NordGen

C&E data: the EURISCO standard

Workshop of the ECPGR Doc & Info WG
20-22 May 2014, Prague

Jonas Nordling, NordGen

Background

EPGRIS3 workshop inBonn May 7 2009
Proposal: ”Inclusion of C&E data in EURISCO – analysis and options”

FP7 applications:
Eurogenebank 2010
Plant Gene Access 2012

ECPGR workshop Prauge, May 2014

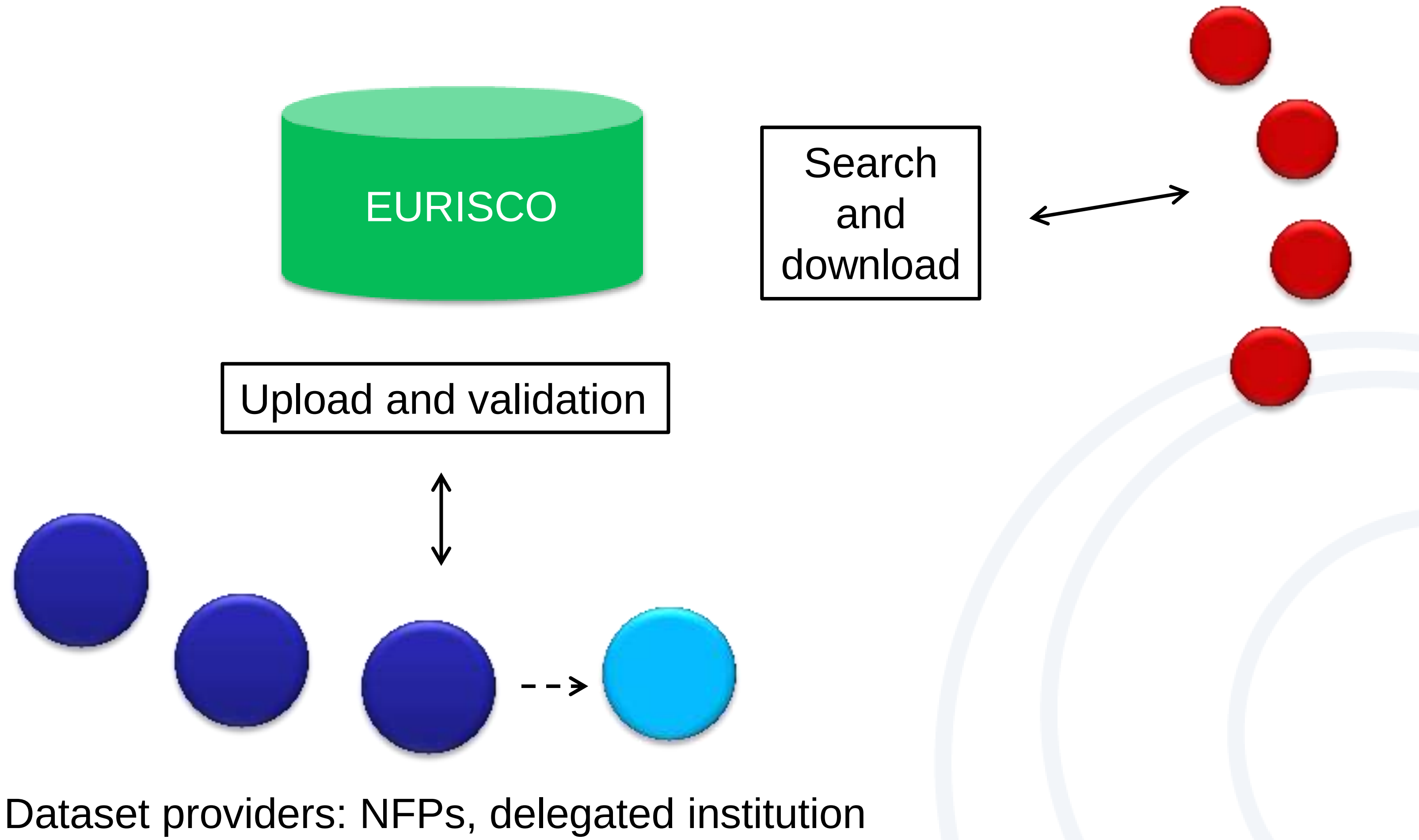
Nature of C&E data

- C&E data concerns measurements on the phenotype
- Phenotype is usually the result of both the genotype and environment, so what is needed to interpret the score?
- What genotype was scored?
- What trait was scored?
- How was it scored?
- When and where was it scored?

