Milestone M4.24

Full text search of CDM-ViBRANT Index

Leading partner: BGBM

Compiled by: Lorna Morris

Date: September 2012

Introduction

For milestone M4.18 we created a simple web-based user interface to enable querying of distinct scientific names in the ViBRANT index database [1]. The ViBRANT index database is a Common Data Model (CDM) based database consisting of taxonomic data imported from a variety of sources (currently Scratchpads, MedChecklist and EuroPlus Med).

We have now extended our web-based user interface to enable full text searching of data.

Implementation

Web service methods were implemented in the CDM to enable full text searching using Lucene [2]. The web service method employed searches for the query string in the Taxon name and in the associated descriptive text.

Lucene allows the building of text documents combining multiple fields, which are distributed in the CDM object graph and thus are searchable very quickly without the need to join multiple tables. Lucene is used to build a query that searches the lucene index for the description element textual description and the Taxon name.

The lucene search returns the indexed document and also a score, which gives an idea of the relevance of each document in the result set.

The HTML/XSLT query interface for searching for distinct scientific names [1] was extended to call the full text search web service. XML generated from the web service query was transformed to generate web pages displaying the search results.

The source code for the web application is available from EDIT's subversion repository:

http://dev.e-taxonomy.eu/trac/browser/trunk/cdmlib/cdmlib-remote-webapp/src/main/webapp

User Interface

The service is available within the BGBM at:

http://160.45.63.201/vibrant_index/vibrant_names.html

The service is available externally (it uses the Cichorieae database [7] as currently licensing restrictions prevent us from making the test data-sets available externally):

http://dev.e-taxonomy.eu/vibrant_index/search/

Screenshots of the user interface describing a simple query scenario are shown in figures 1 and 2.

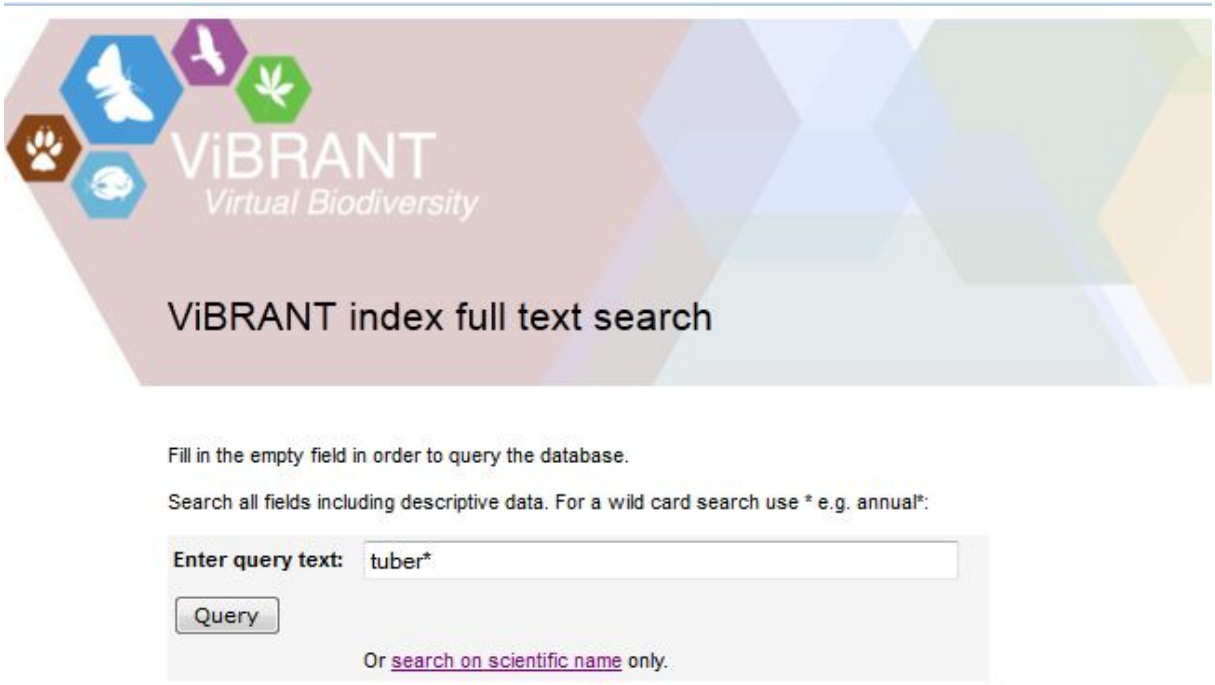

Figure 1: Screenshot of the form for querying the ViBRANT index for a full text search. The example shows a wild-card search, which includes an asterisk after the search term i.e. to search for any term beginning with 'tuber'.

