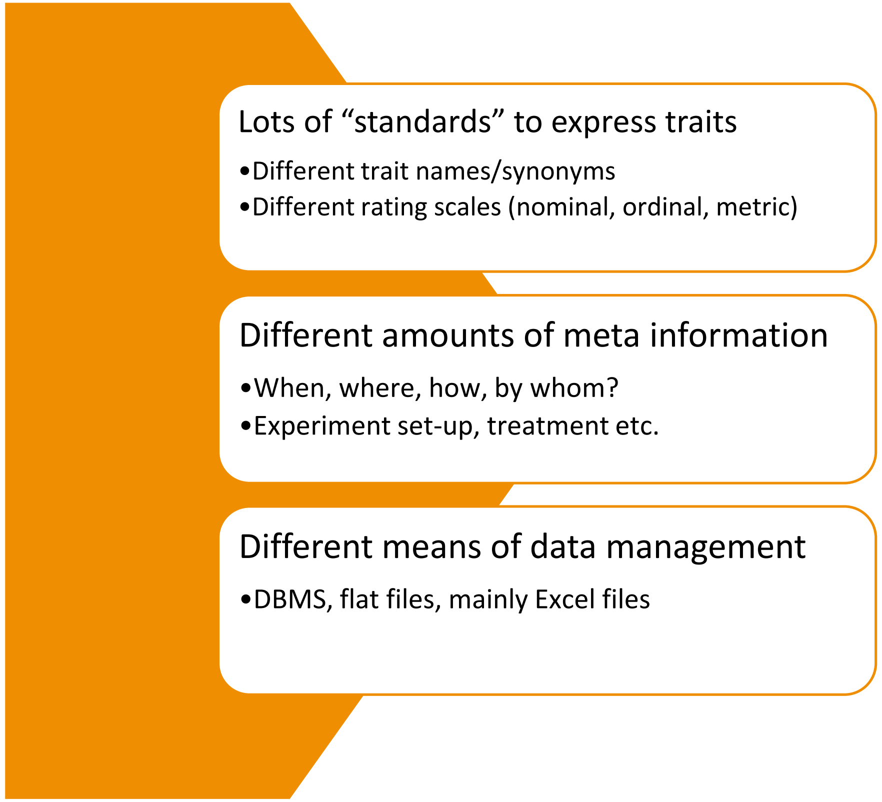
Stephan Weise

Phenotypic data in EURISCO

EURISCO training workshop 2023
12–14 September 2023, Plovdiv, Bulgaria

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

Methods and Descriptors

- Crop-specific definitions of traits, methods etc. like IPGRI descriptor lists
- Often used in parts only and adapted to organisational needs

Exchange Formats

- E.g. Darwin Core germplasm extension (DwC-germplasm; Endresen et al. 2009)
- Great for computer scientists
- Difficult to handle for genebank curators

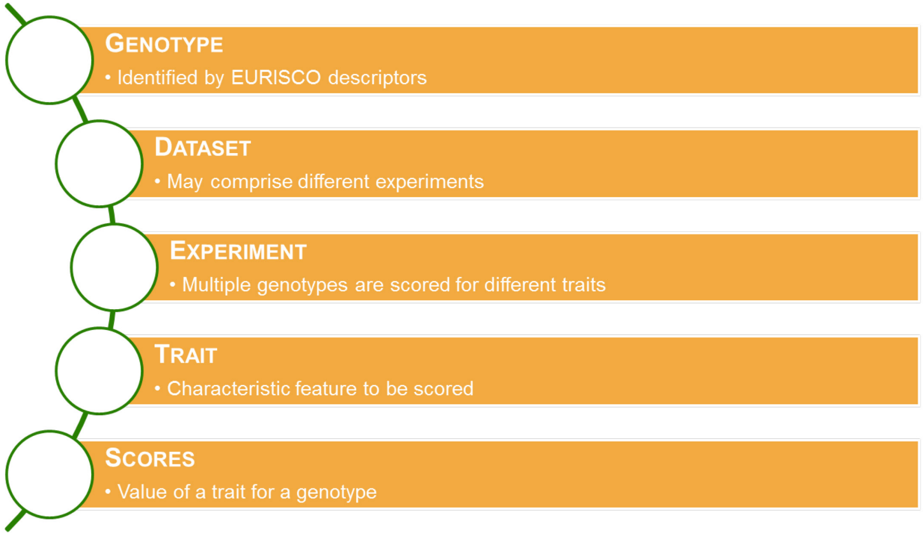
Ontologies

- Help to structure the (phenotypic) world
- Improve interoperability of data
- e.g. Crop Ontology (Arnaud et al. 2012)

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

  → "minimum consensus"
- Data management
  - Highly abstracted, following the single-observation concept (van Hintum et al. 1992)
  - Omitting fine-grained metadata