Cornerstones

- A repository of non-standardized C&E data in the framework of EURISCO.
- It is not feasible to enforce any standardization in terms of experimental design, the use of standard descriptors .
- All data donors should be able to export their data, as they have it, requiring a simple flexible common format.
- The value of C&E data is that high to a user that he/she is willing to invest time in analyzing the data.
- Elements of exchange is
    - Dataset
    - Experiment
    - Trait (Descriptor) with Method
    - Genotype
    - Score

Repository of European C&E data



Users

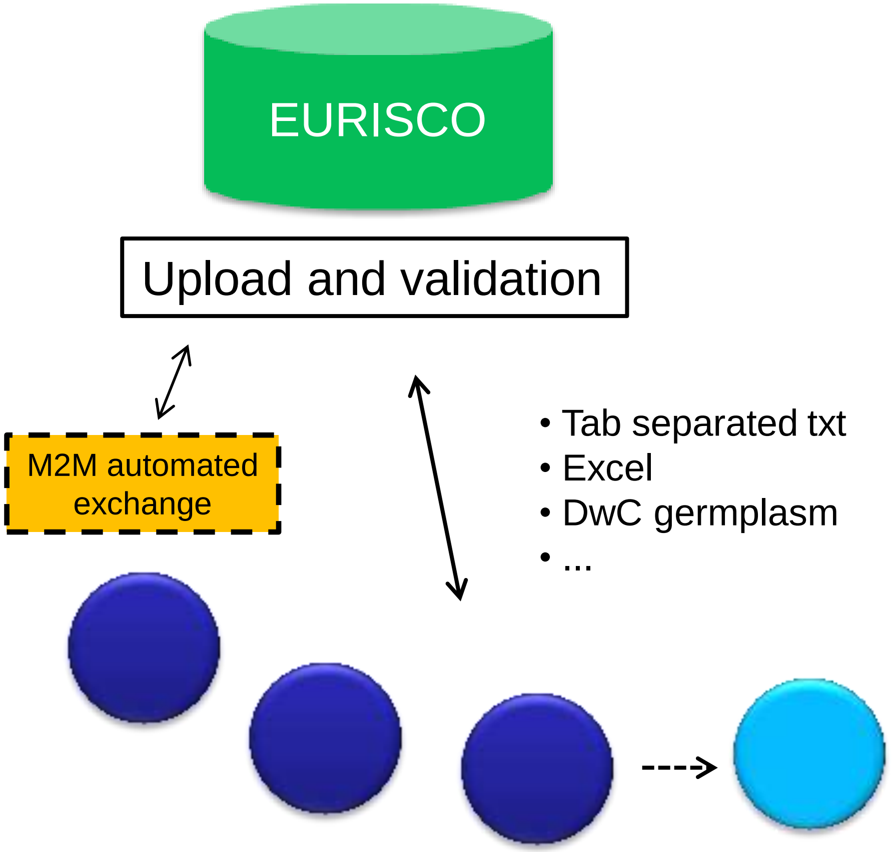
EURISCO

Search and download

Upload and validation

Dataset providers: NFPs, delegated institution

Upload and validation

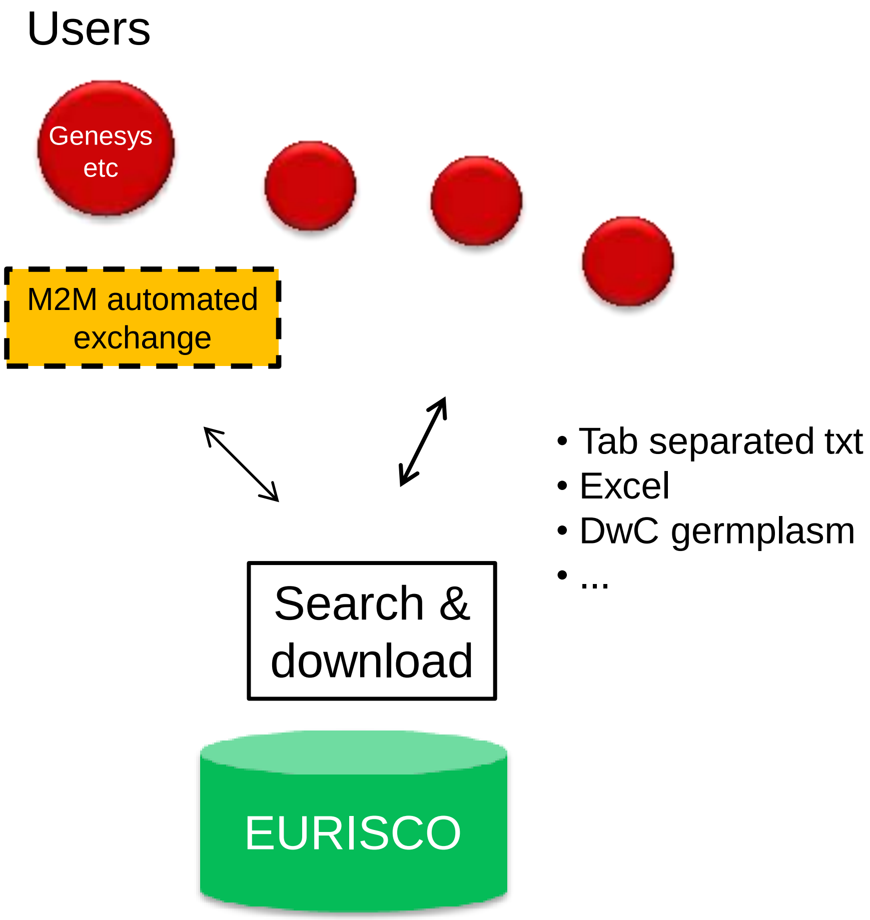


- Tab separated txt
- Excel
- DwC germplasm
- ...

- Registered uploader
- Only accessions in Eurisco
- Not only accessions from own NI
- Only non-confidential data
- Validation of data format
- Error reports back to uploader
- File format to be agreed
- Packages of one or more experiments
- Experiment has unique ID

Dataset providers: NFPs and other registered uploaders, possibly with revision from NFP before commit.

