

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

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

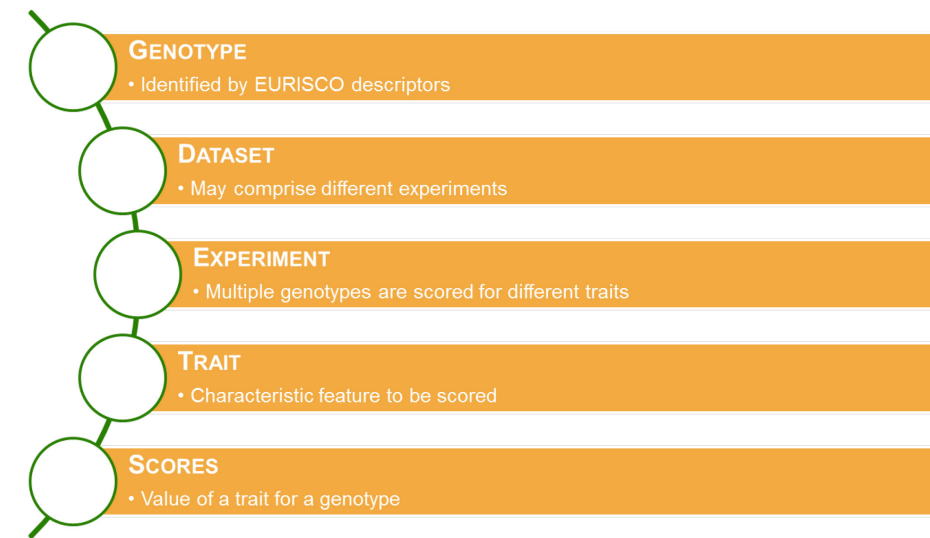
- E.g. Darwin Core germplasm extension (DwC-germplasm; Endresen et al. 2009)
- Great for computer scientists
- Difficult to handle for genebank curators

Ontologies

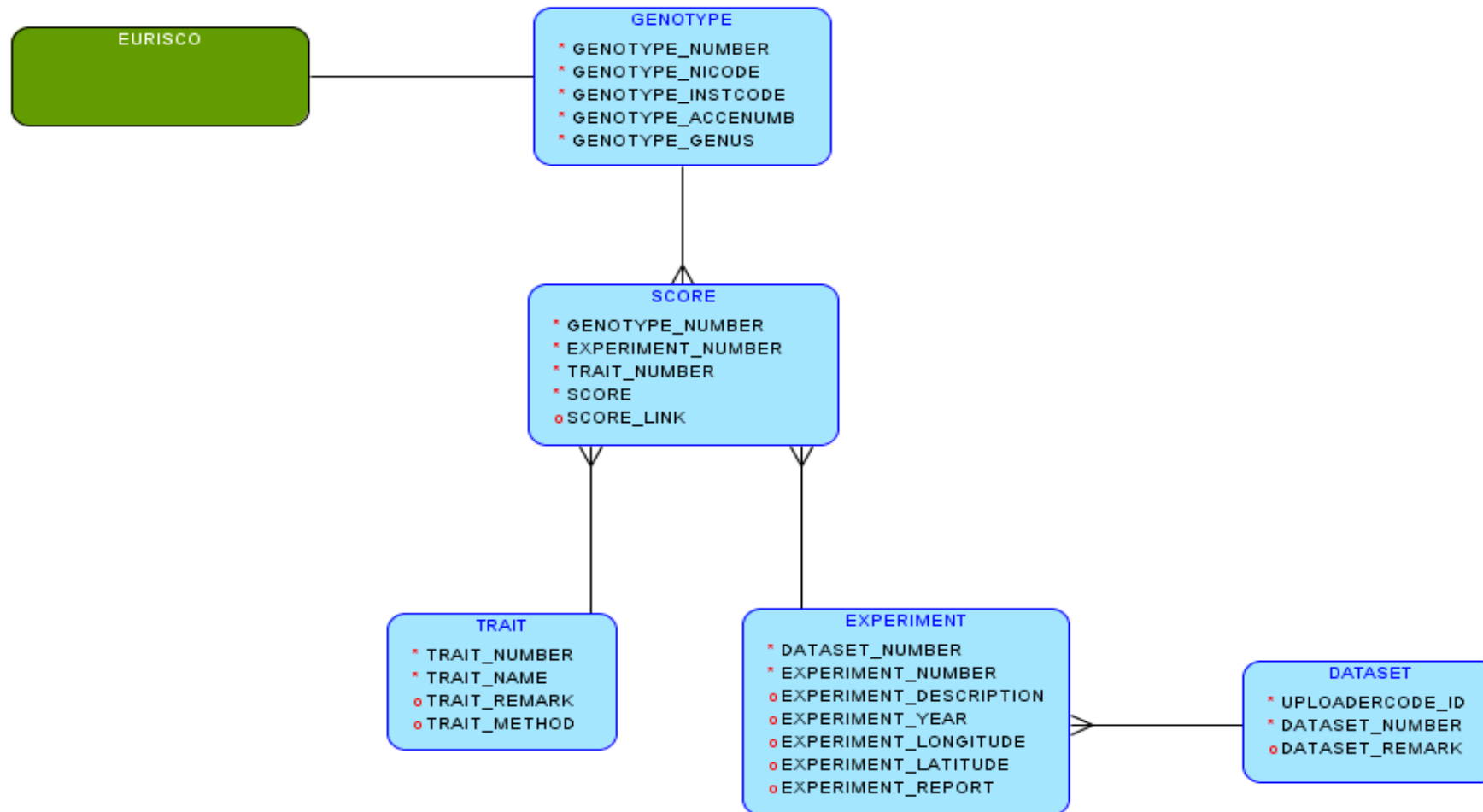
- Help to structure the (phenotypic) world
- Improve interoperability of data
- e.g. Crop Ontology (Arnaud et al. 2012)

