
# Milestone M2.18

## Data search portal providing a single point of entry to all Scratchpad data

**Leading partner**: BGBM

**Compiled by**: Lorna Morris, Andreas Müller, Sybille Bürs

**Date**: 14th August 2013

## Introduction

The aim of this milestone is to import taxonomic data from all available Scratchpads into the CDM via the tools to export Darwin Core Archive (DwC-A) data from Scratchpads and import DwC-A files into the EDIT Common Data Model (CDM). We have previously implemented a DwC-A export module for Scratchpads, which has been further developed by Edward Baker and Simon Rycroft at the NHM [1]. Software to import DwC-A files into the CDM has also been implemented.

The data search portal provides a single point of entry to all Scratchpad data and provides a means for a user to search all Scratchpads indexed in the EDIT Common Data Model (CDM) and link back to the individual Scratchpad entry to view the full dataset. We use Apache Lucene [2] for indexing to allow rapid searching. The portal may be a useful entry point for a user to search for taxonomic data if they don't know which of the individual Scratchpads sites to go to. Harvesting the DwC-A files allows us to validate and check for any inconsistencies in the data and feedback the results to the individual Scratchpad site. Furthermore once the data is in the CDM, all of the CDM web-services can be applied to it, including the workflow-enabled services layer developed in the context of the EU projects i4Life and BioVeL [3]. The imported data can be manipulated in the EDIT Taxonomic Editor where further structure could be added to the data.

Our implementation and results of the harvesting are described below.

## Implementation

There are 3 steps in the harvesting process and creation of the index:

1. Check whether there is a DwC-A (dwca.zip) available at each of the Scratchpad sites. A Java service was implemented, which checks the Scratchpads site list JSON end-point [4] to retrieve the URLs of all the Scratchpad sites. A dwca.zip file is downloaded if available.

2. Data from each dwca.zip is imported into an empty CDM database. For this milestone, we focused on importing data contained in the Taxon core (classficiation.txt) and the reference and description extensions.
3. Trigger Lucene indexing of the entire database. This creates a Lucene index of 5 classes, including Taxon name data, description, specimen and observation data

We have implemented software in Java to import data in DwC-A format into the CDM. The code is available from the EDIT subversion repository:

http://dev.e-taxonomy.eu/svn/trunk/cdmlib/cdmlib-io/src/main/java/eu/etaxonomy/cdm/io/dwca/in/

The code to download to retrieve the dwca.zip files from Scratchpads is also available from the EDIT subversion repository:

http://dev.e-taxonomy.eu/svn/trunk/cdmlib/cdmlib-io/ cdmlib-ext\src\main\java\eu\etaxonomy\cdm\ext\scratchpads

We previously developed a user interface to enable users to search this index [1]. We have developed a toolbox of REST web services to enable a variety of search functionality, including name search, Taxon search and a web-service to generate statistical counts showing the number of objects (e.g. Taxon names, references) contained in the database or in a selected taxonomic classification. Example requests for the statistics service are documented on the EDIT Platform web-site [5]. To request all statistics for a particular Scratchpad classification (e.g. Amaryllidaceae), the following url pattern is used:

<host>/statistics.json?classificationName=Amaryllidaceae (Scratchpads)

# Results

## Harvesting of available Scratchpad data

We downloaded dwca.zip files from 74 scratchpad sites. The zip files were imported into an empty CDM database. A classification was created for each Scratchpad site in the CDM. A total of 61 classifications were created in a single CDM instance (this was lower than the number of available dwca.zip files as 4 classification.txt files were empty and there was an error opening several of the zip files). The data can be browsed after import in the EDIT Taxonomic Editor (figure 1a and 1b). The import was carried out in multiple batches, with 10 zip files in each batch as we found that that the import slowed down the greater the amount of data imported.

After import was complete the Lucene index was created. Lucene indexing creates text documents which combine multiple fields, which are distributed in the CDM object graph and thus are searchable very quickly without the need to join multiple tables. Five CDM classes were indexed: TaxonNameBase, TaxonBase, DescriptionElementBase, DescriptionBase, SpecimenOrObservationBase. These Lucene indexes can be used for name and full text searching and allow rapid retrieval compared to searching the CDM database directly.

The total MySql database size was 997MB and the Lucene index was 3.3GB. It took 70 minutes to generate the Lucene index for this database (with maximum heap set to 3GB, on an Intel Core i3-2100 3.10GHz processor with 4.00GB RAM). The total number of taxonomic names in all classifications is 228,429.