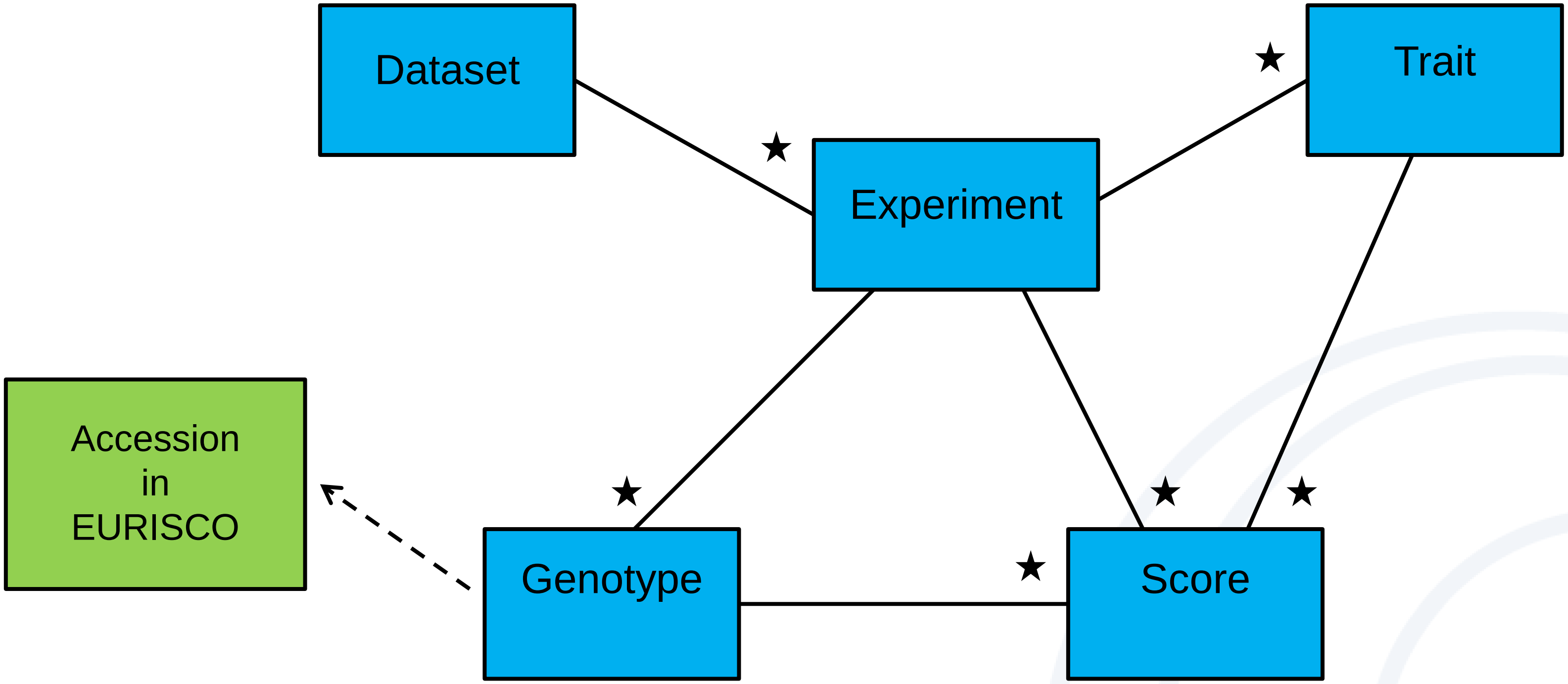
Search and download



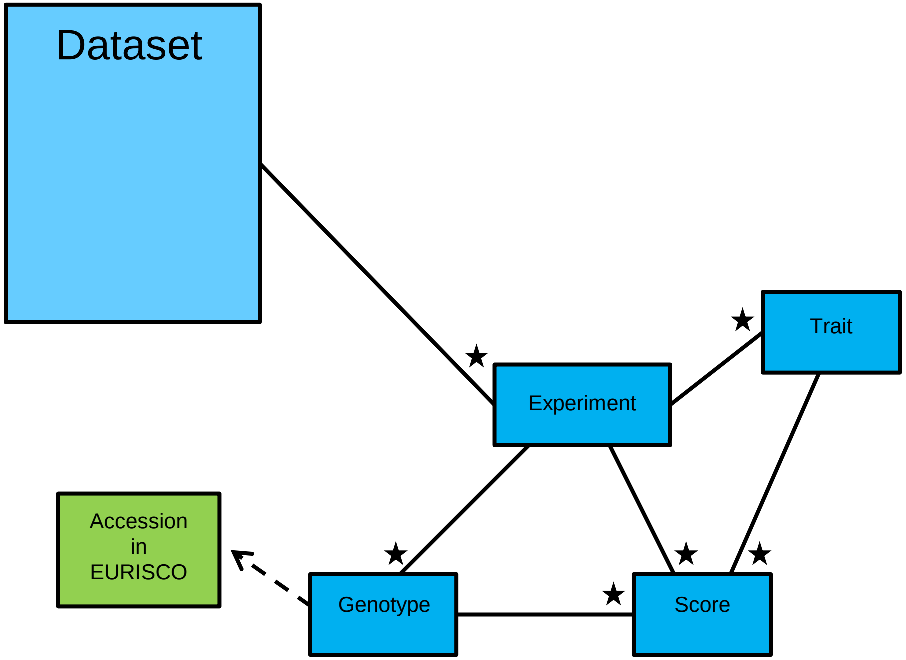
- Tab separated txt
- Excel
- DwC germplasm
- ...

- Possible use-case:
    - Taxonomic selection
    - Trait selection
    - Browse experiments, accessions and scores
    - Download scores for selected accessions with experiment and trait metadata
    - Or use experiment url, a link to original data
- Full text search on several fields will be necessary
- No ontologies for traits required from start. English trait names required. Search can be enhanced with ontologies for traits

