ViBRANT

Virtual Biodiversity

Milestone M4.35

Filter functionality CDM export

Leading partner: BGBM

Compiled by: Andreas Muller, Lorna Morris and Sybille Bürs

Date: June, 30th 2013

Introduction

The aim of this milestone is to add filter functionality to csv export from the CDM. The functionality to export a Darwin Core Archive (DwC-A) from a CDM database has been implemented in an earlier release of the CDM. The export procedure creates a zip file containing multiple csv based text files for different data types and a meta.xml describing the data in each text file. Filter functionality will make it possible to export a specific sub-tree, i.e. a selected Taxon node and all its children from a CDM database. To further improve the interoperability between Scratchpads and the CDM we test import of the CSV data into Scratchpads. Scratchpads do not currently support importing of DwC-A files. However there is an interface to import Excel files [1], therefore we used this interface to import the data into Scratchpads.

CDM export to DwC-A

The CDM supports DwC-A as an export format. DwC-A is primarily a csv export format and therefore suitable for the aim of this milestone. However, to fully meet the requirements the existing export had to be adapted such that it:

- allows unzipped results
- changes the headers of the csv export files according to the Scratchpads the requirments for Scratchpads import
- filters the data such that only those data are exported which can be consumed by Scratchpads (Scratchpads only support a subset of data supported by the CDM) , unsupported datatypes will not be exported to avoid confusion
- change values or even split columns so they can be understood by the scratchpads import format (example: the taxonomic status column needed to be split into 2 columns *Usage* and *Unacceptability reason* where *Usage* defines if a taxon is accepted or not and *Unacceptability reason* is further defining the reason why a taxon is not accepted, also *Unacceptability reason* uses a different vocabulary then DwC-A for the taxonomic status.

A zip/unzip functionality has been added to the DwC-A export which allows the user to specify if the want the export result to be zipped in the DwC-A format or unzipped. Also an additional functionality adaptColumnsToScratchpads has been implemented.

In addition to the above mentioned required adaptations of the DwC-A export format and the DwC-A export routine, the ability to apply content related filters may also be desired. Users may often want to reuse only a certain taxonomic or geographic subset of the existing data (e.g. all data related to a certain genus or to a certain country or continent). However, until now it has been only possible to export an entire CDM database. Exporting a subset that matches conditions as described above was not possible.

Filtering could also be implemented during the subsequent import phase. However, this functionality is not available for the Scratchpads import and doing it manually in the export files would be time consuming and error prone. Also removing the unwanted data after importing them is usually difficult to achieve and will rarely result in a complete deletion of all unneeded data. Therefore it was decided to filter the data during the export phase.

