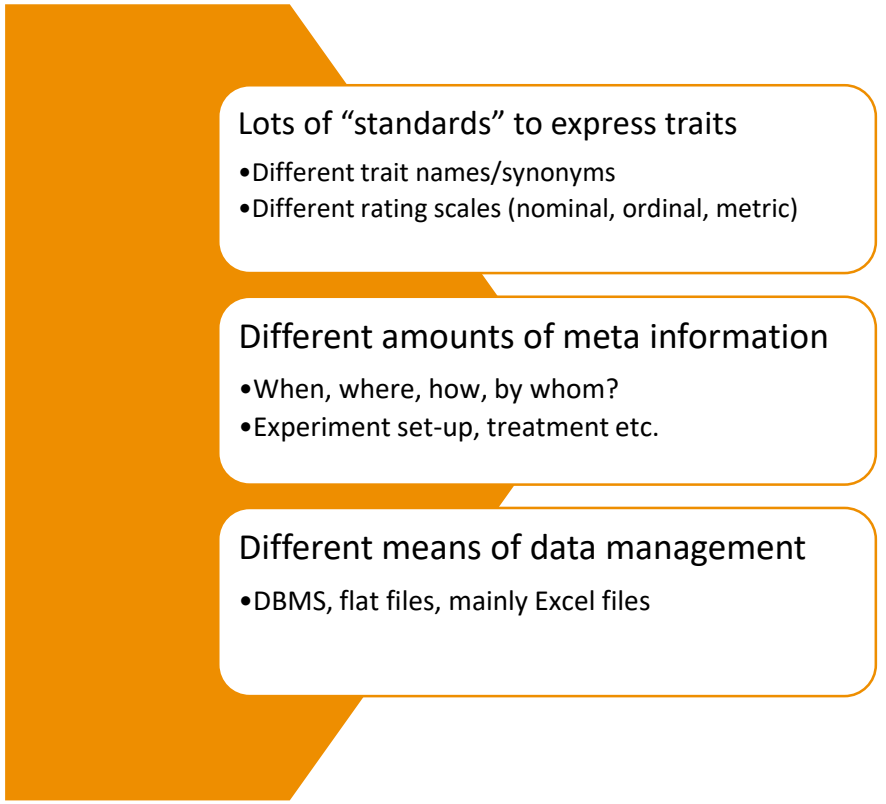
Stephan Weise

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



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

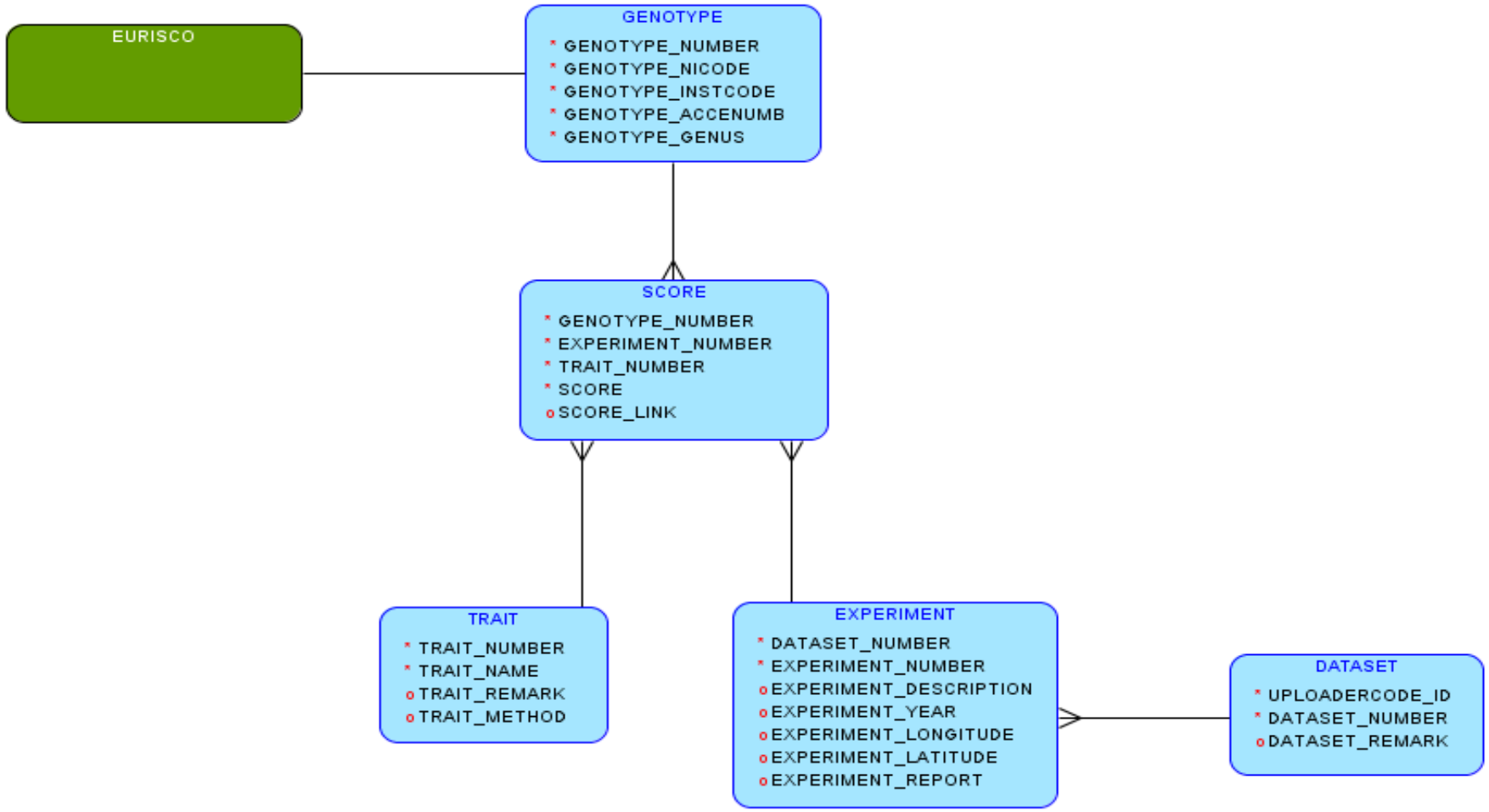
- Help to structure the (phenotypic) world
- Improve interoperability of data
- e.g. Crop Ontology (Arnaud et al. 2012)

