Passport data updates

Presentation & discussion

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

Stephan Weise

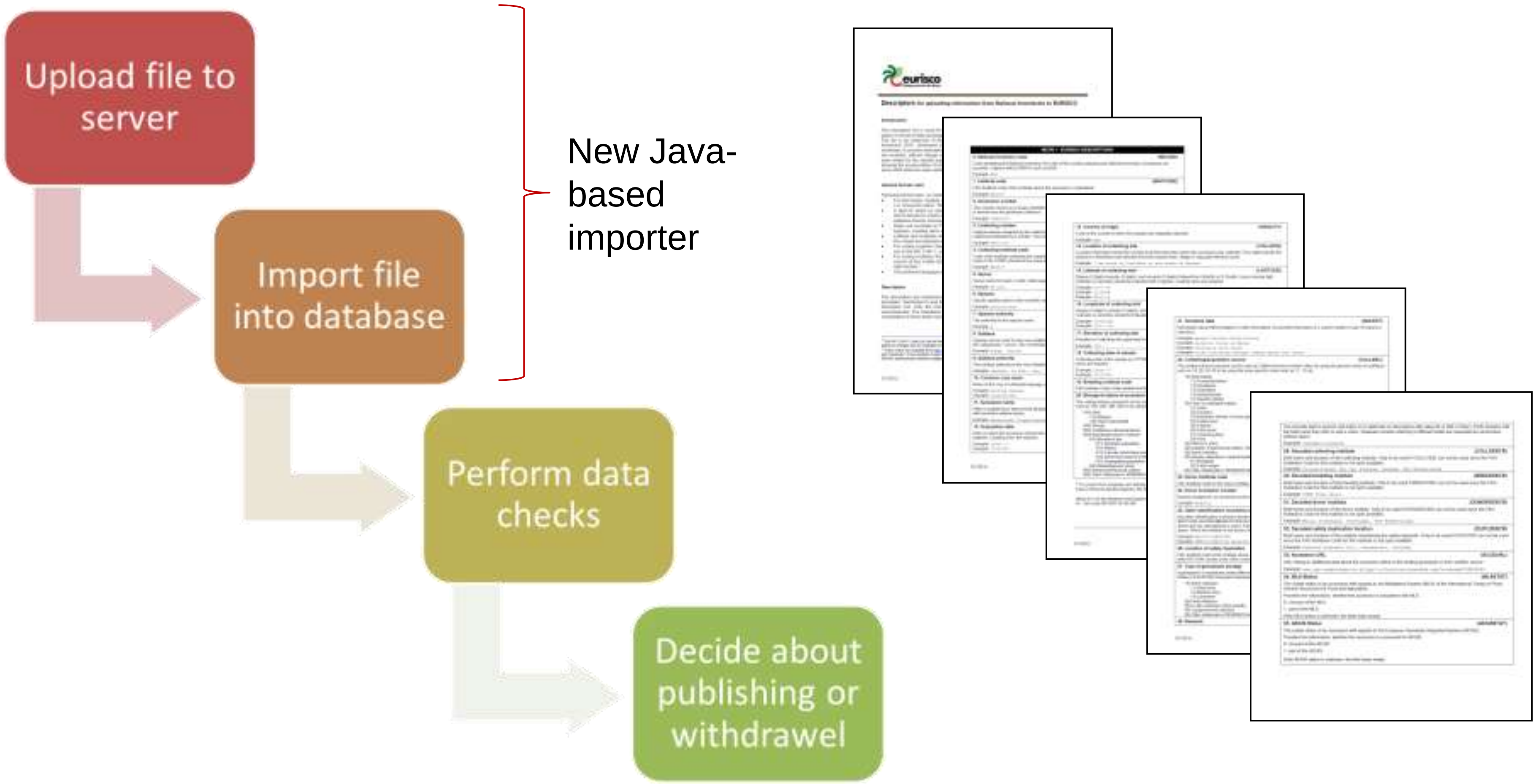
13 September 2017

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

1,042,762 Accessions
378 Institutes
43 Countries
6,233 Genera
41,845 Species
385,446 MLS accessions
20,899 AEGIS accessions

