GABI GAIN: Data Integration in Genomics-based Plant Breeding

Robert Wagner, Pawel Durek and Birgit Kersten

GabiPD Team, Bioinformatics, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14476 Golm

rwagner@mpimp-golm.mpg.de

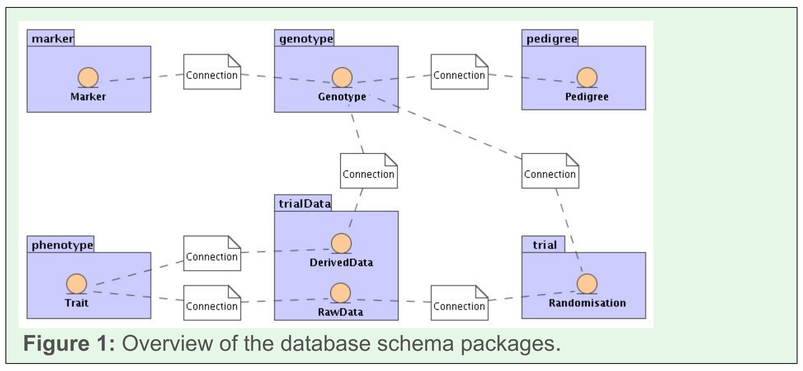
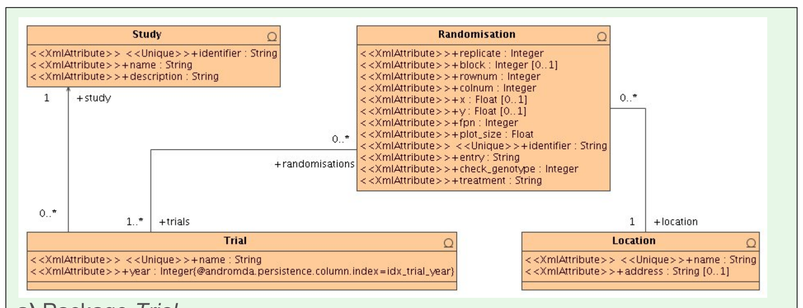


Figure 1: Overview of the database schema packages.



a) Package *Trial*.

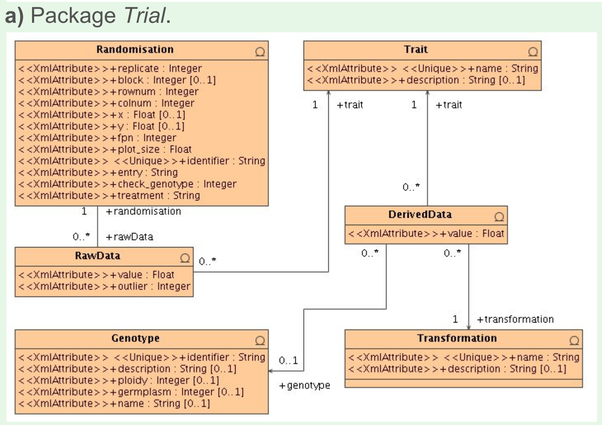
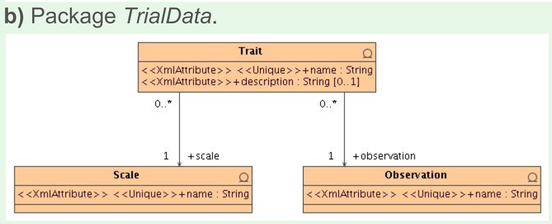
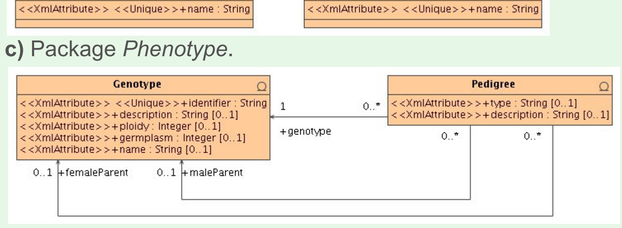


Figure 2: The schema of the database. Classes displayed in more than one of the class diagrams indicate the interfaces between the packages.

b) Package *TrialData*.



c) Package *Phenotype*.



d) Package *Pedigree*.