Data model


Dataset
Trait
★
★
Experiment
Accession
in
EURISCO
★
★
★
Genotype
★
Score

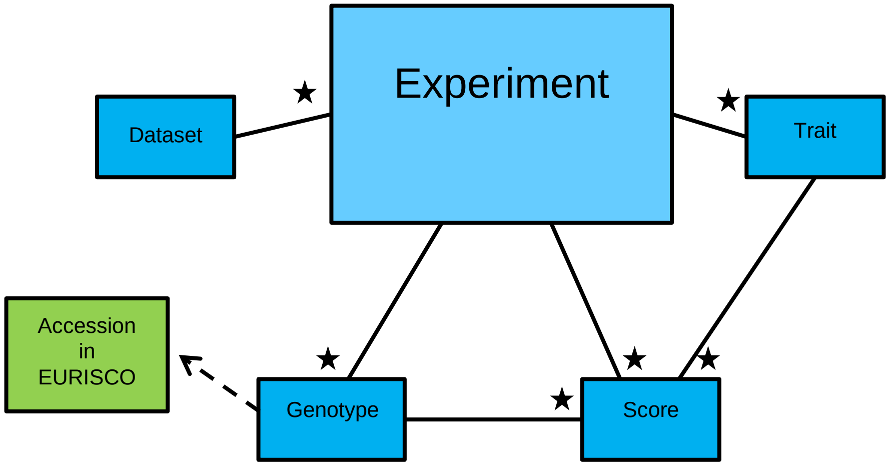
Dataset fields



UPLOADERCODE – ID-Code *
provided by EURISCO, for the registered authorized data provider uploading the data to EURISCO (mandatory)

DATASET_REMARK
any general remark relevant to all scores in the dataset (max 255 alphanumeric)

Experiment fields



EXPERIMENT_NUMBER *
unique number in the dataset for the experiment; this number should be unique and persistent for the data provider (mandatory)

EXPERIMENT_DESCRIPTION
information relevant for the interpretation of the scores in the experiment such as experimental design, experimenter, weather, etc.
(max 255 alphanumeric)

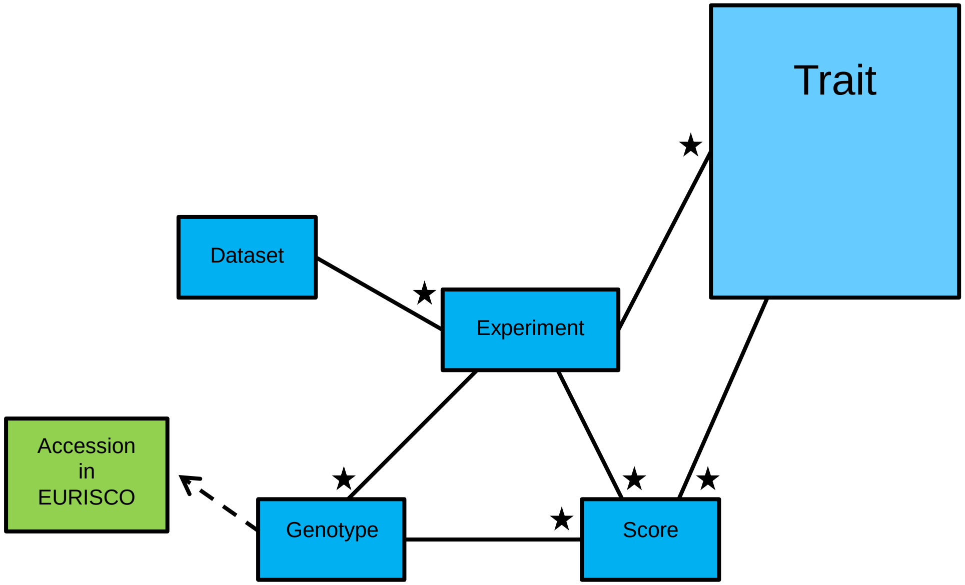
EXPERIMENT_YEAR
the year the experiment was done (started)
(4 numeric)

EXPERIMENT_LONGITUDE
the longitude of the experimental site, provided it was an experiment in the open field
(decimal number)

EXPERIMENT_LATITUDE
the latitude of the experimental site, provided it was an experiment in the open field
(decimal number)

EXPERIMENT_REPORT
a reference to the report of the experiment, either supplied with the data (then only the file name needs to be given, that could be presented as a hot-link in the interface) or the URL of the report or original data (max 100 alphanumeric)