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

Data model for phenotypic data



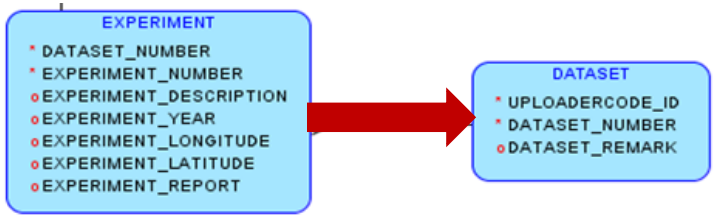
Dataset

- 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

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

- 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

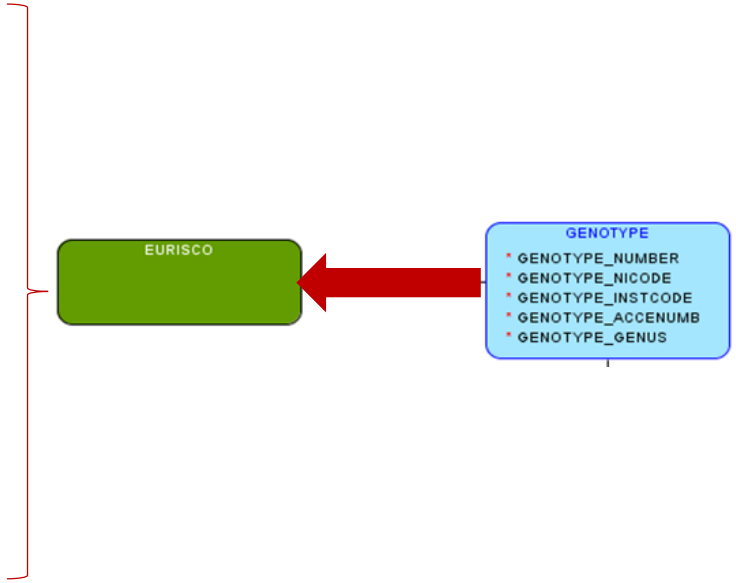
- Describe phenotypic traits and the methods used for scoring
- Fields:
  - TRAIT_NUMBER*:
    - Unique, temporary number of the trait in the dataset
  - TRAIT_NAME*:
    - English name of the trait
  - TRAIT_REMARK:
    - General remark helping to interpret the trait
  - TRAIT_METHOD:
    - English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

