

# **Wageningen workshop report**

## **Meeting Proceedings**

### **2005 Domain Modeling Commissioned Research Task**

Wageningen (WICC) Workshop  
February 14th-16th, 2005

**Generation Challenge Programme**  
*Subprogramme 4 - Informatics*

## Contents

- 1 Executive Summary
- 2 Participants
- 3 Monday, February 14th
  - 3.1 Overview of principles, objectives and proposed strategy for domain modeling task work plan, reviewing of 2004 achievements
  - 3.2 Preliminary Reports from Assigned Editorial Teams
    - 3.2.1 Germplasm Genealogy/Phenotype/Genotype (+ generic model support)
    - 3.2.2 Passport data
    - 3.2.3 Location and environment data
    - 3.2.4 Mapping data
    - 3.2.5 Functional genomics team
      - 3.2.5.1 Basic concepts (semantic entities)
      - 3.2.5.2 Available models
      - 3.2.5.3 Available ontology
      - 3.2.5.4 Use Cases (Functional Genomics)
- 4 Gene Expression Repository and Data Mining Task
  - 4.1 Martin Senger about PML
  - 4.2 Terry Casstevens about GDPC
  - 4.3 End of Afternoon
- 5 Tuesday, February 15th
  - 5.1 What is an ontology?
  - 5.2 Concept v/s Identity v/s Name v/s Definition
    - 5.2.1 Concept
    - 5.2.2 Identity
    - 5.2.3 Names
    - 5.2.4 Definition
  - 5.3 Documentation of Semantic Relationships
  - 5.4 Ontology and Model Driven Architectures
  - 5.5 Choice of XML Model Representations and Tools
  - 5.6 Domain Modeling and the Template Task
  - 5.7 Domain Modeling with the GCP Platform and Network
  - 5.8 Overview of 2004 UML Domain Model
  - 5.9 Editorial team breakout teams

- 6 Wednesday, February 16th
  - 6.1 Editorial team Reports from Breakout Sessions
    - 6.1.1 Germplasm team
    - 6.1.2 Passport team
    - 6.1.3 Phenotype team
    - 6.1.4 Genotype team
    - 6.1.5 Location and Environment Domain Model Team
    - 6.1.6 Mapping team
    - 6.1.7 Functional genomics team
  - 6.2 General Discussion & Domain Modeling Priorities
  - 6.3 Closing comments by Theo Van Hintum
  - 6.4 Key Action Items

### **Executive Summary**

This is a summary report on a Generation Challenge Programme Subprogramme 4 workshop held in the Wageningen International Conference Centre (WICC) for task planning and review for the 2005 commissioned research task SP4-22 entitled "Development of Generation CP domain (data) models". For many of the sessions, associated PowerPoint presentations are available on the GCP virtual workspace. The technical host of the workshop was the task leader Richard Bruskiewich. The local organization was provided by Theo van Hintum and his team at WUR. The full list of scientific participants is given in Appendix A. The primary output of the workshop was the elaboration of Phase I and Phase II activities for the published work plan. Phase I deliverables by the end of April will include the development of a first release Unified Modeling Language (UML) data model in several thematic areas listed herein, and its publication in an online domain model repository. This initial modeling will be undertaken in collaboration with a limited number of domain experts and members of other pertinent GCP tasks. Phase II elaboration of the models will include downstream application of the models to year 2 template, platform and network systems. The next formal meeting relating to this task is tentatively scheduled during the two weeks of May 9th - 20th, 2005 at IRRI (in the Philippines). Editorial teams may wish to schedule theme-specific modeling face-to-face modeling discussions at that time.

## Monday, February 14th

Introduction for the domain modeling task and workshop Theo van Hintum welcomed workshop participants to the workshop and presented his vision of the domain modeling task. He emphasized the central role of domain modeling for achieving the interoperability required in the Generation CP platform and network, the importance of the models for other software development activities in the GCP, but also the possibility to link the GCP with the global community both by the modeling exercise but also by the resulting products. GCP models have the potential to be global standards. Special emphasis was placed on reusing existing models and ontology from the public domain, and involving the global community as much as possible.

### **Overview of principles, objectives and proposed strategy for domain modeling task work plan, reviewing of 2004 achievements**

After discussing the proposed meeting agenda, Richard Bruskiewich made a [<http://cropwiki.irri.org/Wageningen/Monday/DomainModelingTaskOverview.pdf> presentation] of related 2004 activities and a technical overview 2005 task goals, objectives, expected outcomes and proposed strategy, as presented in the task work plan (see presentation).

The key discussion issue arising out of this introduction related to choice of modeling technology: there was some uncertainty about the selection of Protégé and OWL/RDF as the proposed modeling technology, as opposed to the UML models (instantiated as XMI XML descriptions). This technical issue cropped up several times during the meeting (see below). After coffee break, a first review was undertaken of the Unified Modeling Language (UML) specified GCP use case and class diagrams originating from the CIMMYT and IRRI GCP design and implementation meetings in 2004. This review included a brief demo of the GCP "Use Case Database" management tool cross-indexing use cases into a document archive based on Wiki ([cropwiki.irri.org](http://cropwiki.irri.org)) and a project management system based on Gforge ([cropforge.irri.org](http://cropforge.irri.org)).

