# An update on the GERMINATE project, the database and tools

**Jennifer M. Lee**[1,2], David Marshall[2], C. Graham McLaren[3], Michael Ambrose[4], Jo Dicks[4], Noel Ellis[4], Virginia Barnard[4], Daniel Nwankwo[2], Robbie Waugh[2] and Andrew Flavell[1]

1 University of Dundee, Department of Life Sciences, Dundee, DD1 4HN, Scotland UK
2 Scottish Crop Research Institute, Invergowrie, DD2 5DA, Scotland UK
3 International Rice Research Institute (IRRI) Los Baños, Laguna, Philipppines
4 John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK

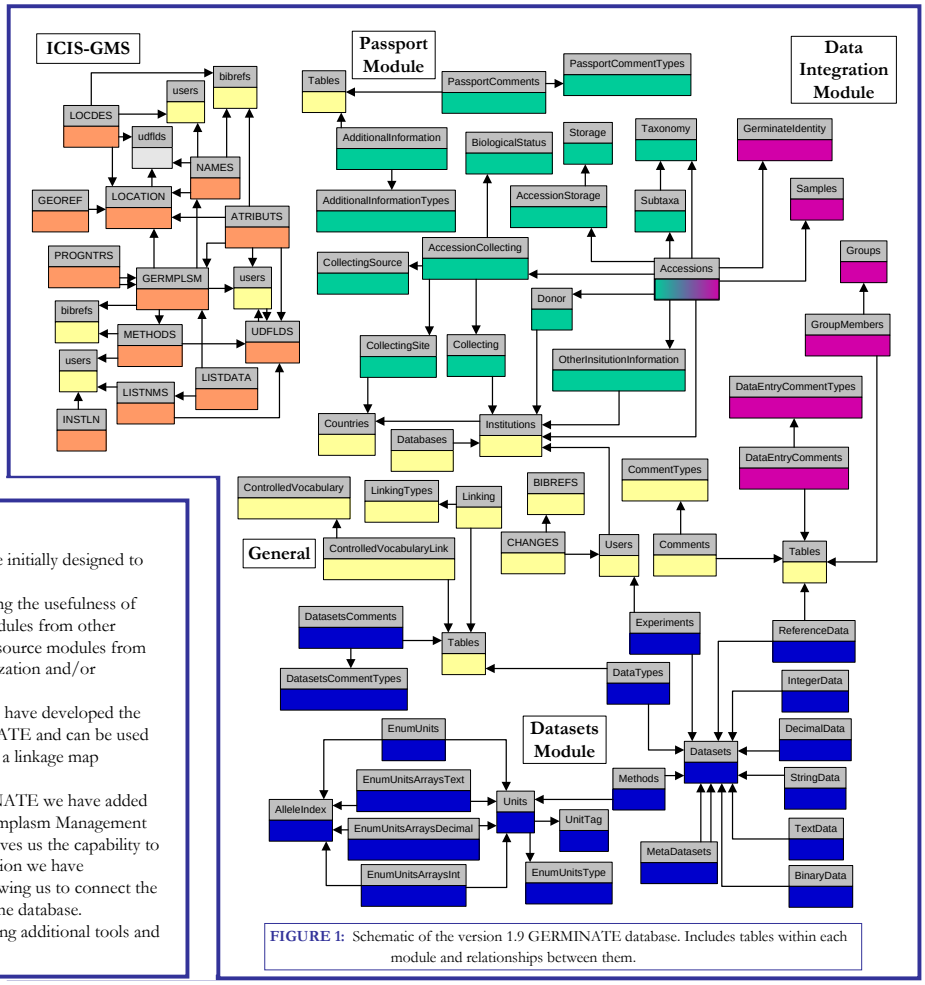


**FIGURE 1:** Schematic of the version 1.9 GERMINATE database. Includes tables within each module and relationships between them.

## ABSTRACT:

The GERMINATE database was developed as an open source generic database initially designed to hold genotype and phenotype data in a genetic resources context (http://germinate.scri.sari.ac.uk/germinate/). We are now in the phase of expanding the usefulness of the database by connecting analysis and visualization tools and through adding modules from other database which will accommodate other types of data. In some cases adding open source modules from other database also means we bring along with it the tools used for analysis, visualization and/or curation.

As a test of the potential of developing and connecting tools to the database we have developed the Genotype Visualization Tool (GVT) which has now been connected to GERMINATE and can be used to display local or chromosomal haplotype information across a set of accession in a linkage map context.

As a proof of concept for including modules from other databases in GERMINATE we have added the International Crop Information Systems (ICIS, http://www.icis.cgiar.org) Germplasm Management System (GMS) into the GERMINATE database. Adding the ICIS-GMS module gives us the capability to visualize and edit genealogy information as well as create and manage lists. In addition we have connected the ICIS-GMS into the Data Integration Module in GERMINATE allowing us to connect the information now stored in the ICIS-GMS module with information in the rest of the database.

