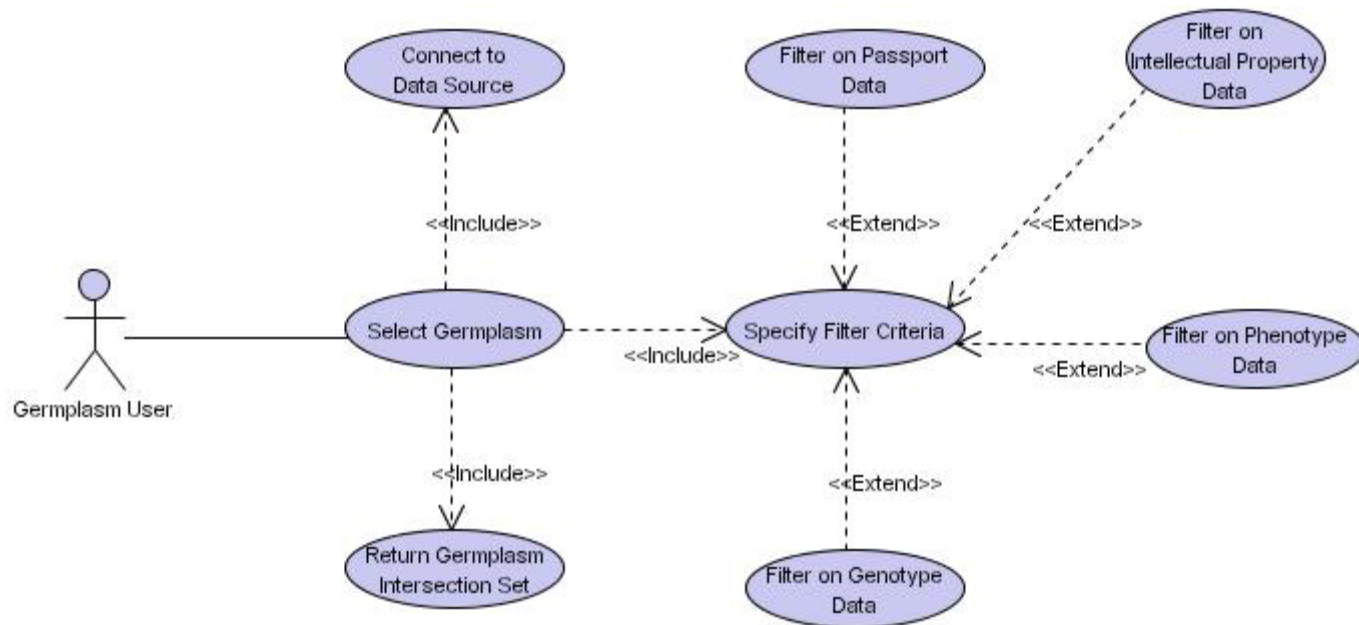
Questions? Discussion?



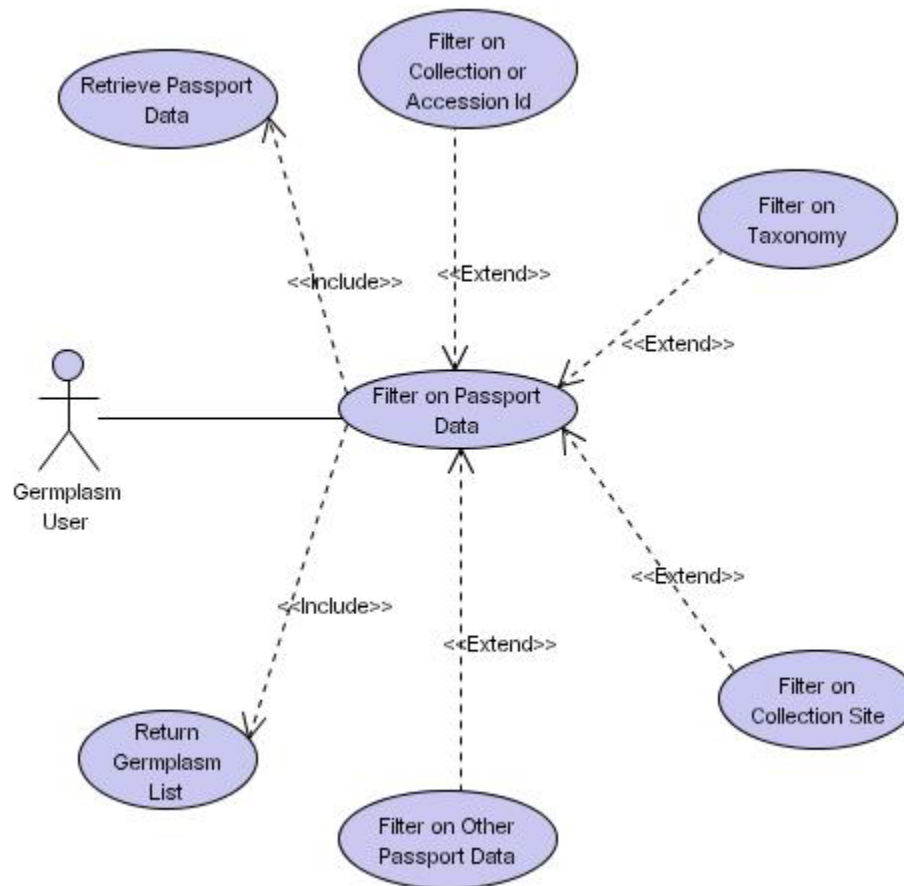


GCP Use Cases (UML from 2004)

