OMICS A Journal of Integrative Biology
Volume 10, Number 2, 2006
© Mary Ann Liebert, Inc.

Generation Challenge Programme (GCP): Standards for Crop Data

RICHARD BRUSKIEWICH,[1] GUY DAVENPORT,[2] TOM HAZEKAMP,[3] THOMAS METZ,[1] MANUEL RUIZ,[4] REINHARD SIMON,[5] MASARU TAKEYA,[6] JENNIFER LEE,[7,8] MARTIN SENGER,[1] GRAHAM McLAREN,[1] and THEO VAN HINTUM[9]

ABSTRACT

The Generation Challenge Programme (GCP) is an international research consortium striving to apply molecular biological advances to crop improvement for developing countries. Central to its activities is the creation of a next generation global crop information platform and network to share genetic resources, genomics, and crop improvement information. This system is being designed based on a comprehensive scientific domain object model and associated shared ontology. This model covers germplasm, genotype, phenotype, functional genomics, and geographical information data types needed in GCP research. This paper provides an overview of this modeling effort.

This paper is part of the special issue of OMICS on data standards.

INTRODUCTION

THE FAST-MOVING FIELDS of comparative genomics, molecular breeding, and bioinformatics have the potential to bring new science to bear on problems encountered by resource-poor farmers bypassed by the earlier wave of innovation of the Green Revolution. These problems include water stresses (both drought and flooding) and biological stresses such as plant diseases. The Generation Challenge Programme (GCP; ⟨www.generationcp.org⟩) aims to exploit advances in molecular biology to harness the rich global heritage of plant genetic resources and contribute to a new generation of stress-tolerant varieties that meet the needs of resource-poor people.

[1] International Rice Research Institute (IRRI), Manila, Philippines.
[2] Centro Internacional de Mejoramiento de MaÌz y Trigo (CIMMYT), Mexico City, Mexico.
[3] International Plant Genetic Resources Institute (IPGRI), Rome, Italy.
[4] Centre International de Recherche Agronomique pour le Développement (CIRAD), Montpellier, France.
[5] Centro Internacional de la Papa (CIP), Lima, Peru.
[6] National Institute for Agrobiological Sciences (NIAS), Tsukuba, Ibaraki, Japan.
[7] University of Dundee, Dundee, Scotland, United Kingdom.
[8] Scottish Crop Research Institute, Dundee, Scotland, United Kingdom.
[9] Centrum voor Genetische Bronnen Nederland, Wageningen Universiteit & Researchcentrum, Wageningen, The Netherlands.

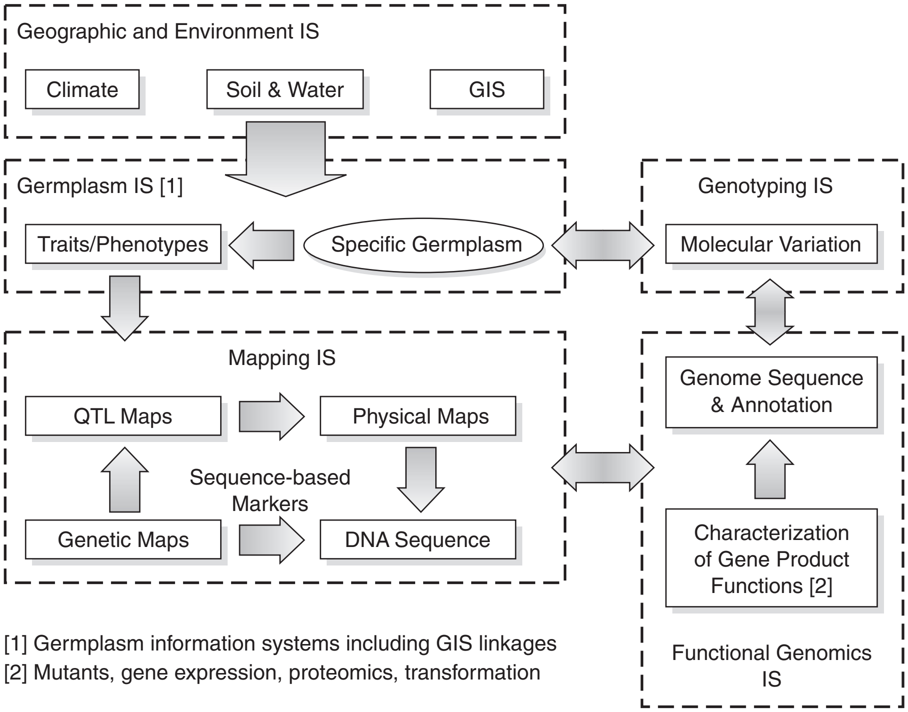
Geographic and Environment IS
Climate
Soil & Water
GIS
Germplasm IS [1]
Traits/Phenotypes
Specific Germplasm
Genotyping IS
Molecular Variation
Mapping IS
QTL Maps
Physical Maps
Sequence-based Markers
Genetic Maps
DNA Sequence
Genome Sequence & Annotation
Characterization of Gene Product Functions [2]
Functional Genomics IS
[1] Germplasm information systems including GIS linkages
[2] Mutants, gene expression, proteomics, transformation

FIG. 1. Scope of Generation Challenge Programme (GCP) crop data types.