### **Preliminary Reports from Assigned Editorial Teams**

The 2005 work plan for this task made proactive provisions for the division of the task into several subdomains for expert modeling by smaller teams. The leadership of these teams were assigned to identified experts in specific GCP partners:

- Germplasm genealogy, phenotype and genotype module data – Richard Bruskiewich, IRRI
- Passport data module – Tom Hazekamp, IPGRI
- Location and environmental data models - Edwin Rojas, CIP
- Mapping (QTL/genetic/sequence) data module - Manuel Ruiz, CIRAD
- Functional genomics data - Masaru Takeya, NIAS

In the afternoon, the invited editorial team leaders presented their preliminary thinking concerning modeling in their assigned domain, seeking to provide an initial answer to the following key questions:

- What are the basic concepts (semantic entities) within your thematic domain(s)?
- What models and ontology are currently publicly available for the theme?
- What international partners should be engaged in the given theme?

- How will public models be incorporated and partners engaged into theme models (strategy, process)?

The highlights of the presentations are as follows (see further details in PowerPoint presentations):

### **Germplasm Genealogy/Phenotype/Genotype (+ generic model support)**

- Germplasm (Genealogy) Model
  - Basic Concepts: germplasm as biological identity; may have multiple names
  - Public models: will start with the ICIS “Genealogy Management System” model including the ICIS germplasm method controlled vocabulary
  - International partners:
    - ICIS community; all other editorial teams; model plant db’s
    - Germplasm derivation ontology under development in Australia (I. Delacy and co.)
  - Strategy & Process:
    - Compile comprehensive ontology model of germplasm from ICIS
    - Review model at ICIS workshop(s)
- Phenotype Data Model
  - Basic Concepts: observables x attributes = traits, trait values, value scales
  - Public models: PATO/EAV model with Plant Ontology and PATO ontologies
  - International partners: Plant ontology, PATO, model organism databases: Gramene, NASC, TAIR, MIPS
  - Strategy & Process:
    - Extend GCP Montpellier phenotype workshop “drought trait” table effort in 2005 SP\* workshops on phenotyping
    - Collaborate with IRFGC (and POC) rice (and Arabidopsis) mutant phenotype documentation efforts, as a model for other crops
- Genotype Data Model
  - Basic Concepts: marker, allele, frequencies, sequence (SNP)
  - Public models: GDPDM, Germinate, ICIS, OMG SNP, MaizeGDB
  - International partners: germinate, GDPC/Gramene, HAP Map project (at CSHL) and OMG
  - Strategy & Process: develop abstract model in partnership with CGIAR center representatives, Germinate and Gramene
- Generic Domain Models

- Basic Concepts: ontology and nomenclature, time (ontology), organizations, persons, publications, intellectual property, statistical study (and associated classes)
- Public models: GO (Chado) models for ontology management, publications, etc.; ICIS Data Management System (DMS) data model for Study et al.
- International partners: ICIS community, GMOD (Chado), other specialist communities (e.g. IP)
- Strategy & Process:
  - Adopt elements of Chado into GCP domain model
  - Capture ICIS DMS in a domain model
  - Develop custom models and ontology (e.g. IP, time) in consultation with experts

### Passport data

- Scope: "To provide the basic information used for the general management of the accession (including registration at the genebank and other identification information) and describe parameters that should be observed when the accession is originally collected."
- Basic concepts:
  - What?: Name, Status, Identifiers
  - Where?: collecting location, pedigree, source
  - When?: collecting & acquisition date
  - Who?: Holding institutes/ collection, safety duplication holding institute, collectors, donors, breeders

Initial 2004 UML model was observed to be incomplete.

- Public models and ontologies available/Public models to be incorporated: Darwin Core, Access to Biological Collections Data, FAO/IPGRI Multi-crop Passport Descriptors

Elaboration of germplasm genealogy, GIS and sub-accession components required.

- International partners: SINGER, EURISCO, GRIN, IPK, CODATA/TDWG Task Group on Access to Biological Collection Data
- Suggested approach:
  - Phase I (- April 2005)
    - Coordinate with IRRRI on pedigree model
    - Coordinate with CIP on location model
    - Consensus with CP partners on Passport data needed (elements and definitions)
    - Consensus Passport data to serve as input for "Template Task"
  - Phase II (May-Dec 2005)

- Build consensus with external partners to develop broad support for formats and models
- Develop full Passport domain model