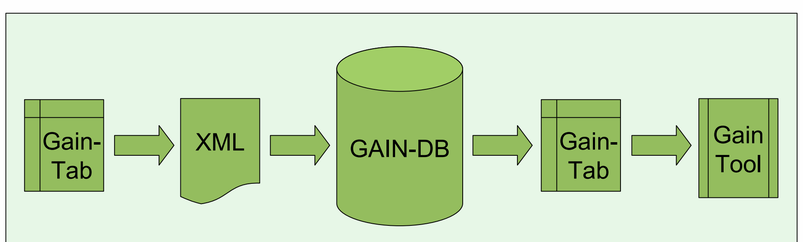


Figure 3: Data flow in the GABI GAIN project.

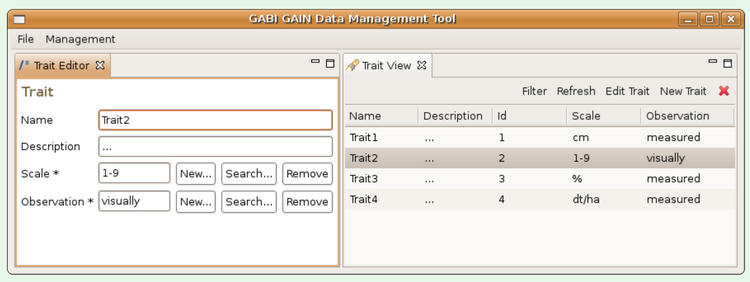


Figure 4: Graphical user interface for the data management.

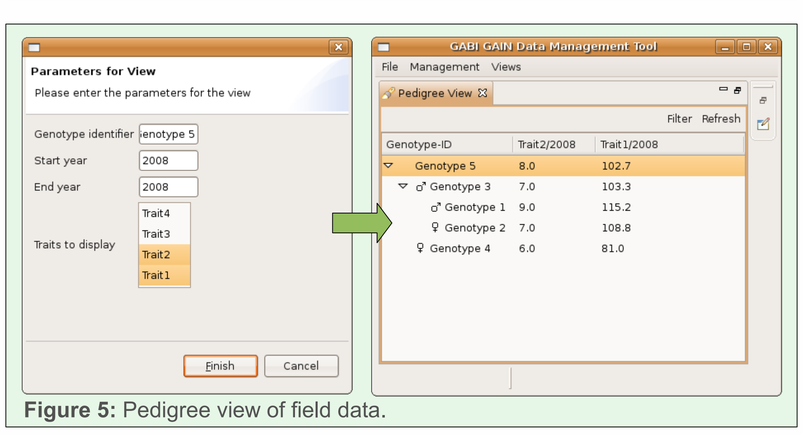


Figure 5: Pedigree view of field data.

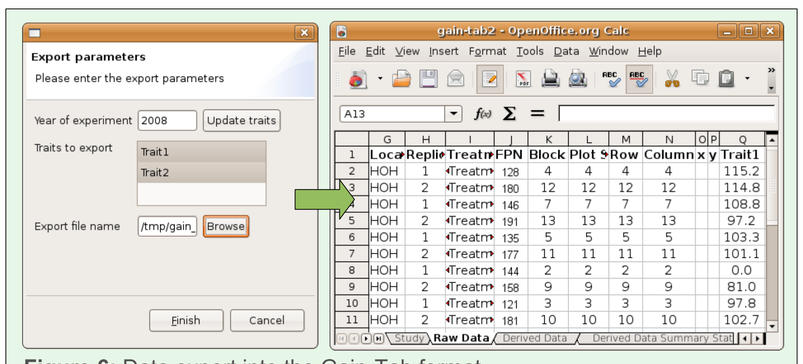


Figure 6: Data export into the Gain-Tab format.

Abstract

Modern plant breeding programs produce a huge amount of genomic and phenotypic data. Therefore efficient data mining becomes indispensable to exploit the generated information and to support the improvement of experimental strategies and the optimization of process design. In the scope of the GABI GAIN project, tools are being developed for the integration, mining and visualization of standardized data produced in genomics-based plant breeding. A data schema has been developed to store integrated breeding data of different crops in an efficient and flexible way. The model has been designed platform independently using the Unified Modeling Language. Hence, it can be realized in all common database management systems. An exchange format, Gain-Tab, has been defined to ease the integration of new data into the database as well as into additional analysis tools developed in the project. Gain-Tab is a table-based format, so that Spreadsheet tools like Excel or OpenOffice can be utilized to easily create files for data exchange. Finally, a graphical user interface is being developed, which provides the possibility of CRUD operations (create, read, update, delete) on the data as well as its selection and visualization. Furthermore importing and exporting from and into Gain-Tab is supported. The tool-package will be provided as a locally installable application for partners within the GABI GAIN project to optimize genomics-based plant breeding.

