Poster B-13

Development of Data Templates for Data Collection, Storage and Database Submission.

Authors:
Miguel Anducho (*Crop Research Informatics Laboratory, CIMMYT.*)
Kyle Braak (*Crop Research Informatics Laboratory, CIMMYT.*)
Richard Bruskiewich (*Crop Research Informatics Laboratory, IRRI.*)
Victoria Carollo (*Center for Computational Biology, Montana State University.*)
Tom Hazekamp (*International Plant Genetic Resources Institute.*)
Andrew Farmer (*National Center for Genomic Resources.*)
David Marshall (*Scottish Crop Research Center.*)
Dave Matthews (*GrainGenes, Cornell University.*)
Ayton Meintjes (*Bioinformatics and Computational Biology Unit, University of Pretoria.*)
Thomas Metz (*Crop Research Informatics Laboratory, IRRI.*)
Jane Morris (*African Centre for Gene Technologies.*)
Manuel Ruiz (*TropDB, CIRAD.*)
Mary Schaeffer (*MaizeGDB, University of Missouri.*)
Theo van Hintum (*Centre for Genetic Resources, The Netherlands.*)
Susan McCouch (*Gramene, Cornell University.*)
Guy Davenport (*Crop Research Informatics Laboratory, CIMMYT.*)

Short Abstract: A large amount of data is generated each year that needs to be made accessible to the scientific community. We have developed database submission templates that define the format of the data and provide sufficient information to allow scientists capture the data, metadata and available controls.

Long Abstract:
A large amount of data is generated each year by the scientific community. These data must be collected with sufficient metadata and controls and have an adequate level of completeness and accuracy that allow it to be utilized and analyzed accurately. The data must also be stored in a machine readable format that allows it to be easily validated and loaded into a database.

We are developing machine readable templates for data captured manually, which provides guidelines on capturing the data, metadata and available controls and defines the level completeness and accuracy required. We are also providing similar guidelines for data captured automatically by scientific equipment, such as genotyping systems, or data generated by analytical software. We have developed data templates for plant accession passports and genotyping data produced in work by the Generation Challenge Program (GCP, www.generationcp.org) and its partners, and are developing additional templates for mapping, QTL, SNP genotyping and plant phenotypic (evaluation) data in collaboration with GrainGenes (wheat.pw.usda.gov), MaizeGDB (www.maizegdb.org) and Gramene (www.gramene.org).

The most commonly used tool for data entry and storage is Microsoft's Excel spreadsheet.

However, due to the large size of GCP datasets it is easy to exceed Excel's internal upper limit on row and columns. Furthermore, Excel’s generic approach to data storage can also be a disadvantage, since data can be placed in any format with little control over what values are allowed. For this reason we are developing both stand-alone and web-based tools for validation and curation both manually entered and automatically captured data, and to convert these data into formats required for database submission, and visualization and analytical software. The templates are provided in both Excel and text formats, which are validated and converted to XML for storage. The software is generic enough to support a range of formats and can conform to most XML schemas. The XML form of the data is then transformed using XSL (Extensible Stylesheet Language) to the various formats required for database submission, visualization and analysis. Where possible our development of data templates and software conforms to community standards and practices. For example, the QTL template uses the trait ontology provided by Gramene and the software is developed within Java and utilizes a number of open source technologies. The templates and software are available on the GCP bioinformatics portal (www.generationcp.org/bioinformatics).