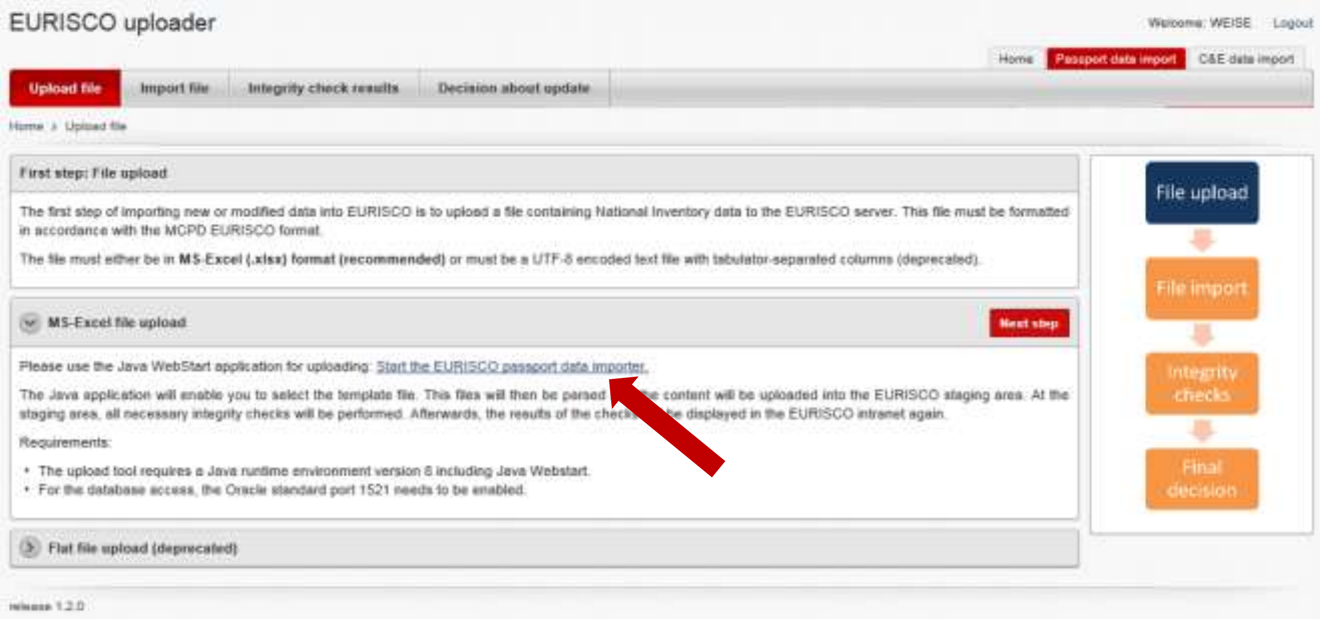
EURISCO intranet II


Upload file to server
Import file into database
Perform data checks
Decide about publishing or withdrawel

New Java-based importer

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



EURISCO uploader

First step: File upload

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).

MS-Excel file upload

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 [illegible] he 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 [illegible] be displayed in the EURISCO intranet again.

Requirements:

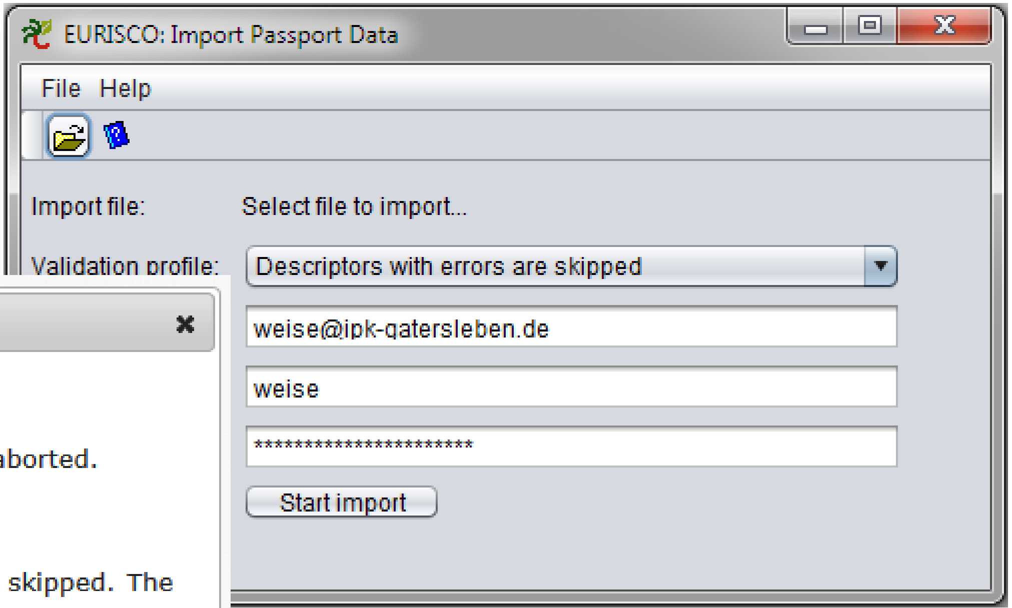
- The upload tool requires a Java runtime environment version 8 including Java Webstart.
- For the database access, the Oracle standard port 1521 needs to be enabled.