### Location and environment data

- **Basic Concepts:** see presentation (biodiversity, DIVA and OpenGIS concepts)
- **Public models:**
  - Open GIS
  - **Ontology:** ISO 3166 Code List of countries, FGDC-CSDGM.- Ontology for Content Standard for Digital Geospatial Metadata (CSDGM) of Federal Geographic Data Committee (FGDC), ISO-19108 Ontology for Geographic Information - Temporal Schema (ISO 19108), ISO-METADATA.-An ontology representing Geographic Information Metadata - (ISO 19115); OGC Ontology for Geography Markup Language (GML3.0) of Open GIS Consortium (OGC).
- **International partners:**
  - OpenGIS Consortium and associated software projects.
  - Global Biodiversity Information Facility
  - Environmental System Research Institute ([www.esri.com](http://www.esri.com))
  - International Institute Geo-Information Science and Earth Observation
  - Center For Research In Water Resources, University of Texas at Austin
  - Carleton University, Department of Geography and Environmental Studies, Ottawa, Canada
- **Strategy & Process:**
  - DIVA + OpenGIS + ? (editor's note: strategy was unclear from presentation)

### Mapping data

- Scope: markers and loci, genetic maps, QTL maps, physical maps and sequence maps (concept overlaps with germplasm, genotyping and functional genomics editorial themes...)
- Basic concepts: marker, locus, allele, chromosome, linkage group, map, map study, germplasm, haplotype, haplotype block, genotype, genetic map, physical map, sequence map, cytogenetic map, contig map, qtl, comparative mapping(?),
- Public models and ontologies available:
  - Mapping data models (schemata): from plant databases – Gramene, TAIR, MaizeGDB, GrainGenes. NCGR Legume IS, UK CropNet, TropGene, OrygenesDb, Germinate, Ensembl, TIGR, etc.
  - Ontology: Sequence Ontology (SO), IPGRI marker descriptors, OBO ontology (but no real mapping data ontology yet)

- International partners: international plant databases and ontology projects
- Strategy & Process:
  - Confirm a consensus high level domain model structure
  - Identify minimal semantic cross linkages between global modules
  - Establish a moderated forum for public feedback about the domain model.
  - Mapping domain meeting (associated to another related domain meeting?)

### **Functional genomics team**

#### **Basic concepts (semantic entities)**

- Gene expression
  - Microarray, MPSS, SAGE, RT-PCR
    - MIAME with related gene and plant ontology
- Mutants
  - Organism
    - Knocked-out genes: Gene Ontology
    - Expressed Traits: Trait Ontology, Plant Ontology
- Proteomics
  - 2D gel, including mass spectroscopic sequencing of spots
  - Post-translational modification
  - 3D Structure
    - X-ray
    - NMR
  - Interaction
    - Protein-protein
    - Protein-nucleotide
- Metabolome
  - High throughput mass-spectroscopic assays
- Biological networks, pathways and systems

#### **Available models**

- NCBI-GEO, EMBL-EBI
- Chado in Generic Model Organism Database Construction set (GMOD)
- MIAME documentation
- International Rice Functional Genomics Consortium (IRFGC): pertinent projects

- NIAS Functional Genomic Databases: RED, Rice Tos17 Insertion DB, Rice Proteome DB

### **Available ontology**

- Biological ontology consortia (GO, POC, OBO)
  - Trait Ontology
  - Environment Ontology (EO), Cereal Plant Growth Stage Ontology (GRO) in Gramene

### **Use Cases (Functional Genomics)**

- LIMS and capture of data
  - Storage of mutant catalog, array/gel image, experimental results, sample treatments, and cross-linking to pertinent data
- Analysis of association genetics across mutations' allelic series and gene/protein expression profile across treatments
- Query for mutant phenotype and gene/protein expression

### **Gene Expression Repository and Data Mining Task**

An additional unscheduled but informative talk was presented by Shoshi Kikuchi of NIAS relating to the SP4-32 task "Development of crop gene expression database and data mining tools" and discussing NIAS (rice) microarray activities. The basic plan for the task was indicated as:

- Start with the NIAS "Rice Expression Database (RED) as a prototype
- Augment the 9K EST rice data with 22K array data
- Obtain data from Arabidopsis as reference
- Enlarge data set with incorporation of other crop EST data (and MPSS? SAGE?)
- Gene expression data to the domain modeling structure

### **Martin Senger about PML**

Later in the afternoon, Martin Senger presented some information on both the OMG life sciences standardization body and a pertinent modeling initiative, the Polymorphism Markup Language (PML; <http://pml.ddbj.nig.ac.jp/>) standards development activity for single nucleotide polymorphism (SNP) data, which is the result of a collaborative effort between a large and experienced team of international SNP database experts and is being published as an OMG standard. The PML is defined by an XML Schema description. The PML covers many kinds of genetic information - SNP, indel and microsatellite, frequency, genotype, haplotype, (biomedical) phenotype – and a wide variety of information around the genetic data - genome, assay, population, gene, protein and publication. It was generally felt by the team that the GCP editorial teams should closely examine PML to review its relevance to the GCP domain modeling activity. Also, Martin Senger suggested that the Consortium consider the OMG

standardization body as a vehicle for promoting the GCP domain model, once the domain model reaches a suitable state of maturity.

