# Data Integration and VISualization (DIVIS) : from large heterogeneous datasets to interpretable visualisations in plant science

O. Thierry[1], R. Boumaza[1], J. Buitink[1], C. Landès[1], O. Leprince[1], M. Orsel[1], P. Santagostini[1], and J. Bourbeillon[1]

[1]IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV, 42 rue Georges Morel, 49071 Beaucouzé Cedex, France prenom.nom@inra.fr

**Keys-words :** Heterogeneous Data Integration, Ontology, Plant Science

The demand by biologists to integrate heterogeneous and large datasets from "omics" and phenotyping activites is rapidly increasing [1, 2, 3]. However, methods automating this approach are still at its infancy and to our knowledge, no operational and user-friendly software yet exists. Experiments are performed independently and resulting data are cross-analysed manually and a-posteriori by scientists [4]. For instance, the biology teams from the IRHS (Institut de Recherche en Horticulture et Semences) in Angers have been accumulating datasets of different natures (transcriptomic, biochemistry, physical measures, sensory analysis, etc.) regarding perennial, annual and biannual plants. These datasets are described using reference ontologies enriched with in-house knowledge and stored in a Laboratory Information Management System (LIMS) which is developed and distributed by the IRHS Bioinformatics team.

The main objective of the DIVIS (Data Integration and VISualization) project is to develop a directly usable prototype of a new data analysis tool, by combining the most promising integration and visualisation approaches that are publicly available using the heterogeneous large scale datasets stored in our LIMS.

As a first step, the tool will download and normalise experimental datasets in respect with samples of similar nature across different scales ranging from the molecule to the organism, types of experiments and experimental designs, which is seldom performed by existing software, in particular in plant biology. The originality of our new integration approach, features the following analysis of the resulting matrix:

1. reduce the number of individuals by regrouping similar samples using a similarity score,
2. calculate this score based on similarity between metadata variables stored in a specifically designed ontology. For each ontology concept, relationships with its neighbours will be associated with similarity indices. These indices will be used to calculate a similarity between individuals associated with these ontology concepts [5],
3. represent each group by an archetype sample,
4. construct graphical representations of the results. The visualisation approach will allow to present data regarding these archetype samples in a multi-layer display separating various subsets of coherent data and to navigate through the results.

In order to validate the methodology and assess how the approach can be adapted to different experimental contexts with an equivalent level of complexity, the tool is developed based on two test datasets acquired as part of matching experiments (including several studies performed on the same samples):

- an apple fruit dataset including descriptors at the variety level (fruit shape, colour or size, tree shape or vigour, etc.) and measures at the fruit level (transcriptomic, biochemical, physical and sensory data),
- a seed dataset containing descriptors at the genotype level (genotype, environmental and climatic data regarding the collection site) and at the seed level (germination kinetics and physical attributes).

So far we have constructed the integrated and normalised data matrices. We are currently designing or reusing relevant ontologies and associate each individual with concepts from these ontologies[1, 5, 6, 7]. The aspects that under consideration are as follows:

- For the apple dataset:
  - reuse existing lists of descriptors associated with apple varieties (fruit colour, fruit shape, etc.) to devise ontologies
- For the seed dataset:
  - design an ontology of the shape (long, short, straight, curved, etc.) of seeds and associate these concepts with actual measurements
  - reuse exisiting colour tables and add relationships between colours to devise a colour ontology associated to HSV values to describe seeds colours
  - reuse the Köppen climate classification to associate climatic data for each genotype collection site to a climate class
  - reuse existing lists of descriptors associated with seed collection sites (topology, pedology, etc.) and plant characteristics for each genotypes (pod shapes, flower colours, leaf shapes, etc.) to devise ontologies

The next stage will be to cluster the individuals according to these ontology concepts.

# References


[1] Solovieva E, Shikanai T, Fujita N, Narimatsu H. GGDonto ontology as a knowledge-base for genetic diseases and disorders of glycan metabolism and their causative genes. Journal of Biomedical Semantics. 2018;9:14. doi:10.1186/s13326-018-0182-0.

[2] Hendler J. Data Integration for Heterogenous Datasets. *Big Data*. 2014;2(4):205-215. doi:10.1089/big.2014.0068.

[3] Fonseca, Frederico T. et al. "Using Ontologies for Integrated Geographic Information Systems." Trans. GIS 6 (2002): 231-257.

[4] Arguello Casteleiro M, Demetriou G, Read W, et al. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. *Journal of Biomedical Semantics*. 2018;9:13. doi:10.1186/s13326-018-0181-1.

[5] Kohler S. Improved ontology-based similarity calculations using a study-wise annotation model. *Database: The Journal of Biological Databases and Curation*. 2018;2018:bay026. doi:10.1093/database/bay026.

[6] Cohen J, Matthen M, Bradford Book, A. (2010). Color Ontology and Color Science.

[7] Hartmann, J, Palma, R, Gómez-Pérez, A. (2009). Ontology Repositories. Handbook on Ontologies.