GERMINATE: A Database For Plant Genetic Resources

Jennifer M. Lee[1], Guy Davenport[2,5], Daniel Nwankwo[3], Theo J.L. van Hintum[4], T.H. Noel Ellis[2], Michael J. Ambrose[2], Jo Dicks[2], David Marshall[3] and Andrew J. Flavell[1]

1 University of Dundee, Department of Life Sciences, Dundee, DD1 4HN, Scotland UK
2 John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK
3 Scottish Crop Research Institute, Invergowrie, DD2 5DA, Scotland UK
4 Centre for Genetic Resources (CGN), P.O. Box 16, 6700AA Wageningen, The Netherlands
5 Current Address: CIMMYT, El Batan, Texcoco, Mexico

ABSTRACT:

GERMINATE (http://bioinf.scri.sari.ac.uk/germinate/) is a general plant database designed to hold data types ranging from molecular data to phenotypic data for any plant species. GERMINATE has been specifically designed to cope with varying ploidy levels found among plants. Its main purpose is to integrate molecular data (e.g. marker data) with descriptive data (e.g. geographical, passport, morphological) for large germplasm or genotype collections. GERMINATE uses a top level Accession entry ID to allow association and querying between disparate sets of data associated with an accession. We have initially tested GERMINATE with pea, wheat, barley, and lettuce at the partner institutions. GERMINATE has now been distributed to interested collaborators for testing in a broader range of species with a wider variety of data. This includes members of the Generation Challenge Program Consortium who are testing GERMINATE for use with species such as rice, potato, and maize. We are also testing various types of interfaces to GERMINATE, from a light-weight perl-cgi interface designed for data retrieval to more complex java based interfaces which would allow analysis as well as data retrieval. We also expect to make the data in GERMINATE available as a web service interface so it will interact well with tools being developed by collaborators. GERMINATE is currently being implemented in the public domain PostgreSQL and is available under the GNU GPL. We envisage that this will form an ideal platform to enable us and other groups to develop a range of interfaces and analysis tools to interact with the database.

DESIGN:

A primary goal of the GERMINATE project is to develop a robust database which can be used for a broad range of species and data. The GERMINATE database has been designed with particular consideration of issues relevant to the genetic resource community. For the database to be most useful to this community the data must be stored in a manner independent of the technology used to generate it. This approach to design means the database will not need to be modified as new technology arises, a common occurrence in genetics. GERMINATE has also been designed to store data in a way that allows retrieval without interpretation, thus allowing users to add their own interpretation to the data for analysis.

Also included in the database is the flexibility to store any number of alleles per locus, thus accommodating plants of any ploidy level, a feature missing from many genetic databases.

GERMINATE links all data by its association with the accession, plant or group used to collect the data. This design permits queries of multiple disparate datasets, allowing for complex queries and analysis.

