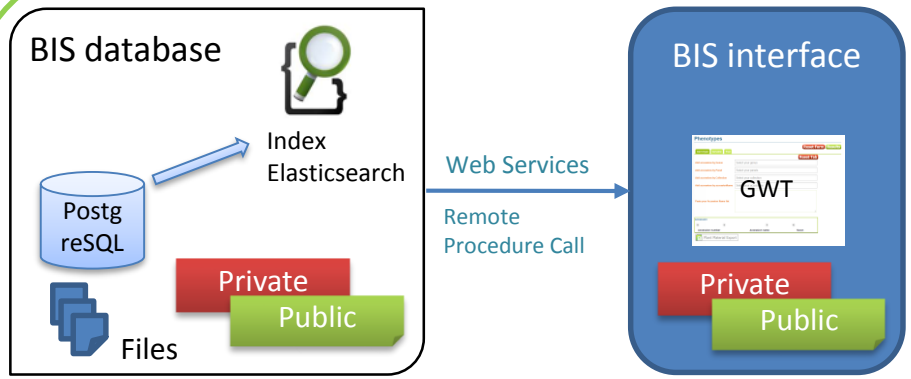



Figure 4 : BIS architecture (database and interface layers)

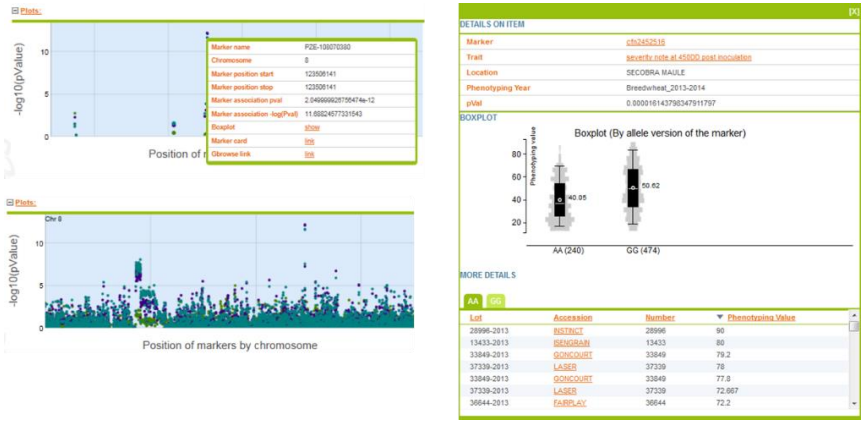



Figure 5 : GWAS graphics from BIS using the Google Web Toolkit technology to generate on-the-fly plots (Manhattan plot and Boxplot)

## *More information*

**www.breedwheat.fr**

*Coordinator :* **Dr. Jacques Le Gouis**, UMR GDEC - jacques.le-gouis@inra.fr

*Project manager :* **Emmanuelle Lagendijk**, INRA Transfert - emmanuelle.lagendijk@paris.inra.fr

*Communication :* **Bernard Bejar**, Céréales Vallée - bernard.bejar@cereales-vallee.org

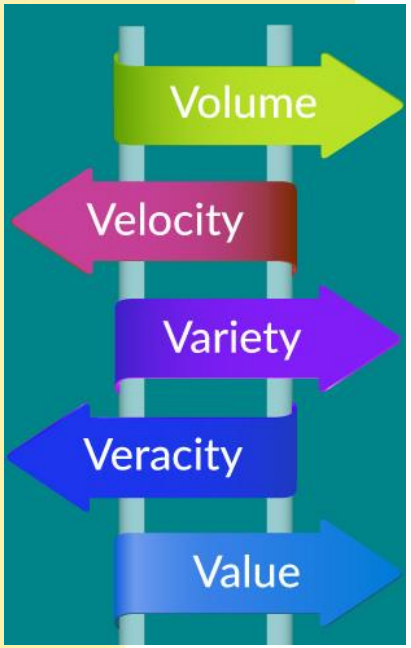
**This project receives funding from the French Government managed by the Research National Agency (ANR) in the framework of the Investments for the Future (ANR-10-BTBR-03), France Agrimer and the French Fund to support Plant Breeding (FSOV).**



# Big data management

July 2017

The "Big data" term is used for datasets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include the capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy of these datasets.
Big data can be described by 5 Vs: Volume, Velocity, Variety, Veracity, Value.




