Phenotypic data in EURISCO

EURISCO training workshop 2021
10–12 November 2021
Dealing with phenotypic data: Great diversity

• Phenotypic data
  • Determines value of germplasm for breeding and research
  • Crop-specific traits and methods
  • Many historical datasets
  • Usually no data from high throughput phenotyping
  • Data has to be aggregated or exchanged between organisations

Lots of “standards” to express traits
  • Different trait names/synonyms
  • Different rating scales (nominal, ordinal, metric)

Different amounts of meta information
  • When, where, how, by whom?
  • Experiment set-up, treatment etc.

Different means of data management
  • DBMS, flat files, mainly Excel files
Dealing with phenotypic data: Existing situation

<table>
<thead>
<tr>
<th>Methods and Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Crop-specific definitions of traits, methods etc. like IPGRI descriptor lists</td>
</tr>
<tr>
<td>• Often used in parts only and adapted to organisational needs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exchange Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>• E.g. Darwin Core germplasm extension (DwC-germplasm; Endresen et al. 2009)</td>
</tr>
<tr>
<td>• Great for computer scientists</td>
</tr>
<tr>
<td>• Difficult to handle for genebank curators</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Help to structure the (phenotypic) world</td>
</tr>
<tr>
<td>• Improve interoperability of data</td>
</tr>
<tr>
<td>• e.g. Crop Ontology (Arnaud et al. 2012)</td>
</tr>
</tbody>
</table>
Dealing with phenotypic data: Current approach

- **Data standardisation**
  - About 600 germplasm collections in Europe, around 400 in EURISCO
  - No standardisation of trait, scale or experimental design
  - Pragmatic approach: Import of existing data as-is to reach critical mass

- **Data exchange**
  - Only standardisation of exchange format
    - As simple as possible
    - As few fields as possible
  \( \rightarrow \) “minimum consensus”

- **Data management**
  - Highly abstracted, following the single-observation concept (van Hintum et al. 1992)
  - Omitting fine-grained metadata
Data model for phenotypic data
Dataset

- Enables to upload multiple experiments at once
- Fields:
  - **UPLOADERCODE***:
    - ID of registered authorised data provider
    - Provided by EURISCO
  - **DATASET_NUMBER***:
    - To link experiments with datasets
    - Unique and persistent for the data provider
  - **DATASET_REMARK**:
    - General remark for all scores in the dataset

<table>
<thead>
<tr>
<th>UPLOADERCODE</th>
<th>DATASET_NUMBER</th>
<th>DATASET_REMARK</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEU271</td>
<td>1</td>
<td>This dataset contains forage grass accessions.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Experiment

• Meta data helping to interpret C&E data
  • Experiment set-up
  • Weather conditions
  • Soil conditions
  • Experiment location
  • …
• Fields:
  • `DATASET_NUMBER*`:
    • Reference to the dataset
  • `EXPERIMENT_NUMBER*`:
    • To link scores with experiments
    • Unique and persistent for the data provider
Experiment

- Fields (cont.):
  - `EXPERIMENT_DESCRIPTION`:
    - Brief English description
    - Information necessary for interpreting the scores, e.g. set-up
  - `EXPERIMENT_START_YEAR`:
    - Year in which the experiment was performed/started
  - `EXPERIMENT_END_YEAR`:
    - Year in which the experiment was ended
  - `EXPERIMENT_LONGITUDE`:
    - Longitude of experimental site
  - `EXPERIMENT_LATITUDE`:
    - Latitude of experimental site
  - `EXPERIMENT_REPORT`:
    - Reference to a report
      - Either report file or report URL
<table>
<thead>
<tr>
<th>DATASET_NUMBER</th>
<th>EXPERIMENT_NUMBER</th>
<th>EXPERIMENT_DESCRIPTION</th>
<th>EXPERIMENT_START_YEAR</th>
<th>EXPERIMENT_END_YEAR</th>
<th>EXPERIMENT_LONGITUDE</th>
<th>EXPERIMENT_LATITUDE</th>
<th>EXPERIMENT_REPORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Characterisation data of Lolium perenne</td>
<td>1999</td>
<td>2000</td>
<td>11.278414</td>
<td>51.826059</td>
<td>http://...</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Characterisation data of Lolium perenne</td>
<td>2000</td>
<td></td>
<td>11.278414</td>
<td>51.826059</td>
<td>http://...</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>Characterisation data of Lolium perenne</td>
<td>2001</td>
<td></td>
<td>11.278414</td>
<td>51.826059</td>
<td>http://...</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>Evaluation data of Lolium perenne (4 replications per accession)</td>
<td>2002</td>
<td></td>
<td>11.278414</td>
<td>51.826059</td>
<td>http://...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Trait