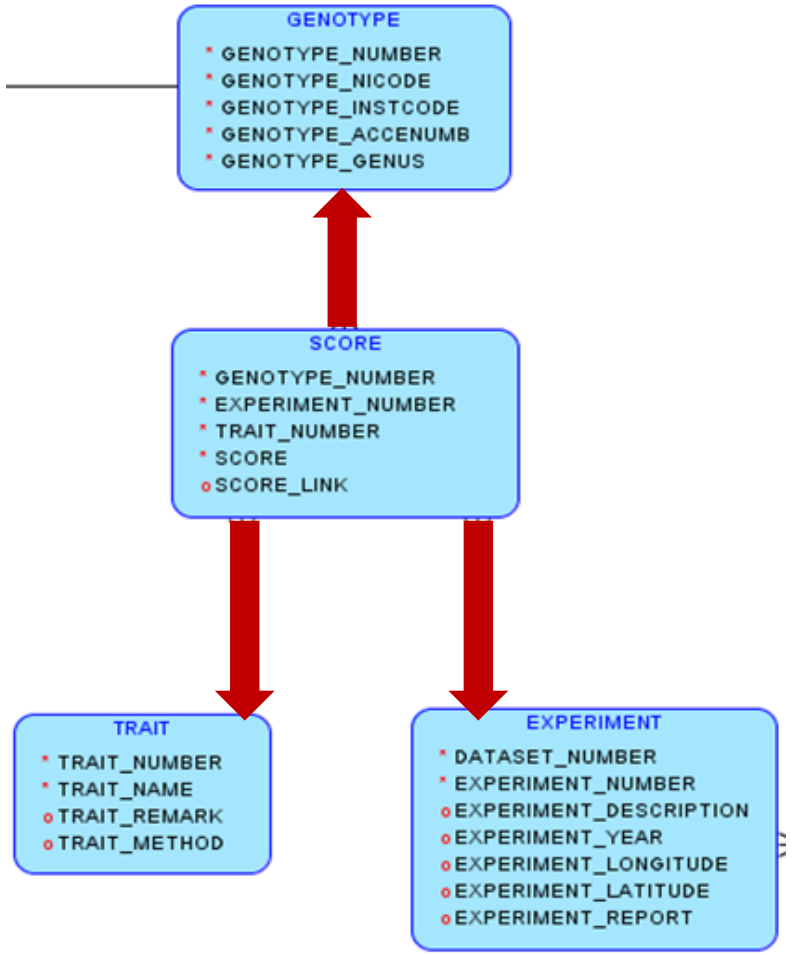


Genotype

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]
…	…	…	…

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://…
…	…	…	…	…
…	…	…	…	…

SCORE

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://…