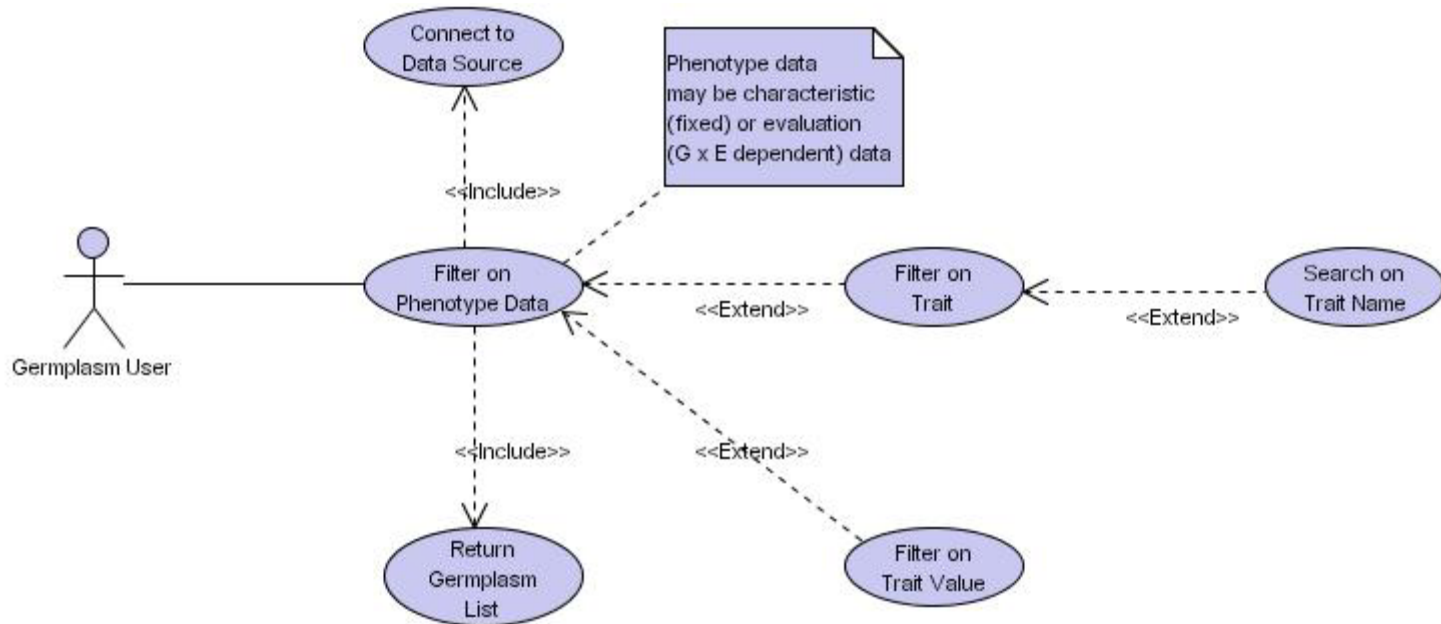
Germplasm Selection Use Case



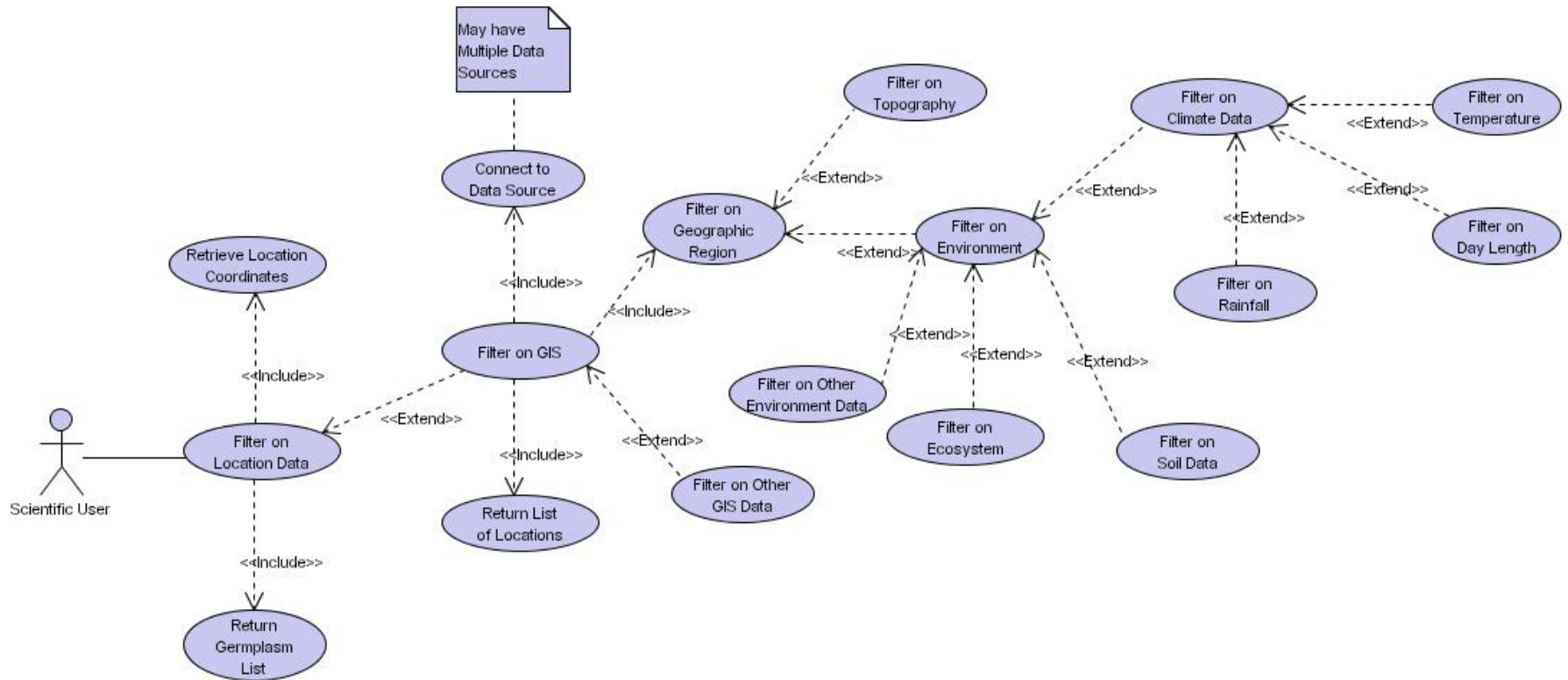
Filter Germplasm on Passport Data



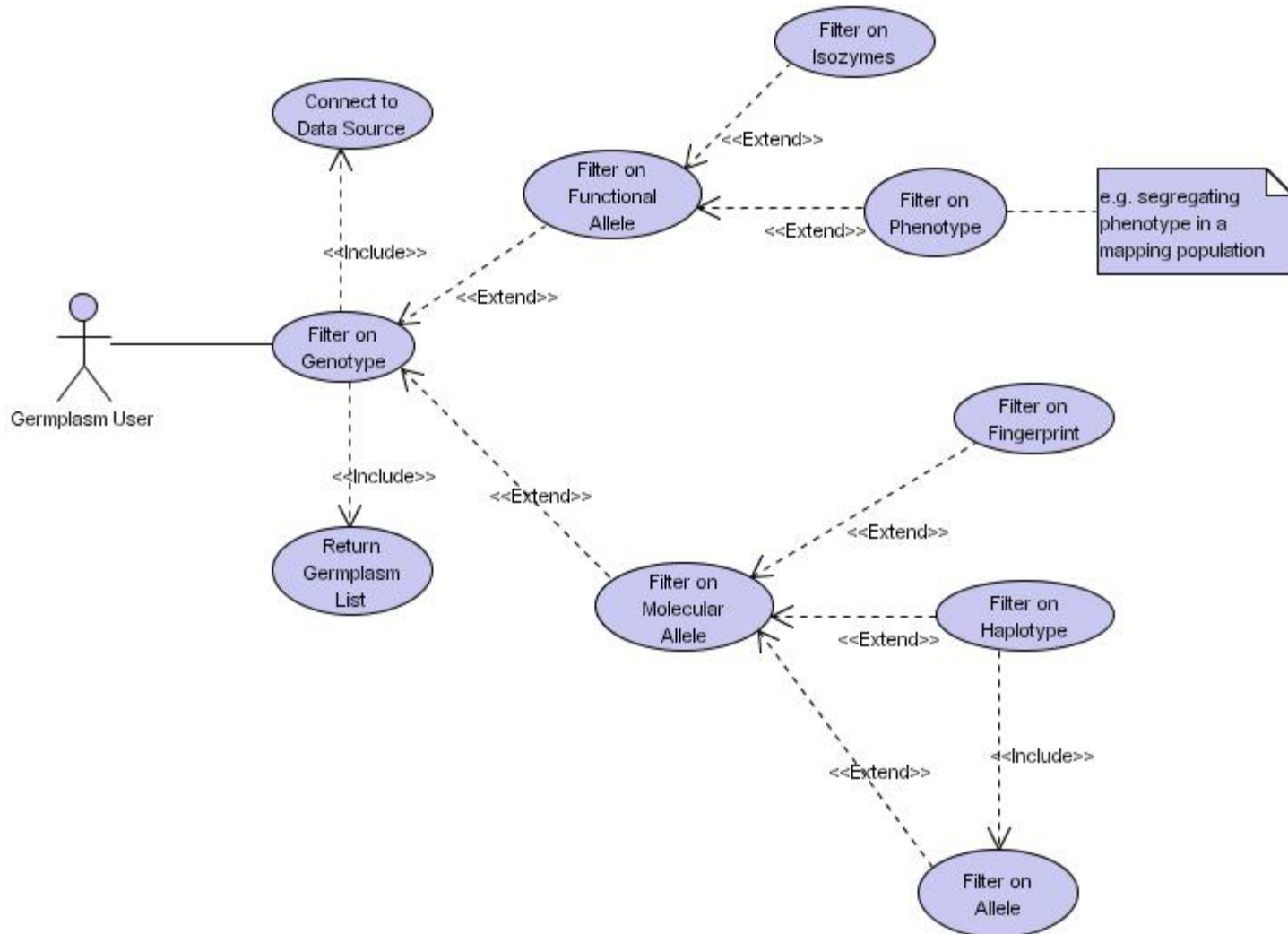