Flat file upload (deprecated)

release 1.2.0

Java-based import II

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



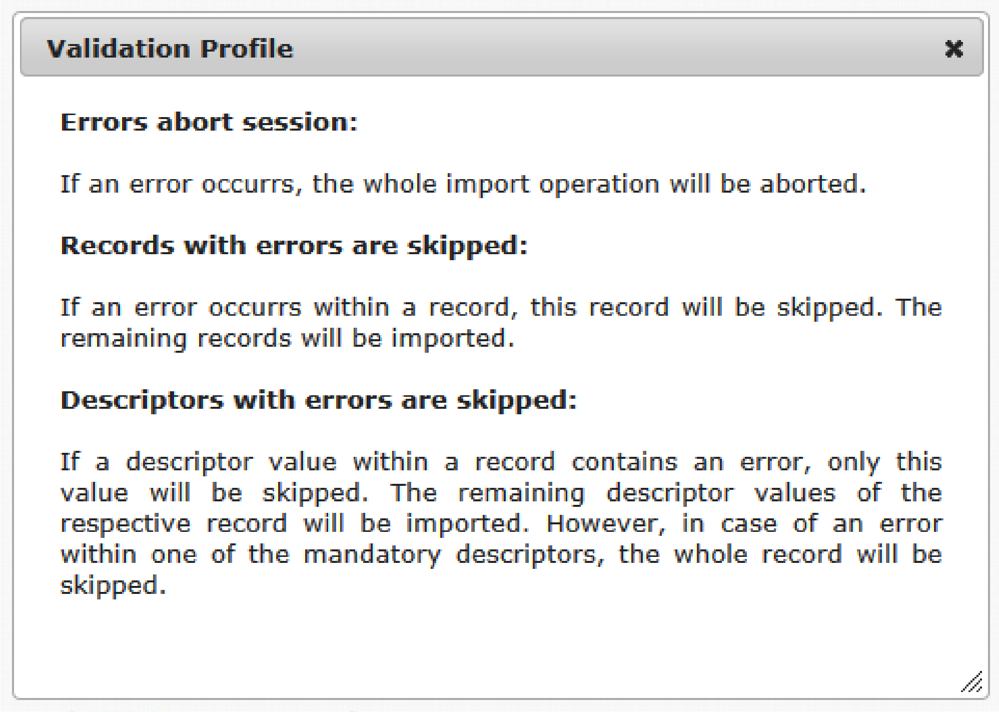
Import file: Select file to import...

Validation profile: Descriptors with errors are skipped

weise@ipk-gatersleben.de

weise

****



Errors abort session:

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

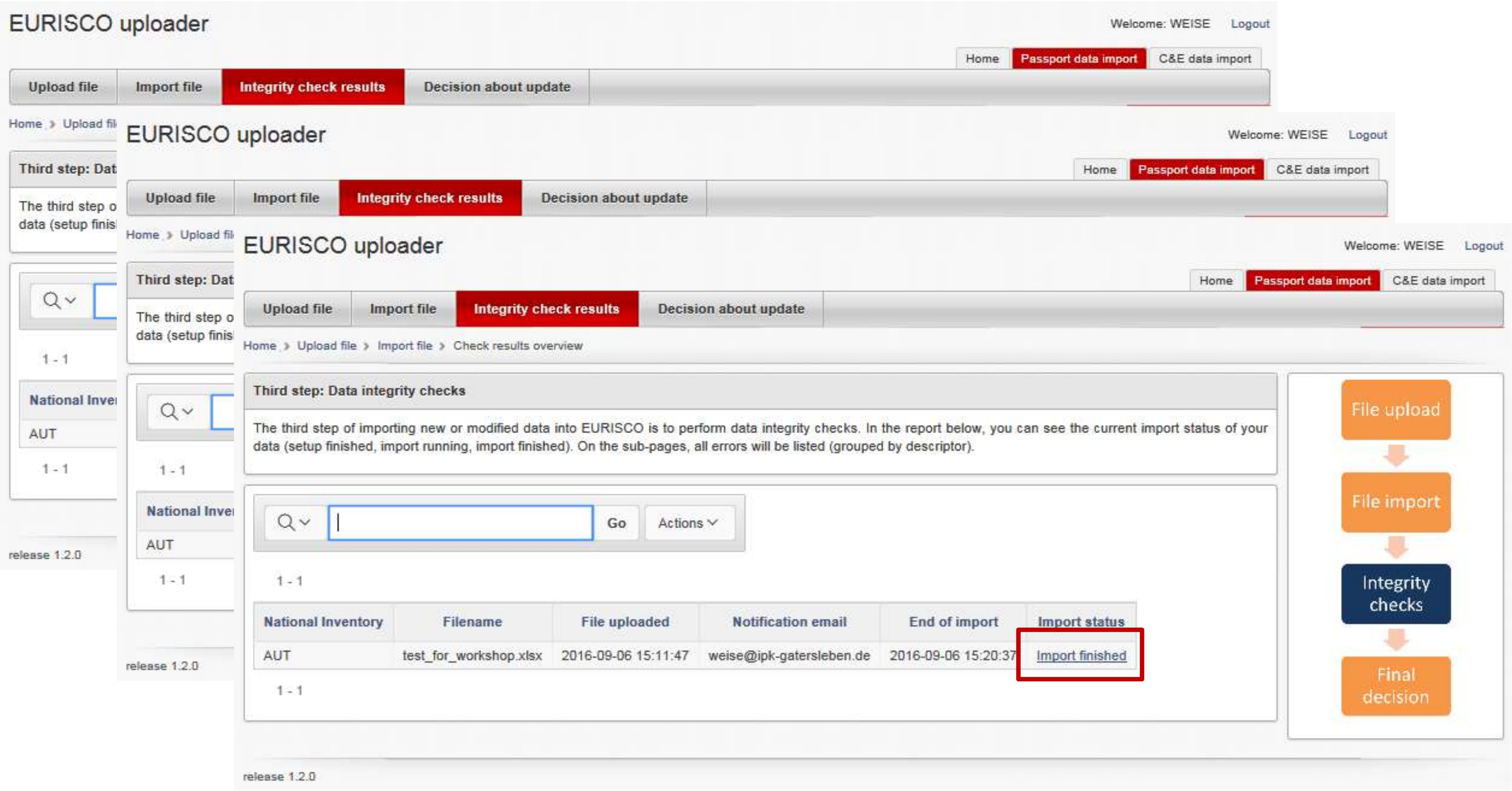
Records with errors are skipped:

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 › 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).