• Describe phenotypic traits and the methods used for scoring

• Fields:
  
  • **Trait_Number***:
    • Unique, temporary number of the trait in the dataset
  
  • **Trait_Name***:
    • English name of the trait
  
  • **Trait_Remark**:
    • General remark helping to interpret the trait
  
  • **Trait_Method**:
    • English description of the used method + scale
<table>
<thead>
<tr>
<th>TRAIT_NUMBER</th>
<th>TRAIT_NAME</th>
<th>TRAIT_REMARK</th>
<th>TRAIT_METHOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sowing date</td>
<td>...</td>
<td>Date</td>
</tr>
<tr>
<td>2</td>
<td>Emerging date</td>
<td>...</td>
<td>Date</td>
</tr>
<tr>
<td>3</td>
<td>Growing before winter</td>
<td>...</td>
<td>Rating value from 1 (min) – 9 (max)</td>
</tr>
<tr>
<td>4</td>
<td>Stem height min</td>
<td>In flowering time, the shortest plant</td>
<td>Measurement [cm]</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Genotype

- All accessions for which C&E data will be uploaded

- Fields:
  - **GENOTYPE_NUMBER**: Unique, temporary number of the genotype in the dataset
  - **GENOTYPE_NICODE**: National Inventory code from EURISCO
  - **GENOTYPE_INSTCODE**: Holding institute code from EURISCO
  - **GENOTYPE_ACCENUMB**: Accession number from EURISCO
  - **GENOTYPE_GENUS**: Genus from EURISCO
  - **GENOTYPE_PUID**: Placeholder for a PUID
<table>
<thead>
<tr>
<th>GENOTYPE_NUMBER</th>
<th>GENOTYPE_NICODE</th>
<th>GENOTYPE_INSTCODE</th>
<th>GENOTYPE_ACCENUMB</th>
<th>GENOTYPE_GENUS</th>
<th>GENOTYPE_PUID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DEU</td>
<td>DEU271</td>
<td>GR 142</td>
<td>Lolium</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>DEU</td>
<td>DEU271</td>
<td>GR 476</td>
<td>Lolium</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>DEU</td>
<td>DEU271</td>
<td>GR 550</td>
<td>Lolium</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>DEU</td>
<td>DEU271</td>
<td>GR 2670</td>
<td>Lolium</td>
<td></td>
</tr>
</tbody>
</table>
Score

- Observed phenotypic values of the accessions
- Fields:
  - **GENOTYPE_NUMBER**: Reference to a genotype
  - **EXPERIMENT_NUMBER**: Reference to an experiment
  - **TRAIT_NUMBER**: Reference to a trait
  - **SCORE**: Observed score
  - **SCORE_LINK**: Link to a publication on accession level
<table>
<thead>
<tr>
<th>GENOTYPE_NUMBER</th>
<th>EXPERIMENT_NUMBER</th>
<th>TRAIT_NUMBER</th>
<th>SCORE</th>
<th>SCORE_LINK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19990313</td>
<td>http://...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>http://...</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td>20020401</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>21</td>
<td>http://...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
## Connecting the templates

<table>
<thead>
<tr>
<th>GENOTYPE_NUMBER</th>
<th>GENOTYPE_IDCODE</th>
<th>GENOTYPE_INSTCODE</th>
<th>GENOTYPE_ACCENUMB</th>
<th>GENOTYPE_GENUS</th>
<th>GENOTYPE_PID</th>
<th>GENOTYPE_CLUSTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DEU</td>
<td>DEU271</td>
<td>GR 142</td>
<td>Lolium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>DEU</td>
<td>DEU271</td>
<td>GR 476</td>
<td>Lolium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>DEU</td>
<td>DEU271</td>
<td>GR 550</td>
<td>Lolium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>DEU</td>
<td>DEU271</td>
<td>GR 2670</td>
<td>Lolium</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Trait

<table>
<thead>
<tr>
<th>TRAIT_NUMBER</th>
<th>TRAIT_NAME</th>
<th>TRAIT_REMARK</th>
<th>TRAIT_METHOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sowing date</td>
<td>...</td>
<td>Date</td>
</tr>
<tr>
<td>2</td>
<td>Emerging date</td>
<td>...</td>
<td>Date</td>
</tr>
<tr>
<td>3</td>
<td>Growing before winter</td>
<td>...</td>
<td>Rating value from 1 (min) – 9 (max)</td>
</tr>
<tr>
<td>4</td>
<td>Stem height min</td>
<td>In flowering time, the shortest plant Measurement [cm]</td>
<td>...</td>
</tr>
</tbody>
</table>