GENOTYPE
- Identified by EURISCO descriptors

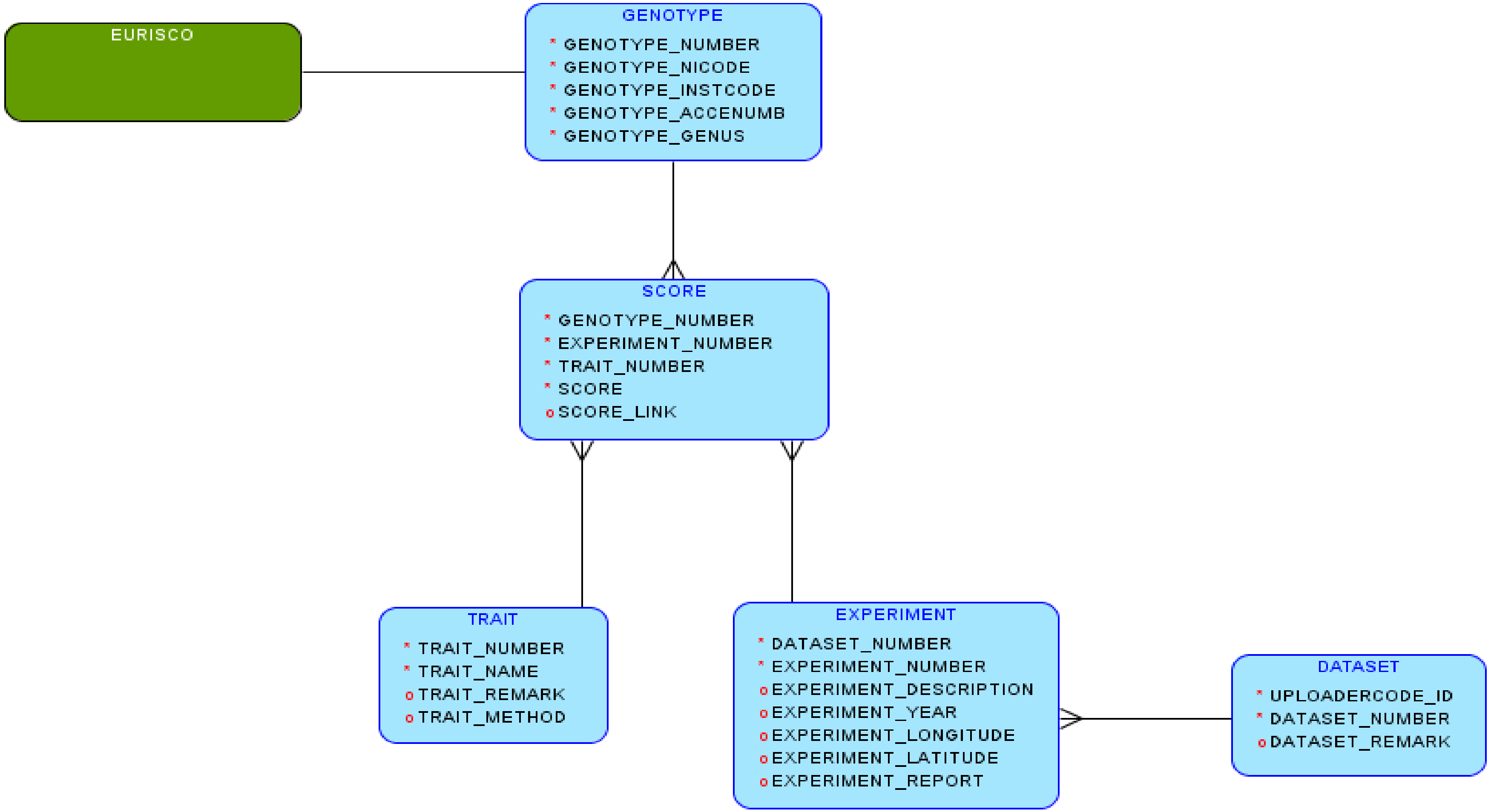
DATASET
- May comprise different experiments

EXPERIMENT
- Multiple genotypes are scored for different traits

TRAIT
- Characteristic feature to be scored

SCORES
- Value of a trait for a genotype

Data model for phenotypic data



EURISCO

GENOTYPE

* GENOTYPE_NUMBER
* GENOTYPE_NICODE
* GENOTYPE_INSTCODE
* GENOTYPE_ACCENUMB
* GENOTYPE_GENUS

SCORE

* GENOTYPE_NUMBER
* EXPERIMENT_NUMBER
* TRAIT_NUMBER
* SCORE

o SCORE_LINK

TRAIT

* TRAIT_NUMBER
* TRAIT_NAME

o TRAIT_REMARK
o TRAIT_METHOD

EXPERIMENT

* DATASET_NUMBER
* EXPERIMENT_NUMBER

o EXPERIMENT_DESCRIPTION
o EXPERIMENT_YEAR
o EXPERIMENT_LONGITUDE
o EXPERIMENT_LATITUDE
o EXPERIMENT_REPORT

DATASET

* UPLOADERCODE_ID
* DATASET_NUMBER

o DATASET_REMARK

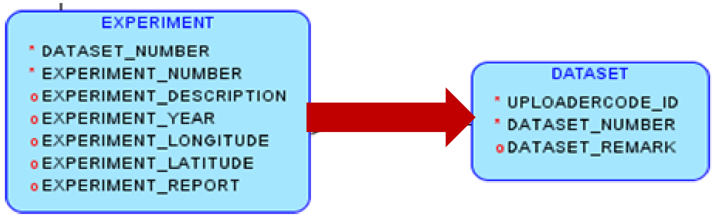
Dataset

- 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

UPLOADERCODE	DATASET_NUMBER	DATASET_REMARK
DEU271	1	This dataset contains forage grass accessions.
…	…	…
…	…	…

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

* DATASET_NUMBER
* EXPERIMENT_NUMBER
o EXPERIMENT_DESCRIPTION
o EXPERIMENT_YEAR
o EXPERIMENT_LONGITUDE
o EXPERIMENT_LATITUDE
o EXPERIMENT_REPORT

DATASET

* UPLOADERCODE_ID
* DATASET_NUMBER
o DATASET_REMARK

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

Experiment

DATASET_NUMBER	EXPERIMENT_NUMBER	EXPERIMENT_DESCRIPTION	EXPERIMENT_START_YEAR	EXPERIMENT_END_YEAR	EXPERIMENT_LONGITUDE	EXPERIMENT_LATITUDE	EXPERIMENT_REPORT
1	1	Characterisation data of Lolium perenne	1999	2000	11.278414	51.826059	http://...
1	2	Characterisation data of Lolium perenne	2000		11.278414	51.826059	http://...
1	3	Characterisation data of Lolium perenne	2001		11.278414	51.826059	http://...
1	4	Evaluation data of Lolium perenne (4 replications per accession)	2002		11.278414	51.826059	http://...
...	...	...	...		...	...	...