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



Data model for phenotypic data



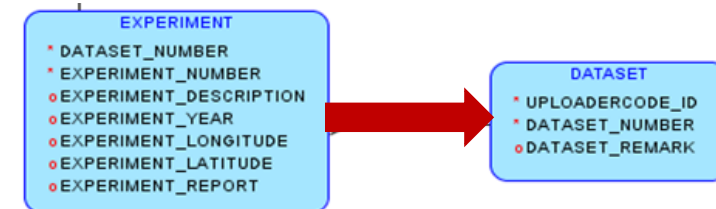
Dataset

- 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

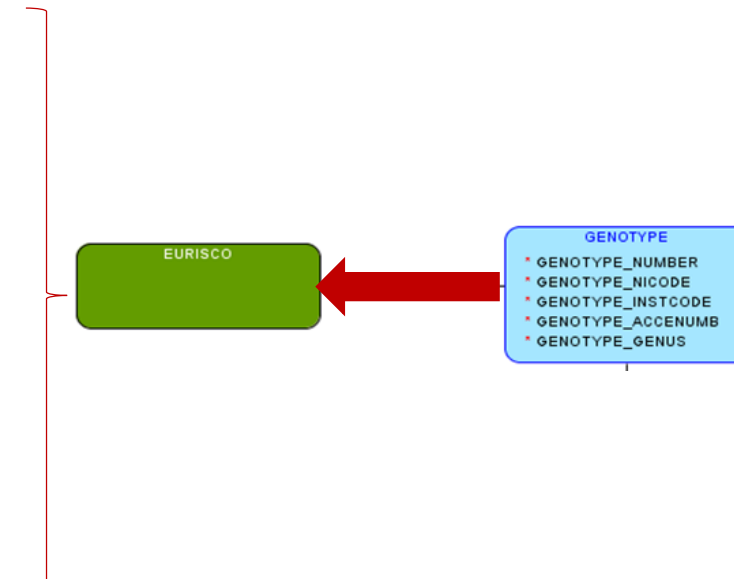
- Describe phenotypic traits and the methods used for scoring
- Fields:
 - **TRAIT_NUMBER***:
 - Unique, temporary number of the trait in the dataset
 - **TRAIT_NAME***:
 - English name of the trait
 - **TRAIT_REMARK**:
 - General remark helping to interpret the trait
 - **TRAIT_METHOD**:
 - English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