The GCP brings together three sets of partners—the centers of the Consultative Group on International Agricultural Research (CGIAR; ⟨www.cgiar.org⟩), advanced research institutes (ARIs), and national agricultural research and extension systems (NARES) in developing countries—to undertake a projected 10-year program of globally integrated scientific research, capacity building, and delivery of products.

One GCP aim is to create an "integrated platform" of molecular biology and bioinformatics tools that will be freely available to researchers and breeders the world over as public goods enabling agricultural scientists, particularly in developing countries, to readily use elite genetic stocks and new marker technologies in their local breeding programmes.

The bioinformatics activities of the GCP are driven by several key observations:

- GCP partners are globally distributed across most of the world's continents,[1] each partner having data sets to share and integrate.
- The research covers a diversity of crop species.
- The research would span a very wide range of scientific data types—germplasm,[2] genomic, phenotype, crop physiological, and geographic information—with some very large crop data sets (Fig. 1).

The subprogram of the GCP dealing with genetic resources, genomics, and crop information systems was conceived to directly meet these challenges and includes, among others, the following components:

- Development of a comprehensive common scientific domain model to describe the data it has to deal with

[1] Except Antarctica, of course.

[2] Defined as the plant material—seeds or clonal stocks—used to preserve and propagate crops.

[3] Using BioMOBY, BioCASE, SoapLab, Taverna, and other pertinent Internet-based integration technologies.

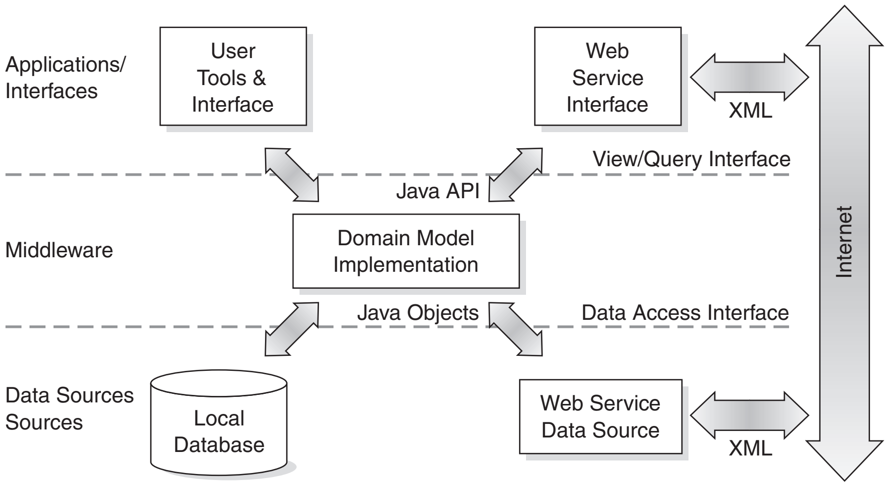

FIG. 2. Generation Challenge Programme (GCP) domain model-driven architecture.

- Development of an Internet-integrated network[3] for crop bioinformatics data exchange
- Development of a data capture and archiving infrastructure
- Integration of existing and new analysis tools into an Internet-connected crop analysis workbench to integrate comparative genomics and convergent functional evidence for candidate genes of agronomic traits, and to facilitate germplasm selection based on specified passport, genotype, phenotype, and source environment criteria

Most software products of GCP bioinformatics research and development are being published in a public project archive called "CropForge" (⟨http://cropforge.org⟩) and are being discussed on a collaborative technical documentation site called "CropWiki" (⟨http://cropwiki.irri.org/gcp⟩).

THE GCP SCIENTIFIC DOMAIN MODEL

To cope with the above scope, diversity, and dispersion of crop information, GCP researchers formulated a vision to specify a common, consensus blueprint of a scientific domain model and associated ontology to guide GCP informatics work. This vision assumes that the GCP will collaboratively build upon other pertinent international initiatives like other established modeling and ontology initiatives, many of which are mentioned in this special *OMICS* journal issue (Ball and Brazma, 2006; Jones et al., 2006; Sansone et al., 2006; Taylor et al., 2006; Whetzel et al., *this issue*). The resulting "model-driven architecture" (⟨www.omg.org/mda/⟩) would be applied to the identification, design/adaptation, implementation, and deployment of consortium information tools and network protocols to deliver robust global public goods of information and software tools specifically for the GCP, and also for global crop research in general (Fig. 2).