Trait fields



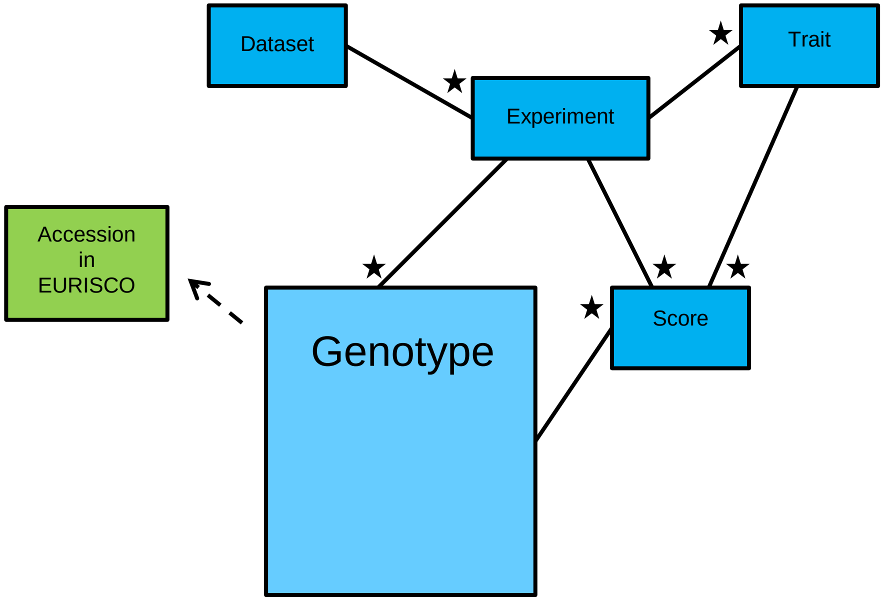
TRAIT_NUMBER *
unique, temporary number for the trait in the dataset (mandatory)

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)

Genotype fields



GENOTYPE_NUMBER *
unique temporary number for the genotype in the dataset (mandatory)

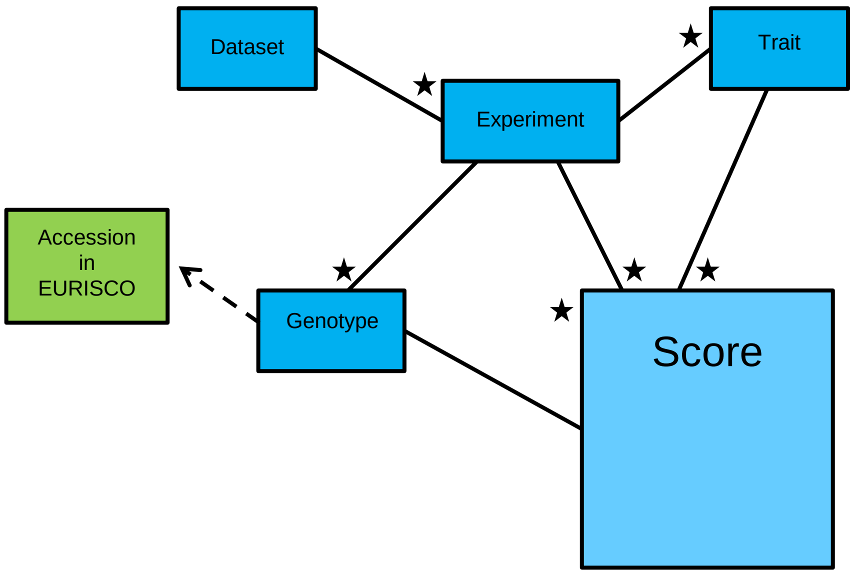
GENOTYPE_NICODE *
Reference to Eurisco field (mandatory)

GENOTYPE_INSTCODE *
Reference to Eurisco field (mandatory)

GENOTYPE_ACCENUMB *
Reference to Eurisco field (mandatory)

GENOTYPE_GENUS *
Reference to Eurisco field (mandatory)

Score fields



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)

Further comments

English recommended for text fields. Trait name required in English.

More fields can be added later on.

Starting to share C&E data may promote standardization, e.g. of descriptors.

Support to data donors may be necessary.

When known descriptor from descriptor lists (upov, ipgri, bioversity etc) are used it should be used as trait name or stated in trait remark

Next step – implementation in Eurisco!