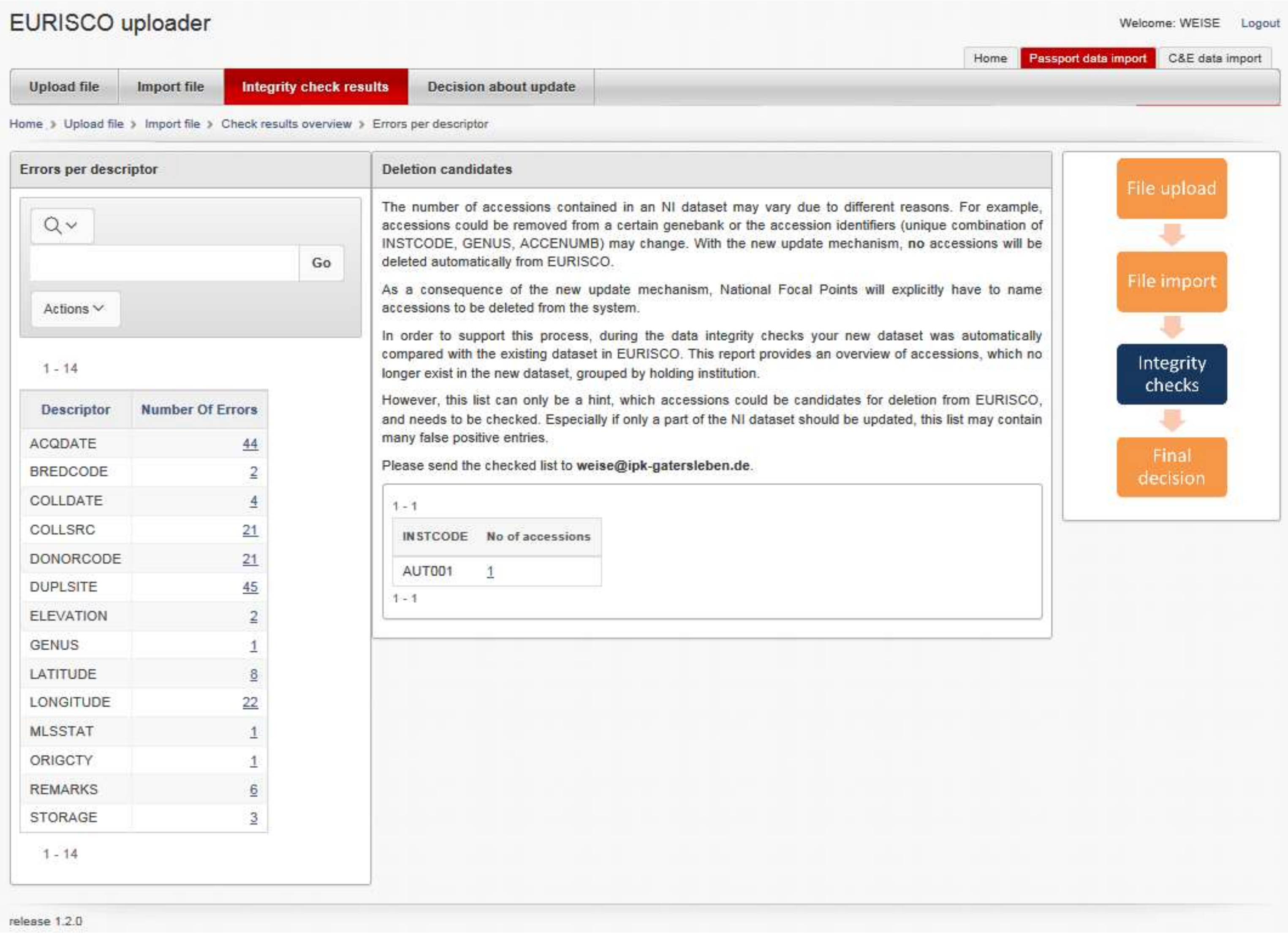
1 - 1

National Inventory	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

1 - 1

release 1.2.0

Integrity checks II – error report



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**.

