**Data and rules for the Challenge 2017**

A common set of wheat data from three different NIR instrument manufacturers will be analyzed. The objective is to evaluate the data with the goal of preprocessing the datasets to match the spectra from all three instrument manufacturers prior to developing ONE regression model for protein that results in equivalent results among the instruments as measured by the reproducibility and accuracy (RMSEP).

The samples correspond to wheat with spectra collected on five instruments per NIR spectrometer manufacturer. The instrument models and instrument serial numbers have been coded. The spectra are in the range and spacing that their respective instrument manufacturers support. The reference protein results are on a 12% moisture basis.

There are 1488 spectra in the calibration data set for 248 samples analyzed on three instruments for manufacturer A and three instruments for manufacturer B.

There are 744 spectra in the test data set for the same 248 samples as in calibration analyzed on a fourth instrument for manufacturer A and B in addition to one instrument for manufacturer C.

There are 450 spectra in the validation data set for an independent set of 150 samples analyzed once each on a new instrument for each manufacturer A, B, and C.

Reference protein values are provided for the calibration and test sets only.

The order of the Calibration and Test samples for all sets and instrument manufacturers is identical. However, the order of the Validation samples was randomized for each instrument manufacturer.

The figure below presents the available datasets.

This year’s challenge will consist in developing the best model for the parameter provided using the calibration data. However, the most important task will be to determine a method of pre-processing the data to minimize the spectral differences among instrument models so that a single calibration will yield both excellent accuracy and excellent reproducibility among "unknown" instruments of each manufacturer without secondary standardization. In addition, the quality of the presentation of the results and the reasoning behind the approach taken will be used to determine the winner. Participants are to:

1) Develop a preprocessing method to match the spectra from all three NIR instrument manufacturers to be used in a common calibration set.

2) Develop the model for protein on the calibration set.

3) Test their model on a test set (reference values provided).

4) Predict a validation set (reference values not provided).

5) Detail the reasoning when selecting pre-treatment methods, regression method, and number of latent variables, etc…

Participants who wish to compete **must submit** their predictions of the calibration, test

and validation sets by **25 January 2017** in an EXCEL file to:

[p.dardenne@cra.wallonie.be](mailto:p.dardenne@cra.wallonie.be)

A ppt or doc file must be added explaining the methodology.

The criterion for deciding the winners is the RMSEP on the 450 validation spectra.

Scheme of the data sets – Challenge 2017 – Paris