…	…	…	…		…	…	…

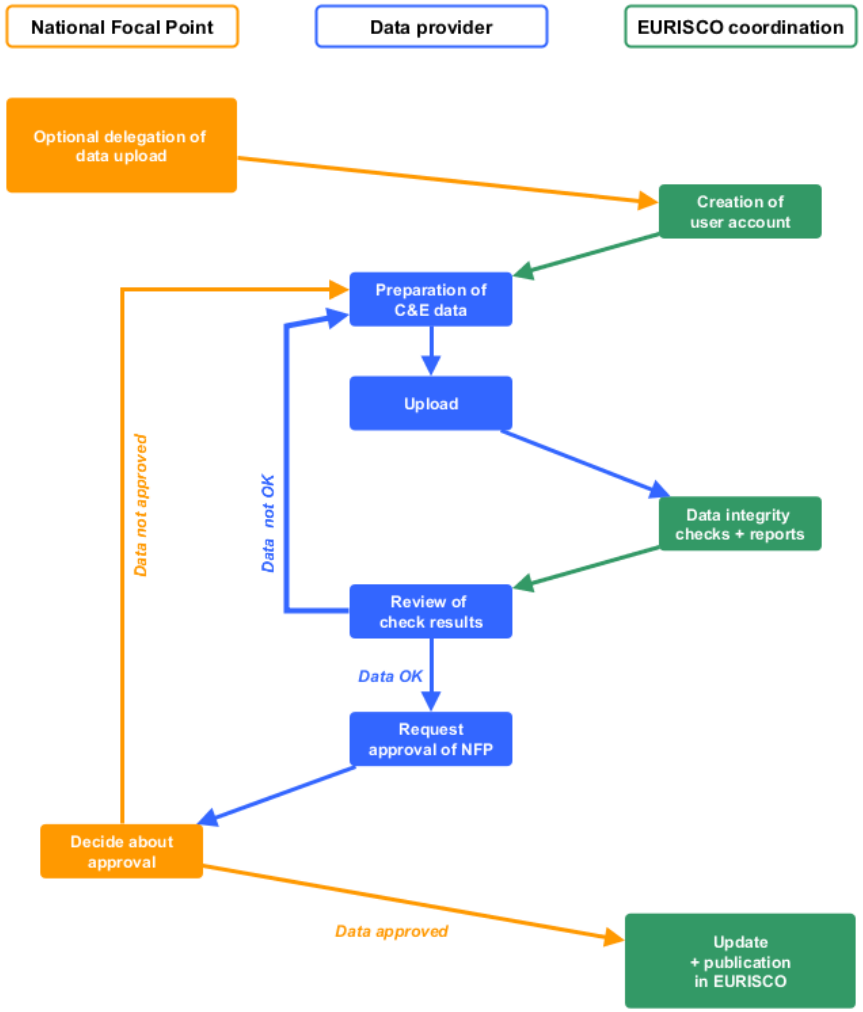
EXPERIMENT

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

DATASET

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

