Project:	EURISCO	Date:	2015-01-15

From:	Stephan Weise (IPK)	To:	Lorenzo Maggioni (ECPGR Secretary) Theo van Hintum (Documentation and Information Working Group)
CC:	Helmut Knüpffer (IPK), Markus Oppermann (IPK)		

This report describes the activities carried out in 2014 for hosting and maintaining the European search catalogue for plant genetic resources (EURISCO), and for coordinating the EURISCO network.

Content

1 EURISCO development

Since 15th April 2014, the responsibilities for hosting and maintaining EURISCO were being moved to the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany. After a period of transfer (April to September 2014), the system is now operating at IPK.

1.1 Transfer to IPK

Due to technical and organisational issues, the existing solutions could not be transferred to the new host as is. Instead, a reengineering from scratch was performed; this regards both to the public EURISCO application (search interface and database structure behind) and to the EURISCO intranet (module for uploading national data sets by the National Focal Points). Thus, it was necessary to find a balance between developments which are visible (search interface) and those which completely took place in the background. For this reason, compared to the previous installation at Bioversity, there will be certain functional limitations of the new system for a little while.

The following activities were carried out:

1.1.1 Frontend (public EURISCO application)

The frontend comprises both the public search interface and the database structure behind.

Due to insufficient documentation of the existing frontend, as a first step the previous search interface was analysed and its functionality was documented. This document formed the basis for the development of the new search interface.

Based on a MySQL database dump received from Bioversity, the database schema of the previous frontend was analysed and successively reengineered using the Oracle 11g DBMS. The new database schema comprises 29 tables, 18 materialised views and 421 indices for performance improvement, and four PL/SQL packages for data download, newsletter etc.

For the transfer, the EURISCO dataset was provided as a MySQL dump. Thus, a migration path to the new database schema (using Oracle DBMS) needed to be developed.

A new search interface was developed using the Oracle Application Express (APEX) technology (Fig. 1). The functionality of the search interface is based on the previous application. Currently, the functionality is limited and will be extended soon.

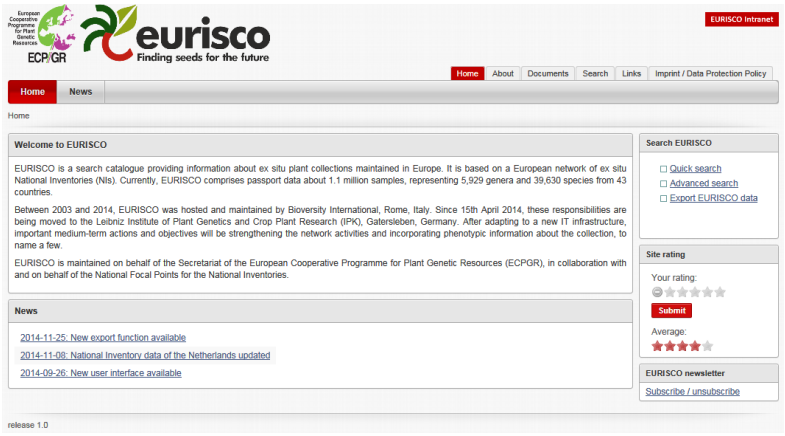

Fig. 1: Screenshot of the new public EURISCO search interface.

As a first completely new feature, a new export function has been implemented in the EURISCO web application, thus enabling the users to create customised export files for each National Inventory (NI) dataset. In addition, the whole EURISCO dataset can be downloaded as a pre-built dump.

1.1.2 Backend (EURISCO intranet)

The EURISCO backend comprises (i) a web application enabling the National Focal Points (NFP) to update their National Inventory datasets, (ii) the necessary database structure, and (iii) all packages and procedures necessary to upload, check and update the datasets.

As in the case of the public EURISCO application, the EURISCO intranet also had to be reengineered from scratch. Based on a textual description of the necessary steps, a new pipeline for updating the NI datasets was developed.

Inserting/updating data in EURISCO takes place in four steps (Fig. 2):

- First, a new data file is uploaded to the server.
- In a second step, the uploaded file is parsed and imported into a staging area.
- Afterwards, comprehensive integrity checks are performed.

- Finally, the results of the import and checking process are reported to the user who can then decide whether to publish or to discard the new data.

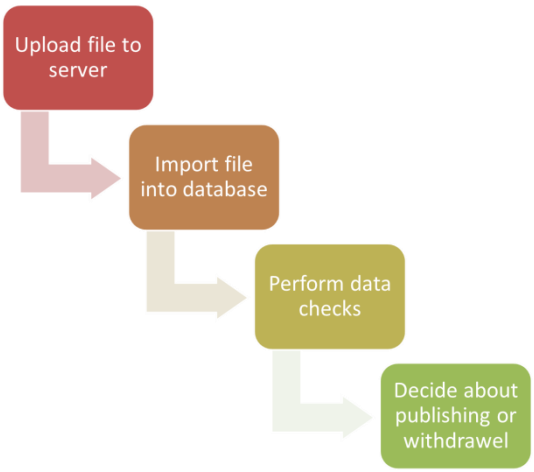

Fig. 2: Procedure of inserting/updating data in the EURISCO intranet.