### **Terry Casstevens about GDPC**

The final speaker for the day was Terry Casstevens who presented an updated report on the Genomic Diversity and Phenotype Connection (GDPC), a genotype and phenotype software platform that maps database schemas onto the GDPC object data model (see GDPC Entity Properties) and exposes it as a web service. The platform also includes a basic model query and browsing tool (see [[http://www.maizegenetics.net/gdpc/browser/browser\\_index.html](http://www.maizegenetics.net/gdpc/browser/browser_index.html) GDPC Browser]). Querying is based on filtering out required data from the model. Model properties are designed to be extensible.

The Genomic Diversity and Phenotype Data Model (GDPDM) is a database schema specifically designed to hold diversity data which the Buckler Lab (now at Cornell University) has implemented in several databases. An adaptor (GDPC Connection) translates this schema to the GDPC model mentioned above.

The GERMINATE schema and the Panzea database have also been adapted to the GDPC data model. Other databases under consideration to be adapted to GDPC are ICIS, GRIN, MaizeGDB, etc.

### **End of Afternoon**

The afternoon closed with some summary discussions during which some key issues were highlighted for further discussion during the meeting, in particular, the issue of choice of modeling technology.

## Tuesday, February 15th

The second day began with a general review of domain and ontology modeling principles by the task leader (R. Bruskiwich). Given their importance to the modeling task, the body of this discussion is duplicated here.

### What is an ontology?

An ontology is simply an organized set concepts (“knowledge representation) about a specified domain. In practical implementations, an “ontology” generally consists of two components:

- An indexed set of controlled vocabulary terms (the “concepts”)
- Information about semantic relationships between these terms

An ontology may be represented as a (computing) graph representation, where nodes of the graph correspond to the controlled vocabulary terms and the (directed) edges of the graph designate term relationships.

### Concept v/s Identity v/s Name v/s Definition

Development of an ontology needs to distinguish between the concept itself, tracking of its identity within the ontology, its range of descriptive name(s), and its semantic meaning.

#### Concept

The concept itself is simply assumed to be a class of thing that can be distinguished from other things similar in some fashion but not identical to itself. A concept is generally defined by its characteristic attributes, behavior and relationships to other things and has independent existence from whatever convention we have for naming it. Instances of classes may represent “one of” terminal concepts in a concept class, e.g. Europe is an instance of the “continent” concept.

#### Identity

In practical ontology management systems, the need to distinguish (track or “bar code”) a concept usually implies the assignment of a unique (albeit otherwise semantically meaningless) accession identifier in the ontology system. (e.g. in GO, this is the role of the GO identifier). Such an identifier permits concepts with the same descriptive name, but different contextual meanings, to be distinguished. A stable accession identifier provides for efficient relationally normalized labeling of data with the concept as well as efficient management of the merging and deprecation of concepts and their relationships in the ontology, as the understanding of a domain of discourse evolves.

#### Names

The descriptive names of a concept may include a common (or “canonical”) descriptive name plus any number of synonyms. Also, in practical systems, although the underlying concept may be relatively immutable, the descriptive name may suffer change as the public understanding of the structure of a domain of discourse changes, which generally

precludes simple use of the name as its identity. This will not be a problem if the accession identification of the concept is used to track all such descriptive names.

### **Definition**

The definition provides a complete semantic description (the “meaning”) of the underlying concept of the term, not otherwise conveyed by its canonical name (or its synonyms). Again, the evolution of understanding of a concept may result in a change in definition but if the conceptual entity described by the definition is tracked by a stable accession identifier, data labeled with the concept will not need to be updated directly.

### **Documentation of Semantic Relationships**

Some ontology initiatives (e.g. GO) document relationships between concept terms as directed acyclic graphs (DAG) that organize terms from the general to the specific concepts. Often such relationships are often restricted to two types:

- Inheritance: B “is a kind of” A, where B is the subclass of the generic class A.
- Compositional: D “is a part of” C, where D is assumed to be a subcomponent of C.

However, additional relationships may be added to enrich the coverage of their respective domains of discourse. Relationships may also be defined between classes rather than instances of concepts. The use of a DAG allows that a concept can share more than one parent (generic) concept. For example, a structural protein may have a molecular function AND be part of one or more sub-cellular structures.