Trait

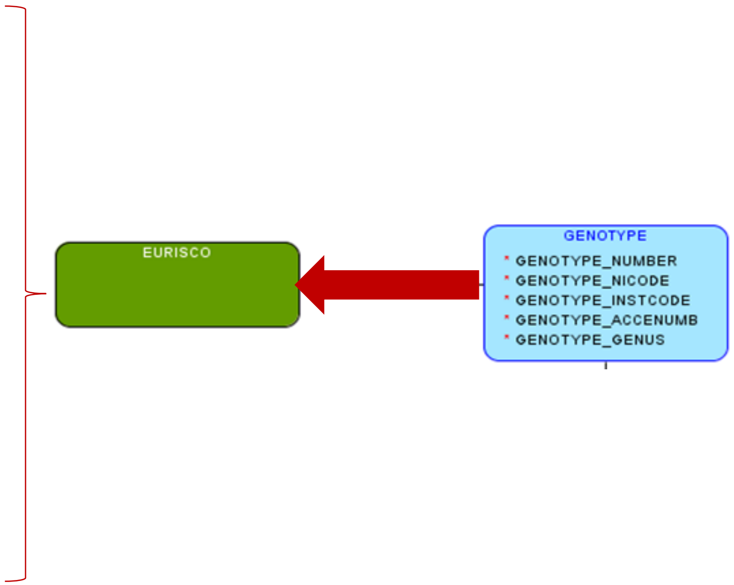
- Describe phenotypic traits and the methods used for scoring
- Fields:
  - TRAIT_NUMBER*:
    - Unique, temporary number of the trait in the dataset
  - TRAIT_NAME*:
    - English name of the trait
  - TRAIT_REMARK:
    - General remark helping to interpret the trait
  - TRAIT_METHOD:
    - English description of the used method + scale

Trait

TRAIT_NUMBER	TRAIT_NAME	TRAIT_REMARK	TRAIT_METHOD
1	Sowing date	…	Date
2	Emerging date	…	Date
3	Growing before winter	…	Rating value from 1 (min) – 9 (max)
4	Stem height min	In flowering time, the shortest plant	Measurement [cm]
…	…	…	…

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

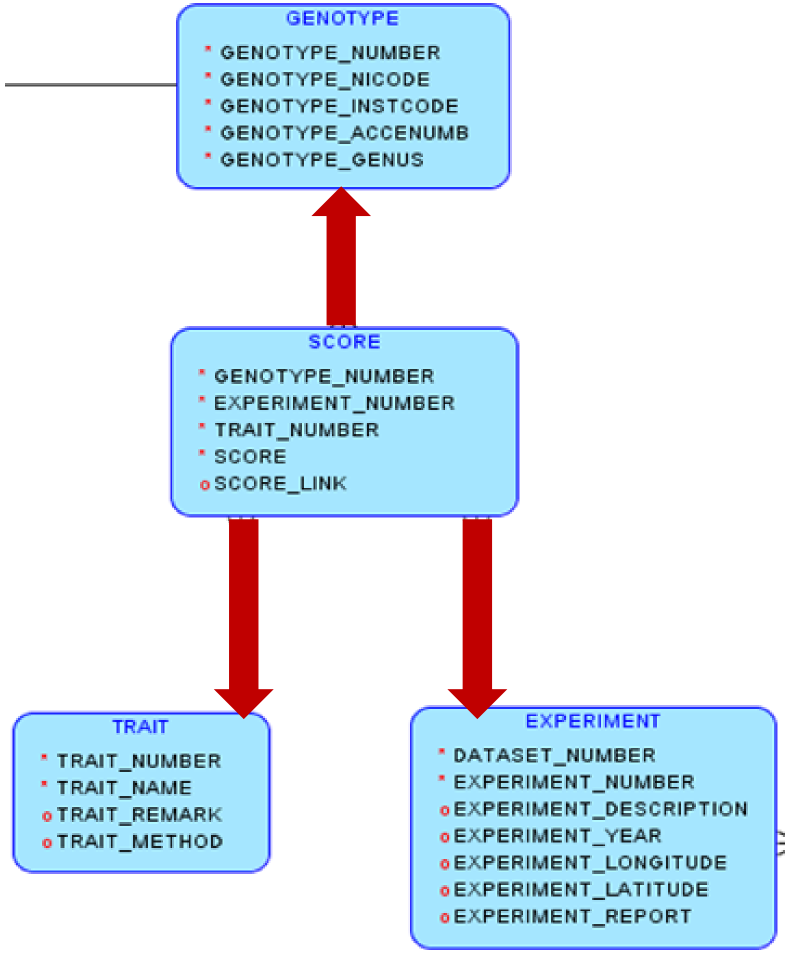


Genotype

GENOTYPE_NUMBER	GENOTYPE_NICODE	GENOTYPE_INSTCODE	GENOTYPE_ACCENUMB	GENOTYPE_GENUS	GENOTYPE_PUID
1	DEU	DEU271	GR 142	Lolium	
2	DEU	DEU271	GR 476	Lolium	
3	DEU	DEU271	GR 550	Lolium	
4	DEU	DEU271	GR 2670	Lolium	

Score

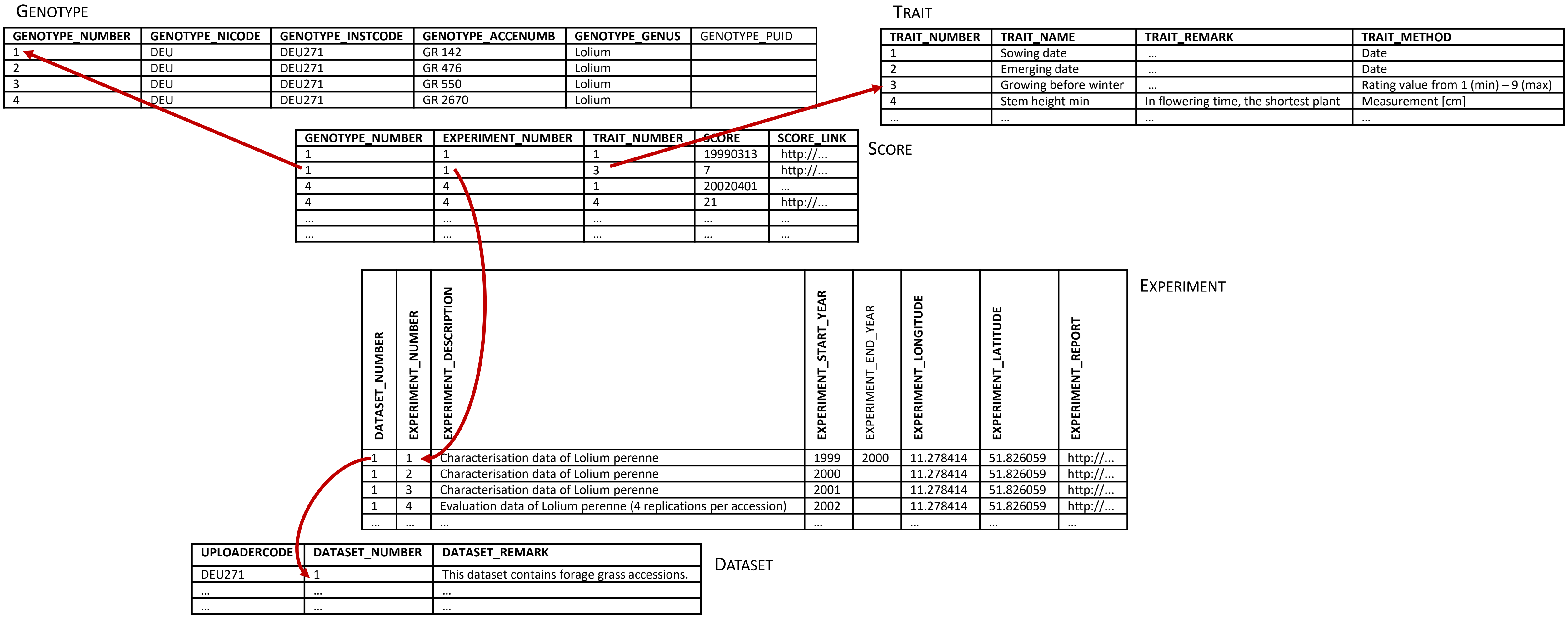
- Observed phenotypic values of the accessions
- Fields:
  - GENOTYPE_NUMBER*:
    - Reference to a genotype
  - EXPERIMENT_NUMBER*:
    - Reference to an experiment
  - TRAIT_NUMBER*:
    - Reference to a trait
  - SCORE*:
    - Observed score
  - SCORE_LINK:
    - Link to a publication on accession level



