

C&E data templates

Presentation & discussion

EURISCO training workshop, 12th to 14th September 2017, Gatersleben

Stephan Weise
13 September 2017



BACKGROUND

C&E data I

- Previously not available in EURISCO
- Of high importance to users of PGR data
 - Determine value of germplasm for breeding and research
- Difficult to handle due to lots of “standards”
 - Different descriptor names/synonyms
 - Different rating scales
 - Nominal, ordinal, metric scale
 - Different amounts of meta information
 - When, where, how, by whom?
 - Experiment set-up, treatment etc.

C&E data II

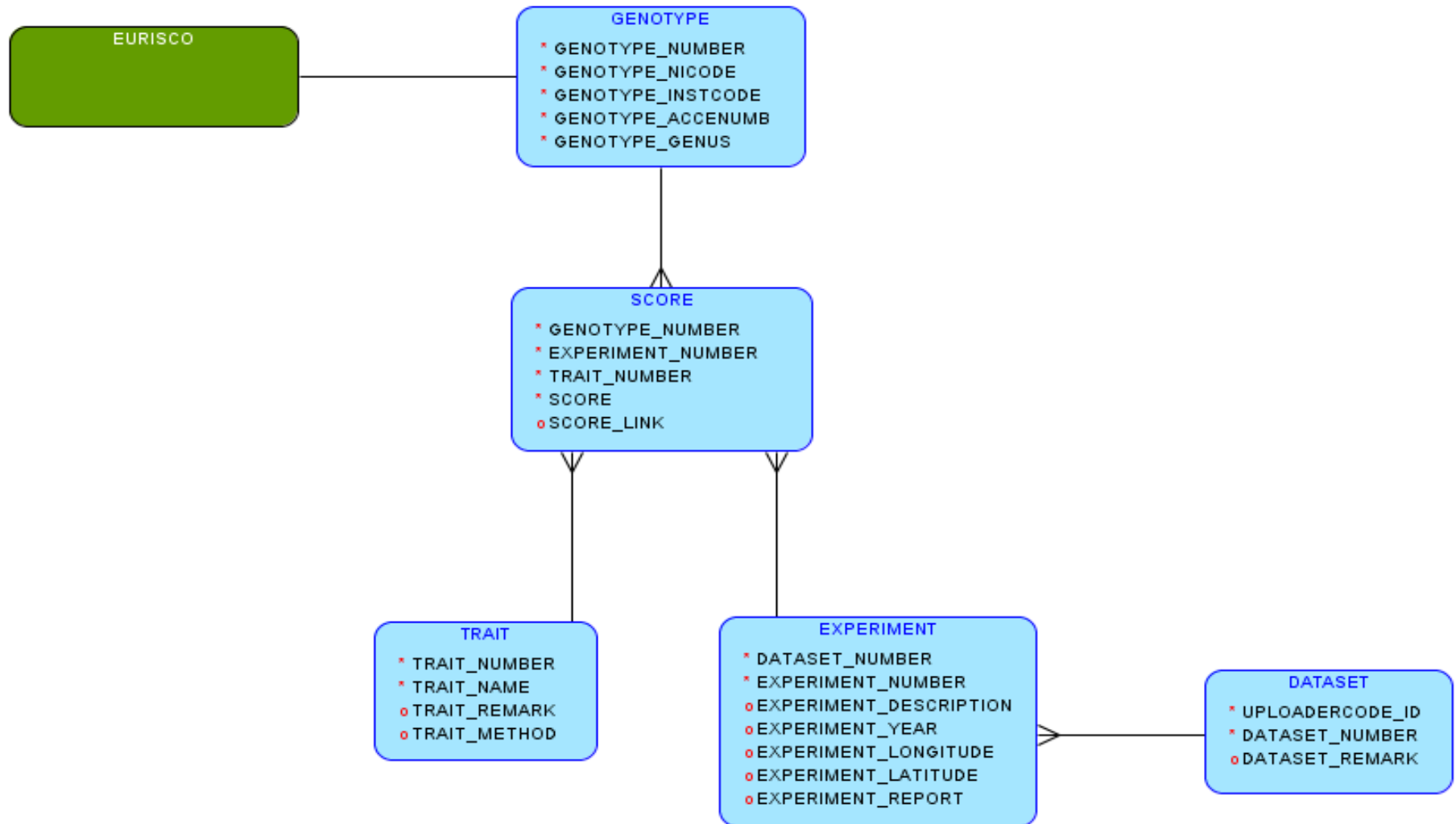
- Discussed in detail in previous years
- Suitable proposal compiled by the ECPGR Doc&Info Network in 2009
 - As simple as possible
 - As few fields as possible
 - “minimum consensus”
- D&I meeting 2014
 - Pragmatic approach:
 - No standardisation of trait, scale or experimental design
 - Only standardisation of exchange format (as simple as possible)
 - Import of existing data as-is → reach critical mass
 - Use proposal of 2009 (+ small adjustments)