Now that these concepts have been demonstrated we are working towards adding additional tools and modules, making GERMINATE useful to a wider range of users.

## INTERFACES:

The ability to perform complex queries on data in the database will depend on what type of interface is used to access the database. The GERMINATE database has been design such that simple to complex interfaces can be used to connect to the database.

- A light-weight Perl-CGI interface which users can perform a range of relatively simple queries. This was originally designed to search passport descriptors and additional functionality has been added as requested. It may now be used to browse all information available for an accession or retrieve datasets.
- Users may also use SQL to perform more complex queries on the databases. We are writing a series of functions to simplify the SQL which would need to be written by a user for some common queries. These functions may then be combined to generate more complex queries.
- Data in GERMINATE is available via the Genomic Diversity and Phenotype Connection (GDPC; http://maizegenetics.net/gdpc/) Web Services. This allows users to use any analysis tools which have been made GDPC aware and data in GERMINATE can also be combined with data in other GDPC enabled databases.
- We are also working with the Generation Challenge Program (GCP; http://generationcp.org) on the middleware model which will be used for the GCP web services. GERMINATE will be available via the GCP web services once the framework is in place.

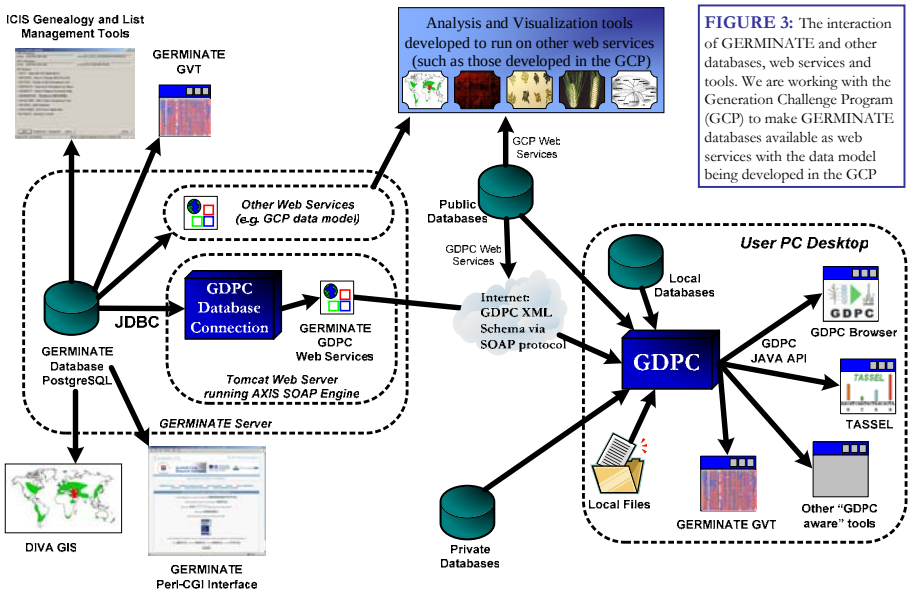


**FIGURE 3:** The interaction of GERMINATE and other databases, web services and tools. We are working with the Generation Challenge Program (GCP) to make GERMINATE databases available as web services with the data model being developed in the GCP

## MODULES:

The GERMINATE database currently contains five modules: Data Integration, Passport, Datasets, ICIS-GMS and a General module which contains tables used by all other modules and facilitates interconnection of modules (Figure 1).

The Data Integration module accommodates the various approaches to data collection used in the plant community.

The tables which comprise the Passport Module in GERMINATE are based on the 2001 FAO/IPGRI Multi-Crop Passport Descriptors (MCPD) (http://www.ipgri.cgiar.org/) with extensions to improve its generality. We consider these descriptors the lowest common denominator between plants, however, this is not an exhaustive list and GERMINATE includes a table to accommodate any additional descriptors used.

The Datasets module is used to hold Genotype and Phenotype data which can accommodate integer, decimal, short and long text, and binary (large object) data. In addition array types for text, integer and decimal types are also available. In order to associate data with any other object in the database a reference data table is used, which is used to link data to the accession or other database entity with which it is associated.

The ICIS-GMS module was taken from the ICIS database (http://.... ) and is used to store and manage genealogy and list information. This module also gave us access to tools for list and pedigree management.

## FUTURE:

- Development of a user friendly data loading interface which will interact with the database.
- Additional access to database
  - Development of Functions and stored queries to assist users using SQL access to GERMINATE databases
  - Coordination of connection of GERMINATE to Generation Challenge Programme defined web services
- Analysis and Visualization tools
  - Connection of existing analysis and visualization tools
  - Development of new tools in collaboration with groups with similar interests and needs
- Further identification of use cases
  - How people would like to use the database.
  - Tools users would like connected to the database.
- Development of community standards in collaboration with other groups around the world

**Acknowledgments:** This work is funded by the BBSRC Grant 94/BEP17084, the Bioinformatics and E-science program and SEERAD FF00589