Filter Germplasm on Phenotype Data



Filter Germplasm on Location Data



Filter Germplasm on Genotype

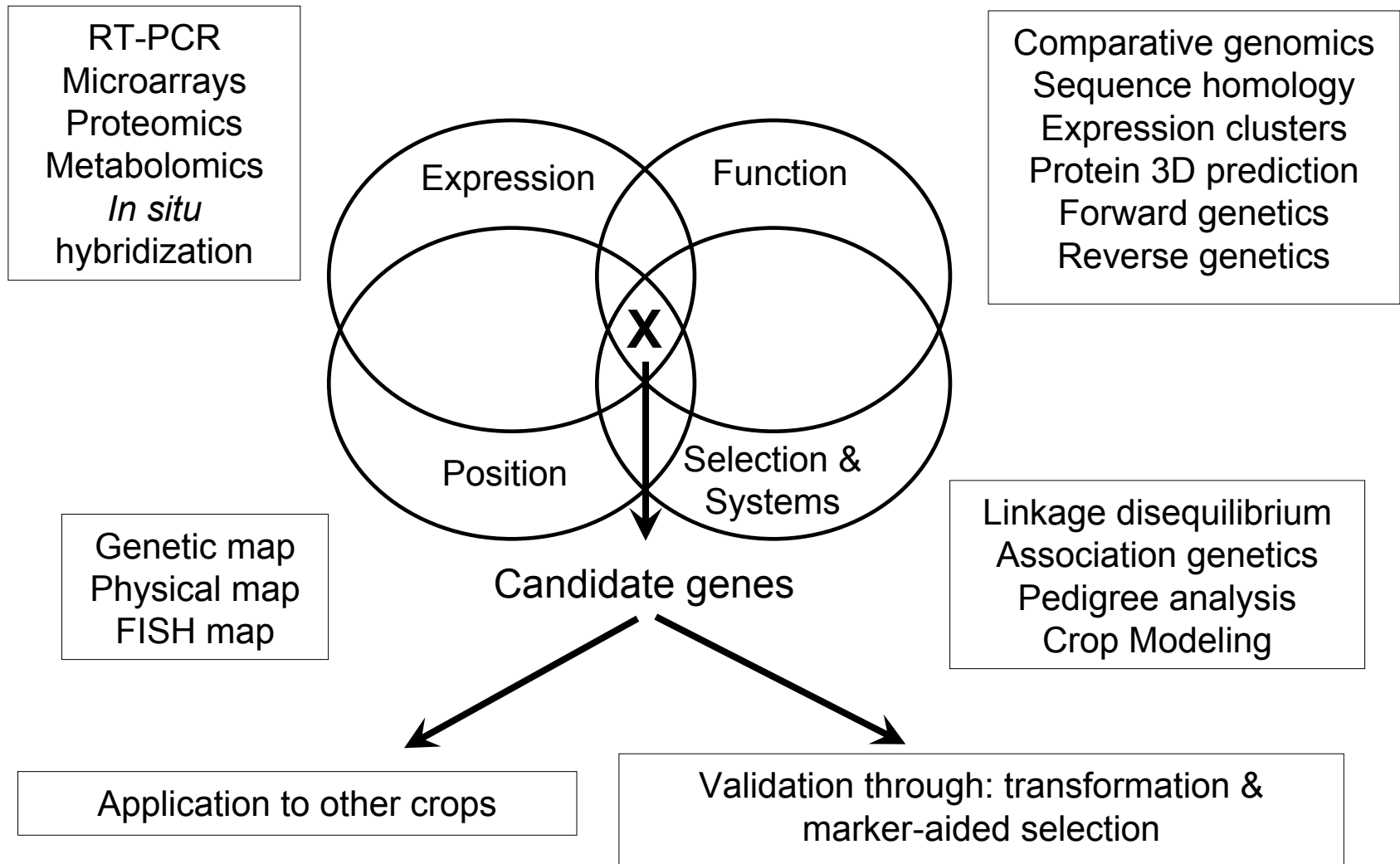


Genetic Diversity Analysis Use Case



Global Subprogramme 2 Use Case(?)

