

Passport data updates

Presentation & discussion

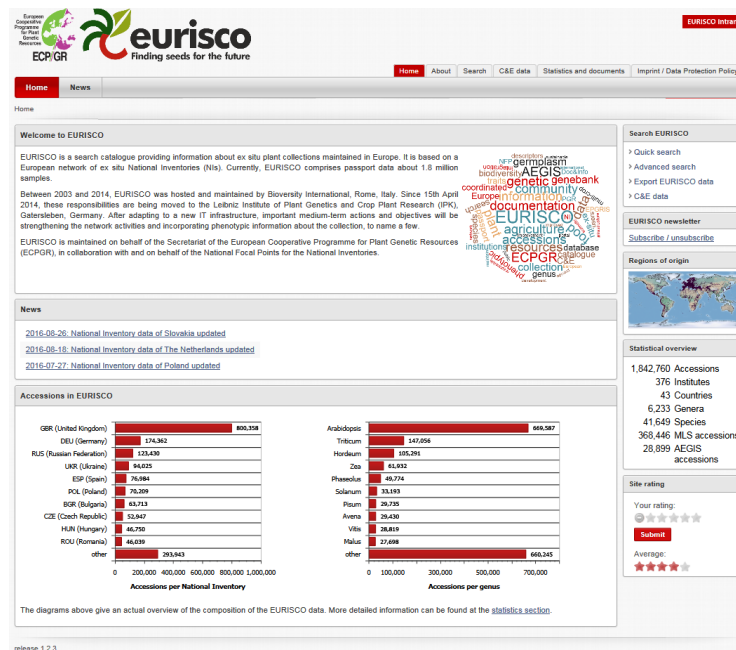
EURISCO training workshop, 12th to 14th October 2016, Angers, France

Stephan Weise
13 October 2016



EURISCO intranet I

- Development of new import component for NIs
 - Web interface with Oracle APEX
 - PL/SQL packages for uploading, checking and updating data
 - Implementation of incremental updates



EURISCO intranet II

Upload file to server

Import file into database

Perform data checks

Decide about publishing or withdrawal

New Java-based importer

The screenshot displays the EURISCO intranet II interface, specifically the 'Descriptors for uploading information from National Inventories to EURISCO' page. The page is organized into a grid of boxes, each representing a different data field or descriptor. Each box contains a title, a brief description of the field, and a list of examples. The fields include:

- 1. National inventory name:** Code identifies the National Inventory, the code of the country issuing the National Inventory. Examples are provided for the format: Country Code (ISO 3166-1) + National Inventory Code (INSTR001).
- 2. Accession number:** Original number assigned by the issuing institution followed by a number. Example: 123456789.
- 3. Accession code:** Code of the institution where the sample was collected. Example: 123456789.
- 4. Collecting institute code:** Code of the institution where the sample was collected. Example: 123456789.
- 5. Origin of origin:** Code of the origin of the material. Example: 123456789.
- 6. Language of collecting site:** Language of the collecting site. Example: 123456789.
- 7. Language of collecting site:** Language of the collecting site. Example: 123456789.
- 8. Collecting date of sample:** Date of collection. Example: 2010-01-01.
- 9. Biological status of accession:** Biological status of the accession. Example: 123456789.
- 10. Breeding institute code:** Code of the breeding institute. Example: 123456789.
- 11. Breeding institute code:** Code of the breeding institute. Example: 123456789.
- 12. Donor institute code:** Code of the donor institute. Example: 123456789.
- 13. Donor accession number:** Accession number of the donor. Example: 123456789.
- 14. Other identification number:** Other identification number. Example: 123456789.
- 15. Location of safety evaluation:** Location of safety evaluation. Example: 123456789.
- 16. Accession URL:** URL of the accession. Example: 123456789.
- 17. MIS Status:** MIS Status. Example: 123456789.
- 18. Remarks:** Remarks. Example: 123456789.

Java-based import I

- So far: Web-based upload
 - Upload via tab-separated text file, UTF-8 encoded
 - Often problems with columns separators + character encoding
- Now: Java-based Excel import
 - Summarises file upload + data import
 - User gets informed when the integrity checks are finished

The screenshot displays the EURISCO uploader interface. At the top, it says "EURISCO uploader" and "Welcome: WEISE Logout". There are navigation tabs: "Home", "Passport data import" (highlighted), and "C&E data import". Below this is a breadcrumb trail: "Home > Upload file".