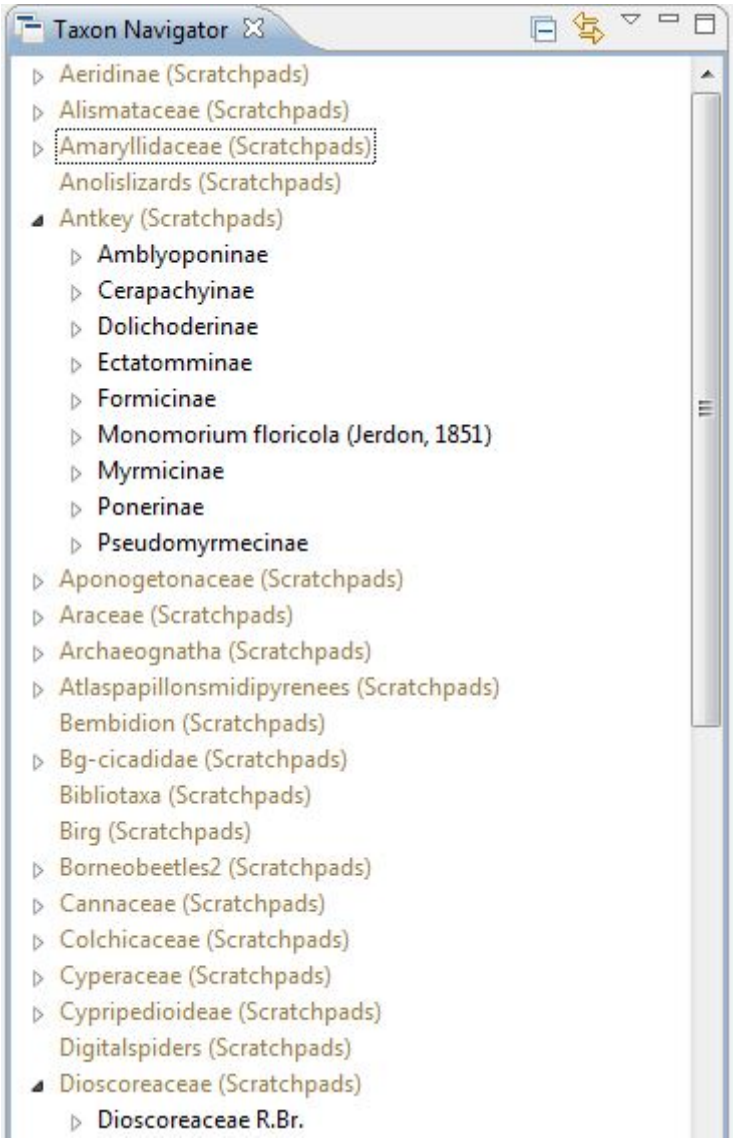



**Figure 1(a) An example of Scratchpads data in the EDIT Taxonomic Editor,** showing the taxonomic navigator panel containing several classifications from the imported Scratchpads.