Overlapping Evidence to Identify Candidate Genes





GCP Use Case/Project Management System

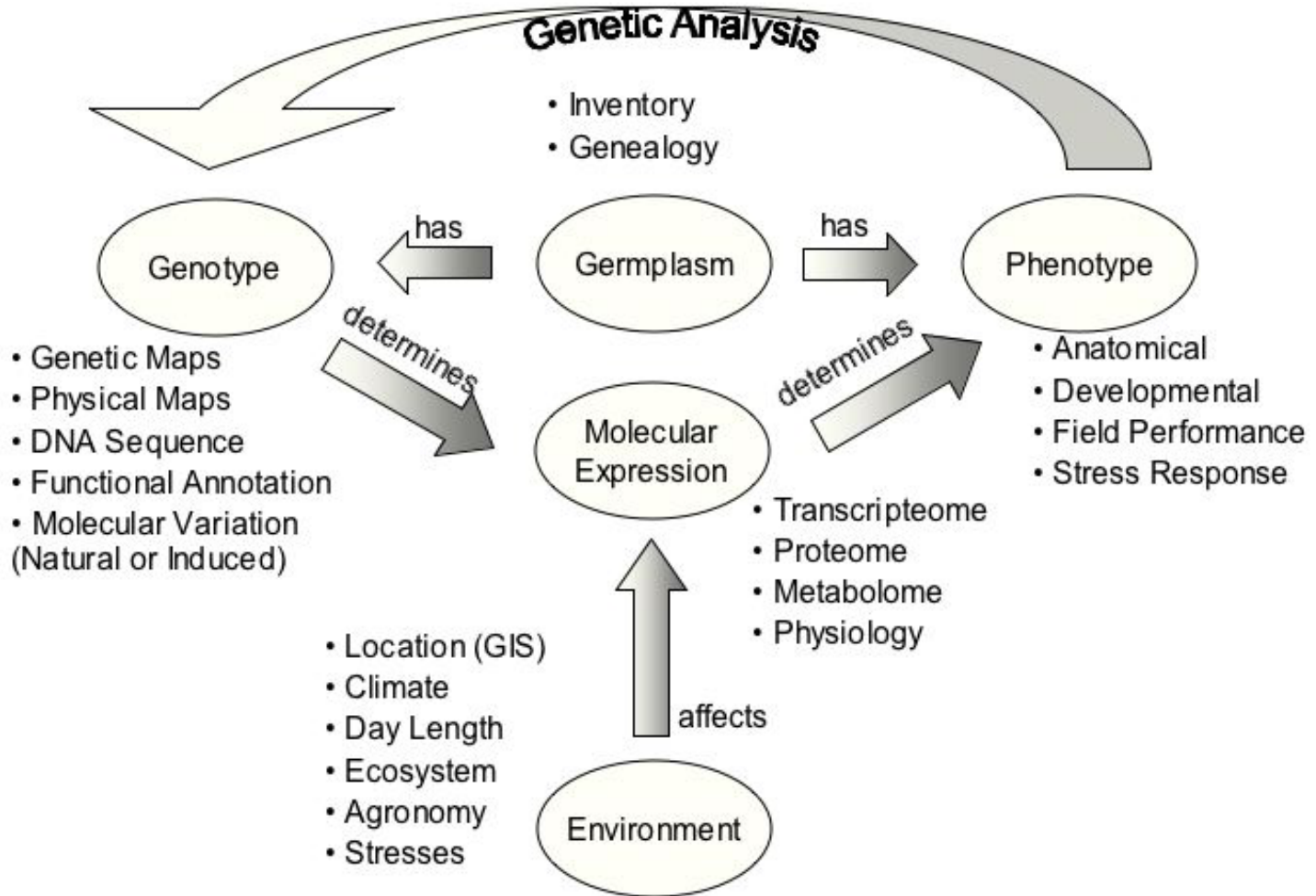
- MySQL database schema
- Standalone Java use case index curation tool
- Read-only WWW use case index browsing interface
- Integrated with detailed WIKI use case documentation and with Gforge ("CropForge") project implementation management system

Demo...