Score

GENOTYPE_NUMBER	EXPERIMENT_NUMBER	TRAIT_NUMBER	SCORE	SCORE_LINK
1	1	1	19990313	http://...
1	1	3	7	http://...
4	4	1	20020401	…
4	4	4	21	http://...
…	…	…	…	…
…	…	…	…	…

Connecting the templates



GENOTYPE

GENOTYPE_NUMBER	GENOTYPE_NICODE	GENOTYPE_INSTCODE	GENOTYPE_ACCENUMB	GENOTYPE_GENUS	GENOTYPE_PUID
1	DEU	DEU271	GR 142	Lolium	
2	DEU	DEU271	GR 476	Lolium	
3	DEU	DEU271	GR 550	Lolium	
4	DEU	DEU271	GR 2670	Lolium	

TRAIT

TRAIT_NUMBER	TRAIT_NAME	TRAIT_REMARK	TRAIT_METHOD
1	Sowing date	…	Date
2	Emerging date	…	Date
3	Growing before winter	…	Rating value from 1 (min) – 9 (max)
4	Stem height min	In flowering time, the shortest plant	Measurement [cm]
…	…	…	…

SCORE

GENOTYPE_NUMBER	EXPERIMENT_NUMBER	TRAIT_NUMBER	SCORE	SCORE_LINK
1	1	1	19990313	http://...
1	1	3	7	http://...
4	4	1	20020401	…
4	4	4	21	http://...
…	…	…	…	…
…	…	…	…	…

EXPERIMENT

DATASET_NUMBER	EXPERIMENT_NUMBER	EXPERIMENT_DESCRIPTION	EXPERIMENT_START_YEAR	EXPERIMENT_END_YEAR	EXPERIMENT_LONGITUDE	EXPERIMENT_LATITUDE	EXPERIMENT_REPORT
1	1	Characterisation data of Lolium perenne	1999	2000	11.278414	51.826059	http://...
1	2	Characterisation data of Lolium perenne	2000		11.278414	51.826059	http://...
1	3	Characterisation data of Lolium perenne	2001		11.278414	51.826059	http://...
1	4	Evaluation data of Lolium perenne (4 replications per accession)	2002		11.278414	51.826059	http://...
…	…	…	…		…	…	…

DATASET

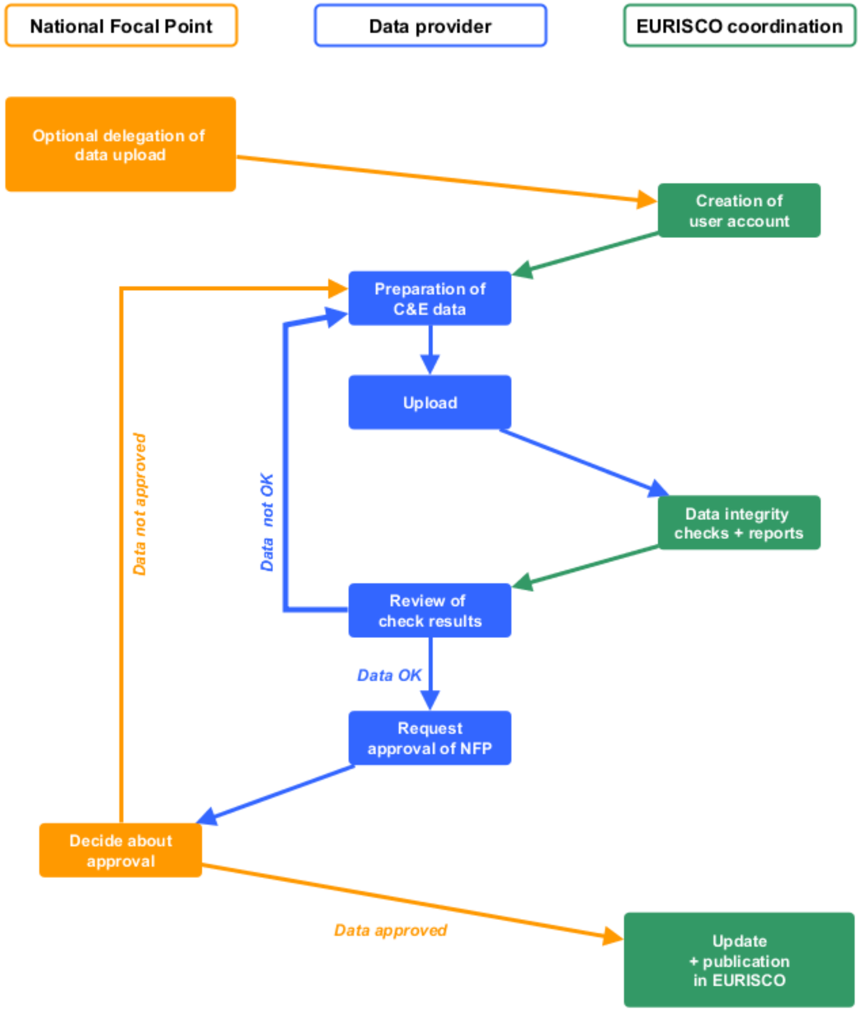
UPLOADERCODE	DATASET_NUMBER	DATASET_REMARK
DEU271	1	This dataset contains forage grass accessions.
…	…	…
…	…	…

