BDGF: a database and web-based information retrieval system for genotype and phenotype

Fábio Danilo Vieira, Danilo Gomes de Moura, Diego Félix da Silva, Roberto Hiroshi Higa, Adhemar Zerlotini

Embrapa Agricultural Informatics, Campinas, SP, Brazil

In recent years, the use of large scale genotyping of tens or hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) to estimate the genomic profile allowed the development of both genotype-phenotype association studies in genomic scale (genome-wide association studies - GWAS) and the introduction of genomic selection technology in breeding programs. However, this situation implies the need of storing large volumes of genotyping, phenotyping and pedigree data from large numbers of animals, a trend that will likely increase over the coming years, given the lower costs for generating the experimental data. In order to effectively integrate such amount of distinctive datasets, it’s advisable to use a robust storage structure, such as a DBMS. Therefore, a major issue to consider is the trade off between normalization and performance during the database modeling stage, as this will have a direct impact on the usability and user experience. In order to get efficient storage and fast queries in this high volume of data, in this work we present the BDGF system (Genotypes and Phenotypes Database). It is based on a data model first proposed by (HIGA, 2015). Nowadays, noSQL has vastly improved our capacity to handle big data, and became integral part of traditional DBMS, such as PosgreSQL. BDGF was completely remodeled so that it has advantages in the use of such technologies. The JSON technology is widely employed in order to allow flexibility to store any phenotype and guarantee immediate query results regardless the number of records. BDGF is designed to support the animal breeding projects of Embrapa, but can be easily adjusted to store data from diverse sources, such as clinical or plant data. Furthermore, the system implements access and security policies to phenotypes, genotypes and pedigree of the animals. The system was developed using webstandards, i18n and free software tools, such as Java, Primefaces, Hibernate and Jboss. BDGF is currently being documented and tested and it's expected to be fully operational within a year.