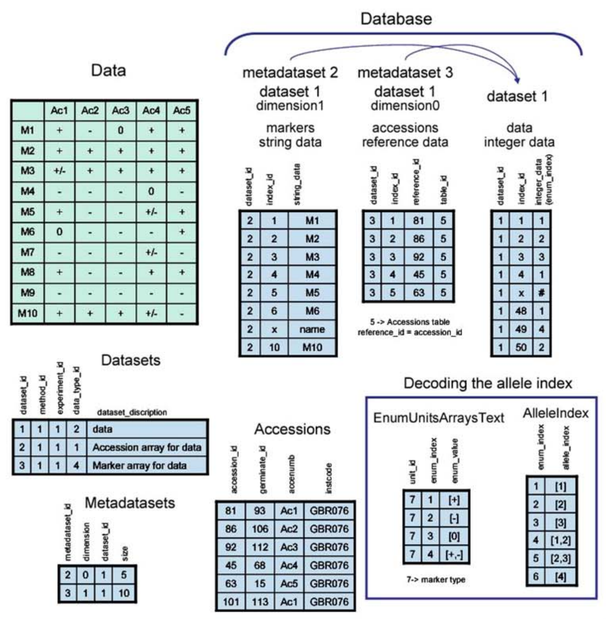


FIGURE 2: An Example of how genetic data is loaded into GERMINATE, showing a set of markers evaluated in a set of accessions. A two dimensional array is stored in GERMINATE where the set of markers and the set of accessions are the metadataset dimensions for the dataset. The primary dataset is the allele values of each marker in each accession. Each dataset and metadataset is entered into the appropriate data table associated with its dataset ID (String Data for Markers, Reference Data for Accessions, and Integer Data for the allele values) and a data index keeps track of the order of each set of data. The metadatasets tables keeps track of the dimension data and will then have two entries associated with the dataset. To keep the database normalized only a single instance of each marker and accession is stored. Therefore, to recreate the dataset a two dimensional grid is set up with accessions as one dimension and markers as the other. The data would then be iterated through to place it in the correct position on the grid. If a user is only interested in a single data point it can be easily accessed by a simple equation which relates the accession, marker and data indices to each other.