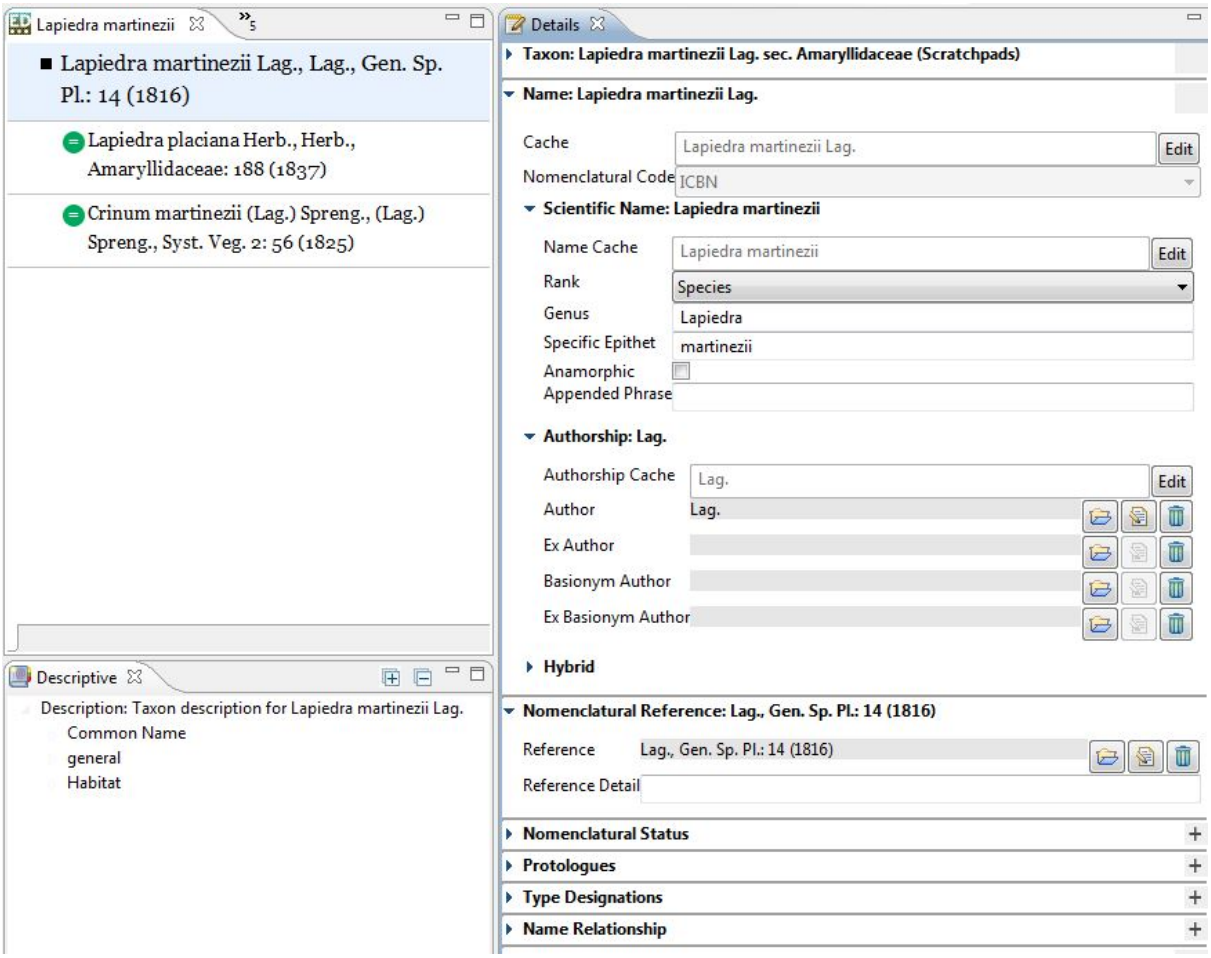



**Figure 1(b) An example of Scratchpads data in the EDIT Taxonomic Editor, for *Lapiedra martinezii* from [http://amaryllidaceae.e-monocot.org/](http://amaryllidaceae.e-monocot.org/).** The panels show the name (indicating 2 synonyms), the details view (showing the nomenclatural reference) and the descriptive view showing the headings for the associated descriptive data.

## Validation of data during import

During the import the DwC-A files were validated and data which was problematic or could not be imported was recorded in the log files. The DwC-A standard is interpreted differently by different people, so this milestone involved discussion between NHM and BGBM in how data could be mapped between Scratchpads and the CDM and the appropriate DwC-A term to use for the data. For example Scratchpads have 2 fields, 