Proceeding for data upload

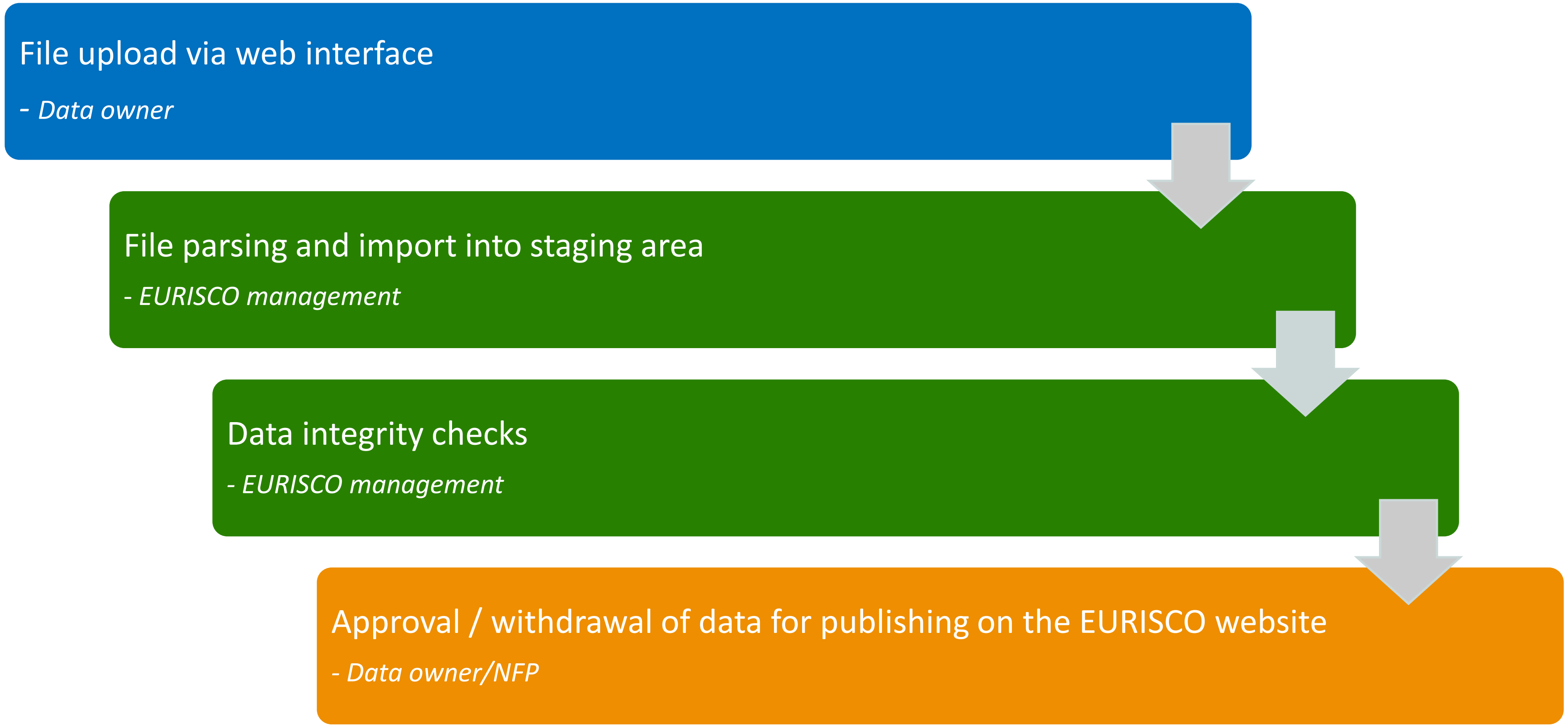
- Prerequisite:
  - Only non-confidential C&E data
  - Only data of accessions listed in EURISCO
- Impact
  - NFPs responsible for data upload (Data Sharing Agreements)