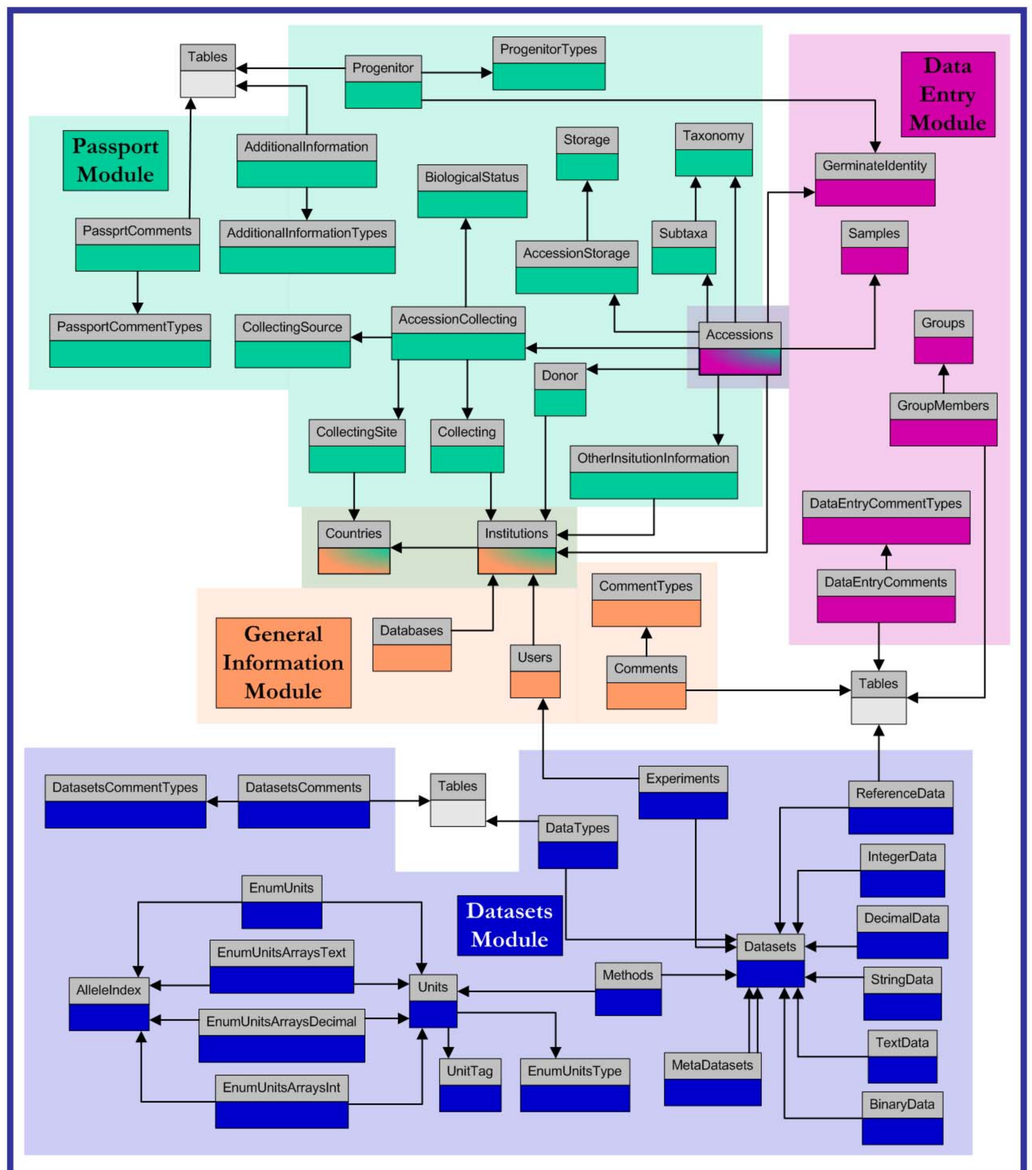


FIGURE 1: Schematic of the GERMINATE modules. Includes tables within each module and relationships between them.

MODULES:

The GERMINATE database is divided into four modules: Data Entry Level module, General Information module, Passport Data module, and the Datasets module (Figure 1). The database is subdivided in such a manner so groups wishing to implement only a part of GERMINATE as an addition to their own database will be able to do so as simply as possible.

The Data Entry Level module accommodates the various approaches to data collection used in the plant community (Figure 1).

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 (Figure 1). This is a crucial feature, as international germplasm collections need to be able to use the database without complicated data adaptation procedures. 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.

Genotype and Phenotype data is held in the datasets module, 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; these arrays are assigned an integer ID which is then stored in the integer data table. The array types allow GERMINATE datasets to be extendable to any ploidy level. 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. (See Figure 2 for an example of how genetic data is loaded in GERMINATE)

The final module, General Information, holds information about Institutions, Countries and Users (Figure 1).

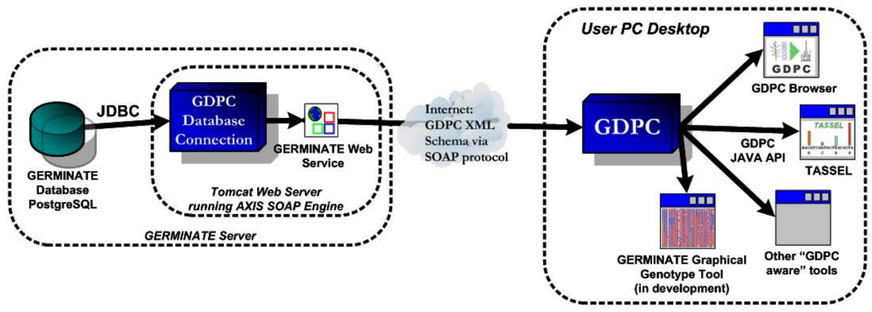
INTERFACES:

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 has been designed for simple queries to the database. Figure 4 shows an image of the passport query interface.

GERMINATE has also been connected to the Genomic Diversity and Phenotype Connection (GDPC) (http://maizegenetics.net/gdpc/). GDPC makes the data in GERMINATE available as web services and allows users to use any analysis tools which have been made GDPC aware. The possibility of extending the versatility of the GDPC interface is currently being explored.

Advanced users also have the option to write SQL queries to retrieve data from GERMINATE. This by-passes potential constraints of the interfaces available.

FIGURE 3: How GDPC is used to connect to the database. This also shows how analysis tools are connected to GERMINATE via GDPC. The Graphical Genotype Tool is currently being developed within the GERMINATE project and will be connected via GDPC. It will be used to display distribution of alleles across taxa and should be extensible to any number of alleles.