DATA MODEL

Data model for C&E data I

- Five components
 - GENOTYPE
 - Identified by EURISCO descriptors
 - DATASET
 - May comprise different experiments
 - EXPERIMENT
 - Multiple genotypes are scored for different traits
 - TRAIT
 - Characteristic feature to be scored
 - SCORE
 - Value of a trait for a genotype

Data model for C&E data II



TEMPLATES

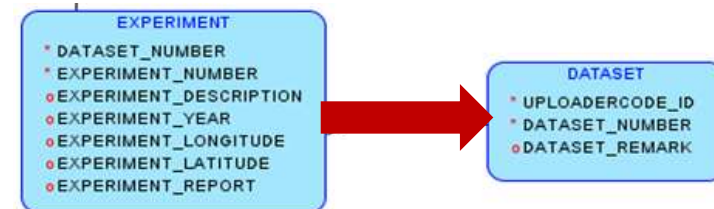
DATASET

- 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 I

- 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 II

- 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 III

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 I

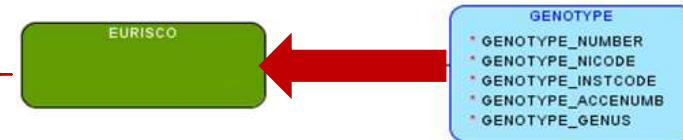
- Describe phenotypic traits and the methods used for scoring
- Fields:
 - **TRAIT_NUMBER***:
 - Unique, temporary number of the trait in the dataset
 - **TRAIT_NAME***:
 - English name of the trait
 - **TRAIT_REMARK**:
 - General remark helping to interpret the trait
 - **TRAIT_METHOD**:
 - English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

TRAIT II

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 I

- 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 II

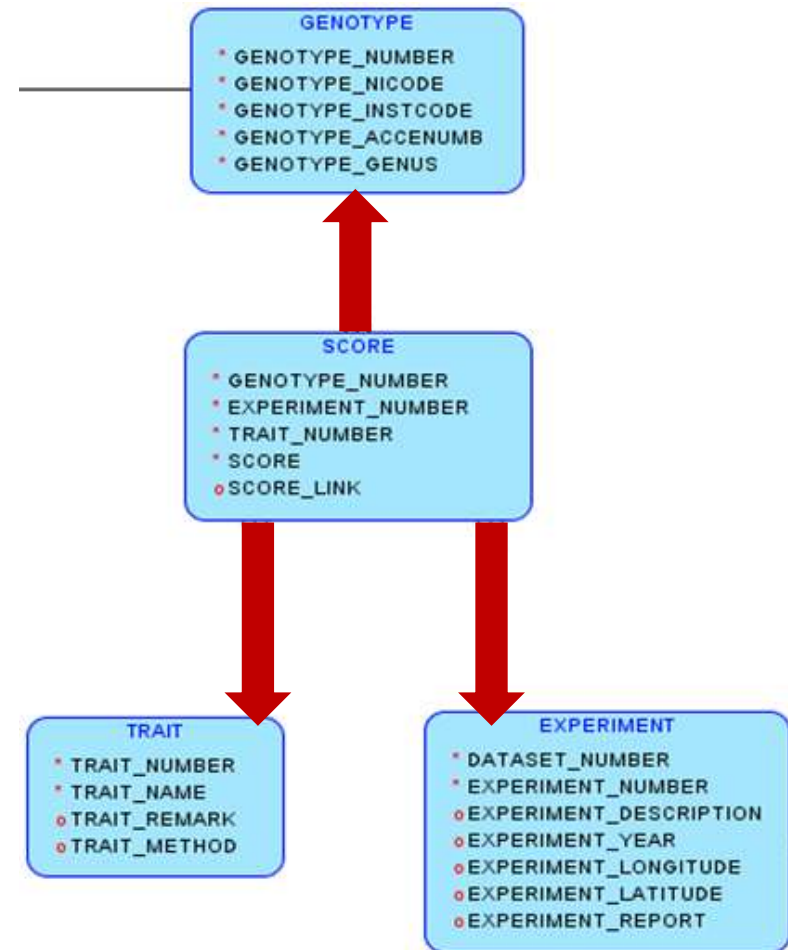
GENOTYPE_NUMBER	GENOTYPE_NICODE	GENOTYPE_INSTCODE	GENOTYPE_ACCENUMB	GENOTYPE_GENUS	GENOTYPE_PUID
1	DEU	DEU271	GR 142	Lolium	
2	DEU	DEU271	GR 476	Lolium	
3	DEU	DEU271	GR 550	Lolium	
4	DEU	DEU271	GR 2670	Lolium	