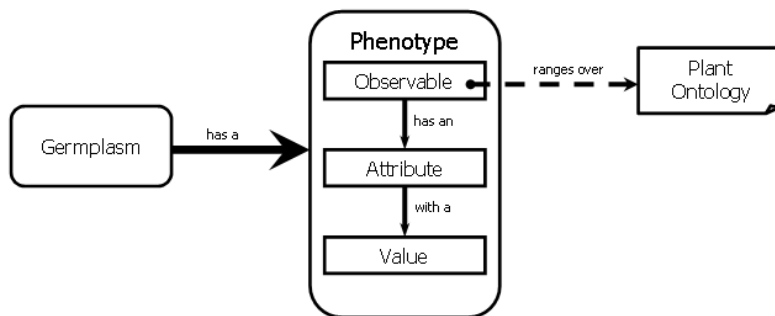
The use of a DAG to represent an ontology, however, is not the only way to represent an ontology. Specifying both direct and transitively reflexive conceptual relationships may be needed, thus breaking the acyclic nature of the DAG (their data structure ends up being a generalized directed graph). Tools such as Protégé allow the construction of such ontology without much difficulty, although the computing on such an ontology is more computationally expensive than DAG processing. “Semantic web” technologies (eXtensible Markup Language (XML) encoded standards of the W3C) such as the Resource Description Framework (RDF) and the Ontology Web Language (OWL) provide suitable encodings of such relationships. Unified Modeling Language (UML) also has this expressive power.

### **Ontology and Model Driven Architectures**

Semantic modeling using ontology can be applied in domain modeling at three levels:

- System architectural level: e.g. black box entities (entity super classes) 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 attributes and methods)
- Attribute value level: attribute values that range over an ontology (e.g. Gene Ontology (GO) term values for a gene product entity)

The following figure represents a portion of the domain model in terms of such ontology levels.



### Choice of XML Model Representations and Tools

XML protocol options – XML Schema, UML/XMI, RDF and OWL – were briefly discussed and OWL/RDF examples given. The Protégé modeling tool was introduced.

A general discussion ensued concerning choice of modeling languages (UML/XMI versus RDF/OWL) and tools (UML versus Protégé) that was somewhat resolved towards the end of the meeting (see action items). XML coding conventions were also raised as an issue to clarify.

### Domain Modeling and the Template Task

A [<http://cropwiki.irri.org/Wageningen/Tuesday/DataTemplateTask.pdf>

presentation of XML model transformation technologies], XSLT, was well formulated by the Template Task leader (Guy Davenport) who also applied XSLT and web technologies to develop a template task prototype data entry tool that was successfully demonstrated to the workshop (see pertinent presentations). The clear conclusion from Guy's presentations was that XSLT and related technology are powerful tools for manipulating XML representations of the domain model into template and other SP4 products. It is also relatively easy to design and deploy basic input and output interfaces to XML specified templates, which may be quickly deployed by open software technology (e.g. Tomcat and Struts/JSP) and viewed on the WWW using built-in web browser XSLT technology. Thus, the template task will be driven by XML and XSLT technology (the first fruits of the domain modeling activities of the GCP SP4).

Guy also mentioned the joint activity of a single entry point web site associated with the GCP web site, for both the data modeling and templating tasks including:

- The list of data types and templates needed in the GCP.

- Requests for additions, amendments and comments from GCP partners and others.
- Link to sections containing progress in both tasks
- Include documentation and current template tools

Additional features of the joint template/domain modeling repository relating more to domain modeling were enumerated later in the afternoon:

- W3C/OMG-like model of formal RFC, with model versioning (in a CVS; possibly a CropForge project?); versioning releases of basic (“high priority”) to more elaborate models. Could pre-publish intentions for future version releases, well in advance of the releases?
- Global table of contents for documented model entities and adopted public ontology/models/standards.
- Publication of model diagrams (e.g. UML) with online primer for novices, concept glossaries (in Wiki?), computable design files (e.g. XMI, OWL/RDF) and links to modeling tools and user interfaces.

Each GCP template on the repository is expected to have:

- XML schema
- Set of recognized input formats
  - Each with a parsers to create an XML instance
  - Single web form for all input formats
- Set of recognized output formats
  - Excel, analytical tools, database import tools and to create easy to read text including web pages
  - Each with a script to transform the XML instance
  - Single web form for all output formats

Each template task team member is expected to deliver a complete package for one template (data) type, using XSLT, etc. technology that Guy will deliver. The package should include: an XML schema, documentation, description of common input formats, required output formats and examples. Prioritization of data types and assignment to template task participants was a discussion item put on the table by Guy. This priorities were not be finalised at the meeting, but will be finalised by early-mid March.

### **Domain Modeling with the GCP Platform and Network**

Alex Cosico of IRRRI gave a talk about the integration of XML models into Java software engineering of the GCP platform, using an ICIS5 prototype as a case study. Java technologies such as Hibernate (for semi-automatic Java bean (object) to relational database mapping) and XML Stream (for Java bean serialization to/from XML) were presented as tools to transform an XML domain model into software and web service artifacts for the GCP platform and network.

## Overview of 2004 UML Domain Model

The UML domain (class) model constructed in 2004 as an outcome of discussions at the CIMMYT and IRRi design workshops was revisited, diagram by diagram to identify errors, refine the scope of concern for the editorial teams and to identify “strata” cross-linkage entity (identities) spanning two adjacent domain model modules (as partitioned by the editorial modeling teams). The following table clarifies the scope of each modeling team entities and primary cross-link entities (as well as generally projected Phase I milestones).