The main content area is titled "First step: File upload". It contains the following text:
The first step of importing new or modified data into EURISCO is to upload a file containing National Inventory data to the EURISCO server. This file must be formatted in accordance with the MCPD EURISCO format.
The file must either be in **MS-Excel (.xlsx) format (recommended)** or must be a UTF-8 encoded text file with tabulator-separated columns (deprecated).

Below this is a section for "MS-Excel file upload" with a "Next step" button. It contains the following text:
Please use the Java WebStart application for uploading: [Start the EURISCO passport data importer](#).
The Java application will enable you to select the template file. This 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.

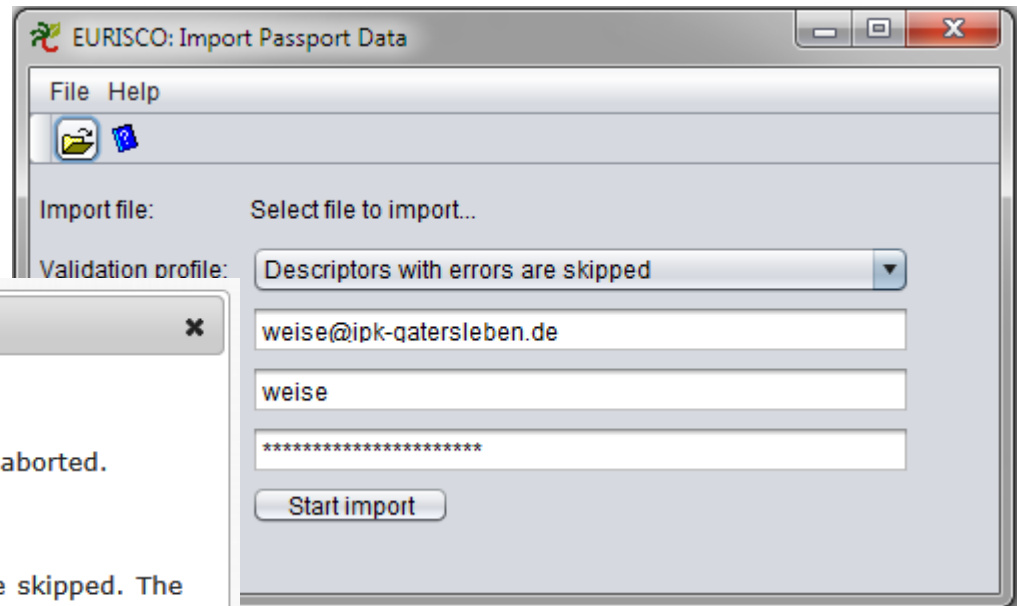
At the bottom of this section is a "Flat file upload (deprecated)" option.

On the right side of the interface, there is a vertical flowchart with four steps: "File upload" (blue box), "File import" (orange box), "Integrity checks" (orange box), and "Final decision" (orange box). A red arrow points from the "Integrity checks" step in the flowchart to the "Start the EURISCO passport data importer" link in the MS-Excel file upload section.

release 1.2.0

Java-based import II

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



Validation Profile

Errors abort session:

If an error occurs, the whole import operation will be aborted.

Records with errors are skipped:

If an error occurs within a record, this record will be skipped. The remaining records will be imported.

Descriptors with errors are skipped:

If a descriptor value within a record contains an error, only this value will be skipped. The remaining descriptor values of the respective record will be imported. However, in case of an error within one of the mandatory descriptors, the whole record will be skipped.