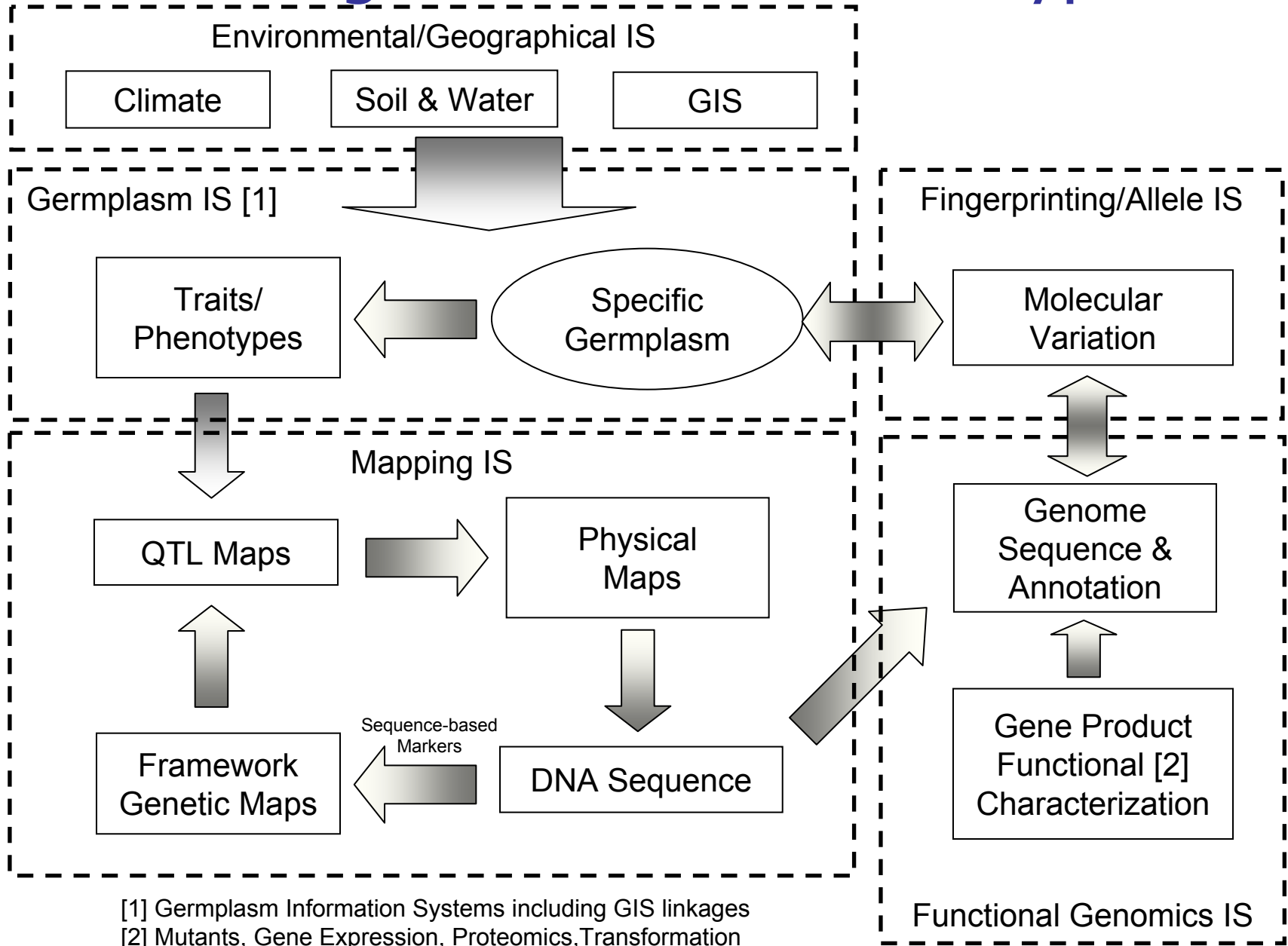


2004 High Level (UML) Domain Models

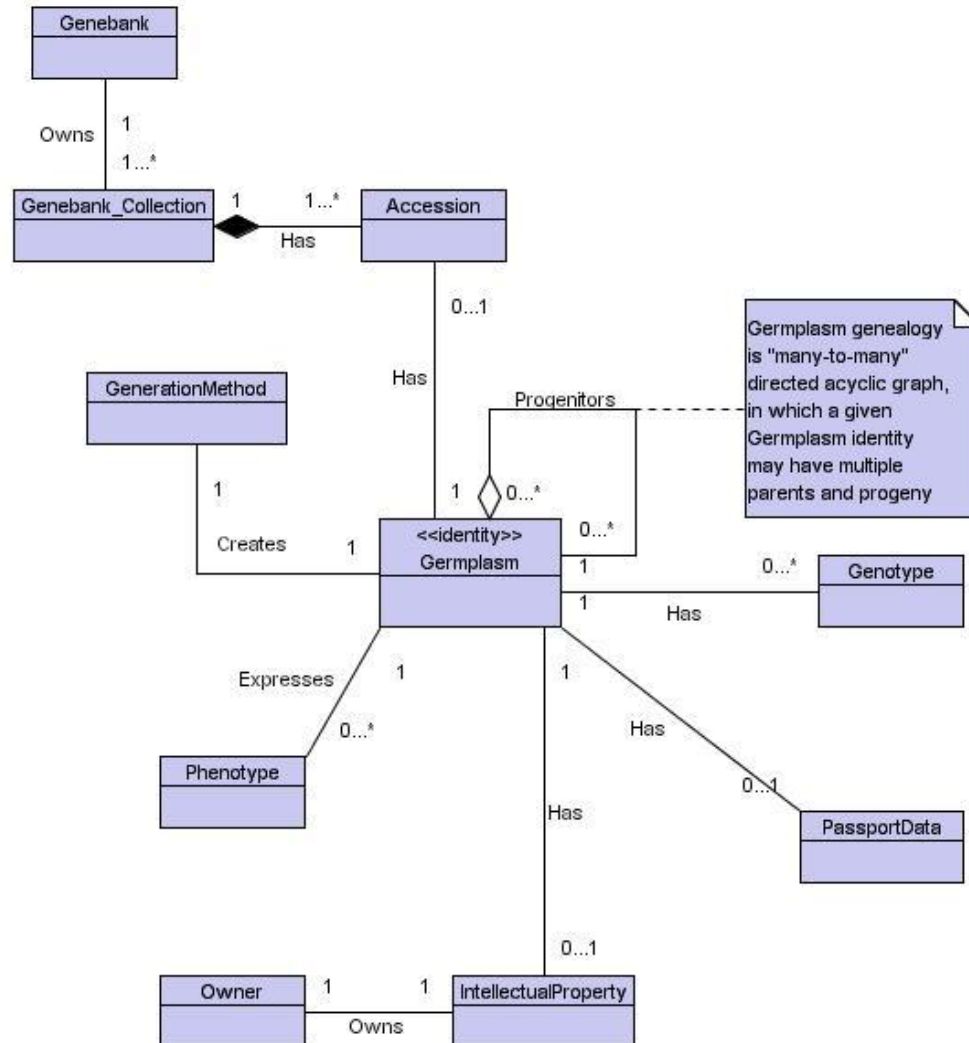
Figure 1 - Crop Biological Concepts, Relationships and Data Types