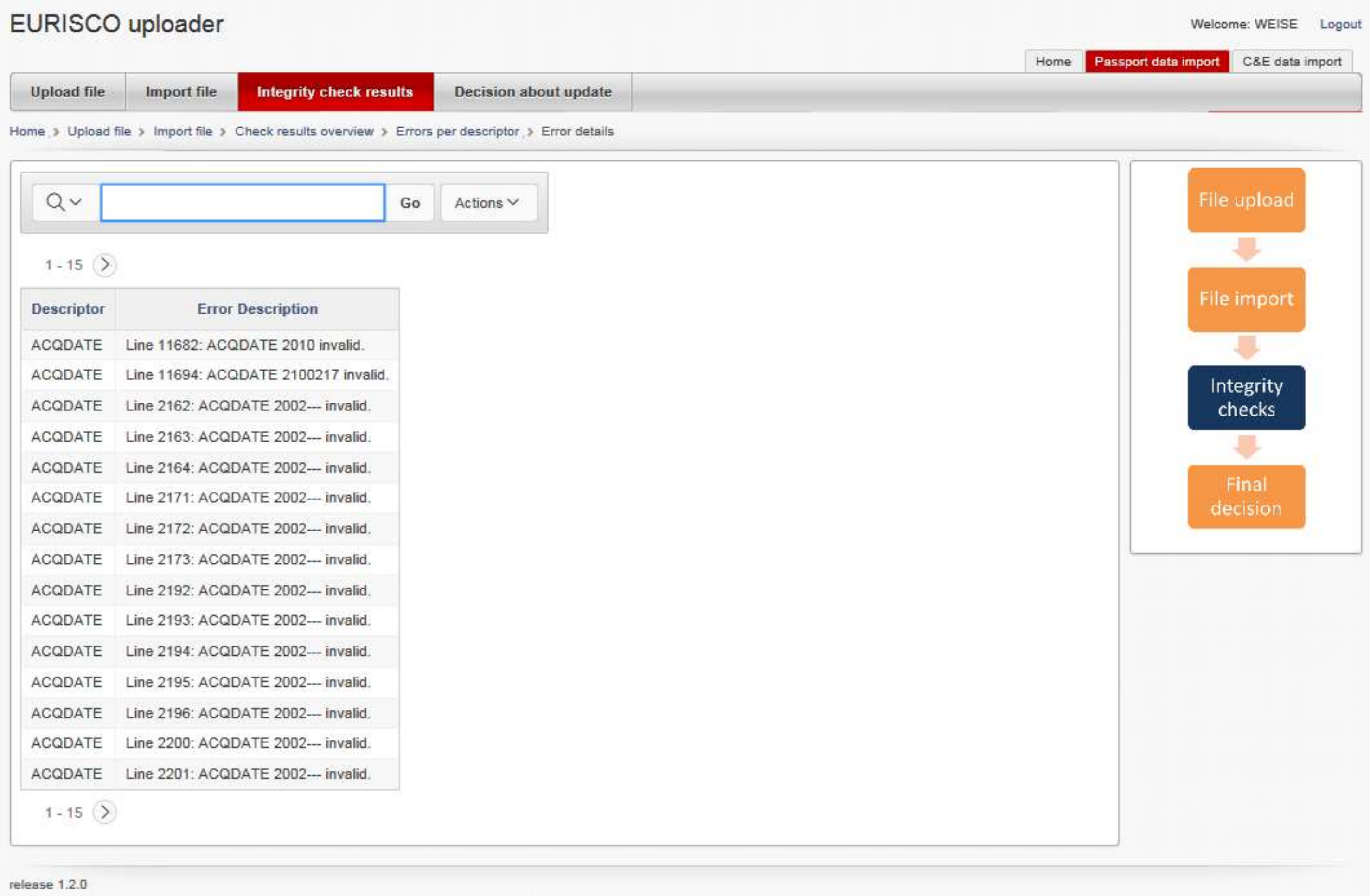
1 - 1

INSTCODE	No of accessions
AUT001	1

1 - 1

Integrity checks III - example

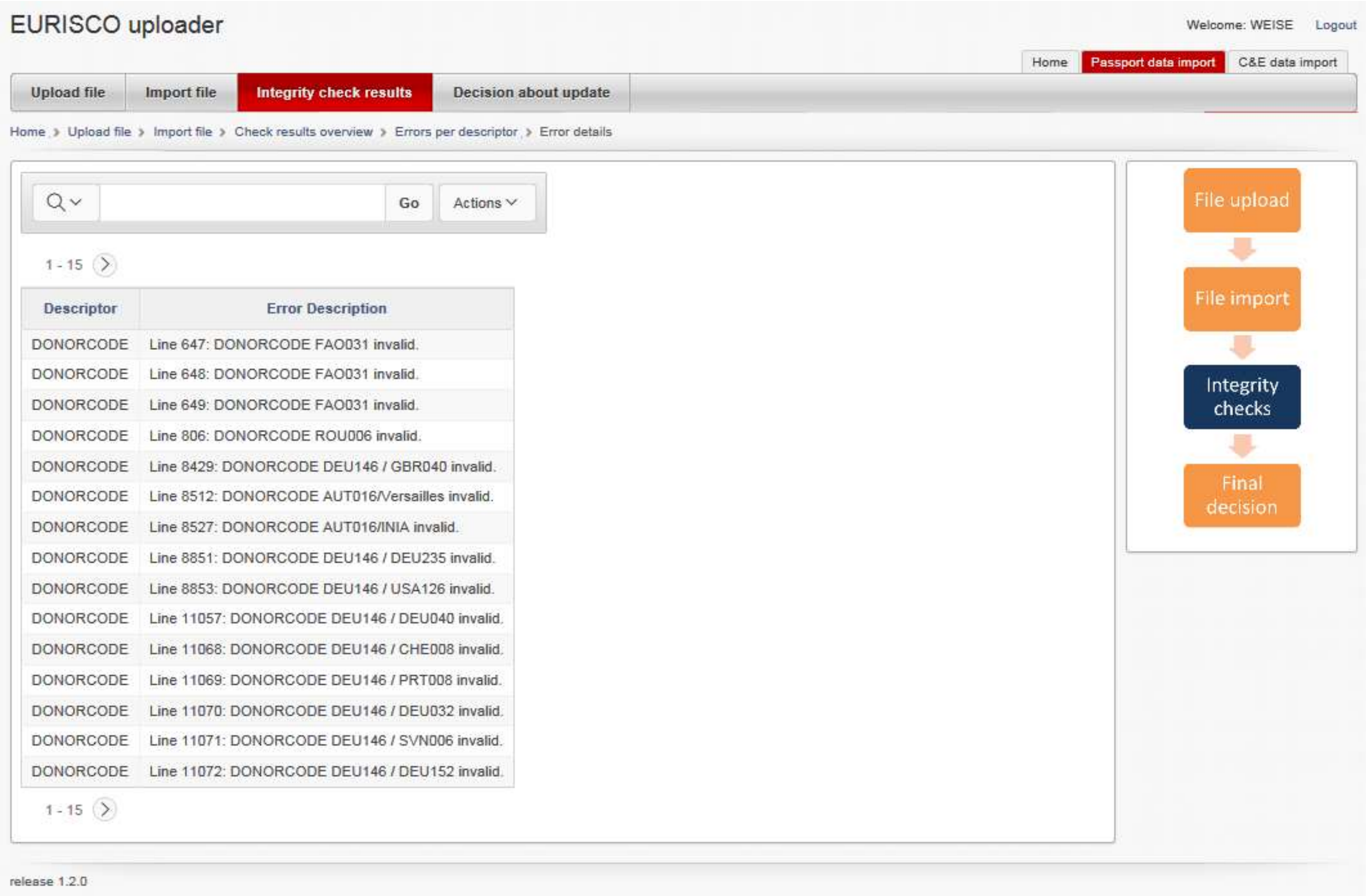
- Wrong date format



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.