The quantity of generated and stored data. Their size determines the value and potential insight- and whether it can actually be considered big data or not.

In this context, the speed at which the data is generated and processed to meet the demands

The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

The quality of captured data can vary greatly, affecting accurate analysis.

It refers to our ability to turn our data into value. It is easy to fall into the buzz trap and embark on big data initiatives without a clear understanding of the scientific value it will bring.

Figure 1: the 5Vs of Big Data

The data generated by BreedWheat that are managed by bioinformaticians from WP5 ("Bioinformatics for gene discovery, data integration and dissemination") can be considered as Big data because of their:
**Volume:** 3.7 billion genotypic data, 1 million phenotypic data (from 2011 to 2014), etc.
**Velocity:** new phenotypic data each year to generate and integrate in the Information System.
**Variety:** germplasms, SNP discovery, genetic maps, genotyping, phenotyping, ontology and genome wide association studies data.
**Veracity:** data are checked and curated by the bioinformaticians of the WP5.
**Value:** data analysis will lead to the development of new varieties.

## The BreedWheat data

So the volume is quite big and the data are diverse, but more important, the data should be linked together, interoperable and displayed easily and quickly. The BreedWheat Information System (BIS) rely on the GnpIS Information System (Steinbach *et al.*, GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database (Oxford) 2013; 2013 bat058. doi: 10.1093/database/bat058). It allows the data integration and is chosen to host all BreedWheat data. BIS also manages the Intellectual Property of the data with some open data and some accessible only by the project partners before opening them to the community.


**Figure 2 : data integrated in the BreedWheat Information System (https://wheat-urgi.versailles.inra.fr/Projects/BreedWheat)**

## The data integration workflow

The data integration starts from data acquisition in the labs and the fields by the researchers and the technicians. Then the data are curated and formatted in dedicated Excel files (one per data type) including their meta-data (linked ontologies, data standards, etc.) with the help of the WP5 bioinformaticians.
The data managers use the Talend tool to extract, transform and load the data from the Excel files to the BIS Information System.
The development team adapts the Information System to the specific needs of the BreedWheat consortium and improves the databases and interfaces.

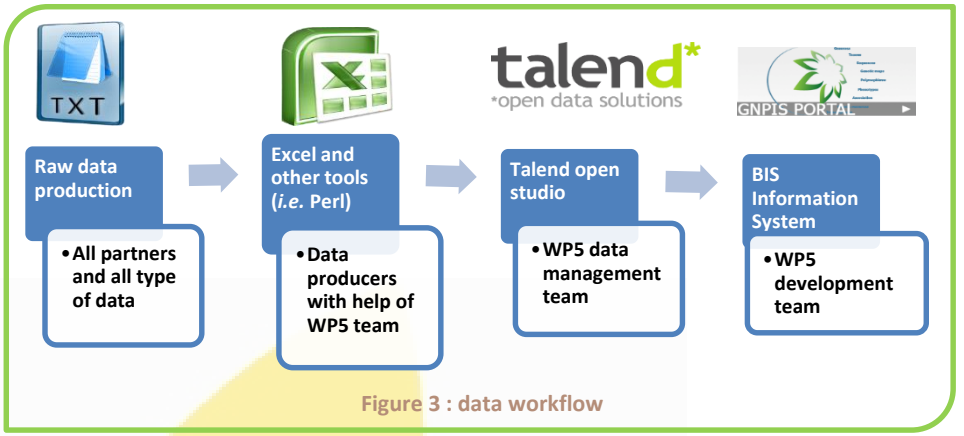



**Figure 3 : data workflow**

## The new technologies

BIS is built on the following technologies:

- ❑ PostgreSQL v9.6 Relational database
- ❑ ElasticSearch v2.3 NoSQL database
- ❑ Google Web Toolkit (GWT) v2.7 framework

The database architecture is an hybrid system using PostgreSQL for the key data to insure the best data quality and consistency, and ElasticSearch for high volume data to allow their storage and fast data query.
BIS interface relies on the GWT framework to generate fast Javascript from Java code. In particular, INRA URGI uses dedicated libraries to display Genome Wide Association Studies (GWAS) plots (gwt.visualization) and synteny browser (gwt.graphics).