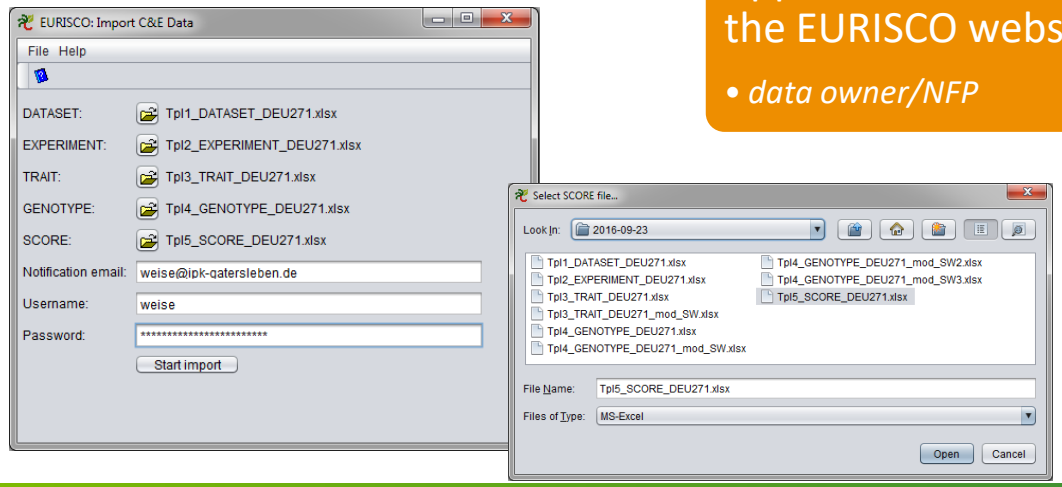
- *data owner*

Data integrity checks

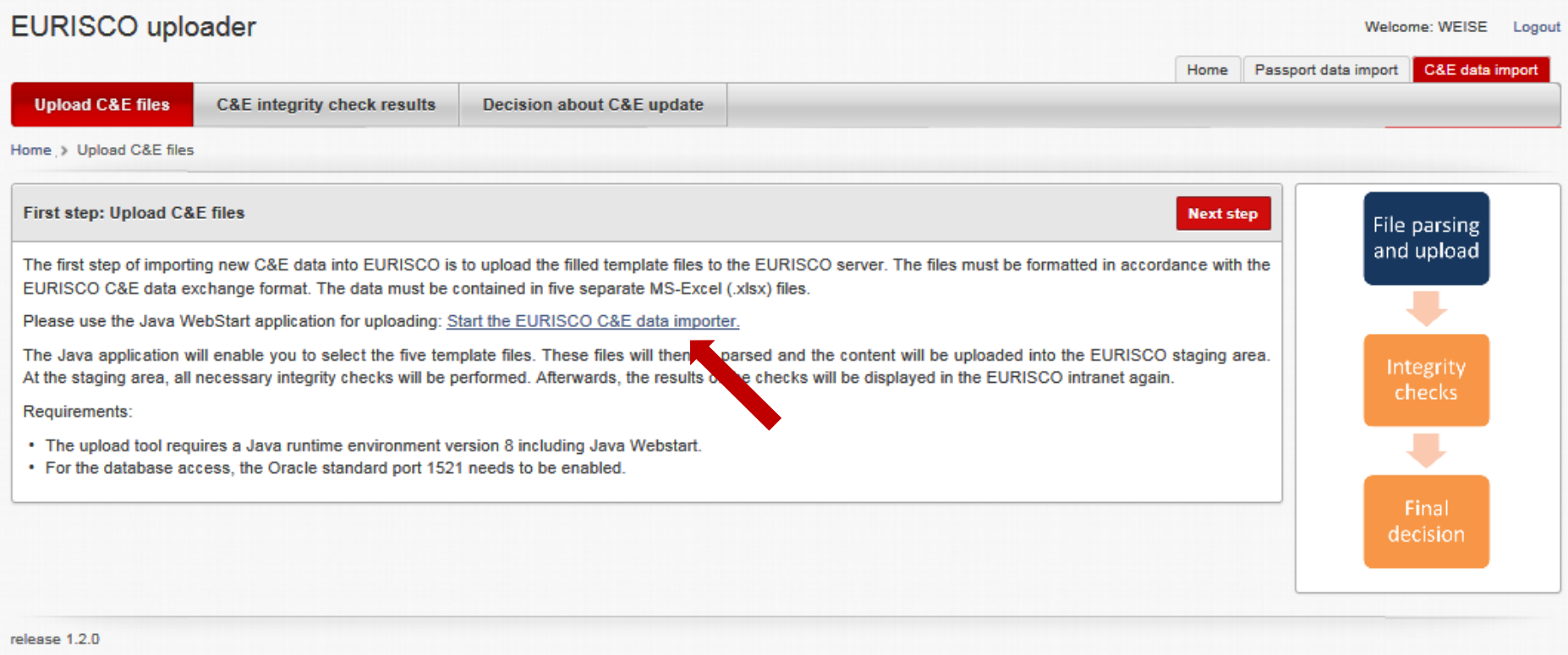
- *EURISCO management*

Approval / withdrawal of data for publishing on the EURISCO website

- *data owner/NFP*