Integrity checks IV - example

- Invalid or multiple donor codes



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

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 / CHE008 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

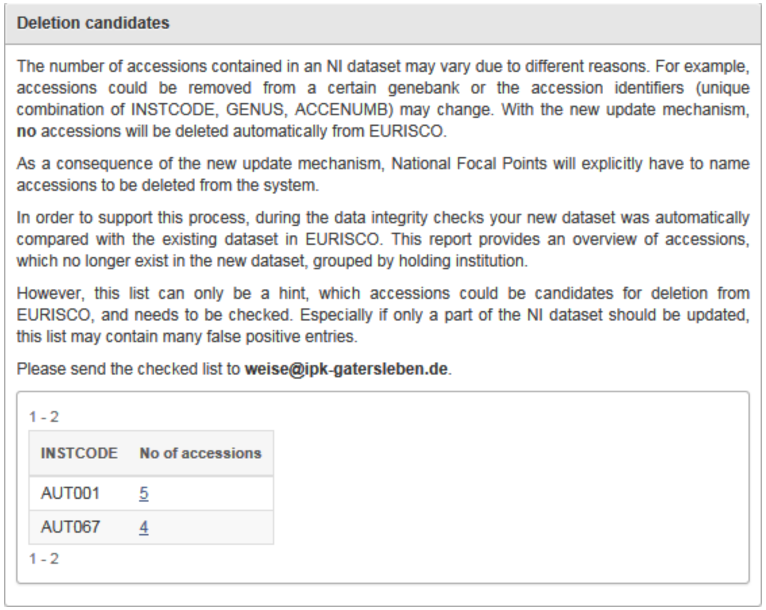
release 1.2.0

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
    - DOI infrastructure of ITPGRFA under preparation
  - 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.

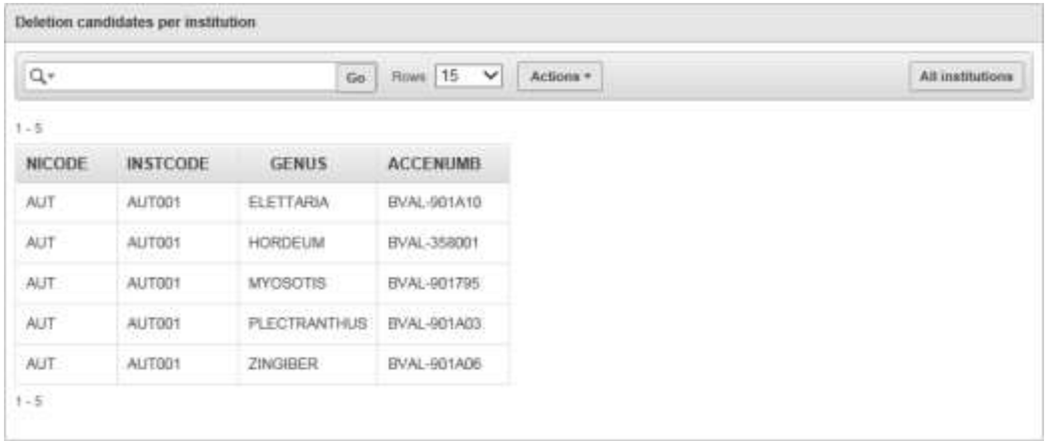
However, this list can only be a hint, which accessions could be candidates for deletion from EURISCO, and needs to be checked. Especially if only a part of the NI dataset should be updated, this list may contain many false positive entries.