Development of this domain model began informally in 2004 based on initial software development meetings reviewing use cases compiled by the project in a series of technical white papers on GCP user needs. A formal project in scientific domain modelling was funded as GCP-commissioned research in 2005. Five subdomain model editorial teams were commissioned to construct germplasm, passport, phenotype, genotype, genomics, and geographic information system (GIS) components of the model in consultation with pertinent end-user and expert communities. Three face-to-face project meetings were convened during the year to review progress and consolidate the subdomain models, a release of which is posted on a GCP website (⟨http://pantheon.generationcp.org⟩).

[3] Using BioMOBY, BioCASE, SoapLab, Taverna, and other pertinent Internet-based integration technologies.

The model is documented in Unified Modelling Language (UML). Computable versions[4] of the UML model are published in the GCP "Pantheon" middleware project in CropForge (⟨http://cropforge.org/projects/pantheon/⟩).

At the heart of the domain model is a generic metadata model upon which other model components are specified. This metadata model defines the general concept of a system "Entity" with "Identifier" and "Feature" components. For specific subdomains, editorial teams generally started with successful extant domain models in their subject area. For example, the germplasm, genotype, and phenotype subdomain model is heavily influenced by the data models of the open-source International Crop Information System (ICIS, ⟨www.icis.cgiar.org⟩; Fox and Skovmand, 1996; Bruskiewich et al., 2003; McLaren et al., 2005).

A significant aspect of the metadata model is the reliance on extensible ontology to define the semantics of the model. For such ontology, where possible, the GCP is simply adopting existing controlled vocabulary and ontology (CVO) standards, such as from the Gene Ontology,[5] Plant Ontology,[6] and Microarray Gene Expression Data Society (MGED) Ontology (currently under review in the light of a Functional Genomics Ontology) (Whetzel et al., *this issue*) consortia. To manage ontology selected for the GCP platform, an online catalog (⟨http://ontology.generationcp.org⟩) was established based on the Generic Model Organism Database Chado[7] schema "cv" module.

GCP funding of the domain modelling activity continues into 2006, during which two primary activities focusing on domain model validation are foreseen. A public review of the current domain model and discussion of associated ontology will be organised during a GCP-sponsored workshop[8] involving GCP scientists and external experts. Furthermore, throughout the year, technology-specific implementations of the model will be designed and coded in the form of data exchange protocols (e.g., data types for BioMOBY[9] and BioCASE[10]) and a GCP analysis workbench.

CONCLUSION

The GCP considers the proper modelling of its data domains as vital for its success, and the activities can be foreseen to continue for the coming few years. Involvement of the larger bioinformatics community will be increasingly important since this will improve the quality of the models and, more importantly, will contribute to the larger common goal of establishing standards for effective data sharing among the global bioinformatics community.

ACKNOWLEDGMENTS

The domain modeling team gratefully acknowledges the inputs of various GCP and external scientists to the discussions about the GCP domain model during various GCP-funded workshops. This work is funded through the Generation Challenge Programme (⟨www.generationcp.org⟩), a consortium funded by several international donors of the Consultative Group for International Agricultural Research (CGIAR; ⟨www.cgiar.org⟩).

[4] Available UML tools sorely lack interoperability for exchange of models, despite attempts at specifying XMI , an XML encoding for UML proposed by the Object Management Group (OMG). After various iterations with diverse commercial UML tools, the GCP domain modelling team is moving toward the open-source Eclipse Modelling Framework (⟨www.eclipse.org/emf⟩) as the common UML format, largely reflecting the GCP team's commitment to the Eclipse Integrated Development Environment for software development, and to the Eclipse Rich Client Platform as a key technology for GCP core workbench implementation.

[5] ⟨www.geneontology.org⟩.

[6] ⟨www.plantontology.org⟩.

[7] ⟨www.gmod.org/schema/⟩.

[8] To be hosted by the African Center for Gene Technology (Pretoria, South Africa) in March.

[9] ⟨www.biomoby.org⟩; Wilkinson et al., 2005.

[10] ⟨www.biocase.org⟩.

REFERENCES

BALL, C.A., and BRAZMA, A. (2006). MGED standards: work in progress. OMICS (*this issue*).