    → 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 four steps



File upload via web interface

- *Data owner*

File parsing and import into staging area

- *EURISCO management*

Data integrity checks

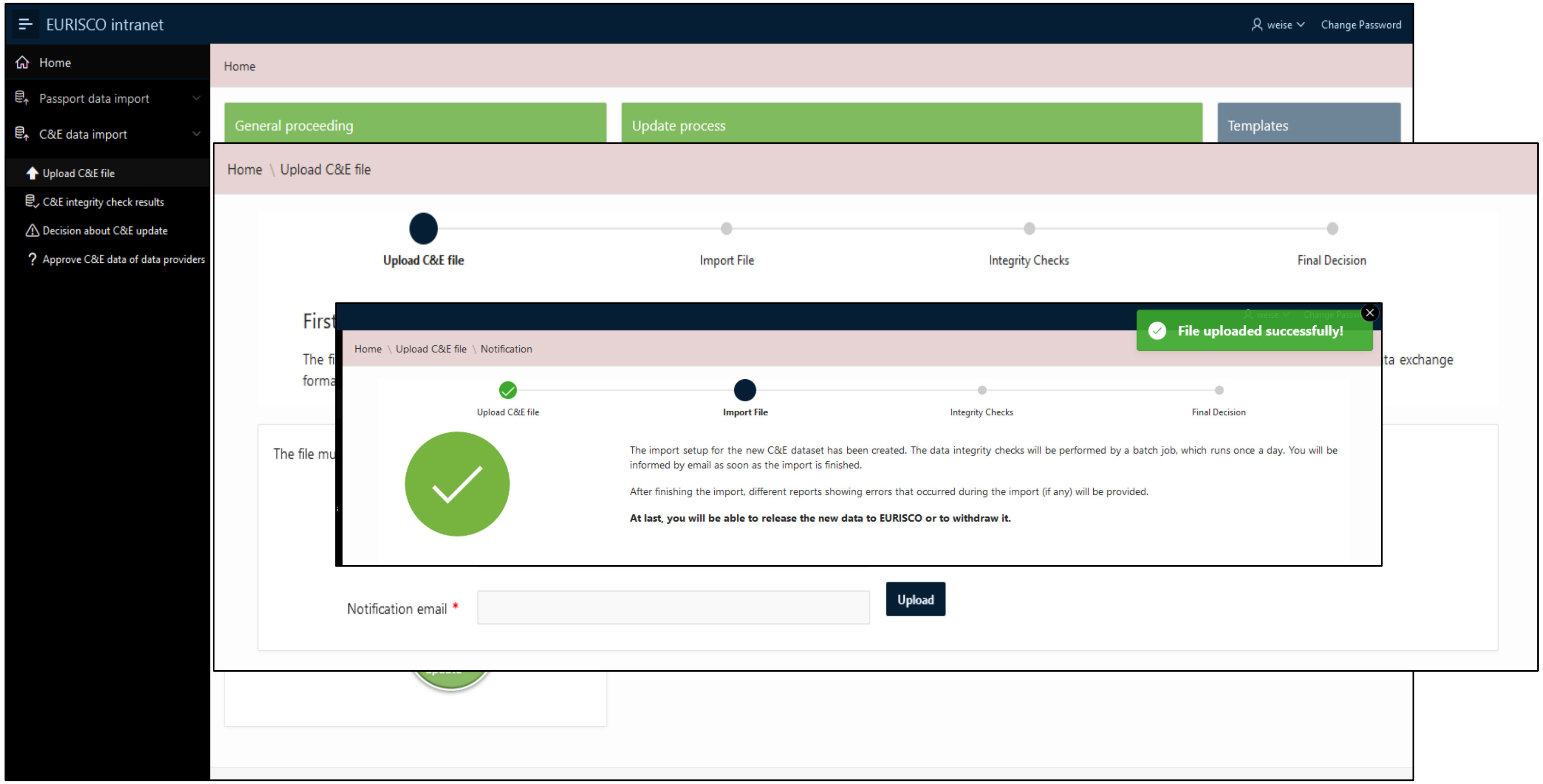
- *EURISCO management*

Approval / withdrawal of data for publishing on the EURISCO website

- *Data owner/NFP*

Upload of phenotypic data files

- Excel template
    - .xlsx format
    - Five sheets:
        - DATASET
        - EXPERIMENT
        - GENOTYPE
        - TRAIT
        - SCORE



File uploaded successfully!

The import setup for the new C&E dataset has been created. The data integrity checks will be performed by a batch job, which runs once a day. You will be informed by email as soon as the import is finished.

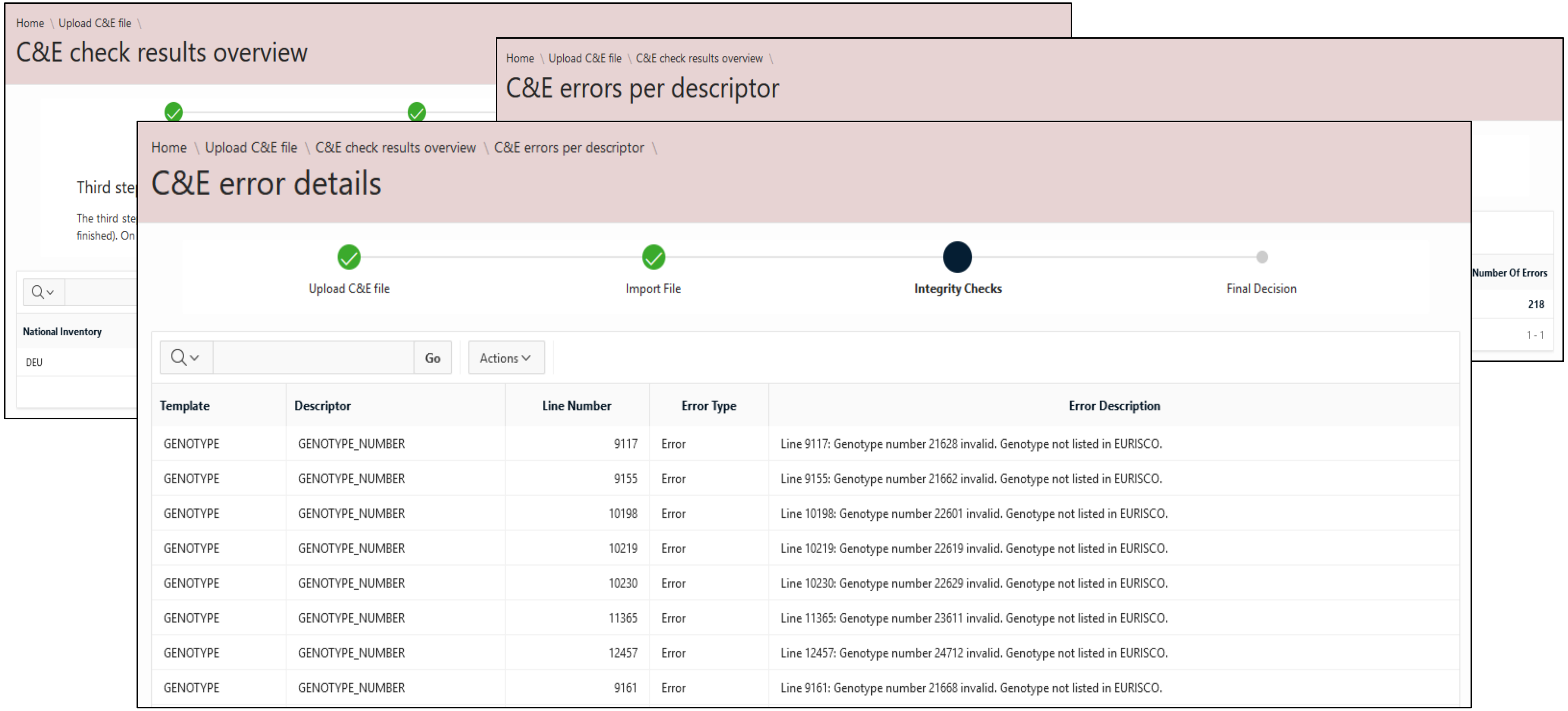
After finishing the import, different reports showing errors that occurred during the import (if any) will be provided.