<b>Module</b>	<b>Leader</b>	<b>Scope</b>	<b><u>Strata Link(s)</u>*</b>	<b>Phase I Milestone</b>
<u>Utility components</u>	Richard (IRRI)	Utility Entities (Study, Time)	Utility entity IDs	Commission domain model repository w/ amended 2004 high level models; base models for germplasm & studies
<u>Germplasm genealogy</u>	Graham (IRRI)	Core germplasm entity, genealogy and derivation methods (ontology)	Germplasm IDs	Formalize (ICIS) germplasm entity description
<u>Passport</u>	Tom (IPGRI)	Passport, Accession, Intellectual property	Germplasm	Base models for Yr. 1 GCP core collection data
<u>Genotype</u>	Thomas (IRRI)	Marker, Genotypes, Alleles	Germplasm, Marker, Allele, Sequence (SNP), Gene Locus	Base models for Yr. 1 markers and genotypes
<u>Phenotype</u>	Richard (IRRI)	Phenotypes (including mutant)	Phenotype (ontology)	Base model for phenotype data anchored to ontology
<u>Location &amp; Environment Data</u>	Edwin (CIP)	Location, Environment (ontology)	Location	Support for Yr. 1 passport data; base models for environment
<u>Mapping</u>	Manuel (CIRAD)	G/P/S maps, genetic and sequence loci (SNP)	Germplasm, Marker, Gene, Locus, Sequence, Phenotype	QTL/genetic map model; meet with functional genomics team concerning common genomics models

<u>Functional Genomics</u>	Masaru Takeya (NIAS)	Gene product (family) function (GO) & expression	Germplasm , Gene Locus, Sequence, Phenotype	Compile and compare available public functional genomic domain models, with special focus on Yr 1 data needs; meet with mapping team to coordinate genomics models
----------------------------	----------------------	--	---	--

### **Editorial team breakout teams**

Subsequent to the general overview of the 2004 UML domain model, each editorial team (or subteam) as noted in the above table, convened a breakout session to discuss their theme areas. The teams were generally as indicated in the preliminary reporting above, except that the IRRI germplasm genealogy, phenotype and genotype team was split into a phenotype and a genotype discussion group (genealogy was considered a “solved” model in ICIS). Meeting participations (aside from the team leaders) were free to choose which team they wished to participate in. The discussions were targeted to identify priority activities and milestones for Phase I (target milestones at end of April) and Phase II modeling for the specified theme, elaborating the preliminary strategy presented the first day.

## Wednesday, February 16th

### Editorial team Reports from Breakout Sessions

One rapporteur per team presented a report. All associated PowerPoint presentations should be available on the virtual web site. A synopsis of key points is as follows:

#### Germplasm team

No team was convened at the meeting to discuss germplasm genealogy, although implicitly, the ICIS Genealogy Management System model is assumed to be a starting point for modeling (see preliminary report above).

#### Passport team

- Model Scoping:
  - Use FAO/IPGRI Multi-crop Passport descriptors (MCPD) as core
  - From SP1 Year 2004 solicit passport data available for core collection
  - Based on feedback develop model comprised of MCDP plus multi crop extensions and crop specific extensions
  - Provide in model “container” for additional passport data
- Strategy, Process and Milestones:
  - Phase I (-April 2005): using the MCPD as the core, consolidate domain model based on feedback from SP1 and consultations with GCP partners, resulting in version 0.2 passport data model ready to feed into “Temple Task”
  - Phase II (May-Dec 2005): consult with external partners (i.e. SINGER, EURISCO, GRIN, IPK, TDWG) to elaborate model to full 1.0 version

#### Phenotype team

- A basic data model (based on ICIS DMS, PATO and other inspired sources) to capture phenotype data is deemed deliverable by the end of April (Phase I), followed by a systematic (long term) documentation of phenotype ontology starting in Phase II. The latter activity is anticipated to be significantly more laborious and challenging, requiring extensive involvement of non-SP4 GCP experts as opportunity presents itself (i.e. SP\* meetings).
- See preliminary report presentation synopsis (simply add “QT Consortium” to the list of possible collaborators)

#### Genotype team

- Proposed Participants:
  - ICIS – IIRI
  - ICARDA (enhanced ICIS)

- GERMINATE
- Gramene (GDPDM) – Cornell
- CIMMYT – Marylin / Guy
- MaizeGDB – Iowa State
- OMG – SNP – HapMap – CSHL
- ICRISAT
- *Objectives:*
  - Discussions and modelling with experts involved
  - Preliminary verification
    - By having a broad range of people/ databases included in discussions
    - Need to ensure representation of fundamentals and range of data which exists and is expected to be produced
    - Model should be extensible
  - Close contact with Mapping group to ensure marker and allele model works for their purposes (editor's note: need to coordinate with functional genomics team as well...)
  - Expand on dictionary of terms – important to database administrators to map model to their databases
- *Process:*
  - Discussion groups via CropWiki
    - Viewable to all, editable by some (developers in task group)
    - Versioning with capability to rollback
    - Discussion separate from result content
    - Track changes from individual users
    - They will know what changed from the last time they looked at it.
  - Capture model with task chosen tools (e.g. UML with XMI export)
    - Phase I: core domain models are to be finalized for year 1 marker and (e.g. Zaragoza template specified) genotype data by discussion among a core team of GCP experts and consideration of pertinent published public domain models.
    - Get process going (the how – WIKI)
    - Environment for discussion
      - Testing and deciding on software
  - Preliminary model and dictionary of terms
    - Input from external groups
    - Interaction with mapping group
    - Interaction with Guy's template task