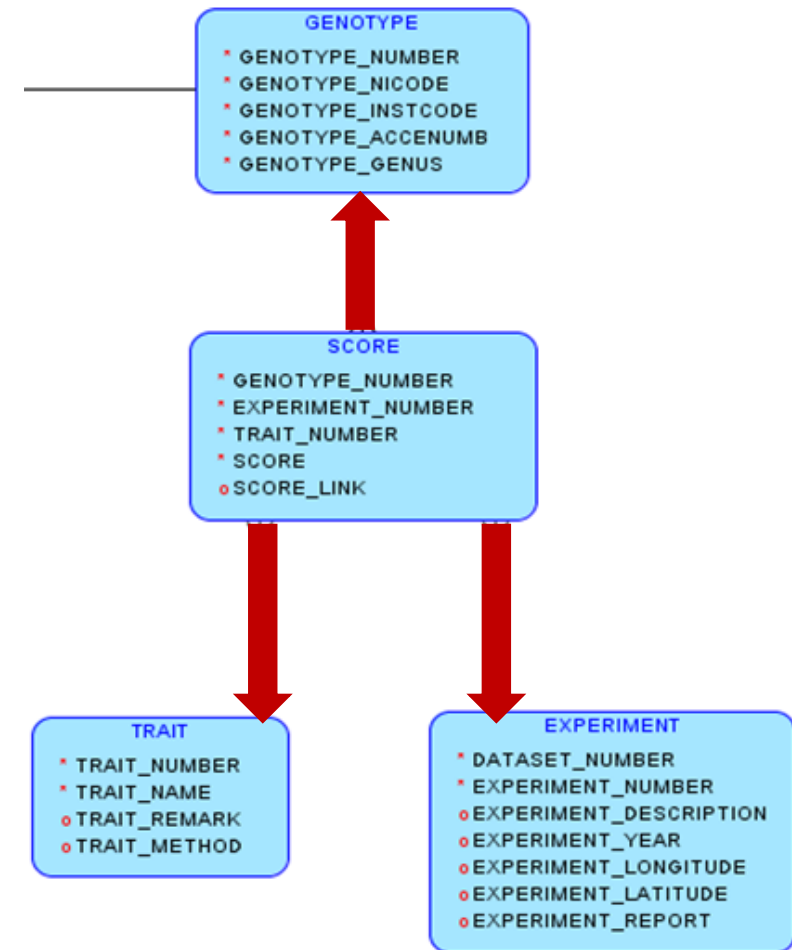


Genotype

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://...
...