At last, you will be able to release the new data to EURISCO or to withdraw it.

Next steps (background process)

- Parsing of Excel file
  - Data temporarily written into staging area
- Data integrity checks
  - Error logs written
  - Error reports generated
- Data provider will be informed by email

Review integrity check results



Home \ Upload C&E file \

C&E check results overview

Third ste

The third ste finished). On

National Inventory

DEU

Home \ Upload C&E file \ C&E check results overview \

C&E errors per descriptor

Number Of Errors

218

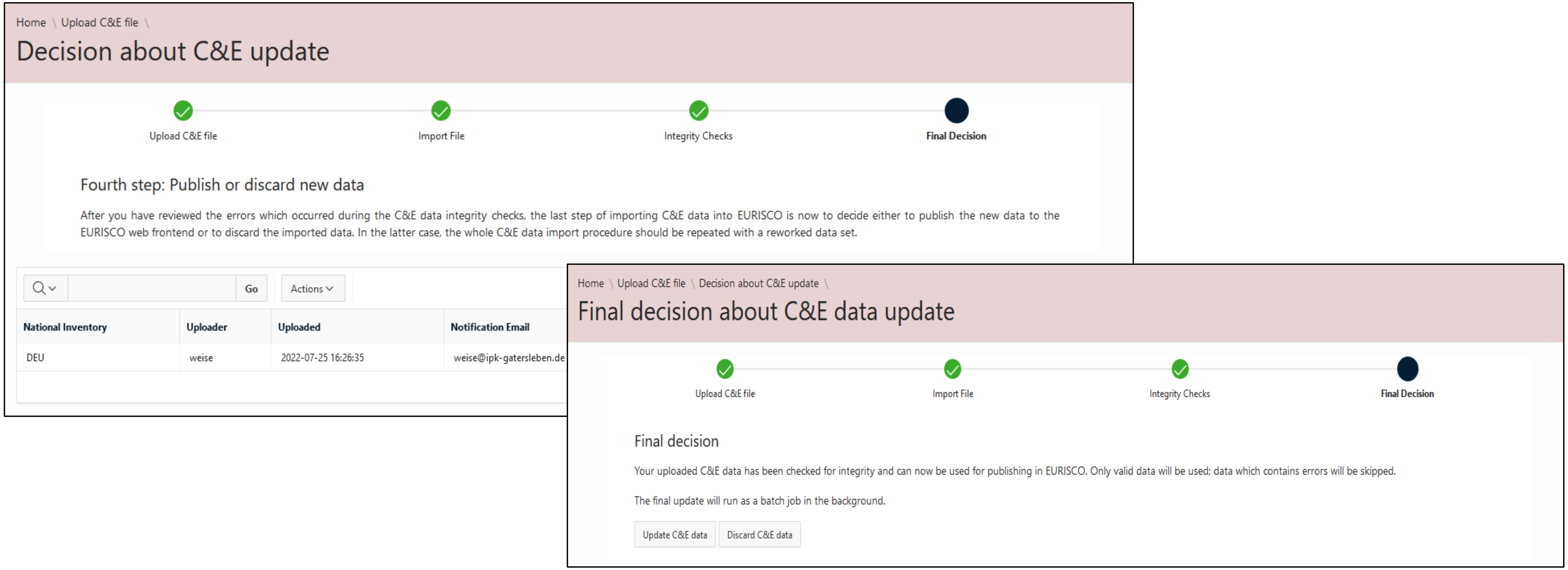
1 - 1

Home \ Upload C&E file \ C&E check results overview \ C&E errors per descriptor \

C&E error details

Template	Descriptor	Line Number	Error Type	Error Description
GENOTYPE	GENOTYPE_NUMBER	9117	Error	Line 9117: Genotype number 21628 invalid. Genotype not listed in EURISCO.
GENOTYPE	GENOTYPE_NUMBER	9155	Error	Line 9155: Genotype number 21662 invalid. Genotype not listed in EURISCO.
GENOTYPE	GENOTYPE_NUMBER	10198	Error	Line 10198: Genotype number 22601 invalid. Genotype not listed in EURISCO.
GENOTYPE	GENOTYPE_NUMBER	10219	Error	Line 10219: Genotype number 22619 invalid. Genotype not listed in EURISCO.
GENOTYPE	GENOTYPE_NUMBER	10230	Error	Line 10230: Genotype number 22629 invalid. Genotype not listed in EURISCO.
GENOTYPE	GENOTYPE_NUMBER	11365	Error	Line 11365: Genotype number 23611 invalid. Genotype not listed in EURISCO.
GENOTYPE	GENOTYPE_NUMBER	12457	Error	Line 12457: Genotype number 24712 invalid. Genotype not listed in EURISCO.
GENOTYPE	GENOTYPE_NUMBER	9161	Error	Line 9161: Genotype number 21668 invalid. Genotype not listed in EURISCO.

Final decision



Home \ Upload C&E file \

Decision about C&E update

Fourth step: Publish or discard new data

After you have reviewed the errors which occurred during the C&E data integrity checks, the last step of importing C&E data into EURISCO is now to decide either to publish the new data to the EURISCO web frontend or to discard the imported data. In the latter case, the whole C&E data import procedure should be repeated with a reworked data set.

National Inventory	Uploader	Uploaded	Notification Email
DEU	weise	2022-07-25 16:26:35	weise@ipk-gatersleben.de

Home \ Upload C&E file \ Decision about C&E update \

Final decision about C&E data update

Final decision

Your uploaded C&E data has been checked for integrity and can now be used for publishing in EURISCO. Only valid data will be used; data which contains errors will be skipped.

The final update will run as a batch job in the background.