'Usage' and 'Unacceptability Reason' where synonym data is recorded, however in the DwC-A standard the data is represented by one term (taxonomicStatus [6]). Edward Baker at the NHM is modifying the DwC-A export module to ensure this data appears in future DwC-A exports from Scratchpads.

A common problem detected during CDM import was that a taxon had an ambiguous taxon status, which occurred when the taxon pointed (via the acceptedNameUsageId [7]) to another taxon which was also 'not accepted'. Another problem was that a taxon or reference pointed to a taxon (id) that did not exist. In these cases we ignore import of these rows. In the reference.txt file some of the rows had an empty taxon id, these references could be filtered from the Scratchpad view so they are not exported.

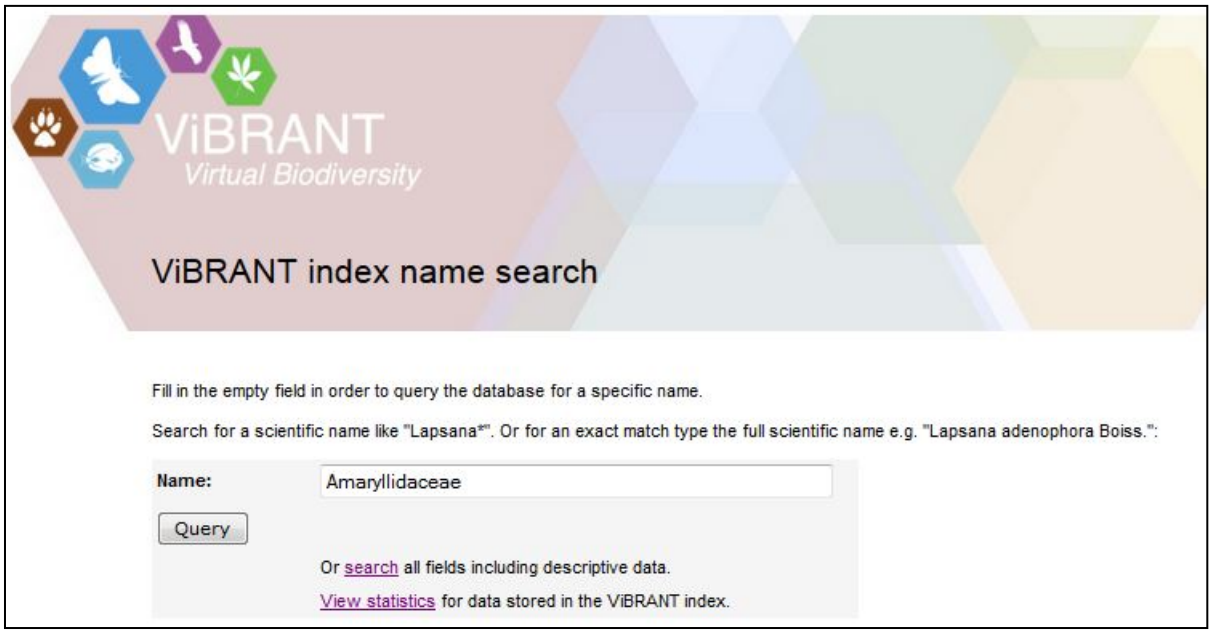
## User Interface for CDM –ViBRANT index