BRUSKIEWICH, R., COSICO, A., EUSEBIO, W., et al. (2003). Linking genotype to phenotype: the International Rice Information System (IRIS). Bioinformatics **19,** i63–i65.

FOX, P.N., and SKOVMAND, B. (1996). The International Crop Information System (ICIS)—connects genebank to breeder to farmer's field. In *Plant Adaptation and Crop Improvement*. (M. Cooper and G.L. Hammer, eds. (CAB International), Wallingford, Oxfordshire, UK, pp. 317–326.

JONES, A.R., PIZARRO, A., SPELLMAN, P., et al. (2006). FuGE: Functional Genomics Experiment object model. OMICS (*this issue*).

McLAREN, C.G., BRUSKIEWICH, R.M., PORTUGAL, A.M., et al. (2005). The International Rice Information System (IRIS): a platform for meta-analysis of rice crop data. Plant Physiol **139,** 637–642.

SANSONE, S.-A., ROCCA-SERRA, P., TONG, W., et al. (2006). A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group. OMICS (*this issue*).

TAYLOR, C.F., HERMJAKOB, H., JULIAN, Jr., R.K., et al. (2006). The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). OMICS (*this issue*).

WHETZEL, P.L., BRINKMAN, R.R., CAUSTON, H.C., et al. (2006). Development of FuGO: an ontology for functional genomics investigations. OMICS (*this issue*).

WILKINSON, M, SCHOOF, H, ERNST, R., et al. (2005). BioMOBY successfully integrates distributed heterogeneous bioinformatics web services: the PlaNet exemplar case. Plant Physiol **138,** 1–13.

Address reprint requests to:
Dr. Richard Bruskiewich
International Rice Research Institute (IRRI)
DAPO Box 7777
Metro Manila, Philippines

E-mail: r.bruskiewich@cgiar.org

This article has been cited by:

1. Julien Wollbrett, Pierre Larmande, Frédéric de Lamotte, Manuel Ruiz. 2013. Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases. *BMC Bioinformatics* **14**:1, 126. [CrossRef]
2. Mathieu Rouard, Sebastien Carpentier, Stephanie Bocs, Gaëtan Droc, Xavier Argout, Nicolas Roux, Manuel RuizRole of Bioinformatics as a Tool 194-216. [CrossRef]
3. Daniel Puente-Rodríguez. 2010. Biotechnologizing Jatropha for local sustainable development. *Agriculture and Human Values* **27**:3, 351-363. [CrossRef]
4. R. Shrestha, E. Arnaud, R. Mauleon, M. Senger, G. F. Davenport, D. Hancock, N. Morrison, R. Bruskiewich, G. McLaren. 2010. Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB Plants* **2010**:0, plq008-plq008. [CrossRef]
5. C. Graham McLaren, Thomas Metz, Marco van den Berg, Richard M. Bruskiewich, Noel P. Magor, David ShiresChapter 4 Informatics in Agricultural Research for Development **102**, 135-157. [CrossRef]
6. R. Serraj, A. Kumar, K.L. McNally, I. Slamet-Loedin, R. Bruskiewich, R. Mauleon, J. Cairns, R.J. HijmansImprovement of Drought Resistance in Rice **103**, 41-99. [CrossRef]
7. Richard Bruskiewich, Martin Senger, Guy Davenport, Manuel Ruiz, Mathieu Rouard, Tom Hazekamp, Masaru Takeya, Koji Doi, Kouji Satoh, Marcos Costa, Reinhard Simon, Jayashree Balaji, Akinnola Akintunde, Ramil Mauleon, Samart Wanchana, Trushar Shah, Mylah Anacleto, Arllet Portugal, Victor Jun Ulat, Supat Thongjuea, Kyle Braak, Sebastian Ritter, Alexis Dereeper, Milko Skofic, Edwin Rojas, Natalia Martins, Georgios Pappas, Ryan Alamban, Roque Almodiel, Lord Hendrix Barboza, Jeffrey Detras, Kevin Manansala, Michael Jonathan Mendoza, Jeffrey Morales, Barry Peralta, Rowena Valerio, Yi Zhang, Sergio Gregorio, Joseph Hermocilla, Michael Echavez, Jan Michael Yap, Andrew Farmer, Gary Schiltz, Jennifer Lee, Terry Casstevens, Pankaj Jaiswal, Ayton Meintjes, Mark Wilkinson, Benjamin Good, James Wagner, Jane Morris, David Marshall, Anthony Collins, Shoshi Kikuchi, Thomas Metz, Graham McLaren, Theo van Hintum. 2008. The Generation Challenge Programme Platform: Semantic Standards and Workbench for Crop Science. *International Journal of Plant Genomics* **2008**, 1-6. [CrossRef]