Next steps again (background process)

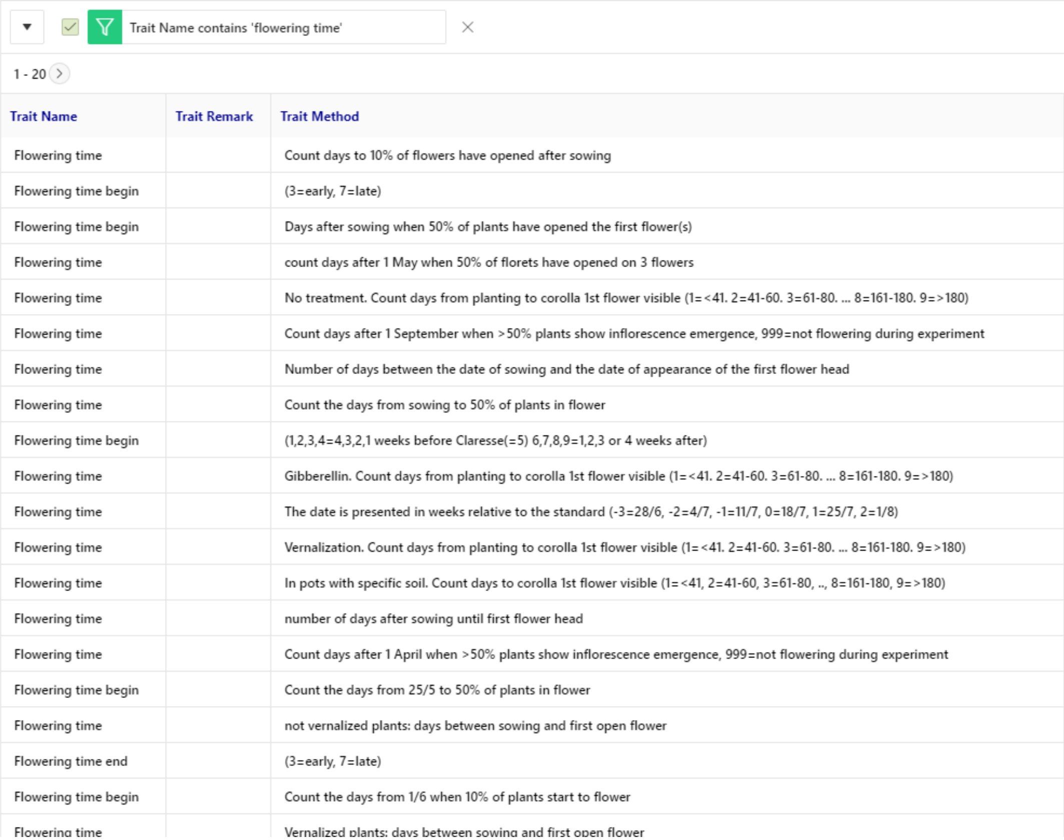
- New dataset will be applied to EURISCO stage schema
  - Existing phenotypic data will **not** be overwritten
  - Existing phenotypic data may be removed on **request**

- 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
- 2,726,998 records
- 91,366 accs. with phenotypic data
- 21 countries
- 73 phenotypic datasets
- 3,919 experiments
- 9,730 traits

- Increasingly accepted as repository, but limited comparability

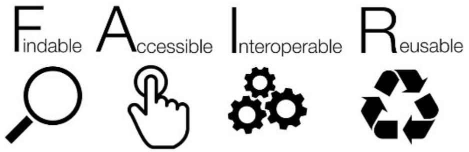


Trait Name	Trait Remark	Trait Method
Flowering time		Count days to 10% of flowers have opened after sowing
Flowering time begin		(3=early, 7=late)
Flowering time begin		Days after sowing when 50% of plants have opened the first flower(s)
Flowering time		count days after 1 May when 50% of florets have opened on 3 flowers
Flowering time		No treatment. Count days from planting to corolla 1st flower visible (1=<41. 2=41-60. 3=61-80. ... 8=161-180. 9=>180)
Flowering time		Count days after 1 September when >50% plants show inflorescence emergence, 999=not flowering during experiment
Flowering time		Number of days between the date of sowing and the date of appearance of the first flower head
Flowering time		Count the days from sowing to 50% of plants in flower
Flowering time begin		(1,2,3,4=4,3,2,1 weeks before Claresse(=5) 6,7,8,9=1,2,3 or 4 weeks after)
Flowering time		Gibberellin. Count days from planting to corolla 1st flower visible (1=<41. 2=41-60. 3=61-80. ... 8=161-180. 9=>180)
Flowering time		The date is presented in weeks relative to the standard (-3=28/6, -2=4/7, -1=11/7, 0=18/7, 1=25/7, 2=1/8)
Flowering time		Vernalization. Count days from planting to corolla 1st flower visible (1=<41. 2=41-60. 3=61-80. ... 8=161-180. 9=>180)
Flowering time		In pots with specific soil. Count days to corolla 1st flower visible (1=<41, 2=41-60, 3=61-80, .., 8=161-180, 9=>180)
Flowering time		number of days after sowing until first flower head
Flowering time		Count days after 1 April when >50% plants show inflorescence emergence, 999=not flowering during experiment
Flowering time begin		Count the days from 25/5 to 50% of plants in flower
Flowering time		not vernalized plants: days between sowing and first open flower
Flowering time end		(3=early, 7=late)
Flowering time begin		Count the days from 1/6 when 10% of plants start to flower
Flowering time		Vernalized plants: days between sowing and first open flower

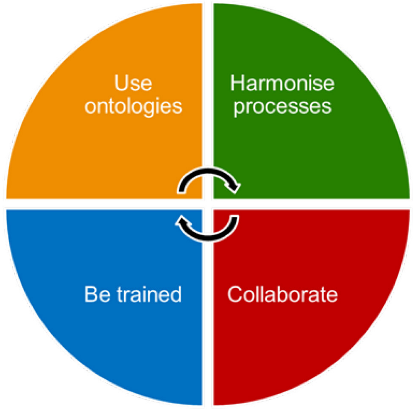
as of 2023-09-01

Dealing with phenotypic data: Towards FAIR data

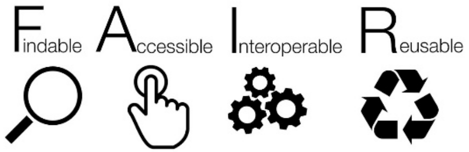
- Data harmonisation
  - Experiment set-up, treatment etc.
  - Reach MIAPPE-compliance (Krajewski et al. 2015)
- Better structuring
  - Traits/methods/scales
    - Development of common vocabularies/approaches
    - Improve comparability
      - Mapping onto ontology terms
      - Ontology of choice: 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



SangyaPundir



AGENT/EVA as a blueprint



SangyaPundir

- 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 evaluation in EVA/AGENT

see example Excel files