We will make the VIBRANT-CDM index interface available externally at the following URL after the next CDM release (September 2013) providing there are no licensing issues with the Scratchpad data displayed:

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

Screenshots showing examples of searching the ViBRANT index are shown in figure 2, 3 and 4.

**a**



b

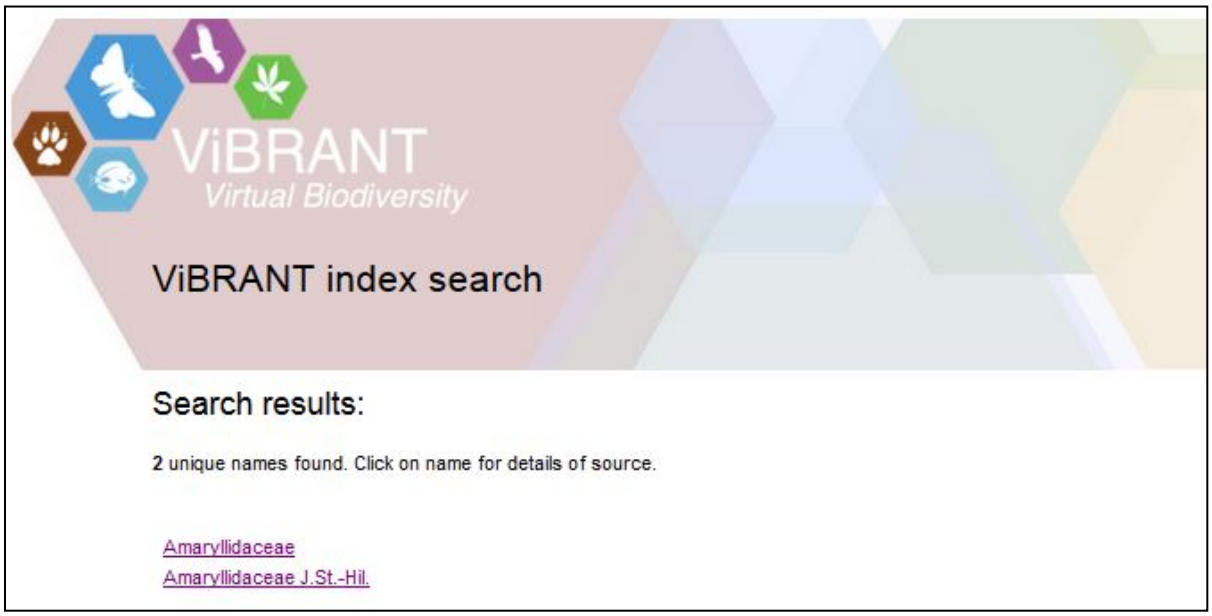
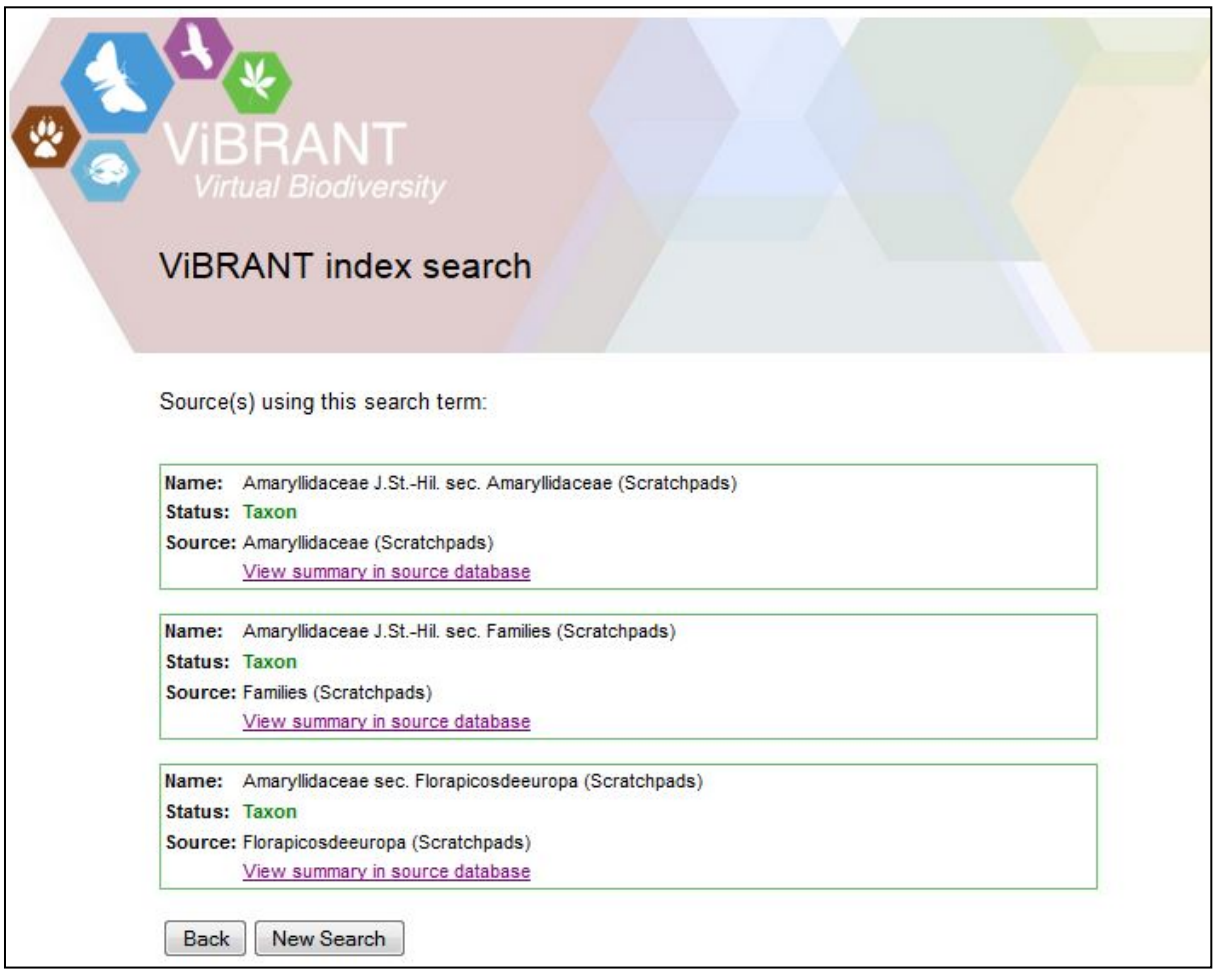



**Figure 2 (a)** Screenshot showing querying of the Scratchpads ViBRANT index with 'Amaryllidaceae'. **(b)** Result shows 2 hits. Clicking on a link to a distinct name provides further taxonomic data and information of the sources where that taxonomic name is found. **(c)** (Below) shows that the name 'Amaryllidaceae' is found in 3 different sources, linking to 3 different Scratchpad sites (http://amaryllidaceae.e-monocot.org/ http://florapicosdeeuropa.myspecies.info http://families.e-monocot.org)

c


ViBRANT
Virtual Biodiversity
ViBRANT index search
Source(s) using this search term:
Name: Amaryllidaceae J.St.-Hil. sec. Amaryllidaceae (Scratchpads)
Status: Taxon
Source: Amaryllidaceae (Scratchpads)
View summary in source database
Name: Amaryllidaceae J.St.-Hil. sec. Families (Scratchpads)
Status: Taxon
Source: Families (Scratchpads)
View summary in source database
Name: Amaryllidaceae sec. Florapicosdeeuropa (Scratchpads)
Status: Taxon
Source: Florapicosdeeuropa (Scratchpads)
View summary in source database
Back
New Search