During the reengineering of the EURISCO backend, the most important procedures for performing integrity checks and for updating were developed. However, as it used to be in the previous system, the integrity of taxonomic names is not yet checked. The development of checking procedures for the taxonomic nomenclature is to be planned for the future.

Once the user has decided to publish the new data (after successfully performing integrity checks), the dataset is synchronised from the backend database schema to the frontend database schema.

The EURISCO intranet web application was developed using the Oracle Application Express technology; for the data management, an Oracle 11g DBMS was used. The new database schema comprises 28 tables, 55 triggers, two views, 287 indices, and seven PL/SQL packages with 77 procedures for importing, checking and updating new datasets.

1.1.3 Data

After reengineering the target database schema, a first import was made in order to test the newly developed data migration path and data cleansing procedures. This (test) dataset was then used to develop both the new public EURISCO search interface and the new EURISCO intranet.

After successful tests, a productive import with the up-to-date dataset was performed. Since end of September 2014, the URL http://eurisco.ecpgr.org/ redirects to the new EURISCO system at IPK.

1.1.4 Documentation and planning of the next steps

In parallel to the implementation of the certain modules, all developments were documented.

The future development steps were planned and a work plan for 2015 was created (see section 2).

1.2 Extension for C&E data

Because of the implementation of the incremental updates (section 1.3), the discussion about the C&E data exchange format was postponed to 2015.

However, the respective proposals by Theo van Hintum (CGN)[1] and Jonas Nordling (NordGen)[2] were viewed and first ideas were developed.

1.3 New upload mechanism

During the EURISCO transfer process, the development of a new upload mechanism with only limited functionality was planned, which needs to be extended.

1.3.1 Direct import of data from MS Excel files

While preparing tabulator-separated text files in EURISCO format, often encoding errors occur. Thus, it was planned to test existing software libraries regarding their applicability for importing MS Excel files containing passport data and C&E data, respectively. Here, the focus will be on SAX-like parsers, thus enabling data imports with low resource requirements. The development of the MS Excel importer for passport data was started and will be continued in 2015.

1.3.2 Incremental updates

In the previous system, no incremental updates of EURISCO data took place. Instead, only full replacements on NI level were performed. This means, in case of an update the whole data set of an NI was first deleted and then replaced by the reworked whole data set of this NI. This proceeding resulted in two disadvantages:

- This solution was not efficient, especially if only a small number of accession records needed to be updated while the other data remained unchanged.
- All data belonging to an NI needed to be updated at once (all or nothing). It was not possible to successively update data of different collections (e.g. gene banks), which all belonged to the same NI.

Moreover, EURISCO will be extended for the inclusion of C&E data soon. Since C&E data is directly connected with the passport data of an accession, the above proceeding would imply that for each update of the passport data also all C&E data needs to be deleted and reimported.

Thus, although it was originally not planned for 2014, incremental updates were given a priority for implementation. Henceforward, National Focal Points should only send data that was really changed — additional accessions and existing accessions with modified content. New accessions will then be inserted while existing accessions will be updated. If necessary, the update will of course also work with the full data set of an NI. Following this procedure, no accessions will be deleted automatically. In the future, National Focal Points will explicitly have to name accessions to be deleted by a simplified EURISCO format.

2 EURISCO coordination

In 2014, the main focus of the activities was on the transfer of the EURISCO system to its new host (section 1). However, a number of activities regarding the EURISCO coordination were performed, which are described in the following:

[1] Inclusion of C&E data in EURISCO – analysis and options
[2] C&E data: the EURISCO standard

For 2015, a EURISCO regional training workshop is planned, with focus on ECPGR member countries from South-Eastern Europe. A proposal for financing was prepared and submitted to the ECPGR Activity Grant Scheme (Phase IX). End of December 2014, the proposal was accepted.

All National Focal Points and National Coordinators as well as all members of the ECPGR working groups were contacted by email. A short report about the status of the EURISCO transfer was given and all recipients were asked for their input for the future development of EURISCO. The aim is to shift the development of EURISCO towards a demand driven development in order to improve the acceptance and usability of EURISCO.

Direct, personal communication took place with several National Focal Points and National Coordinators (e.g. support for updates, provision of specific database queries and special data export formats, discussion about future developments).

Currently, the user's feedback for the new public search portal is collected. All feedback regarding improvements or possible extensions will be checked for feasibility; a priority list for the implementation will be created together with the EURISCO steering committee.

The EURISCO coordination provided substantial legwork to the preparation of project proposals (H2020 and others) in order to acquire additional funding for the future development of EURISCO.

A presentation on EURISCO was given at the workshop "SEEDNet – the Way Ahead" (Ljubljana, Slovenia, 5-6 November 2014).

Finally, an activity report for 2014 and a work plan for 2015 were prepared.