Inclusion of C&E data in EURISCO analysis and options

Theo van Hintum

Centre for Genetic Resources, the Netherlands

C&E data in EURISCO

- the presentation
  - introduction
    - conceptual issues
    - experiences
  - proposal
    - concept
    - elements
    - upload mechanism
    - download mechanism
    - implementation
  - concluding remarks

C&E data in EURISCO

- introduction - conceptual issues
  - all think it’s important – it didn’t happen
  - C&E data
    - scores of genotypic traits
    - characterization: highly heritable, easily observable traits
      - flower color, row number, flowering time, number of shoots
    - evaluation: more difficult to observe traits requiring specific experiments and/or equipment to determine
      - protein content, grain yield, resistance to a specific pathotype
    - line between C and E is very vague – treat as one category C&E
    - molecular fingerprinting data are not considered C&E data

C&E data in EURISCO

- introduction - conceptual issues (contn'd)
  - C&E: measurements on the phenotype
    - model for phenotype: $p_{ij} = g_i + e_j + ge_{ij} + \varepsilon_{ij}$
      - we are interested in g_i – but cannot know it
  - proper interpretation of a score (p_{ij}) requires info about
    - genotype (one or more plants of an accession)
    - trait (property that was scored)
      - plant height, pl-ln, plantlengte, C204, length in vegetative stage
    - method (scale, precision, heterogeneity-handling)
    - experiment (conditions, treatment, design, environment)

C&E data in EURISCO

- introduction - conceptual issues (contn'd)
  - extreme options for exchanging scores
    - *heritability* : only use highly heritable traits, standardize scale
      - effect e_j, ge_{ij} and ε_{ij} low
      - typical characterization traits (row number, crop type)
    - *standardization* : standardize experiment - include standards, prescribe design, control environment (irrigation, soil, disease control)
      - effect e_j, ge_{ij} constant and ε_{ij} low
      - registration and breeding trials
    - *interpretation* : use raw scores, also exchange context data
      - statistical and/or heuristic analysis is needed to look over experiment boundaries

C&E data in EURISCO

- introduction – experiences (brief)
  - C&E data rarely available on genebank websites
    - even more rarely searchable
  - obtaining C&E data from genebanks is very difficult
    - low level of computerization
    - labor involved in the required standardization
    - IP issues?
  - CCDBs use different approaches
    - none of them the ‘silver bullet’

C&E data in EURISCO

- introduction – main messages
  - C&E data are important to the user but complicated in nature
  - big challenge: to get data from the source (genebank)
    - don’t require too much manual input
    - create a one-time solution that can work from then on

C&E data in EURISCO

- proposal – concept
  - assumptions
    - it is not feasible to enforce any standardization in terms of experimental design, the use of standards or even the scale of measurement
    - all (potential) data donors should be able to export their data, as they have it, in a common format, provided that this is a flexible format
    - the value of C&E data is that high to a user that (s)he is willing to invest time in analyzing the data
  - principle
    - create a C&E data repository
      - create a data exchange format that is able to cope with unstandardised C&E data
      - describe genotype, trait, method and experiment

C&E data in EURISCO

- proposal – elements
  - genotype
    - concerns accessions already registered in EURISCO
    - identification via unique key of EURISCO (combined key consisting of the fields NICODE, INSTCODE, ACCENUMB and GENUS)
  - trait
    - no agreed descriptor list or ontology exists (work on controlled vocabulary or ontology as source of inspiration)
    - accept the names as used by the data providers - ask is to provide English name of trait

C&E data in EURISCO

- proposal – elements (contn'd)
  - method
    - brief description, in English, of the way the trait was scored
      - scale that was used
      - additional info such as 'the average of five random spikes'
  - experiment
    - brief description, in English, of relevant aspects of the experiment:
      - 'on sandy soil in the Netherlands', 'during multiplication', 'from a randomized complete block experiment in triplo', 'start of growing season was dark and humid', etc.