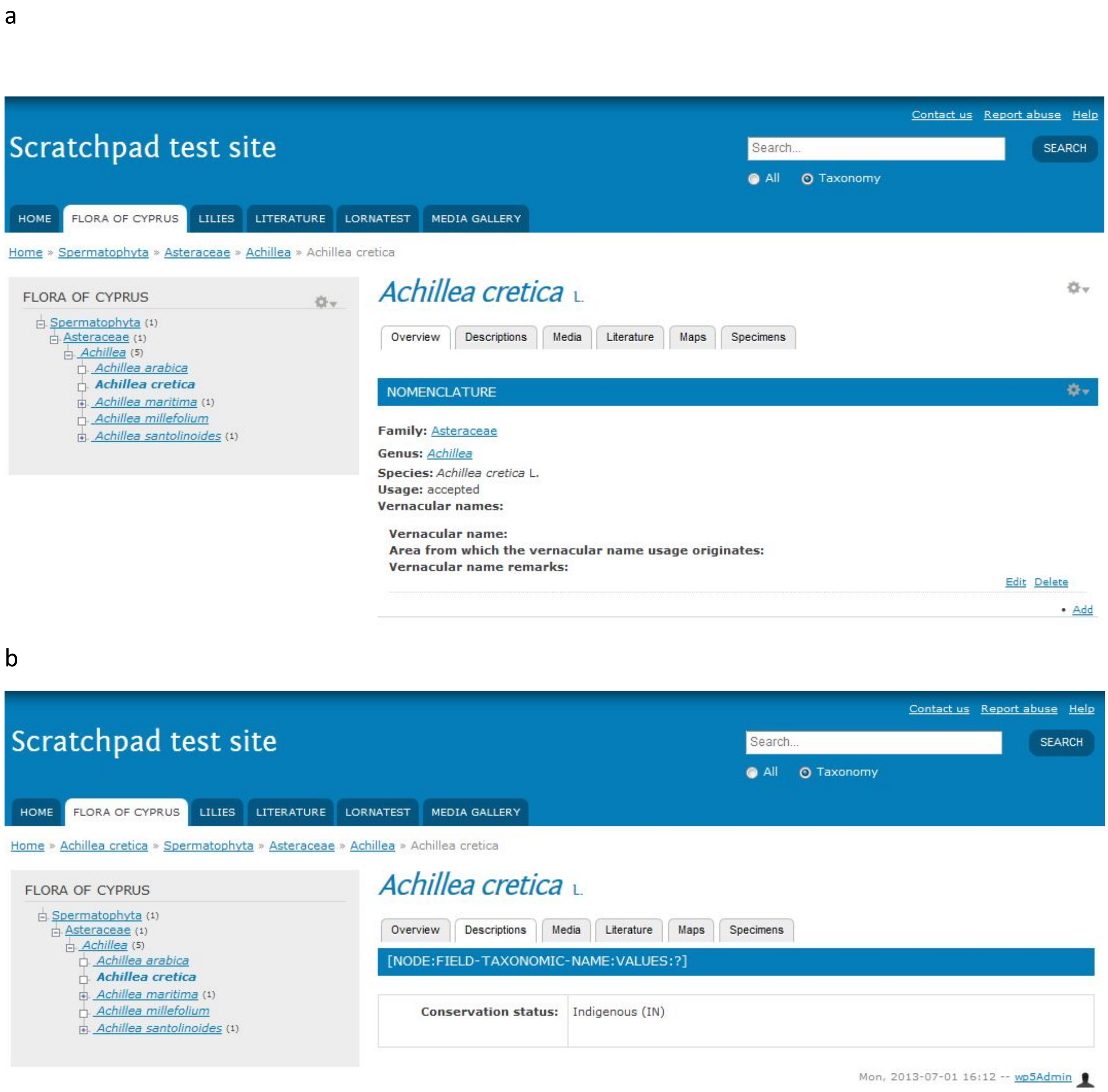
Therefore a filtering mechanism for the Dwc-A export has been implemented in the CDM. To make this filtering as generic as possible a filter interface *ICdmExportFilter* and an implementation for this interface *ICdmExportFilterBase* has been developed which can be added to any DwC-A export configurator (*DwcaTaxExportConfigurator*) and can be reused also by other CDM export routines.
As the CDM DwC-A export routine uses separate classes for the creation of each csv file the filter class can be applied to each of these classes, so not only the core csv file holding the taxonomic backbone but also the so called Dwc-A extension files are filtered in the same way and the result will be consistent.

The filter mechanism is also extendable, so whenever new filter requirements are raised by users they can be added to the existing filter class without having to refactor the whole system.
DwC-A export functionality can be invoked either by using the Taxonomic Editor export functionality or via command line. For the time being the additional functionality has only been added to the command line export functionality. A new indexing structure for the taxonomic classification has recently been implemented in the CDM. This indexing structure when used in the taxonomic filtering mechanism will result in a much better performance. To avoid “unpleasant” user experience due to missing indexing mechanisms it was decided to publish the new functionality in conjunction with the new index in the CDM release planned for the beginning of August 2013. However, the functionality can already be used or tested using the command line option.