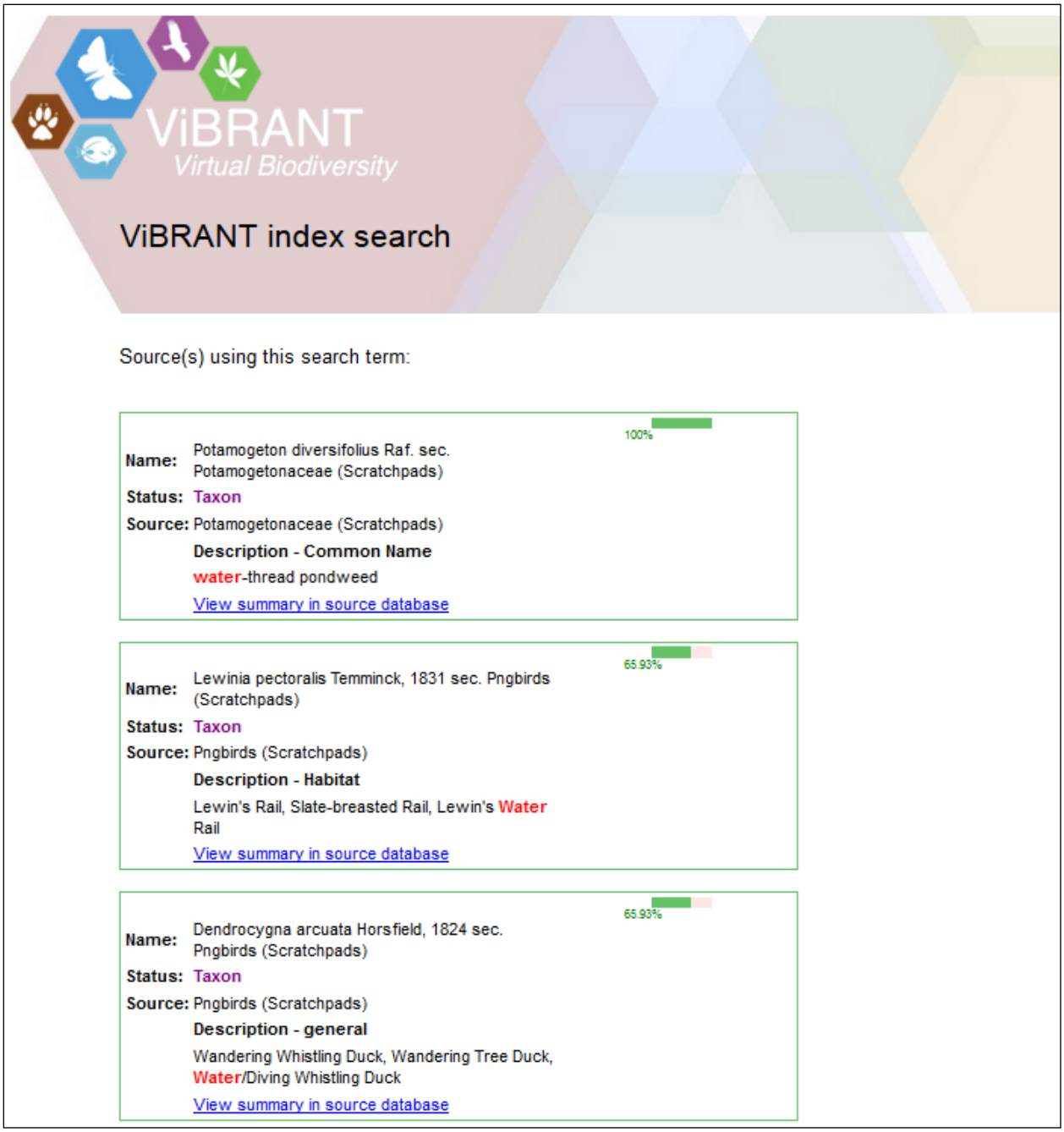



**Figure 3. Screenshot showing results from calling the full text search web service**. The example shows the top 3 hits, from searching the ViBRANT index database with the query 'Water'. The red highlighted text shows the search query in the descriptive text. The 3 hits are from 2 different Scratchpad sites, and clicking the link takes the user to the individual site for further information.