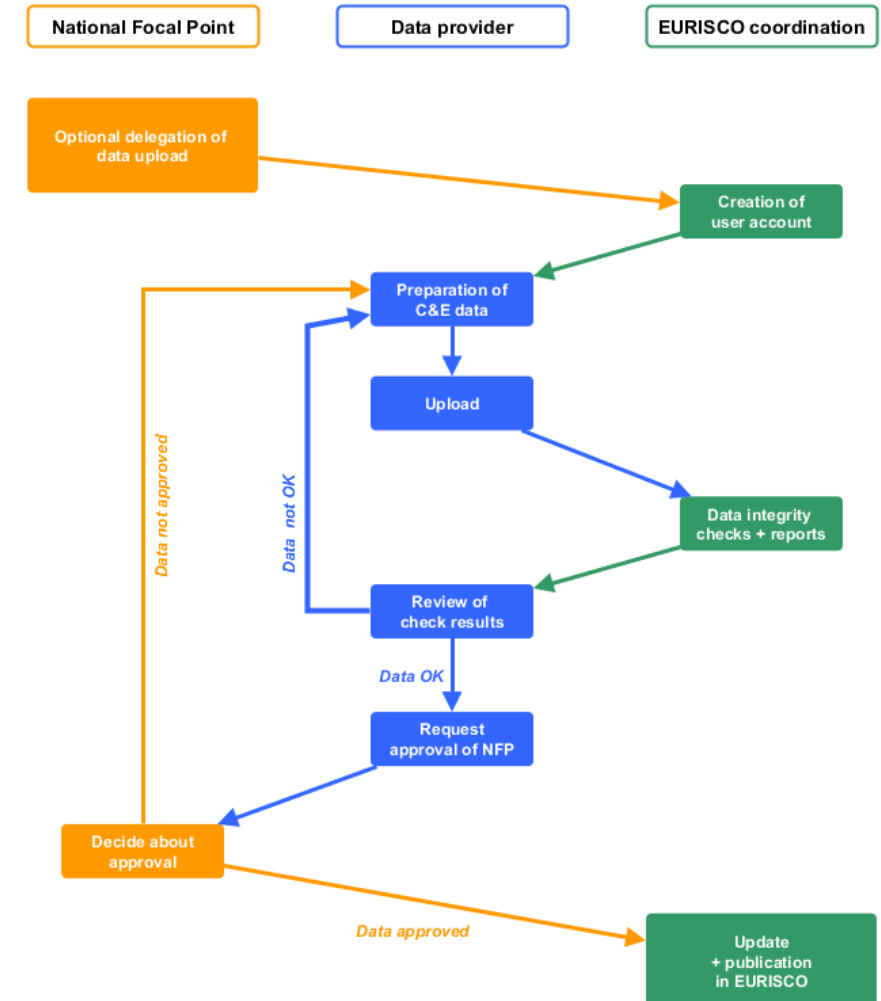
EXPERIMENT

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

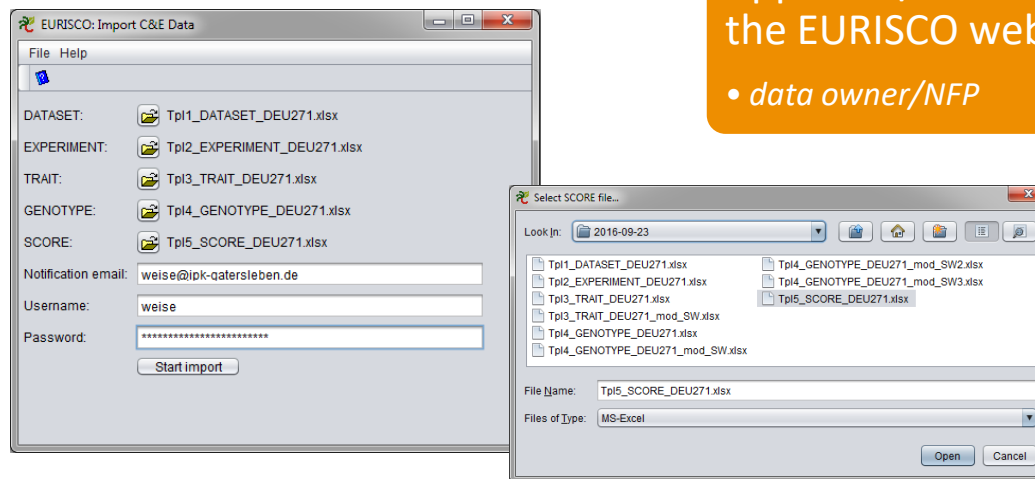
- *data owner*

Data integrity checks

- *EURISCO management*

Approval / withdrawal of data for publishing on the EURISCO website

- *data owner/NFP*



Upload of phenotypic data files

EURISCO uploader

Welcome: WEISE Logout

Home Passport data import **C&E data import**

Upload C&E files C&E integrity check results Decision about C&E update

Home > Upload C&E files

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

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.