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

1 - 2

INSTCODE	No of accessions
AUT001	5
AUT067	4

1 - 2



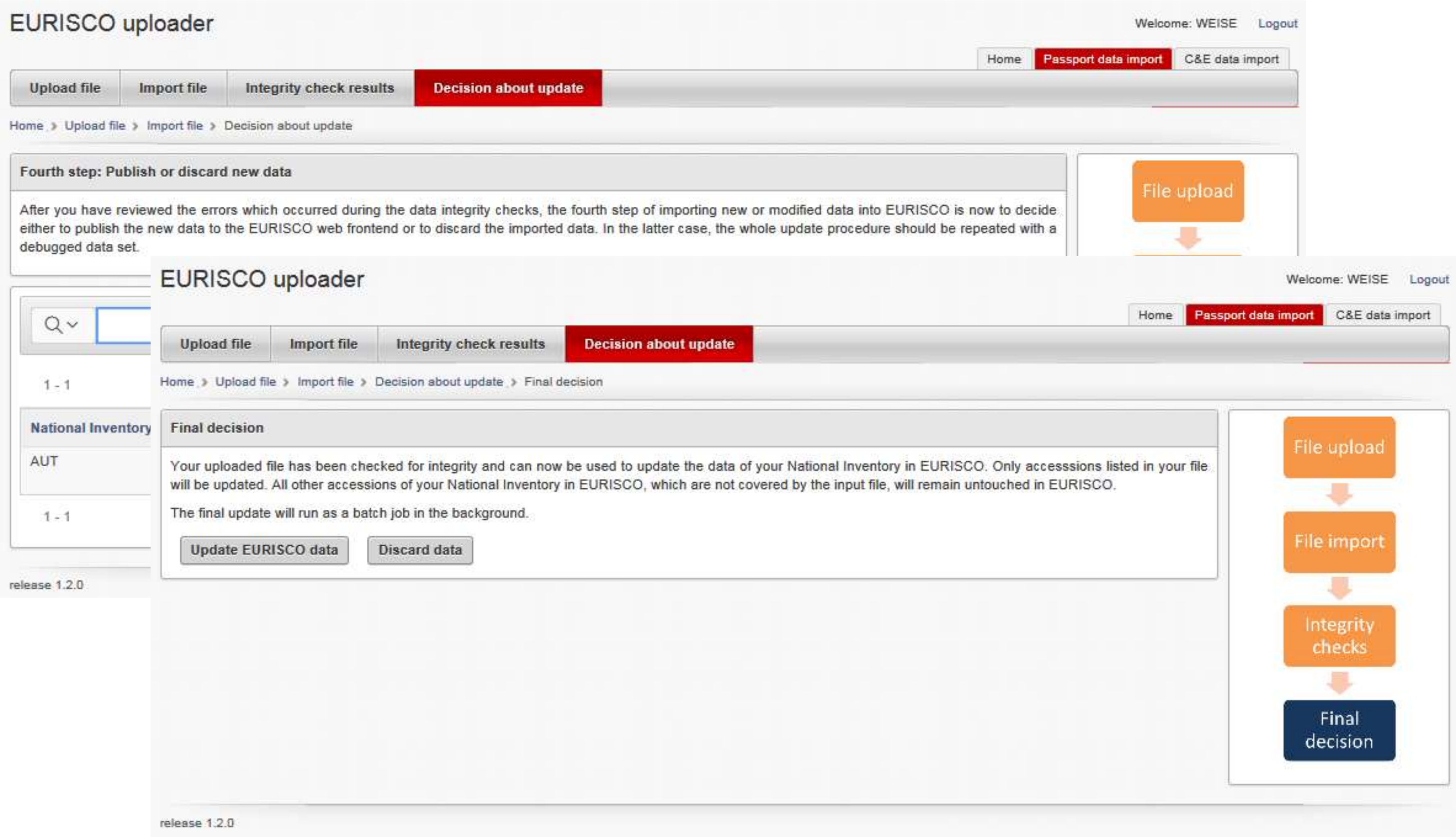
Deletion candidates per institution

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 › 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.

EURISCO uploader

Welcome: WEISE Logout

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

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 accesssions 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.

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

Migration to v2.1 of MCPD I

- Current data exchange format (May 2012)
  - MCPD v1
  - 8 additional, EURISCO-specific descriptors

- In the meantime, evolution of MCPD to v2.1 (Dec 2015)

→ Adaptation of the EURISCO exchange format
  - Harmonisation with MCPD 2.1
  - 4 additional descriptors

Migration to v2.1 of MCPD II

- New descriptors
  - PUID
    - Persistent unique identifier, e.g. DOI
  - COLLINSTADDRESS
    - Address of collecting institute
  - COLLMISSID
    - Identifier of collecting mission
  - DECLATITUDE
    - Latitude in decimal degrees
  - DECLONGITUDE
    - Longitude in decimal degrees
  - COORDUNCERT
    - Uncertainty of coordinates in metres
  - COORDDATUM
    - Geodetic datum or reference system, e.g. WGS84
  - GEOREFMETH
    - Referencing methos, e.g. GPS
  - HISTORIC
    - Accession maintenance status

Migration to v2.1 of MCPD III

- Modified descriptors
  - COLLCODE: multiple values allowed
  - DUPLSITE: multiple values allowed
  - BREDCODE: multiple values allowed

  - COLLNAME: replaces COLLDESCR
  - BREDNAME: replaces BREDDESCR
  - DONORNAME: replaces DONORDESCR
  - DUPLINSTNAME: replaces DUPLDESCR