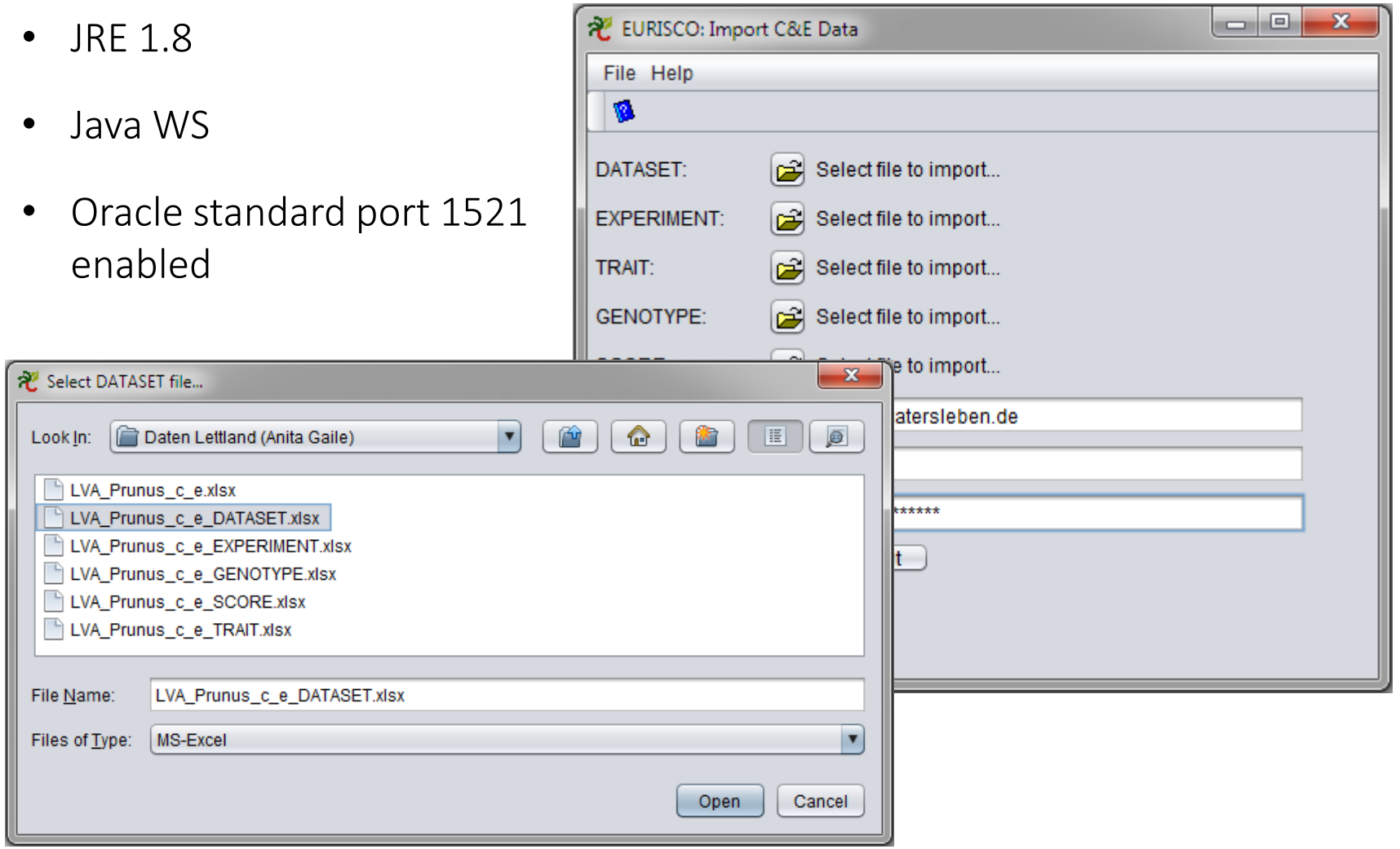


Upload of phenotypic data files

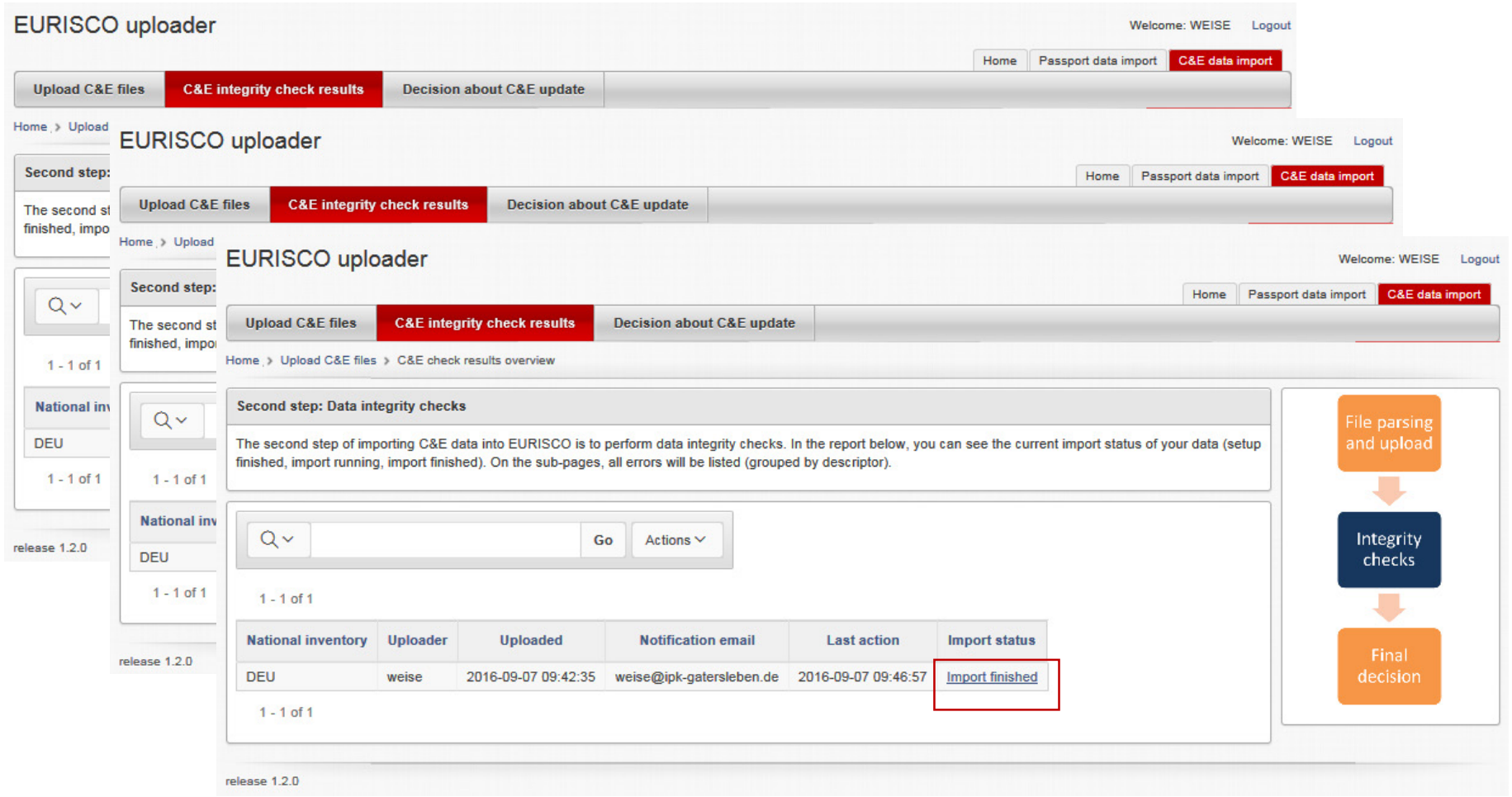


Upload of phenotypic data files

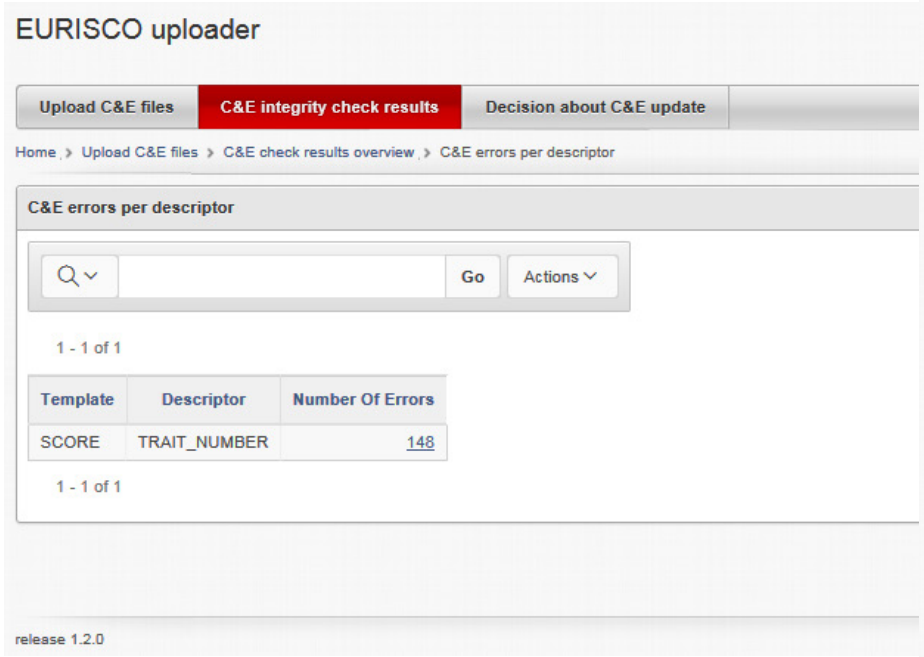
- JRE 1.8
- Java WS
- Oracle standard port 1521 enabled