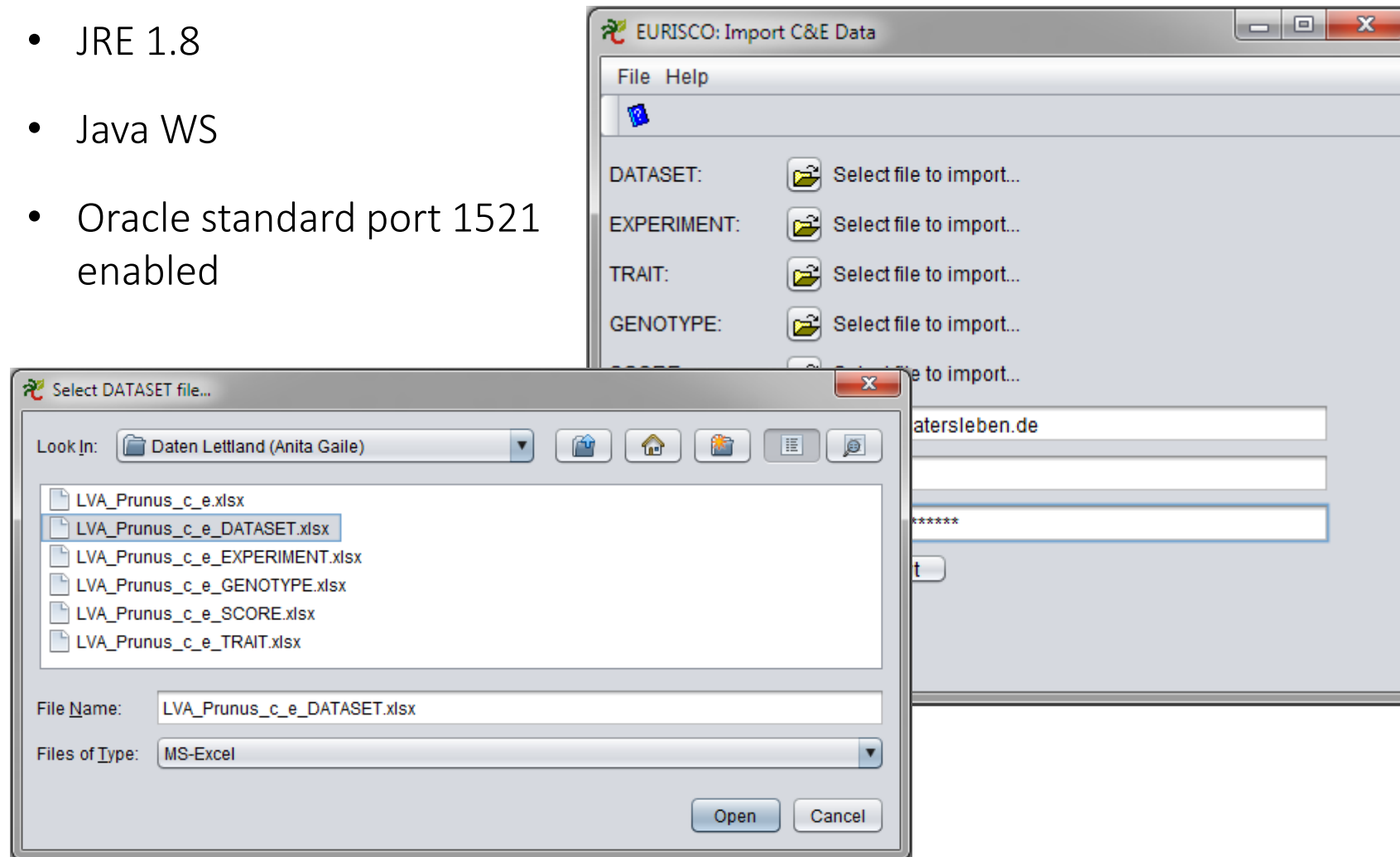
release 1.2.0



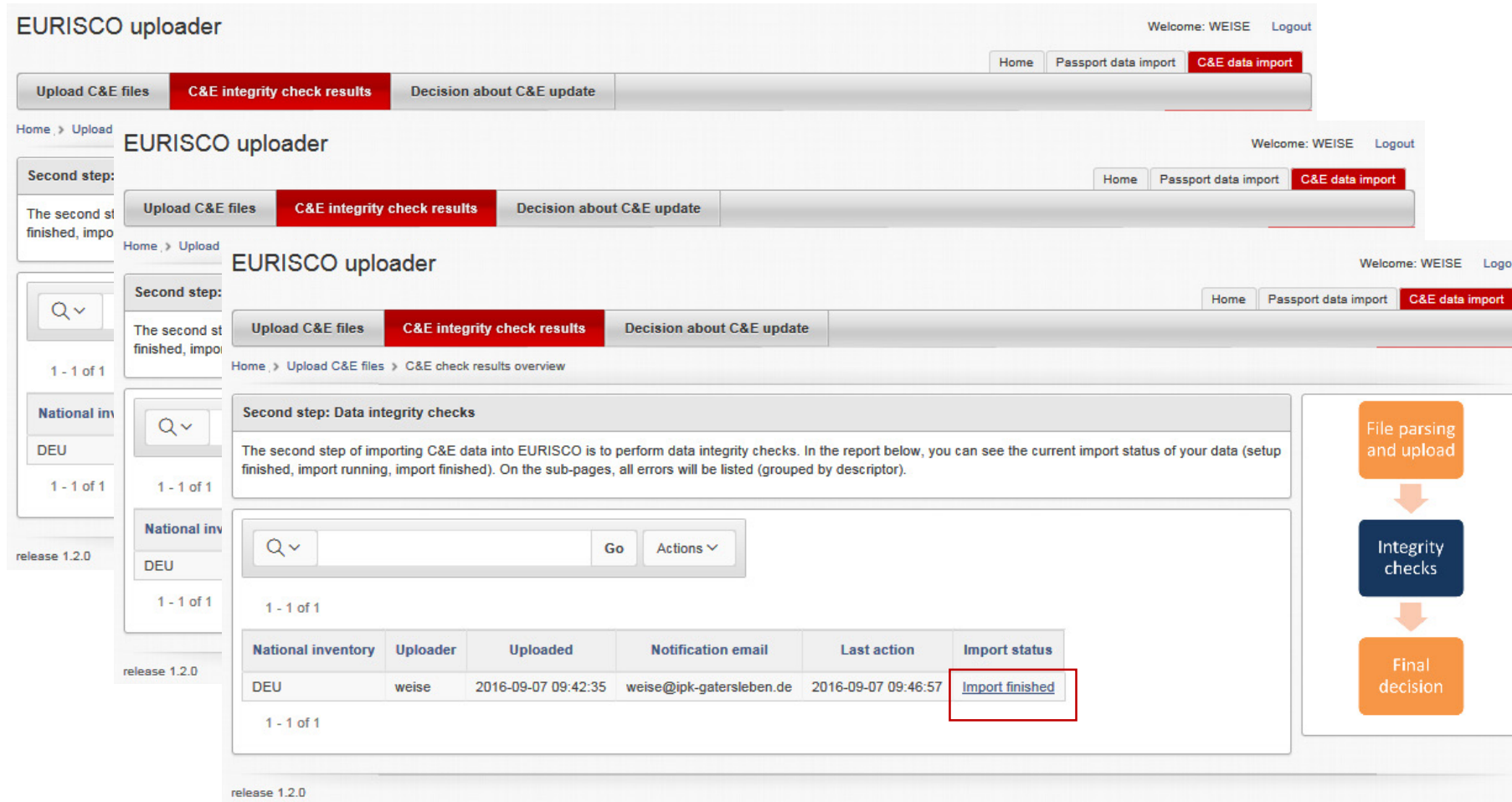
```
graph TD; A[File parsing and upload] --> B[Integrity checks]; B --> C[Final decision];
```

Upload of phenotypic data files

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



Integrity checks



EURISCO uploader

Welcome: WEISE Logout

Home Passport data import **C&E data import**

Upload C&E files **C&E integrity check results** Decision about C&E update

Home > Upload

EURISCO uploader

Welcome: WEISE Logout

Home Passport data import **C&E data import**

Upload C&E files **C&E integrity check results** Decision about C&E update

Home > Upload

EURISCO uploader

Welcome: WEISE Logout

Home Passport data import **C&E data import**

Upload C&E files **C&E integrity check results** Decision about C&E update

Home > Upload C&E files > C&E check results overview

Second step: Data integrity checks

The second step of importing C&E data into EURISCO is to perform data integrity checks. In the report below, you can see the current import status of your data (setup finished, import running, import finished). On the sub-pages, all errors will be listed (grouped by descriptor).