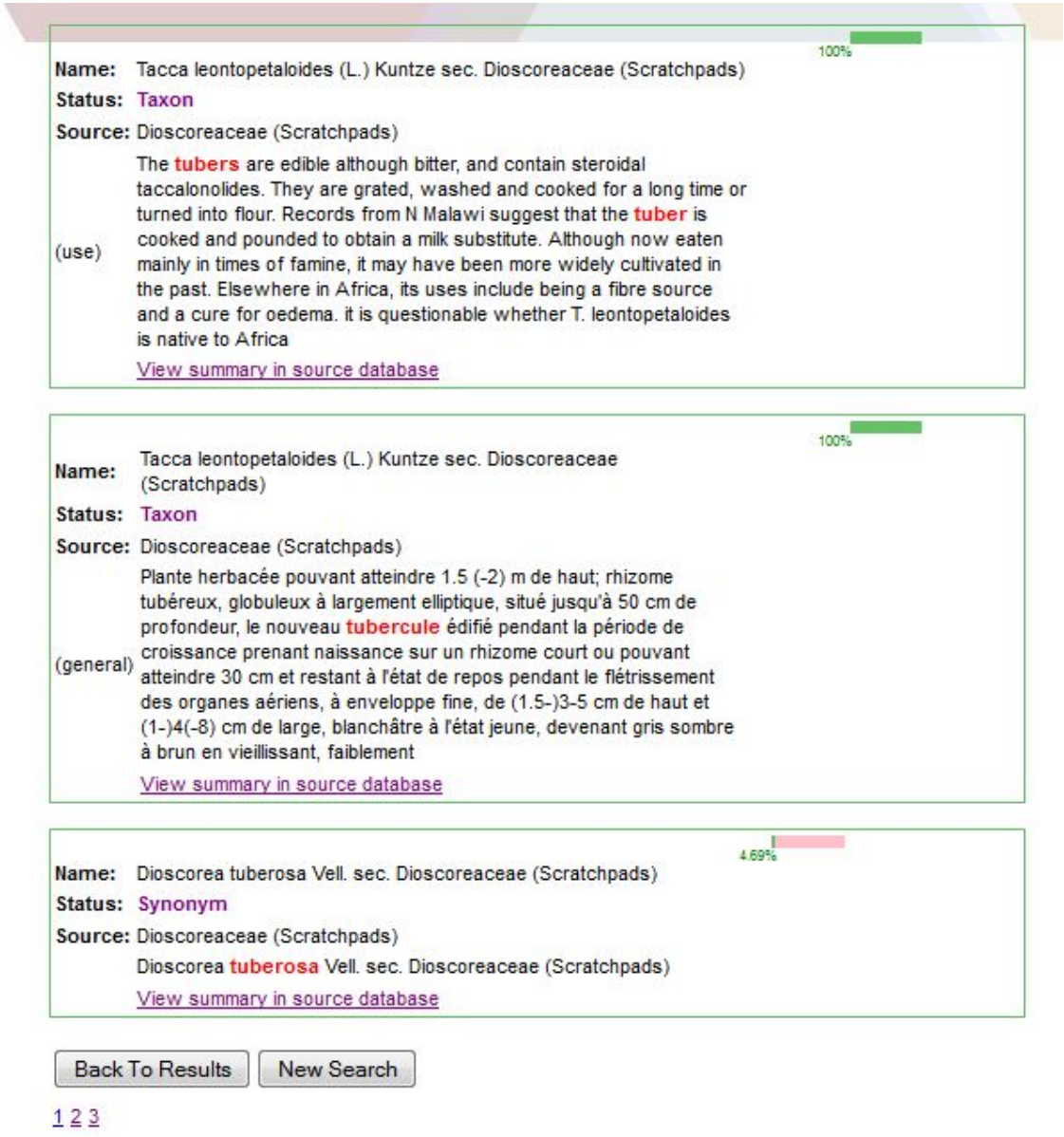

Figure 2: Screenshot showing page 3 of the results returned from the search shown in figure 1. The user is presented the results in order of their score and the search term is highlighted in red. If the search term is found in the description field the type of descriptive data is shown in brackets (e.g. usage, general description).

In figure 2 the same Taxon has returned twice as the search term 'tubule' has been found in 2 different descriptions (one of type:use the other of type:general). This is represented in the maximum score value of 100% for these hits. We are currently using lucene 2.4 but since lucene 3.2 it is possible to group results and therefore when we upgrade to this version of lucene we can reflect the grouped results in our user interface. The third hit in figure 2 shows that the search term has been found in the taxonomic name field.

Testing

To test the interface we used a database containing data for 3 different source databases (Euro+Med PlantBase [5] and Med-Checklist [4] and the Dioscoreaceae [3] dataset exported from Scratchpads2).

Further work

We plan to import taxonomic data from all available Scratchpads2 sources into the CDM ViBRANT index database via the tools to export and import DwC-A data from Scratchpads [6].

Further work is needed to create the Scratchpads URLs that link the Taxon back to the Scratchpads2 source. Source URLs for individual Taxons require a term identifier:

http://dioscoreaceae.e-monocot.org/taxonomy/term/1776

The term identifier is absent in the imported DwC-A file, so further modifications need to be made to the DwC-A export to make this available [6]. The DwC-A export files currently use the UUID to link data between the Taxon and associated extension data. For the additional identifiers we can use the 'Alternative Identifiers' Extension (http://rs.gbif.org/terms/1.0/Identifier). We can map DwC-A terms to this view using the flexible mapping module [6] or just provide this extension in the default meta.xml that accompanies the module.

We plan to extend the user interface to enable the user to filter the results e.g. on description type (Scratchpads description type data include for example: general description, distribution, conservation). We have corresponding Feature types in the CDM and therefore this information is already available in the CDM import of the Scratchpad data.

References

1. http://vbrant.eu/sites/vbrant.eu/files/ViBRANT_M4.18_%E2%80%94_Human_Interface_for_CDM-ViBRANT_index.pdf
2. http://lucene.apache.org/
3. http://dioscoreaceae.e-monocot.org/dwca.zip
4. http://ww2.bgbm.org/mcl/home.asp
5. http://ww2.bgbm.org/EuroPlusMed/query.asp
6. http://vbrant.eu/sites/vbrant.eu/files/ViBRANT_M4.23.pdf
7. http://wp6-cichorieae.e-taxonomy.eu/portal/