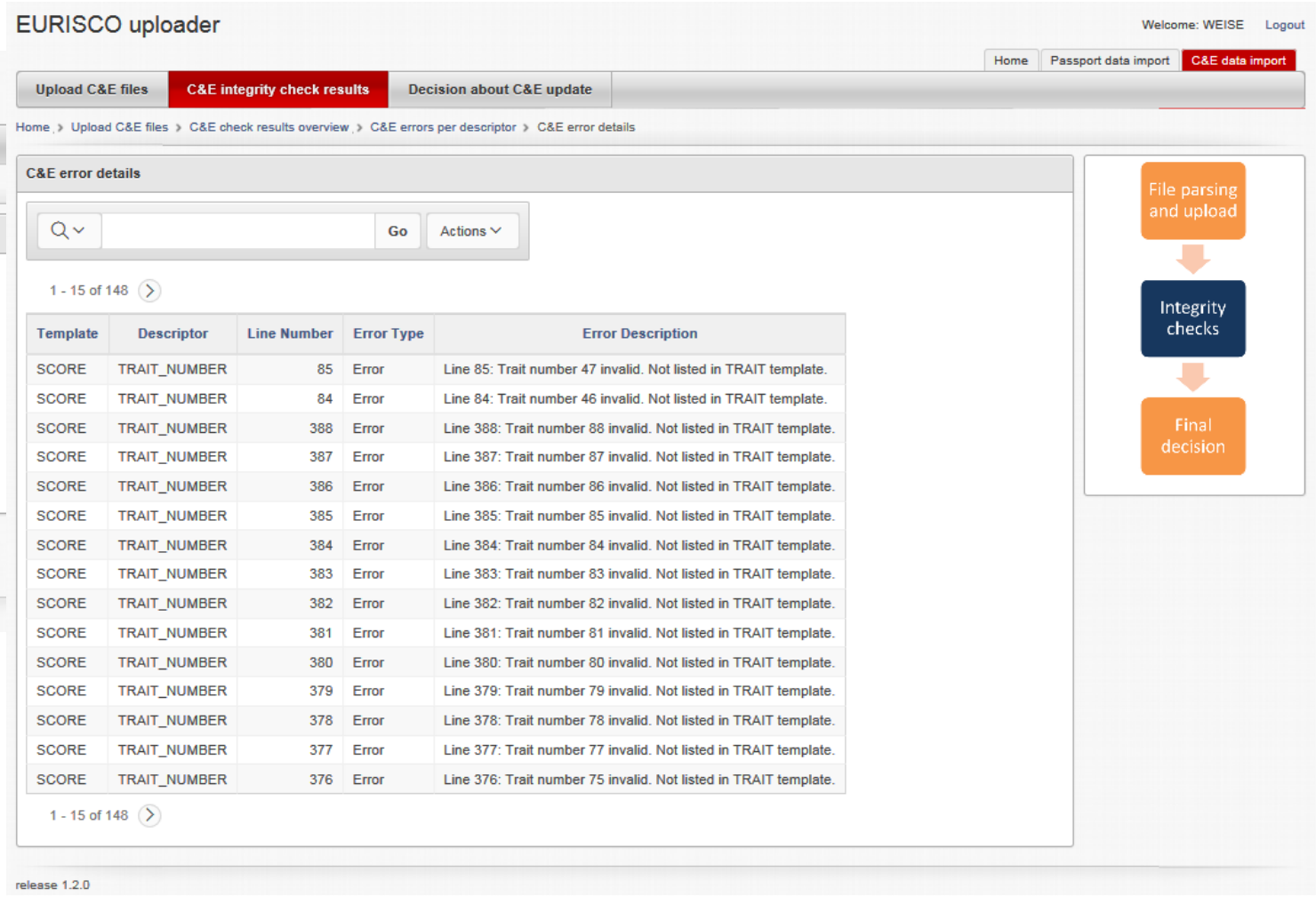


Integrity checks



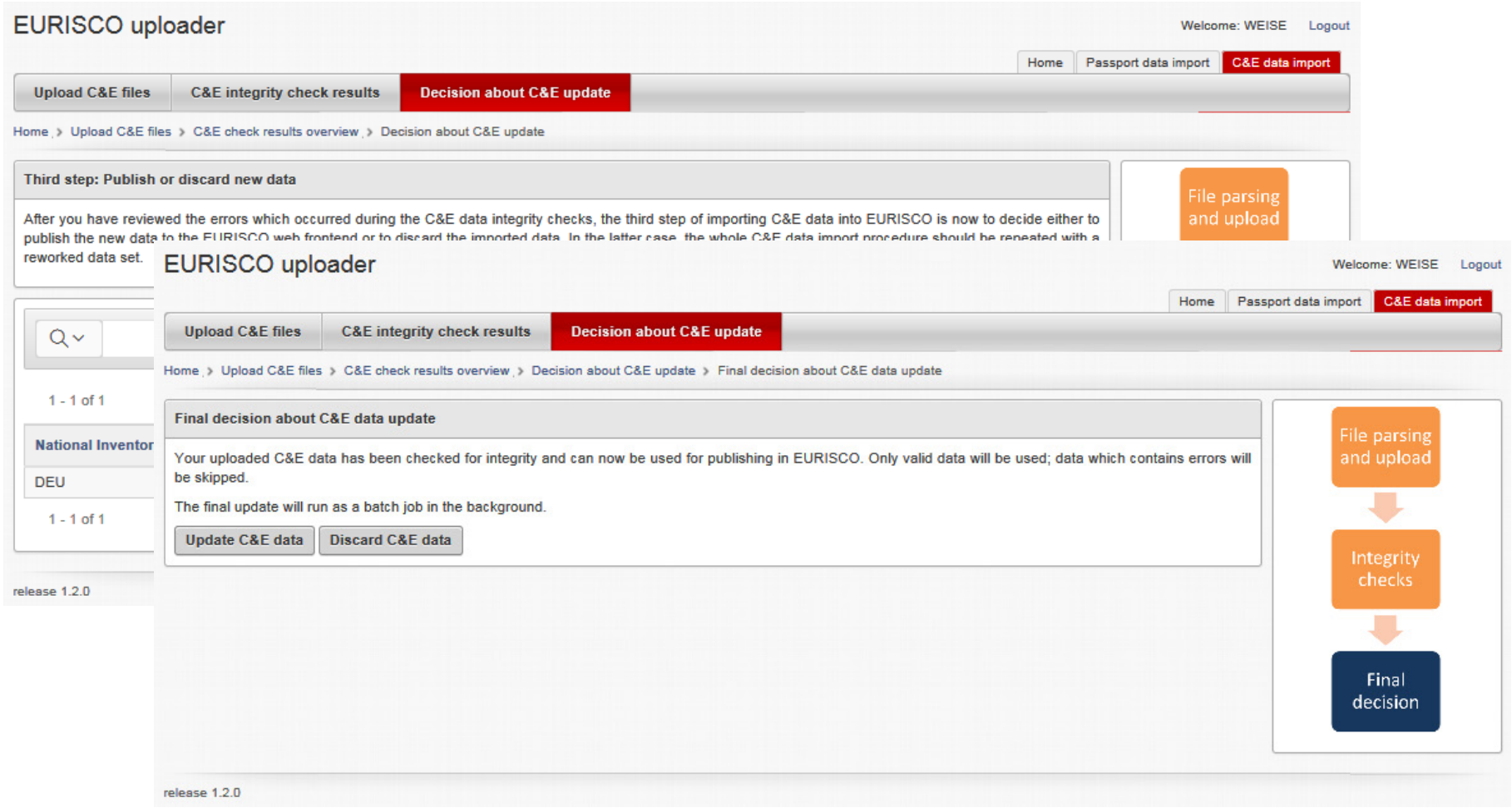
Integrity checks





- Undefined trait number

Final decision



Next steps (background process)