Integrity checks I

EURISCO uploader

Welcome: WEISE Logout

Home Passport data import C&E data import

Upload file Import file Integrity check results Decision about update

Home > Upload file

EURISCO uploader

Welcome: WEISE Logout

Home Passport data import C&E data import

Upload file Import file Integrity check results Decision about update

Home > Upload file

EURISCO uploader

Welcome: WEISE Logout

Home Passport data import C&E data import

Upload file Import file Integrity check results Decision about update

Home > Upload file > Import file > Check results overview

Third step: Data integrity checks

The third step of importing new or modified 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

National Inventory

AUT

1 - 1

1 - 1

1 - 1

1 - 1

1 - 1

1 - 1

1 - 1

release 1.2.0

1 - 1

release 1.2.0

National Inventory	Filename	File uploaded	Notification email	End of import	Import status
AUT	test_for_workshop.xlsx	2016-09-06 15:11:47	weise@ipk-gatersleben.de	2016-09-06 15:20:37	Import finished

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

Integrity checks II – error report

EURISCO uploader Welcome: WEISE Logout

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

Upload file Import file **Integrity check results** Decision about update

Home > Upload file > Import file > Check results overview > Errors per descriptor

Errors per descriptor

1 - 14

Descriptor	Number Of Errors
ACQDATE	44
BREDCODE	2
COLLDATE	4
COLLSRC	21
DONORCODE	21
DUPLSITE	45
ELEVATION	2
GENUS	1
LATITUDE	8
LONGITUDE	22
MLSSTAT	1
ORIGCTY	1
REMARKS	6
STORAGE	3

1 - 14

Deletion candidates

The number of accessions contained in an NI dataset may vary due to different reasons. For example, accessions could be removed from a certain genebank or the accession identifiers (unique combination of INSTCODE, GENUS, ACCENUMB) may change. With the new update mechanism, no accessions will be deleted automatically from EURISCO.

As a consequence of the new update mechanism, National Focal Points will explicitly have to name accessions to be deleted from the system.

In order to support this process, during the data integrity checks your new dataset was automatically compared with the existing dataset in EURISCO. This report provides an overview of accessions, which no longer exist in the new dataset, grouped by holding institution.

However, this list can only be a hint, which accessions could be candidates for deletion from EURISCO, and needs to be checked. Especially if only a part of the NI dataset should be updated, this list may contain many false positive entries.

Please send the checked list to weise@ipk-gatersleben.de.

1 - 1

INSTCODE	No of accessions
AUT001	1

1 - 1

```

graph TD
    A[File upload] --> B[File import]
    B --> C[Integrity checks]
    C --> D[Final decision]
    
```

Integrity checks III - example

- Wrong date format

EURISCO uploader

Welcome: WEISE Logout

Home Passport data import C&E data import

Upload file Import file **Integrity check results** Decision about update

Home > Upload file > Import file > Check results overview > Errors per descriptor > Error details

Q Go Actions

1 - 15 >

Descriptor	Error Description
ACQDATE	Line 11682: ACQDATE 2010 invalid.
ACQDATE	Line 11694: ACQDATE 2100217 invalid.
ACQDATE	Line 2162: ACQDATE 2002--- invalid.
ACQDATE	Line 2163: ACQDATE 2002--- invalid.
ACQDATE	Line 2164: ACQDATE 2002--- invalid.
ACQDATE	Line 2171: ACQDATE 2002--- invalid.
ACQDATE	Line 2172: ACQDATE 2002--- invalid.
ACQDATE	Line 2173: ACQDATE 2002--- invalid.
ACQDATE	Line 2192: ACQDATE 2002--- invalid.
ACQDATE	Line 2193: ACQDATE 2002--- invalid.
ACQDATE	Line 2194: ACQDATE 2002--- invalid.
ACQDATE	Line 2195: ACQDATE 2002--- invalid.
ACQDATE	Line 2196: ACQDATE 2002--- invalid.
ACQDATE	Line 2200: ACQDATE 2002--- invalid.
ACQDATE	Line 2201: ACQDATE 2002--- invalid.

1 - 15 >

release 1.2.0

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


Integrity checks IV - example

- Invalid or multiple donor codes

EURISCO uploader Welcome: WEISE Logout

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

Upload file Import file **Integrity check results** Decision about update

Home > Upload file > Import file > Check results overview > Errors per descriptor > Error details