FUTURE:

The main goals for the future are the further development of a flexible, user-friendly interface to the database and to develop a broad set of analysis and visualization tools to maximise the accessibility and usefulness of the stored data. The development of tools is currently underway within the GERMINATE project. Among the tools we have planned for use with GERMINATE will include the Graphical Genotype Tool currently being developed. We have also been in contact with CIP about using DIVA (http://diva.riu.cip.cgiar.org/index.php) for GIS data in GERMINATE, this will be done after the next release of DIVA is available. It will also be possible via GDPC to use analysis tools which have been GDPC enabled and recognized the objects returned by GDPC and possibly to enable existing tools to work with GDPC.

The data in GERMINATE will be made available as web services to broaden the applications which can make use of the data. Web services are available through GDPC. Additional objects which can be made available as web services may be included via GDPC or as a separate middleware layer depending on the needs of the users and programs.

We also plan to create a data loading interface which will compare data being submitted to that already in the database and flag potential conflicts which may exist. In addition, we hope that the future development of GERMINATE will see database standards come to realization, and be accepted and expanded upon by those involved with development and utilization of GERMINATE.

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

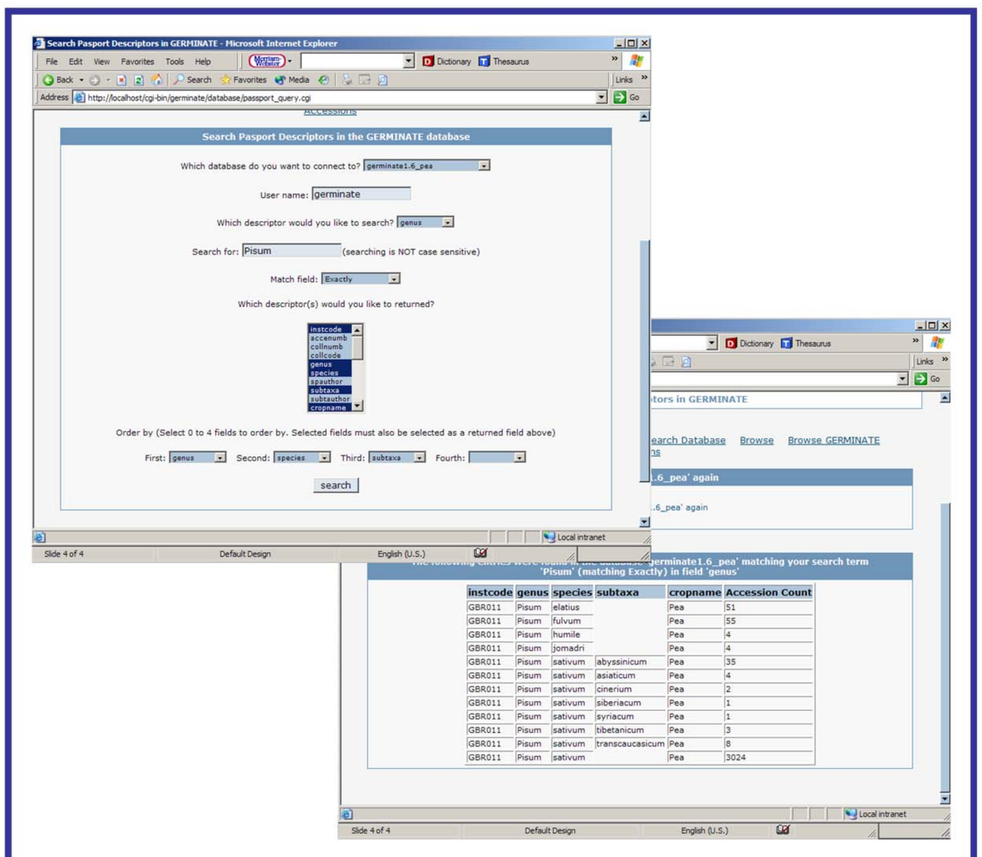


FIGURE 4: The GERMINATE Perl-CGI Interface. A Passport Query Example.