This work is supported by the German Federal Ministry of Education and Research BMBF (GABI-FUTURE grant 0315072C).

Data Model

The data schema has been modeled using the Unified Modeling Language (UML). It comprises 17 classes unified in six packages. Each package contains classes that share the same purpose (Fig. 1). Classes may have attributes and associations to other classes. Stereotypes and cardinalities define constraints. No features specific to certain database management systems (DBMS) or programming languages have been used, thus making the model platform-independent.

Package *Trial* (Fig. 2a) contains the description of the trial, information on the randomization parameters as well as the characteristics of the location. In the *TrialData* package (Fig. 2b) the result data of a trial as well as transformations between the data are situated. Package *Phenotype* (Fig. 2c) includes information on traits, their units, the scales as well as the kind of measurement. The *Genotype* package describes genotypes and their corresponding species. The information about inheritances between genotypes is defined in the Pedigree package (Fig. 2d). Finally, the *Marker* package describes markers, their types and alleles.

SQL statements for data definition (DDL) have been generated from the UML model to create the database in the DBMS. Classes have been converted into tables, associations into foreign keys, stereotypes and cardinalities into table and column constraints.

Data Exchange

In order to facilitate data exchange a format based on the Extensible Markup Language (XML) has been defined. The use of XML permits application of standard parsers and validation tools to ensure data integrity without the implementation of large amounts of software code. However, XML is not the common format for data produced in GABI GAIN and not the expected input format for most of the analysis tools used in the project. Furthermore, XML is hard to read and to create for the normal user. Hence, a table based format, Gain-Tab, has been additionally defined which may easily be generated using spreadsheet tools. Converters have been implemented to transform data from Gain-Tab into XML in order to avoid duplication of implementation work for database access code. Since the underlying relational database returns data as a set of rows, the export of data into Gain-Tab may easily be achieved on the direct way. Figure 3 illustrates a typical data flow: A data-set in Gain-Tab format is converted into XML. Subsequently the data is parsed, validated and inserted into the database. After possible modifications and aggregations data is exported into Gain-Tab and may be used in additional analysis tools developed in the GABI GAIN project.

The XML Schema definition for the XML format has been created automatically from the contents of the UML model. Each class in the model is represented by an element and a corresponding complex type in the XML-Schema. Attributes of the classes have been transferred to XML-Schema attributes within the complex types, whereas associations have become elements.

Data Visualization

Access to the data is provided by a prototype of a graphical user interface (GUI). In the current state it allows CRUD (create, read, update, delete) operations on data for a particular class (Fig. 4). After selecting a class the user is able to search for certain entries, modify their contents and insert or delete values. In case references to other classes are necessary, the GUI displays possible contents for selection. For each class a help page is provided with an explanation of the fields and the expected data types.

Several data views have been defined to facilitate common tasks raised in genomics-based plant breeding. One view, for instance, provides data corresponding to a genotype for different years and traits. In addition the values for the pedigree of the genotype are displayed in a tree-like manner (Fig. 5).

Data can be exported into Gain-Tab for further analysis with other tools developed in GABI GAIN. A dialog guides the user through this process (Fig. 6). After entering the year of the data to be exported, a list of traits used in corresponding experiments is displayed. Upon choosing the relevant traits and the name of the destination file, the data is subsequently written in Gain-Tab format.

Likewise, the import of data into the database follows an easy procedure. The user simply selects the Gain-Tab file, whereupon the import process is started. Problems with the format, missing or unacceptable values are reported. Finally, the number of imported rows is displayed.

Large parts of the GUI have been generated from the entity classes within the UML model. Stereotypes and tagged values define views. Dependencies between entity classes and service classes result in methods to transform objects between the different software layers.

Technologies

The tool is being developed using the paradigm of the model-driven software development (MDSD). This generic approach allows performing changes in the model without the necessity to reimplement major components of the interface, since lots of the software code is generated automatically from the UML model. The model has been developed using *MagicDraw*, the code has been generated by *AndroMDA* and *maven.* The target language is *Java*, and for data storage the HSQL DBMS is currently used. For the development of the software layers *Hibernate* has been used for the data access layer, *Spring* for the business layer and *Eclipse Rich Client Platform* for the presentation layer.