- 1st check of model
  - Should be a part of work environment so it is incorporated into the data model (to avoid duplication)
  - Capture richness of templates
- PML SNP standard
  - Look for overlap between model and genotype model
  - Follow documentation
  - Platform Independent Modeling & detailed descriptions of objects
- Phase II:
  - Feedback from a wider group of community experts noted above
  - Platform Dependent Implementations tests with various databases/data involved:
    - Existing databases implement model
    - Template task data
  - Document how they implemented it

### **[edit] Location and Environment Domain Model Team**

([Powerpoint here](#))

- Modeling Milestones
  - Develop generic and extensible model for location/environment modeling
  - Define a minimum set of environment attributes, example: temperature, rain, radiation, altitude and soil type
- Strategy
  - Compare and contract different location models (DIVA-GIS and other GIS tools, IRRI genebank and other genebanks)
  - Find minimum set of entities and attributes to represent location/environment for GCP
  - Define entity relationships and cardinality in UML model
- Outputs
  - Location model in XMI with entities, attributes and operations
  - Environment model in XMI with entities, attributes and operations
  - Database schema for location and environment model with linkage to passport database schema
  - Linkage of database schema for location and environment with pre-existing global dataset (example: gazetteer for locations, temperature minimum and maximum for each month, altitude, type of soil)
  - Use cases for location: How to reuse pre-existing global datasets ?  
Coordinates quality control with pre-existing global datasets ?

- Task Leader's note:
  - Phase I modeling should generate a basic set of location data models based on public standards (e.g. OpenGIS) through a collaborative discussion between CIP, IRRI (i.e. Isaiah Mukema), IPGRI and other GCP partners who have pertinent location data modeling expertise. Core classes of environmental attributes (e.g. soil, climate) to be modeled by ontology should also be identified. Special emphasis can be made on the design of a core location data model usable for by the Passport editorial team for passport location data. In this light, GCP core collection passport location data can serve as a sample data set for modeling insights.
  - Phase II should extend the base model (the DIVA model forms a good basis for this) and engage a larger community of discussion concerning location and environment ontology. In particular, environment ontology should be developed collaboratively with GCP affiliated crop physiologists (task leader note: this editorial team should establish contact with François Tardieu (Agropolis-INRA), Scott Chapman (at CSIRO?) and other crop modelers/physiologists for assistance in this activity).

### **Mapping team**

- OMG consortium: genomics map specifications (Philip Lijnzaard), public schemata (e.g. CMAP) and PML/SNP model (editor's emphasis...) will be compared to 2004 GCP UML (map and locus) domain model:
  - QTL/Genetic map
  - Addition of Comparative Mapping entities
  - Will check the concepts and relationships
- Genetic and QTL Map domain model will be attempted first:
  - Text description for biologists (UML diagram support)
    - Concepts definitions
    - Relationships UML rules
  - Forum deposit and feedback from database experts (Gramene, LIS, etc.), GCP bioinformaticians and biologists
    - Same process will be applied to physical maps, annotated (sequence) maps and comparative maps
- *Milestones:*
  - Forum UML deposit, March
  - Revised UML March-April
  - Public UML QTL/genetic map model, April 2005
  - UML model for genome sequence map model for review with NIAS (functional genomics editorial team) and IRRI prior or during IRRI hackathon (May) to elaborate work on core genomic data models and cross-linkages.

- Protégé version of QTL/Genetic Map model May-June 2005
- Protégé version of Genome/Sequence Map model end 2005

### **Functional genomics team**

- The domain model is indexed around the gene entity and will cover various “omics” data sets with a special emphasis on comparative biology. Modeling will be a collaboration between NIAS, IRRI, CIRAD, public plant researchers (e.g. NASC/PlaNET) and public ontology consortia (GO, PO, MGED). Domain modeling will be use case driven in consultation with team noted above.
- Functional genomics modeling is deemed very mature in the public domain due to the high degree of global activity in human and model organism genomic-focused bioinformatics. It is anticipated that very little reinvention of the wheel will be necessary. Rather, a map of public functional genomics models for core functional genomics data may be constructed focusing on identification of pertinent public model cross-linkages and model deficiencies (the latter relative to GCP needs), in particular, with respect to comparative crop biology (orthology).
- Phase I to the end of April will focus on gene expression, mutant characterization and proteomics domain modeling.
- Phase II will focus on systems biology domain modeling: metabolomics, pathways, etc.
- As noted above, a joint genomics (sequence/gene) domain modeling meeting with the mapping team will be convened by the end of Phase I.