### Score

<table>
<thead>
<tr>
<th>GENOTYPE_NUMBER</th>
<th>EXPERIMENT_NUMBER</th>
<th>TRAIT_NUMBER</th>
<th>SCORE</th>
<th>SCORE_LINK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>http://...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>http://...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>http://...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>http://...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Dataset

<table>
<thead>
<tr>
<th>DATASET_NUMBER</th>
<th>EXPERIMENT_NUMBER</th>
<th>EXPERIMENT_DESCRIPTION</th>
<th>EXPERIMENT_START_YEAR</th>
<th>EXPERIMENT_END_YEAR</th>
<th>EXPERIMENT_LONGITUDE</th>
<th>EXPERIMENT_LATITUDE</th>
<th>EXPERIMENT_REPORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Characterisation data of Lolium perenne</td>
<td>1999</td>
<td>2000</td>
<td>11.278414</td>
<td>51.826059</td>
<td>http://...</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Characterisation data of Lolium perenne</td>
<td>2000</td>
<td>11.278414</td>
<td>51.826059</td>
<td>http://...</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>Evaluation data of Lolium perenne (4 replications per accession)</td>
<td>2001</td>
<td>11.278414</td>
<td>51.826059</td>
<td>http://...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Uploadercode

<table>
<thead>
<tr>
<th>DATASET_NUMBER</th>
<th>DATASET_REMARK</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEU271</td>
<td>This dataset contains forage grass accessions.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

---

# FuturePlants

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)  
www.ipk-gatersleben.de
Proceeding for data upload

• Prerequisite:
  • Only non-confidential C&E data
  • Only data of accessions listed in EURISCO

• Impact
  • NFPs responsible for data upload (Data Sharing Agreements)
    ➔ May nominate users for (sub) accounts for data uploads
    ➔ NFPs must approve data before publication

• Data formatting
  • According to exchange format in MS Excel (.xlsx) files

• Upload via EURISCO intranet
Data upload in three steps

File parsing and upload via Java tool
- data owner

Data integrity checks
- EURISCO management

Approval / withdrawal of data for publishing on the EURISCO website
- data owner/NFP
Upload of phenotypic data files

**EURISCO uploader**

**First step: Upload C&E files**

The first step of importing new C&E data into EURISCO is to upload the filled template files to the EURISCO server. The files must be formatted in accordance with the EURISCO C&E data exchange format. The data must be contained in five separate MS-Excel (.xlsx) files.

Please use the Java WebStart application for uploading: Start the EURISCO C&E data importer.

The Java application will enable you to select the five template files. These files will then be parsed and the content will be uploaded into the EURISCO staging area. At the staging area, all necessary integrity checks will be performed. Afterwards, the results of the checks will be displayed in the EURISCO intranet again.

**Requirements:**
- The upload tool requires a Java runtime environment version 8 including Java Webstart.
- For the database access, the Oracle standard port 1521 needs to be enabled.

**Release 1.2.0**
Upload of phenotypic data files

- JRE 1.8
- Java WS
- Oracle standard port 1521 enabled
Integrity checks
Integrity checks

- Undefined trait number
Final decision
Next steps (background process)

• New dataset will be applied to EURISCO stage schema
  • Existing phenotypic data will **not** be overwritten
  • Existing phenotypic data may be removed on **request**

• EURISCO stage will be synchronised to the EURISCO web schema (time lag!)
  • Not in main business hours
  • Rebuild of materialised views
  • News message on EURISCO webpage
Dealing with phenotypic data: Data overview

- Extension available since 2016
- 2,682,962 records
- 90,627 accs. with phenotypic data
- 17 countries
- 69 phenotypic datasets
- 3,867 experiments
- 9,453 traits

- Increasingly accepted as repository, but limited comparability

as of 2021-10-29
Dealing with phenotypic data: Towards FAIR data

• Data harmonisation
  • Experiment set-up, treatment etc.
  • Reach MIAPPE-compliance (Krajewski et al. 2015)

• Better structuring
  • Traits/methods/scales
    • Development of common vocabularies/approaches
    • Improve comparability
      • Mapping onto ontology terms
      • Ontology of choice: Crop Ontology (Arnaud et al. 2012)
      • Crux: Sustainability of ontologies

• Provide training + helpdesk

• Additional activities together with various partners, e.g. AGENT or ECPGR-EVA
AGENT/EVA as a blueprint

- Current limitations
  - EURISCO data exchange format represents a “minimum consensus”
  - Difficult to compile files manually
  - Very limited reproducibility and comparability
- AGENT/EVA approach
  - Simplification of data collection → one column per trait to support manual recording
  - Distinction in two types of data
    - Simplified format for historic data → available, but no dedicated importer yet
    - More sophisticated template for new data → under development