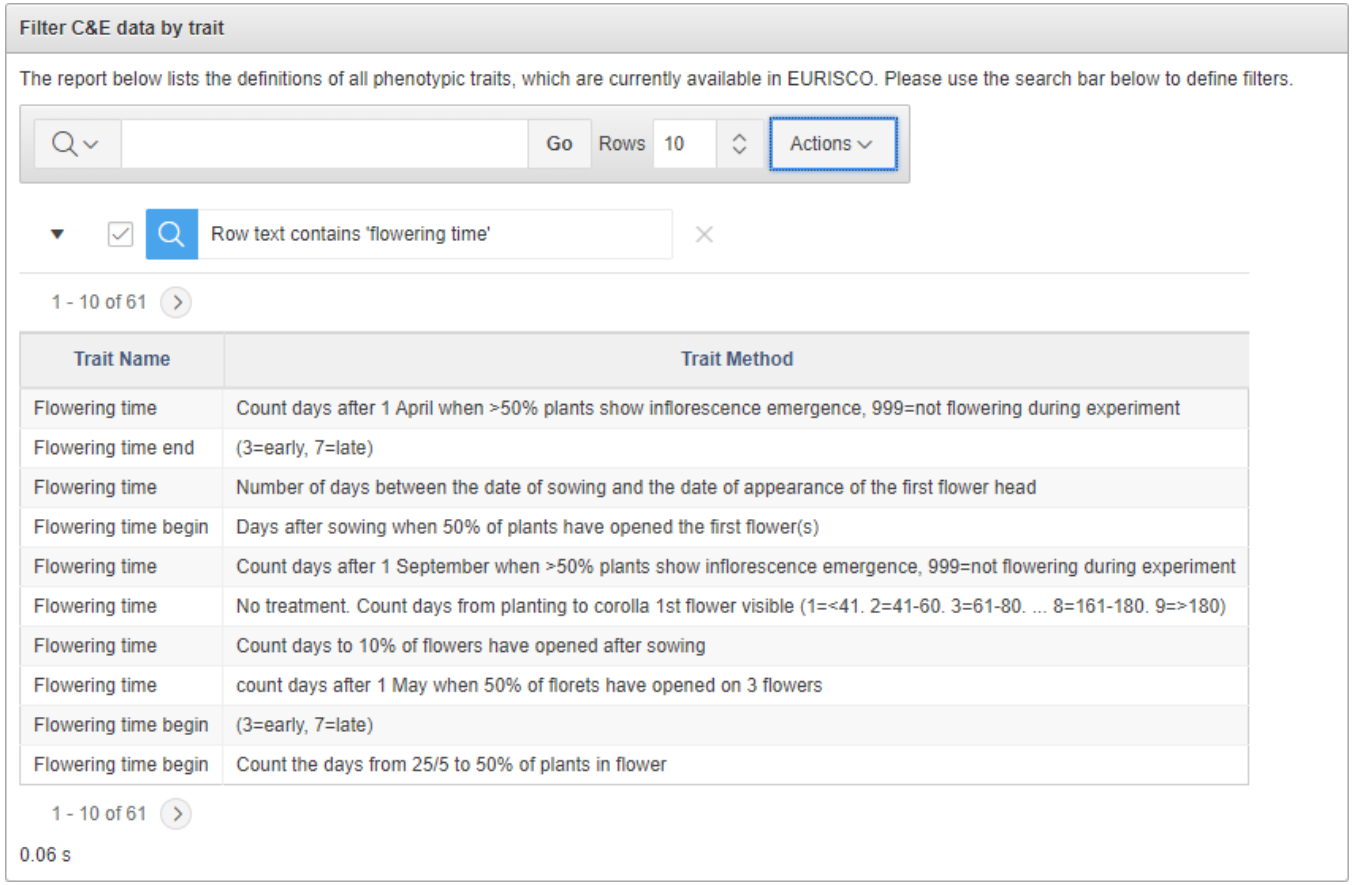
- New dataset will be applied to EURISCO stage schema
  - Existing phenotypic data will **not** be overwritten
  - Existing phenotypic data may be removed on **request**

- 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

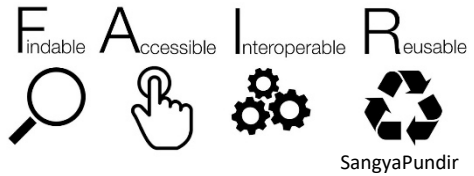
- Extension available since 2016
- 2,682,962 records
- 90,627 accs. with phenotypic data
- 17 countries
- 69 phenotypic datasets
- 3,867 experiments
- 9,453 traits

- Increasingly accepted as repository, but limited comparability



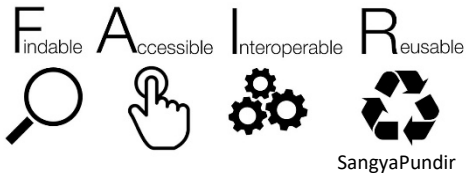
as of 2021-10-29

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