GCP Integration across Data Types

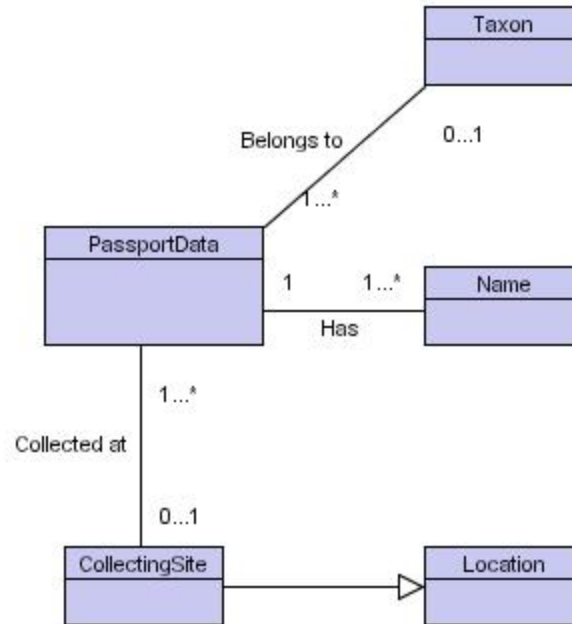


Germplasm Data Model

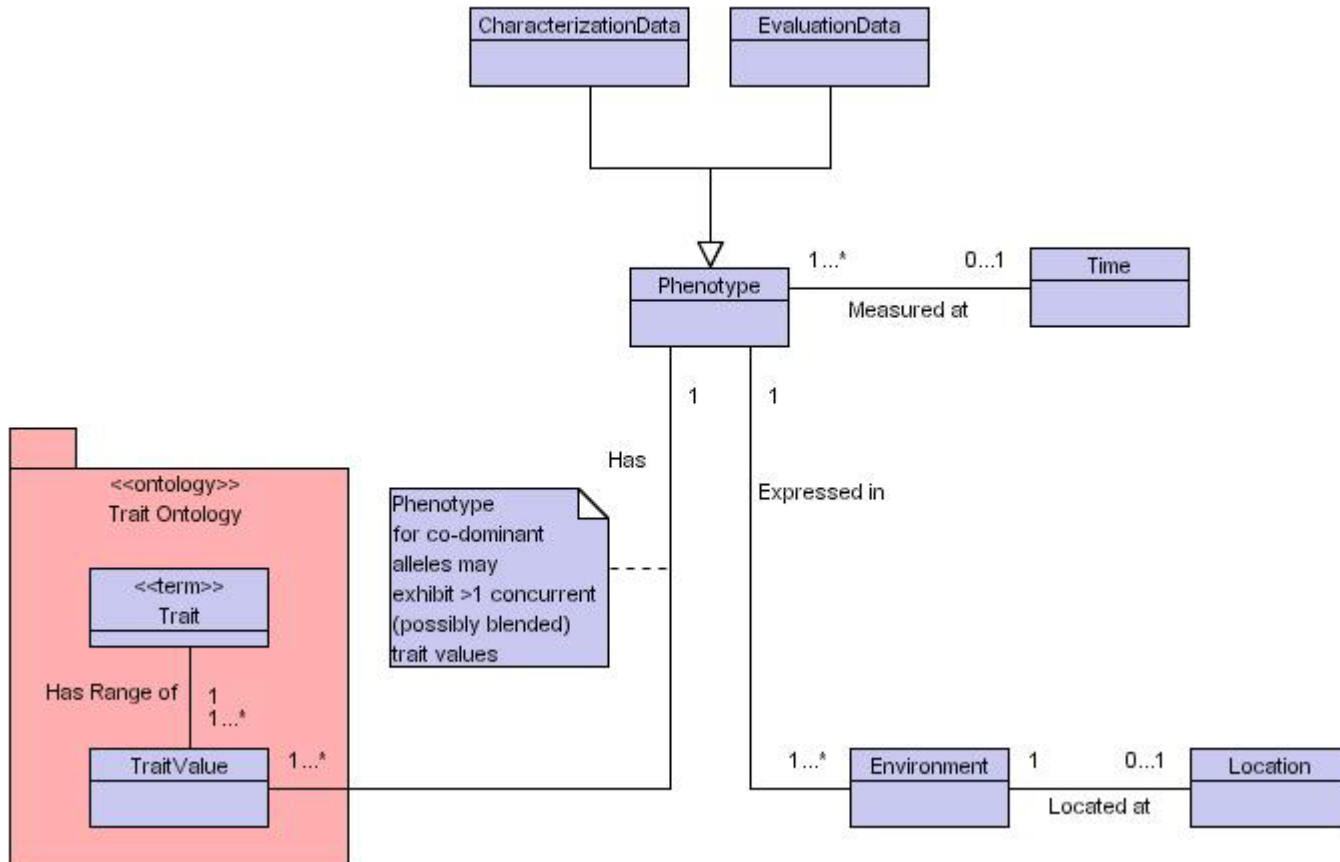


Passport Data Model

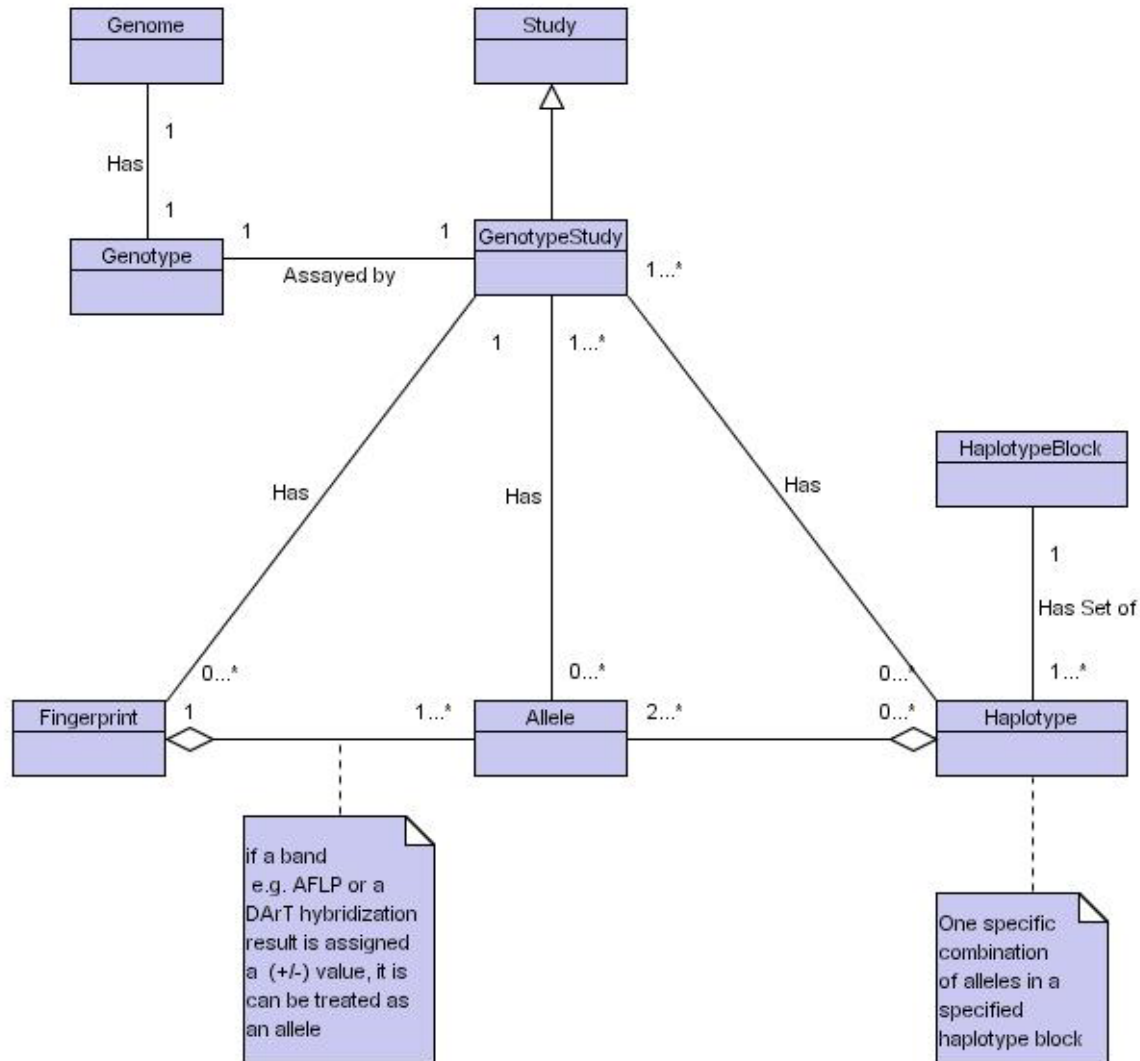
Passport
Data Model



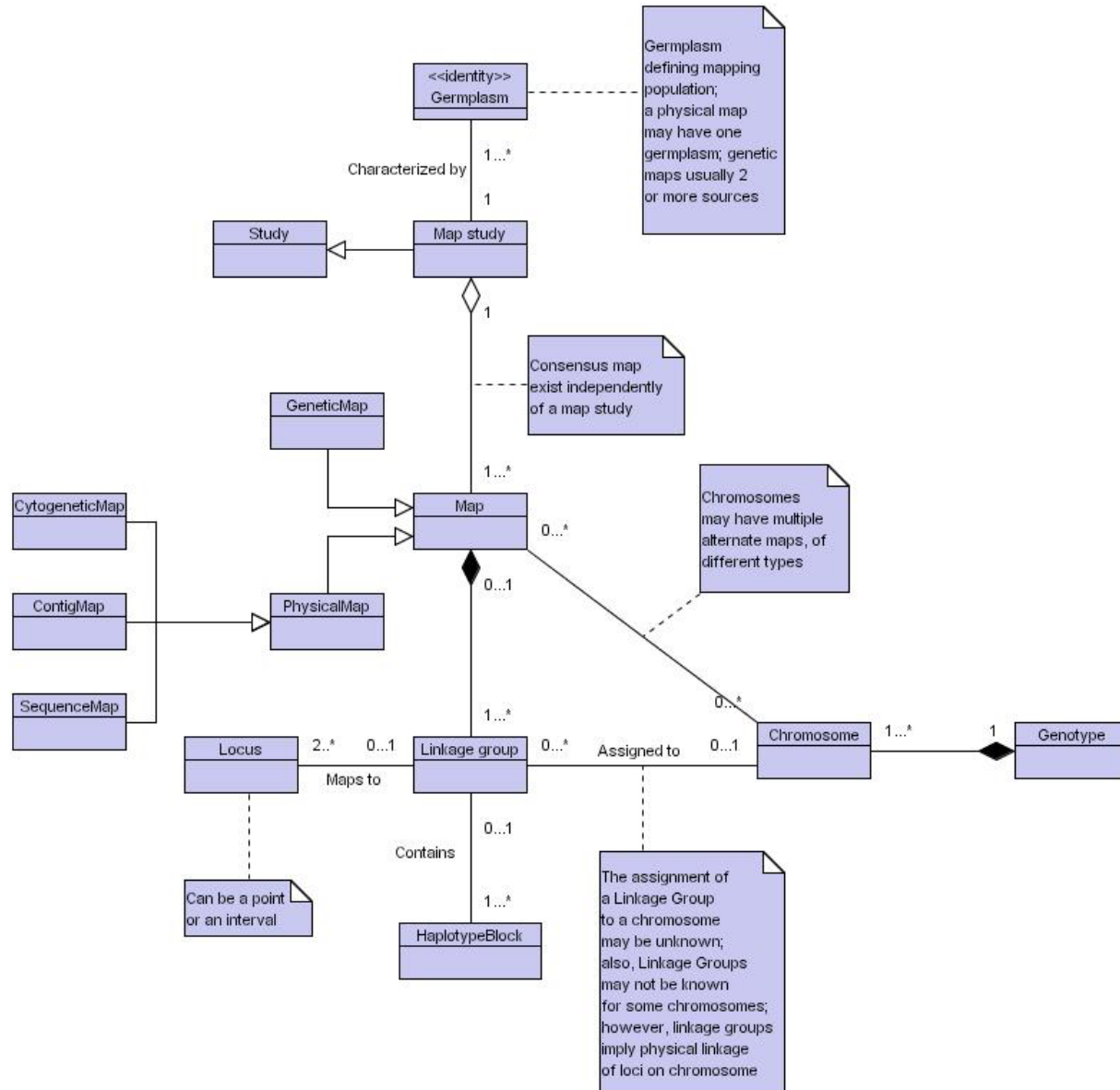
Phenotype Data Model



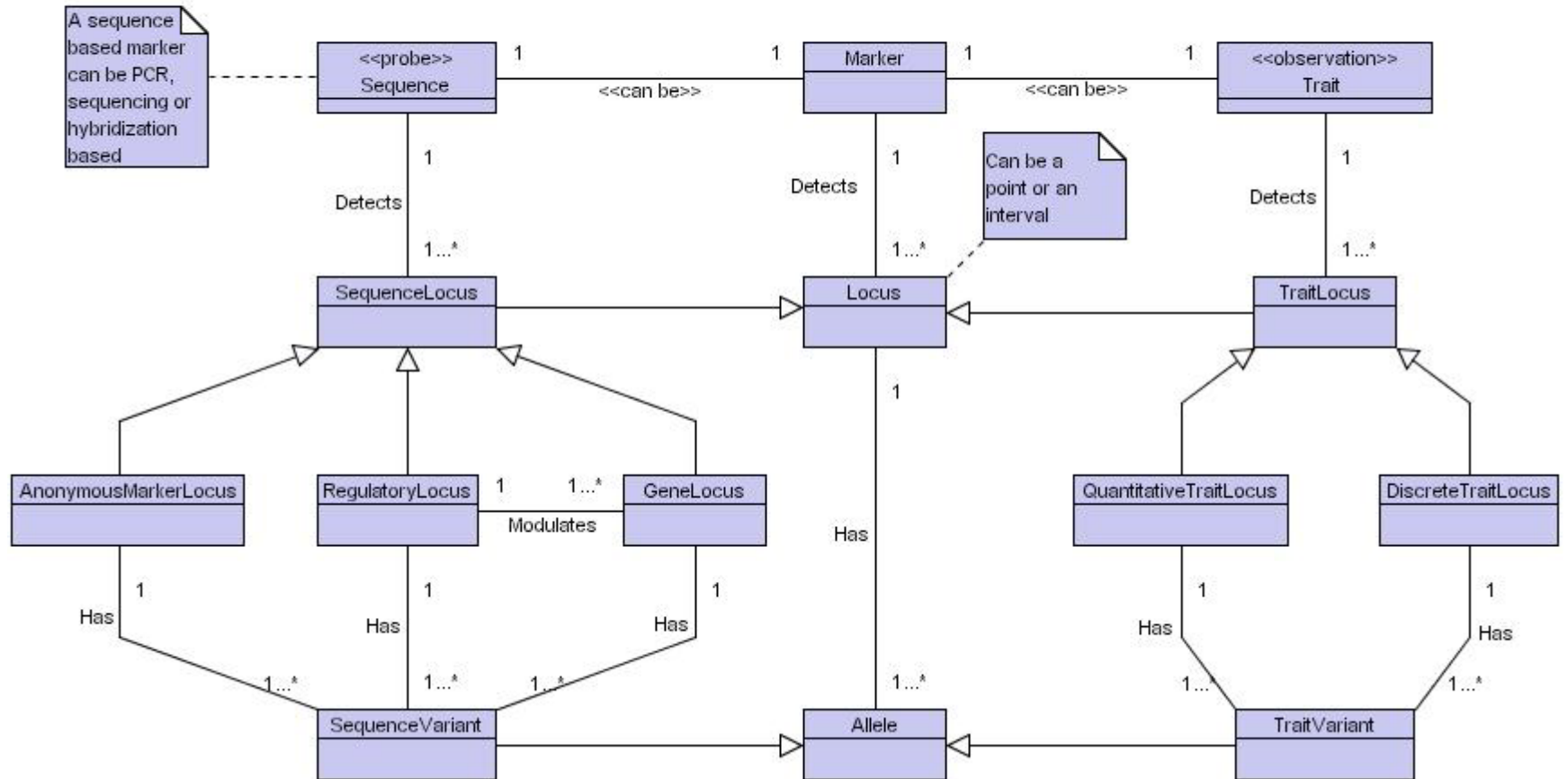