Q Go Actions

1 - 15 >

Descriptor	Error Description
DONORCODE	Line 647: DONORCODE FAO031 invalid.
DONORCODE	Line 648: DONORCODE FAO031 invalid.
DONORCODE	Line 649: DONORCODE FAO031 invalid.
DONORCODE	Line 806: DONORCODE ROU006 invalid.
DONORCODE	Line 8429: DONORCODE DEU146 / GBR040 invalid.
DONORCODE	Line 8512: DONORCODE AUT016/Versailles invalid.
DONORCODE	Line 8527: DONORCODE AUT016/INIA invalid.
DONORCODE	Line 8851: DONORCODE DEU146 / DEU235 invalid.
DONORCODE	Line 8853: DONORCODE DEU146 / USA126 invalid.
DONORCODE	Line 11057: DONORCODE DEU146 / DEU040 invalid.
DONORCODE	Line 11068: DONORCODE DEU146 / CHED08 invalid.
DONORCODE	Line 11069: DONORCODE DEU146 / PRT008 invalid.
DONORCODE	Line 11070: DONORCODE DEU146 / DEU032 invalid.
DONORCODE	Line 11071: DONORCODE DEU146 / SVN006 invalid.
DONORCODE	Line 11072: DONORCODE DEU146 / DEU152 invalid.

1 - 15 >

release 1.2.0

File upload

↓

File import

↓

Integrity checks

↓

Final decision

Incremental updates I

- Why incremental updates?
 - So far: Full replacement
 - Delete whole dataset + reimport data afterwards
 - Even if only a couple of rows have been modified
 - Not possible to update parts of data (e.g. single genebank collection)
 - That's why: From full replacement to real update
 - Only incremental data needs to be updated
 - Necessary: Unique identifiers
 - Currently: Combination of NICODE, INSTCODE, ACCENUMB and GENUS
 - Important for managing C&E data
 - Cannot exist without passport data

Incremental updates II

- Deletion candidates
 - Check of new dataset against existing data
 - List of accessions not contained in the new dataset
 - Not deleted automatically
 - False positive hits in case of partial update!!!

Deletion candidates

The number of accessions contained in an NI dataset may vary due to different reasons. For example, accessions could be removed from a certain genebank or the accession identifiers (unique combination of INSTCODE, GENUS, ACCENUMB) may change. With the new update mechanism, no accessions will be deleted automatically from EURISCO.

As a consequence of the new update mechanism, National Focal Points will explicitly have to name accessions to be deleted from the system.

In order to support this process, during the data integrity checks your new dataset was automatically compared with the existing dataset in EURISCO. This report provides an overview of accessions, which no longer exist in the new dataset, grouped by holding institution.

However, this list can only be a hint, which accessions could be candidates for deletion from EURISCO, and needs to be checked. Especially if only a part of the NI dataset should be updated, this list may contain many false positive entries.

Please send the checked list to weise@ipk-gatersleben.de.

1 - 2

INSTCODE	No of accessions
AUT001	5
AUT067	4

1 - 2

Deletion candidates per institution

Q Go Rows 15 Actions

1 - 5

NICODE	INSTCODE	GENUS	ACCENUMB
AUT	AUT001	ELETTARIA	BVAL-901A10
AUT	AUT001	HORDEUM	BVAL-358001
AUT	AUT001	MYOSOTIS	BVAL-901795
AUT	AUT001	PLECTRANTHUS	BVAL-901A03
AUT	AUT001	ZINGIBER	BVAL-901A06

1 - 5

Final decision

EURISCO uploader

Welcome: WEISE Logout

Home Passport data import C&E data import

Upload file Import file Integrity check results **Decision about update**

Home > Upload file > Import file > Decision about update

Fourth step: Publish or discard new data

After you have reviewed the errors which occurred during the data integrity checks, the fourth step of importing new or modified 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 update procedure should be repeated with a debugged data set.

File upload



EURISCO uploader

Welcome: WEISE Logout

Home Passport data import C&E data import

Upload file Import file Integrity check results **Decision about update**

Home > Upload file > Import file > Decision about update > Final decision

1 - 1

National Inventory

AUT

1 - 1

Final decision

Your uploaded file has been checked for integrity and can now be used to update the data of your National Inventory in EURISCO. Only accessions listed in your file will be updated. All other accessions of your National Inventory in EURISCO, which are not covered by the input file, will remain untouched in EURISCO.

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

Update EURISCO data

Discard data

File upload



File import



Integrity checks



Final decision

release 1.2.0

Next steps (in background)

- Updated dataset will be applied to EURISCO stage schema
- EURISCO stage will be synchronised to the EURISCO web schema (Time lag!)
 - Not in main business hours
 - Rebuild of materialised views
 - Creation of new full dump (MS Access)
 - News message on EURISCO webpage