Import of CDM data into Scratchpads

Importing of the individual text files from CDM Dwc-A export via the Scratchpads import menu needs to be carried out in a specific order. First a classification needs to be imported [2]. There is a mapping between some of the fields required in this file to the core classification file in the DwC-A. We modified the DwC-A classification file to correspond to the required format in order to enable import of the data into Scratchpads. Each Taxon (row) in the classification file has a unique identifier (uuid) from the CDM database. After this taxonomy data has been imported into Scratchpads, we can then import other data to link to the Taxon names imported, via the uuid. For example, from the Scratchpad import menu, we can select a Nodes import with the content type set to 'Taxon Description' and import the description data, consisting of one of more fields of different description types (e.g. general description, ecology) linked to the taxon uuid. Figure 1 shows a screenshot of Flora of Cyprus data imported from the CDM (2 files were imported, a taxonomy file filtered for the genus Achillea) and a file containing the description data for these taxa). Similarly we can import the Specimen data text file, which contains a uuid to link it to the taxonomic name. The Scratchpads data type 'Location' contains distribution data, e.g. ISO-country code, coordinates. This file does not link directly to the taxon uuid, it links to the Specimen (via a location Id in the specimen text file). As our implementation is based on a DwC-A text files from the CDM, we do not have this information, without first querying the Scratchpads database, so we leave this field blank in our import.

Scratchpads do not allow linking to external images. To link images to taxonomic names the images have to be first imported into the Scratchpad, and then a prepopulated template file can be exported which the user can then edit to link image names with taxonomic names. Therefore we didn't implement the import of images and the linking to taxa. We also didn't support the import of bibliographic data, as this can't be done in Scratchpads by csv import, but requires other formats such as Bibtex or EndNoteXML. The core classification file can however contain the author and title of the nomenclatural reference.


a

Scratchpad test site

Contact us Report abuse Help

Search... SEARCH

● All ⊙ Taxonomy

HOME FLORA OF CYPRUS LILIES LITERATURE LORNATEST MEDIA GALLERY

Home » Spermatophyta » Asteraceae » Achillea » Achillea cretica

FLORA OF CYPRUS

- Spermatophyta (1)
  - Asteraceae (1)
    - Achillea (5)
      - Achillea arabica
      - **Achillea cretica**
      - Achillea maritima (1)
      - Achillea millefolium
      - Achillea santolinoides (1)

Achillea cretica L.

Overview Descriptions Media Literature Maps Specimens

NOMENCLATURE

Family: Asteraceae
Genus: *Achillea*
Species: *Achillea cretica* L.
Usage: accepted
Vernacular names:

Vernacular name:
Area from which the vernacular name usage originates:
Vernacular name remarks:

Edit Delete

- Add

b

Scratchpad test site

Contact us Report abuse Help

Search... SEARCH

● All ⊙ Taxonomy

HOME FLORA OF CYPRUS LILIES LITERATURE LORNATEST MEDIA GALLERY

Home » Achillea cretica » Spermatophyta » Asteraceae » Achillea » Achillea cretica

FLORA OF CYPRUS

- Spermatophyta (1)
  - Asteraceae (1)
    - Achillea (5)
      - Achillea arabica
      - **Achillea cretica**
      - Achillea maritima (1)
      - Achillea millefolium
      - Achillea santolinoides (1)

Achillea cretica L.

Overview Descriptions Media Literature Maps Specimens

[NODE:FIELD-TAXONOMIC-NAME:VALUES:?]

Conservation status:	Indigenous (IN)

Mon, 2013-07-01 16:12 -- wp5Admin

Figure 1: Screenshot showing CDM data from the Flora of Cyprus (http://www.flora-of-cyprus.eu/) database (genus Achillea) after import into Scratchpads. The taxonomic tree is on the left (a) shows the overview data in the right-hand panel (from the imported classification csv file), (b) shows the associated Description data (from the imported description csv file).

Further work

If there is user demand a Drupal module to automatically import a DwC-A into Scratchpads could be developed using the Drupal Feeds module. This would allow CDM users to view and manipulate a filtered selection of their CDM data within Scratchpads and open up the possibility of integrating this data with data from other sources within Scratchpads.

Further output filter functionality will be created using the generic filter classes in the CDM depending on user requirements.

Finally it is planned to also allow MS Excel based export from the CDM. Excel is the primary import format for Scratchpads, via the import in the Scratchpads administration menu. Excel files can be easily created from csv files. However, for convenience a direct export into Excel files is the preferred solution.

References

1. http://help.scratchpads.eu/w/Import
2. http://help.scratchpads.eu/w/Import_your_own_classification