Genotype Data Model



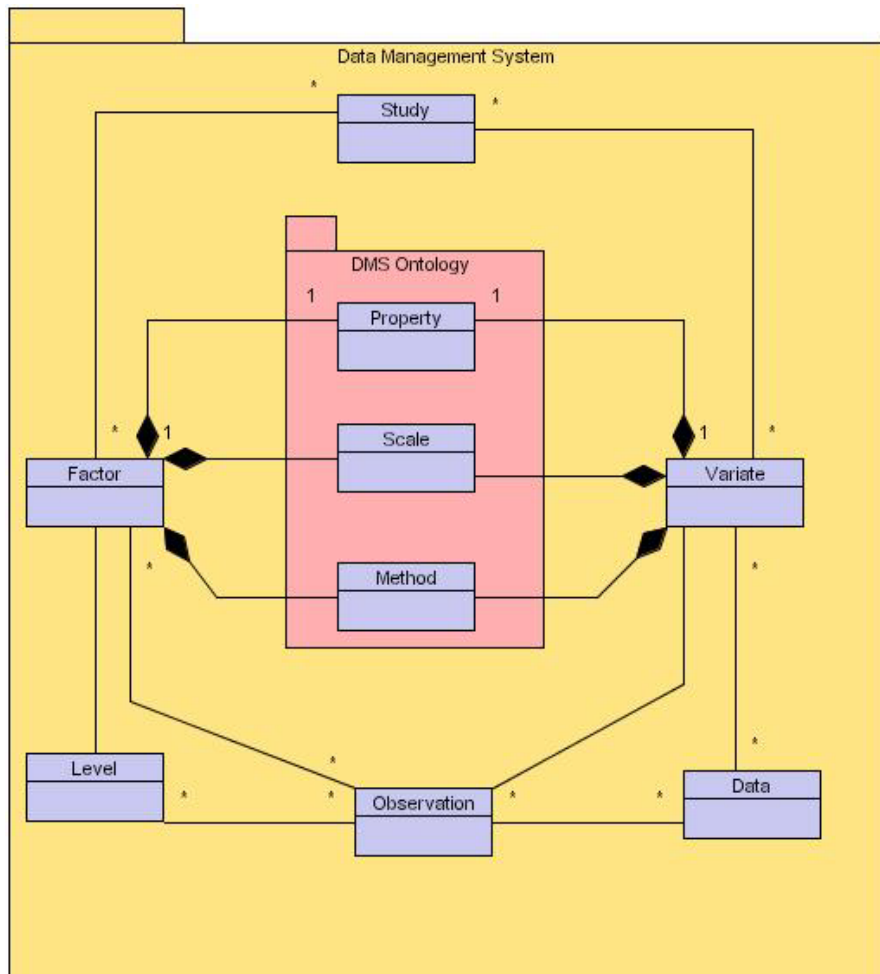
Map Data Model



Locus Data Model



Experimental Study Data Model(?)



DMS from ICIS
(for review by team)

Study types:

- Map studies
- Genotype studies
- Phenotype studies
- LIMS data

Pertinent ontology:

- Marker ontology
- Phenotype ontology

Functional Genomics

Domain Models & Ontology

- Sequence maps:
 - Gene and sequence ontology
 - Schemata: Ensembl, Chado
- Gene expression data:
 - MGED: MIAME, MAGE-ML
 - Plant ontology
 - Schemata: SGD/Longhorne, Barleybase, ?
- Mutant stocks:
 - Plant and phenotype/trait ontology
 - Schemata: ICIS(?) + sequence schemata (FST)



Questions? Discussion?