a

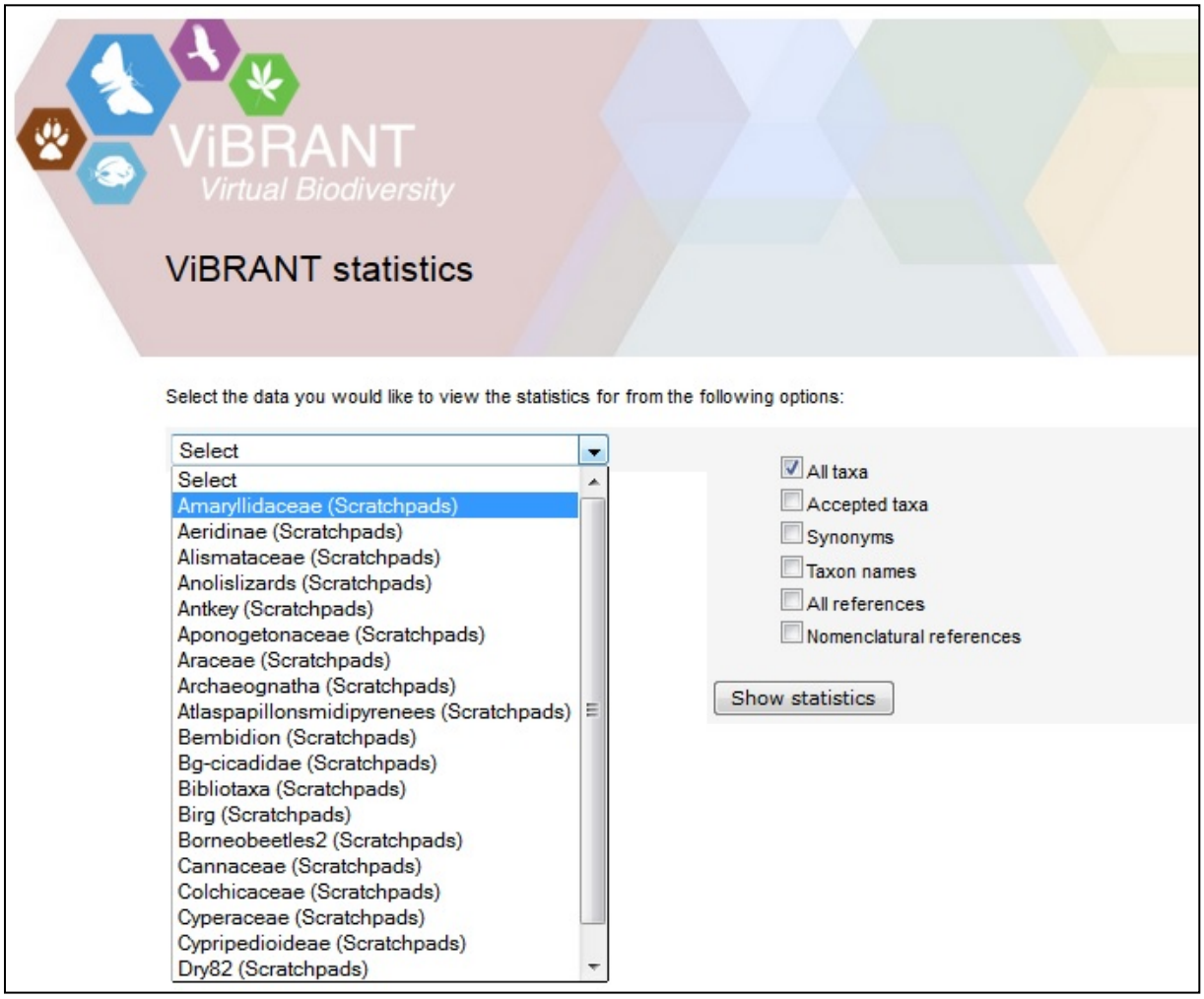
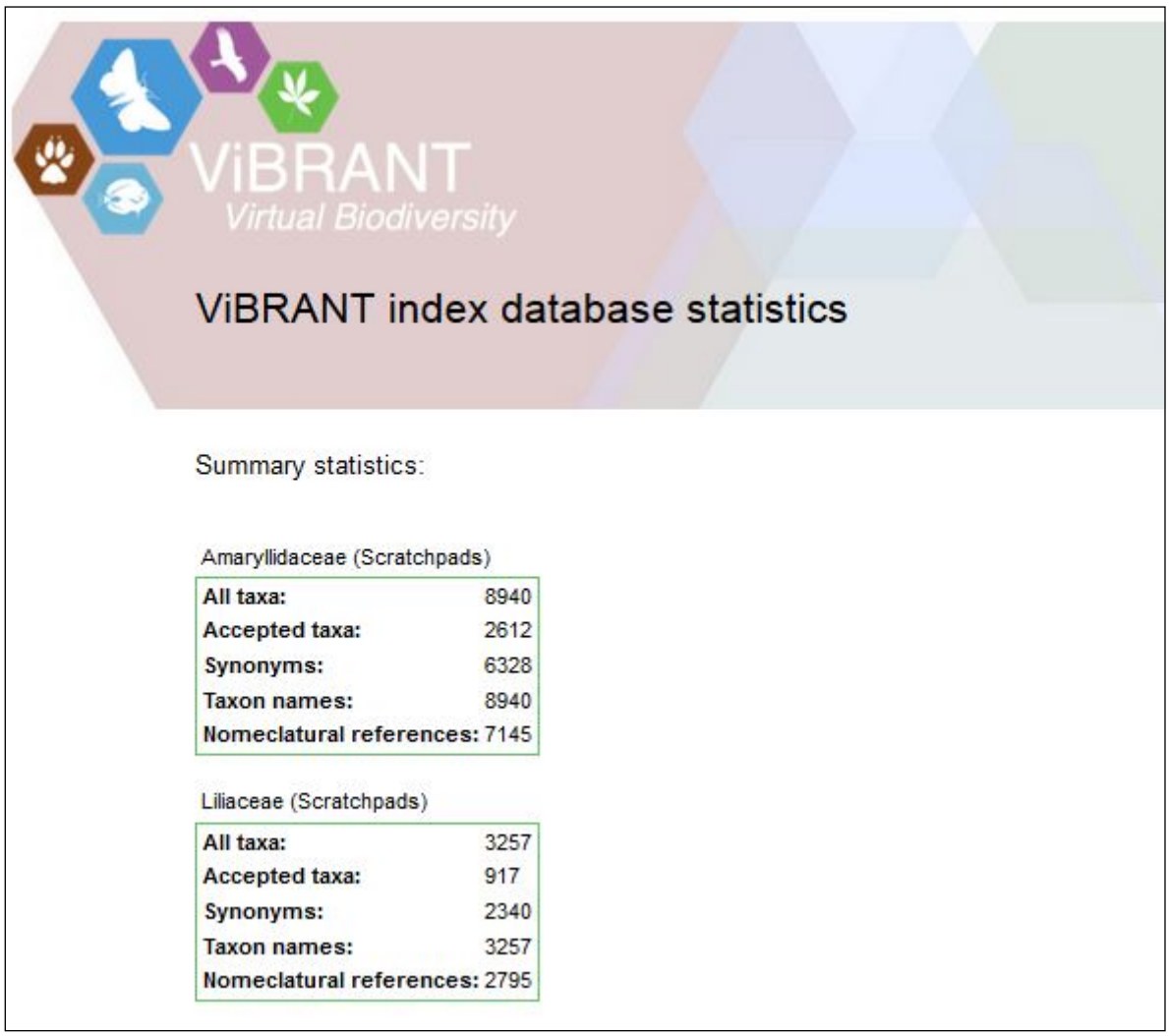