SCORE I

- 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 II

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://...
...
...

Putting all this together

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	

GENOTYPE

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]
...

TRAIT

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

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

Upload of C&E data files I

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

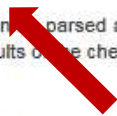
The first step of importing new C&E data into EURISCO is to upload the filled template files to the EURISCO server. The files must be formatted in accordance with the EURISCO C&E data exchange format. The data must be contained in five separate MS-Excel (.xlsx) files.

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.

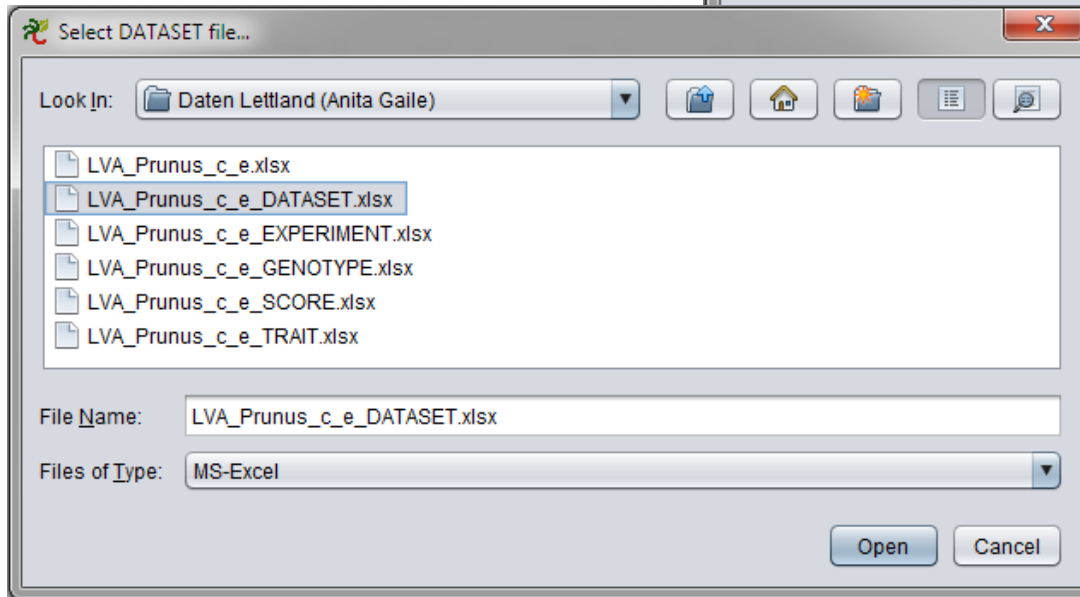
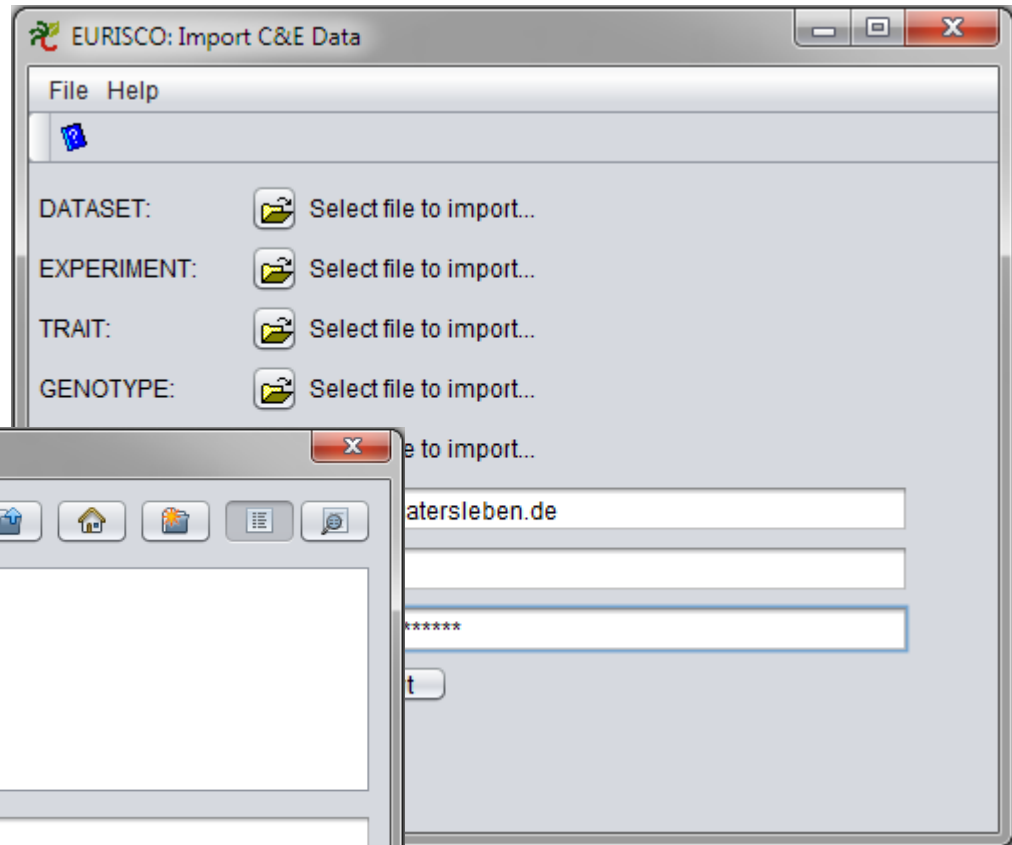
Requirements:

- 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.



Upload of C&E data files II

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



Integrity checks I

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**

Second step:

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

The second step finished, import running

Home > Upload

EURISCO uploader

Welcome: WEISE Logout

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

Second step:

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

The second step finished, import finished

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).


Search: [Q] [v] Go Actions [v]

1 - 1 of 1

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



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


Integrity checks II

EURISCO uploader

Welcome: WEISE Logout

Upload C&E files C&E integrity check results

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

EURISCO uploader

Welcome: WEISE Logout

Home » Passport data import C&E data import

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

C&E errors per descriptor

Q v

1 - 1 of 1

Template	Descriptor	Number Of t
SCORE	TRAIT_NUMBER	

1 - 1 of 1

release 1.2.0

C&E error details

Q v Go Actions v

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

File parsing and upload

Integrity checks

Final decision

- Undefined trait number

Final decision

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 > Decision about C&E update

Third step: Publish or discard new data

After you have reviewed the errors which occurred during the C&E data integrity checks, the third 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.

File parsing
and 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 > Decision about C&E update > Final decision about C&E data update

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.

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

Update C&E data Discard C&E data

File parsing
and upload

Integrity
checks

Final
decision

release 1.2.0

Next steps (in background)

- New dataset will be applied to EURISCO stage schema
 - Existing C&E data will not be overwritten
 - Existing C&E 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