### **General Discussion & Domain Modeling Priorities**

<http://cropwiki.irri.org/Wageningen/Wednesday/DomainModelingPriorities.pdf>

Throughout the workshop, several discussions highlighted key issues for the task. Some of these issues are mentioned here:

- As part of both Phases of the work, all editorial teams are to consider the identification, elaboration and full documentation (in the GCP Use Case Database) of a basic set of thematic domain model specific use cases for implementation of basic web services (and other simple platform functionality), as one of their target deliverables from Phase I modeling.
- The scope of domain models for each editorial theme likely remain to be delineated more precisely by editorial teams, to clarify what level of model granularity is appropriate and what components primarily needed (as oppose to be optional) to the GCP.
- GCP should consider possible publication of the resulting domain models as an OMG standard.
- Choice of domain model representations (e.g. UML/XMI versus RDF/OWL) were extensively discussed. Martin Senger contacted an external modeling expert, Philip Lord, for some guidance on choice of technologies. Although Protégé and

RDF/OWL were highlighted as a more powerful and flexible tool and language (respectively) for semantic modeling, and appropriate for consideration as a core modeling tool and format for domain modeling, the collective discussion at the workshop decided to continue to use UML modeling (and its associated XML export format, XMI) for domain modeling over the next few months (i.e. at least Phase I of the task) given the clarity and familiarity of UML graphical representations of the domain model to the GCP task team, and the relative lack of experience of the team with Protégé and RDF/OWL.

- The use of UML/XMI as the initial domain model representation does not exclude the possibility of a future switch to alternative representation (e.g. RDF/OWL) and parallel use of alternative tools (e.g. Protégé) for pure ontology development in the project.
- The template task is using XML Schema for basic template development, but this too could be driven by alternative XML technologies later.
- XSLT technologies demonstrated to provide a flexible, powerful and (relatively) quick way to apply for necessary XML inter-conversion, and thus, could potentially be applied to transform XMI into XML Schema and/or RDF/OWL representations of pertinent portions of the models, for template, web services and platform deployment purposes.
- XML domain models can be used directly with GCP platform as a specification for Java bean/schema maps, using technology like XML Stream and Hibernate.

### **Closing comments by Theo Van Hintum**

Theo Van Hintum thanked the participants in the meeting for a productive workshop and highlighted a final set of concerns about task execution to which he wished to see some attention given:

- Harmonization of domain modeling standards - nomenclature of entity names, tags, etc. - and tools will be required. Theo suggested quick formulation and circulation of a policy document by the task leader (Action by R. Bruskiwich)
- Communication between domain model editorial teams seems well defined (e.g. proposed web repository) but editorial team strategies for communication to and engagement of end users (and external collaborators?) is less well formulated and requires clarification. Short term deliverables and feedback from users assured through template task, but longer term outputs need to be more clearly formalized. Martin Senger suggested that the best strategy to achieve this is to put the domain modeling products (and their related software artifacts) into the hands of the users for usage that will validate their utility (Action by all Editorial Team Leaders, in collaboration with the Task Leader)
- Clear and careful thinking about the user interface design (i.e. target audience) of the proposed domain modeling web repository will be essential. Theo suggested that communication experts be consulted. Guy Davenport emphasized the need for the repository to be interactive. Other commentators emphasized the need for curation and moderation.

## Key Action Items

Towards the end of the workshop, the following items were highlighted as action items for Phase I.

- Meeting proceedings (this report, by Task Leader) combined with workshop PowerPoint presentations are to be posted to GCP virtual workspace (Action by Task Leader, by Mid-March).
- Model versioning and nomenclature/style policy/guideline document to be drafted and circulated among meeting participants for discussion and agreement. This document should also discuss the harmonization of tools through a table compare/contrast table of suggested (UML) domain modeling tools (e.g. exhibiting XMI compliance) and possibly, further assessment of Protégé (Action by Task Leader, by Mid-March)
- Domain model repository to be established by end of April including cross-indexing to use case database entries and related SP4 information. An amended version of the 2004 UML (XMI) high level domain model to be posted as the version 0.1 release of the GCP domain model, with indications of editorial scope and inter-module “strata” linkages (Action by Task Leaders for this task, template task and use case infrastructure task, by April)
- All editorial teams to initiate and coordinate work on promised Phase I outputs for delivery by end of April, in collaboration with template task. Projected outputs to include linkages to the design of related data templates and the enumeration of a basic core set of operational use cases interoperating with domain models, appropriate for immediate implementation as platform and web service components (Action by all Editorial Team Leaders).
- The next formal meeting relating to this task is tentatively scheduled during the two weeks of May 9th - 20th, 2005 at IRR1 (in the Philippines). Editorial teams may wish to schedule theme-specific modeling face-to-face modeling discussions at that time.