C&E data in EURISCO

- proposal – elements (contn'd)
  - C&E data uploaded in packages consisting of one or more experiments with possibly a generic methodological remark
    - e.g. the convention for handling variation within accessions
  - one experiment contains *n* genotypes and *m* traits (with their method) and of course *n* x *m* scores
    - easy to implement in relational database

C&E data in EURISCO

- proposal – elements (contn'd)
  - upload implemented in any format
    - xml, xls, csv
    - upload files, webservice
  - five elements
    - DATASET
    - EXPERIMENT
    - TRAIT
    - GENOTYPE
    - SCORE

C&E data in EURISCO

- proposal – elements (contn'd)
  - DATASET containing
    - NICODE – see EURISCO (mandatory)
    - DATASET_REMARK – any general remark relevant to all scores in the dataset (max 255 alphanumeric)

C&E data in EURISCO

- proposal – elements (contn'd)
  - EXPERIMENT containing
    - EXPERIMENT_NUMBER – unique number in the dataset for the experiment; this number should be unique for the NI (mandatory)
    - EXPERIMENT_DESCRIPTION – information relevant for the interpretation of the scores in the experiment such as experimental design, location, experimentor, weather, etc. (max 255 alphanumeric)
    - EXPERIMENT_YEAR – the year the experiment was done (started) (4 numeric)
    - EXPERIMENT_REPORT – a reference to the report of the experiment, either supplied with the data (then only the file name needs to be given) or the URL of the report (max 100 alphanumeric)

C&E data in EURISCO

- proposal – elements (contn'd)
  - TRAIT containing
    - TRAIT_NUMBER – unique number for the trait in the dataset (mandatory)
    - TRAIT_NAME – English name of the trait (max 50 alphanumeric, mandatory)
    - TRAIT_REMARK - any general remark that helps interpret the trait (max 255 alphanumeric)
    - TRAIT_METHOD – a description of the method for measuring and the scale used (max 255 alphanumeric)

C&E data in EURISCO

- proposal – elements (contn'd)
  - GENOTYPE containing
    - GENOTYPE_NUMBER – unique number for the genotype in the dataset (mandatory)
    - GENOTYPE_INSTCODE – see EURISCO (mandatory)
    - GENOTYPE_ACCENUMB – see EURISCO (mandatory)
    - GENOTYPE_GENUS – see EURISCO (mandatory)

C&E data in EURISCO

- proposal – elements (contn'd)
  - SCORE containing
    - GENOTYPE_NUMBER – key to GENOTYPE (mandatory)
    - EXPERIMENT_NUMBER – key to EXPERIMENT (mandatory)
    - TRAIT_NUMBER– key to TRAIT (mandatory)
    - SCORE – actual score (max 10 alphanumeric, mandatory)

C&E data in EURISCO

- proposal - upload mechanism
  - aligned with the current EURISCO upload mechanism
    - responsibility of NI focal point
    - report about the replace and insert actions should be send to the uploader

C&E data in EURISCO

- proposal - download mechanism
  - not obvious - needs much attention
  - use-case oriented
    - different users should be identified and their needs should be described and accommodated
    - two major user-groups:
      - the bulk user, such as CCDB managers creating or maintaining a crop specific PGR portal and scientists doing a large survey
      - the trait searcher, a breeder or scientist who is looking for a specific trait

C&E data in EURISCO

- proposal - download mechanism (contn'd)
  - complicating factor: EURISCO doesn't have a standardized division in crops
    - any user starts by selecting accessions, for example of all *Triticum* and *Aegilops*, in all spelling and format versions currently featured in EURISCO
    - after selecting the accessions, the user should only be confronted with the C&E data on those accessions
  - next step: selection of traits and experiments
    - trait names are not standardized this might involve long lists of trait-names, and might require a search interface
    - after selecting the trait(s), the user should be allowed to select the experiments that (s)he would like to get access to

C&E data in EURISCO

- proposal - download mechanism (contn'd)
  - final step: downloading the data
    - could take many shapes, including download of entire experiments or download of matrices with accession times trait/experiment combinations
    - display of selected data in the selected format might be a problem because of the size of the information
    - required format should be selected (xls, xml, csv)
    - output should be generated, with appropriate meta information (decoded codes, a readme for the use and interpretation) and made available in a downloadable shape (in a zip file or on a html page with clickable files)

C&E data in EURISCO

- proposal – implementation
  - steps
    - create ownership in the community of genebanks for the approach to follow
    - agree on and define (the elements of) the mechanisms in detail
    - get commitment of a few large potential data donors to supply their data in the testing phase (NGB, CGN, BLE)
    - build required software and test upload mechanism
    - support potential new data donors by approaching them on a personal level, organizing training workshops and/or technical visits
    - improve on interface and download format in collaboration with selected users
    - promote resulting database via publications and/or presentations PGR community, plant scientists and breeders

C&E data in EURISCO

- concluding remarks
  - creating a EURISCO C&E repository is do-able
    - provided support of genebank community
    - provided support of EURISCO
  - a EURISCO C&E repository positive for standardization
    - C&E data themselves
    - trait ontology
    - C&E methodology

C&E data in EURISCO

we have been talking about C&E data for too long

let’s get some work done …