1 - 1 of 1

National inventory

DEU

1 - 1 of 1

release 1.2.0

1 - 1 of 1

release 1.2.0

1 - 1 of 1

release 1.2.0

National inventory	Uploader	Uploaded	Notification email	Last action	Import status
DEU	weise	2016-09-07 09:42:35	weise@ipk-gatersleben.de	2016-09-07 09:46:57	Import finished

1 - 1 of 1

release 1.2.0

1 - 1 of 1

release 1.2.0

1 - 1 of 1

release 1.2.0

File parsing and upload

Integrity checks

Final decision

Integrity checks

EURISCO uploader

Upload C&E files **C&E integrity check results** Decision about C&E update

Home > Upload C&E files > C&E check results overview > C&E errors per descriptor

C&E errors per descriptor

1 - 1 of 1

Template	Descriptor	Number Of Errors
SCORE	TRAIT_NUMBER	148

1 - 1 of 1

release 1.2.0

- Undefined trait number

EURISCO uploader

Welcome: WEISE Logout

Home Passport data import **C&E data import**

Upload C&E files **C&E integrity check results** Decision about C&E update

Home > Upload C&E files > C&E check results overview > C&E errors per descriptor > C&E error details

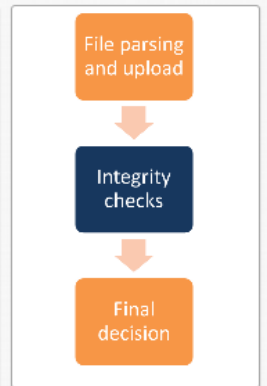
C&E error details

1 - 15 of 148

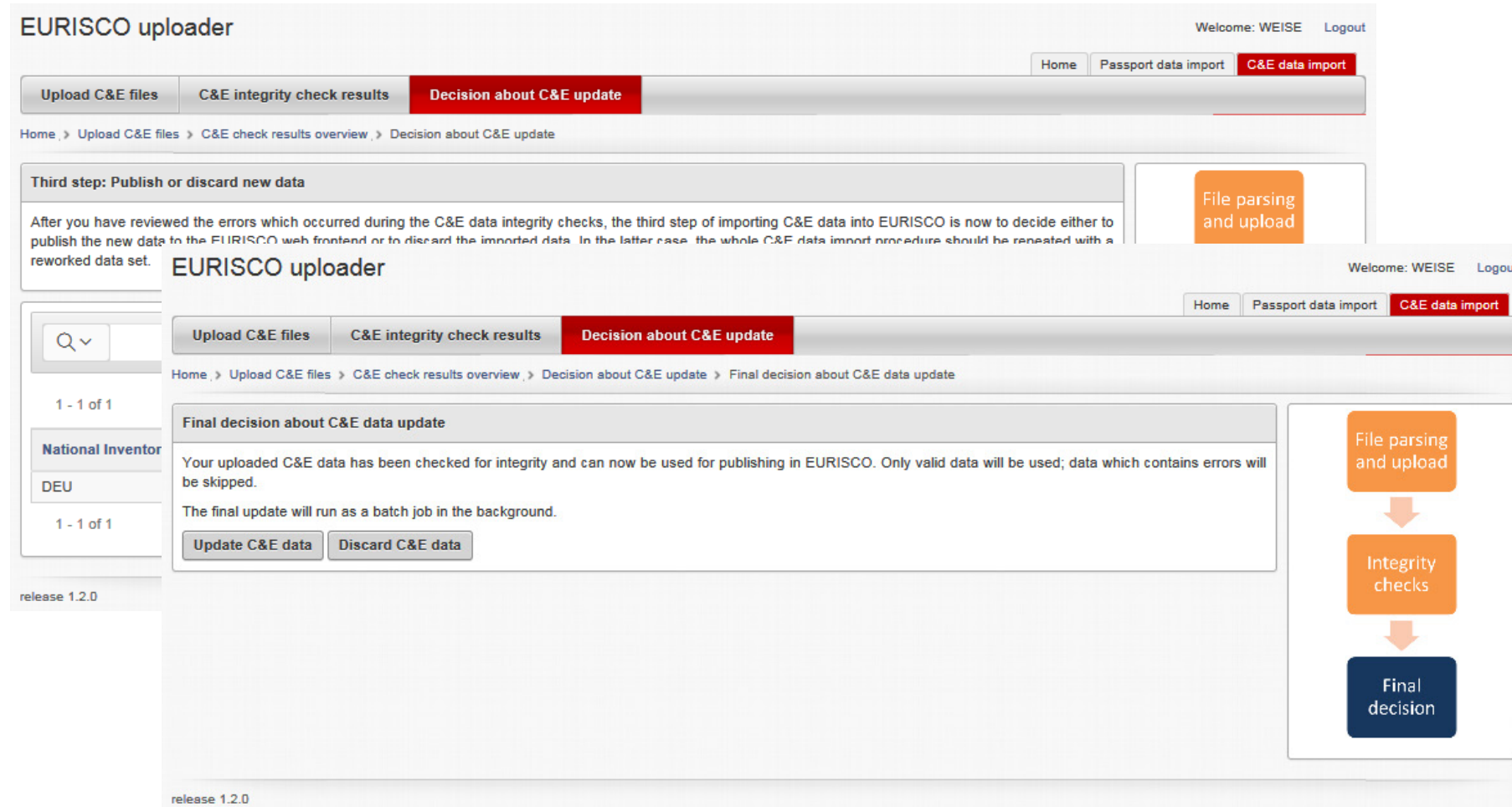
Template	Descriptor	Line Number	Error Type	Error Description
SCORE	TRAIT_NUMBER	85	Error	Line 85: Trait number 47 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	84	Error	Line 84: Trait number 46 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	388	Error	Line 388: Trait number 88 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	387	Error	Line 387: Trait number 87 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	386	Error	Line 386: Trait number 86 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	385	Error	Line 385: Trait number 85 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	384	Error	Line 384: Trait number 84 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	383	Error	Line 383: Trait number 83 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	382	Error	Line 382: Trait number 82 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	381	Error	Line 381: Trait number 81 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	380	Error	Line 380: Trait number 80 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	379	Error	Line 379: Trait number 79 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	378	Error	Line 378: Trait number 78 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	377	Error	Line 377: Trait number 77 invalid. Not listed in TRAIT template.
SCORE	TRAIT_NUMBER	376	Error	Line 376: Trait number 75 invalid. Not listed in TRAIT template.