**Figure 4. Screenshots showing results from calling the CDM Statistics web service**. (a) Screenshot showing interface for querying for statistics for Scratchpads data contained in the ViBRANT index database. The drop-down menu displays the classifications available (each classification is created from a single Scratchpad site) (b) Result of running the statistics service on the Scratchpad Amaryllidaceae and Liliaceae classifications, contained in the CDM.

b




## Further work

Now that the Scratchpads data is in the CDM it is possible to apply any of the CDM REST based web-services to the data. In addition to the web services described above, a fuzzy name search has been implemented. Fuzzy matching enables a name to be found even if it is spelled incorrectly, and the resulting hits can be ordered by score according to how closely they match the search string. If there is demand for this functionality for the ViBRANT index portal, it will be added to the user interface. Further work will be carried out on the statistics service (for milestone M4.40) e.g. to count descriptions and descriptive source references.

DwC-A files are de-normalised so that in the text file there is repetition of data in certain fields. During import we carried out some de-duplication of data, for example before adding a rights statement to a description, we check if it has already been added. Further de-duplication of data could be carried out, for example if a reference has already been created from the import of one Scratchpad we should not create it again.

To keep the Scratchpad search portal up-to-date we would need to update the data and the Lucene index whenever the DwC-A files are updated on the source sites. If there is user demand for this service we would employ a cron job, to download and import the data into the CDM whenever a new dwca.zip file is available.

For full text search, if the search term is found within the description field we need to display any Copyright statement associated with this description if this is present in the DwC-A file. We will discuss further with the Scratchpads team at the NHM any licensing restrictions, for example whether we need to show the individual Scratchpads site license. This currently isn't contained within the DwC-A file, but one possibility is to create an additional file (eml.xml) containing the metadata describing the Scratchpad and include the site license here.

## References


1. http://vbrant.eu/sites/vbrant.eu/files/ViBRANT_M4_24.pdf
2. http://lucene.apache.org/]
3. http://wp5.e-taxonomy.eu/cdmlib/rest-api-name-catalogue.html
4. http://scratchpads.eu/explore/sites-list/json
5. http://wp5.e-taxonomy.eu/cdmlib/rest-api-statistics.html
6. http://rs.tdwg.org/dwc/terms/taxonomicStatus
7. http://rs.tdwg.org/dwc/terms/acceptedNameUsageID