fine-tuning the

DuplicateFinder

DuplicateFinder is a tool to support you in your search for duplicates and most appropriate accessions, it doesn't do the job for you. It will still be difficult and time consuming to find the duplicates. We just hope DuplicateFinder will help you in this effort and save you a little time.

Introduction

Before you continue, please first read the 'short manual'. For those who are experienced in VBA, this manual gives some possibilities for fine tuning the software.

Exclude specific ACCENAME content from the search process

A soundex is created on the column ACCENAME. Amongst others, this soundex is used to match accessions. Sometimes the content of this field is too unspecific (e.g. MESTNYI or landrace) and should be excluded from the matching process. Therefore, unhide the sheet noSndx (right-click a sheet-tab, chose unhide) and in the column noSOUNDEX you can add names that should be ignored.

For advanced VBA users only:

Parameters for the calculation of the Similarities

The macro 'DoCalcSimilarities' in the module 'CalcSimlarities' calculates the similarities between accessions, using the information stored in the (hidden) sheet DATA3. The hidden sheet 'SimRules' gives an overview of the parameters used in the calculations. Adapting the parameters in that sheet has no effect. The parameters need to be changed in the macro itself.

Make Duplicate Groups

In the macro 'ClusterDuplicationGroups' in the module 'Cluster' the threshold for grouping is set to 30% (SimilarityThreshold = 0.3). You may prefer a higher threshold (e.g. 0.5), this will result in smaller groups since (clusters of) accessions have to be more similar to be joined.

For further clustering of (clusters of) accessions default the MINIMUM similarity option is used, which will create relative small groups. Optionally you can choose MAXIMUM (-->

large groups) or AVERAGE similarity (--> medium sized groups) by deactivating the default (put a ' in front of the line) and activate your choice (by removing the ' in the front of the specific line) in de Do-loop.

NB: the hidden sheet ‘MCPD Codes’ is used by the macro’s to check which DESCRIPTORS / SAMPSTAT / COLLSRC / ORIGCTY / INSTCODE / STORAGE values are allowed.