1 - 15 of 148

release 1.2.0



Final decision



The screenshot displays the EURISCO uploader interface at the 'Final decision about C&E data update' stage. The breadcrumb trail is: Home > Upload C&E files > C&E check results overview > Decision about C&E update > Final decision about C&E data update. The main content area contains the following text: 'Final decision about C&E data update', '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.', and 'The final update will run as a batch job in the background.' Below this text are two buttons: 'Update C&E data' and 'Discard C&E data'. A sidebar on the left includes a search bar, 'National Inventor', and 'DEU'. A flowchart on the right illustrates the process: 'File parsing and upload' (orange box) leads to 'Integrity checks' (orange box), which leads to 'Final decision' (dark blue box). The interface also shows navigation tabs for 'Upload C&E files', 'C&E integrity check results', and 'Decision about C&E update' (highlighted in red). The user is logged in as 'WEISE'.

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

Filter C&E data by trait

The report below lists the definitions of all phenotypic traits, which are currently available in EURISCO. Please use the search bar below to define filters.

Rows 10

1 - 10 of 61

Trait Name	Trait Method
Flowering time	Count days after 1 April when >50% plants show inflorescence emergence, 999=not flowering during experiment
Flowering time end	(3=early, 7=late)
Flowering time	Number of days between the date of sowing and the date of appearance of the first flower head
Flowering time begin	Days after sowing when 50% of plants have opened the first flower(s)
Flowering time	Count days after 1 September when >50% plants show inflorescence emergence, 999=not flowering during experiment
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 to 10% of flowers have opened after sowing
Flowering time	count days after 1 May when 50% of florets have opened on 3 flowers
Flowering time begin	(3=early, 7=late)
Flowering time begin	Count the days from 25/5 to 50% of plants in flower

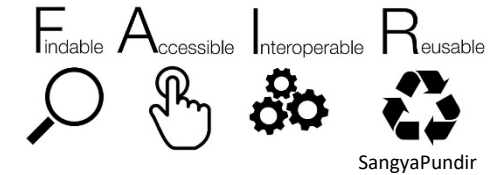
1 - 10 of 61

0.06 s

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

