

# Annex 4

## The Import module of DaRWIN

<b>1. INTRODUCTION</b>	<b>2</b>
<b>2. METHODOLOGY</b>	<b>2</b>
2.1 Reworked importation procedure	2
Figure 1: The new import procedure in 3 steps using XLS templates	3
2.2 Lower complexity when importing external specimen data	4
2.3 Additional importation templates and verification interfaces	5
<b>3. INFRASTRUCTURE</b>	<b>7</b>
<b>4. RESULTS AND RECOMMENDATIONS</b>	<b>7</b>
<b>4.1 The DaRWIN side changes</b>	<b>7</b>
4.1.1 Taxonomy	7
4.1.2 Localities	10
4.1.3 Lithostratigraphy	11
4.1.4 Embedded multimedia files	13
4.1.5 Links to remote multimedia files	14
<b>4.2 Input Templates</b>	<b>15</b>
4.2.1 Need for an import tool	15
4.2.2 RMCA Excel template	16
a. Sheets	16
b. Buttons	18
c. Forms	19
d. Export	22
4.2.3 LibreOffice template	24
a. Sheets	24
b. Buttons	25
c. Forms	25
d. Export	29
e. Taxonomy check	29
<b>4.3 Integration of previous databases and Import of data in DaRWIN</b>	<b>29</b>
4.3.1 Mapping of the RBINS MISTA database	30
4.3.2 The RBINS Geology Collection	30
4.3.3 The RBINS Paleontology Collection	30
4.3.4 RMCA zoology	30
4.3.5 Mapping of RMCA wood biology data	31

## 1. INTRODUCTION

The existing import module of DaRWIN was developed by RBINS in the framework of a previous project. The procedure was based on a huge XLS file, exporting an XML file with all data (Sampling location, Taxonomy and specimens data). This XML file was then imported by DaRWIN with several levels of data checking.

This was extremely complex thanks to the size of the XLS file and the number of fields and frustrating for users because the import was always blocked somewhere.

The import procedure is nevertheless extremely useful to add new specimens to the existing database(s). Curators and research scientists already have many specimens encoded in smaller databases or use XLS files. They know how to use spreadsheets which are common softwares for users.

The manual encoding of data into the DaRWIN database is estimated to be between 5000 and 10000 specimens / year / encoder. This process is thus very slow and not efficient.

The import procedure from existing databases and/or spreadsheets allows to import up to 10 times more specimens in the same time period for a trained FTE “import” encoder. This is why we decided to completely review the import processes.

## 2. METHODOLOGY

### 2.1 Reworked importation procedure

The workflow, database logic and interfaces to import external data from files into DaRWIN have been extensively reworked and expanded.

The import is now divided in 3 steps:

- Taxonomy
- Location(s)
- Specimens data

This segmented procedure allows to simplify the validation of the data. For Taxonomy and Locations files, it is possible to use external data validation using web services of authority databases.

A reference manual on the current importation procedure, written by Marielle Adam, is available on the gitHub repository of the project:

[https://github.com/naturalsciences/natural\\_heritage\\_darwin/blob/STABLE\\_2020/doc/import%20user%20manual.docx](https://github.com/naturalsciences/natural_heritage_darwin/blob/STABLE_2020/doc/import%20user%20manual.docx)

Number of the action	Description of the action	KPI	Deliverable	2015	2016	2017	2018
SO1_OO1_A2	To select the optimal tools for collection management (with easy import functionalities)	An analysis is available in the first half of 2016		X	X		

	A	B	C	D	E	F
	phylum	class	order	family	genus	species
2	CHORDATA	ACTINOPTERYGII	SILURIFORMES	Claridae	Tanganikallabes	Tanganikallabes alluapera Wright & Bailey,2012
3	CHORDATA	ACTINOPTERYGII	SILURIFORMES	Claridae	Tanganikallabes	
4	CHORDATA	ACTINOPTERYGII	SILURIFORMES	Claridae	Tanganikallabes	Tanganikallabes mortiauxi Poll,1943
5	CHORDATA	ACTINOPTERYGII	SILURIFORMES	Mochokidae	Synodontis	
6	CHORDATA	ACTINOPTERYGII	SYMBRANCHIFORMES	Mastacembelidae	Mastacembelus	Mastacembelus tanganicae (Günther,1893)
7	CHORDATA	ACTINOPTERYGII	SILURIFORMES	Claroidae	Lophiobagrus	Lophiobagrus cyclurus (Worthington & Ricardo,1936)
8	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Neolamprologus	Neolamprologus toae (Poll,1949)
9	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Aulonocranus	Aulonocranus dewindti (Boulenger,1899)
10	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Chalinochromis	Chalinochromis trichardi Poll,1974
11	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Cyprichromis	
12	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Cretodus	
13	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Neolamprologus	Neolamprologus fasciatus (Boulenger,1898)
14	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Neolamprologus	Neolamprologus tetrocephalus (Boulenger,1899)
15	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Neolamprologus	Neolamprologus toae (Poll,1949)
16	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Neolamprologus	Neolamprologus niger (Poll,1956)
17	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Neolamprologus	Neolamprologus niger (Poll,1956)
18	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Telmatochromis	Telmatochromis bifrenatus Myers,1936
19	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Telmatochromis	Telmatochromis dhonti (Boulenger,1919)
20	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Pseudosimochromis	Pseudosimochromis curvifrons (Poll,1942)
21	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Petrochromis	Petrochromis fasciolatus Boulenger,1914
22	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Petrochromis	
23	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Uphthamotilapia	Uphthamotilapia nasuta (Poll & Matthes,1962)
24	CHORDATA	ACTINOPTERYGII	PERCIFORMES	Cichlidae	Aulonocranus	Aulonocranus dewindti (Boulenger,1899)

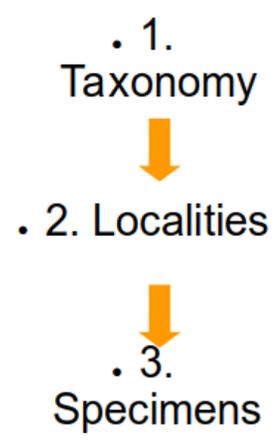


Figure 1: The new import procedure in 3 steps using XLS templates

As described in the reference manual, the complete workflow to import specimen data in DarWIN is now splitted into three steps:

1. Importing the taxonomy
2. Importing localities (collecting stations)
3. Importing specimen data

Specimen data has to be imported last:

- The binding between the specimen and the taxa is done by the *full scientific name* (word containing the scientific name and the authorship information, without considering the rank )
- The binding between the specimen and the locality is done via the station number, which is then supposed to be unique. It is possible to disambiguate duplicate station numbers for one specific record or the whole dataset in the validation interface.

## 2.2 Lower complexity when importing external specimen data

Another major change in DaRWIN has been the simplification of the procedures and workflow to import external data into DaRWIN. The initial version only accepted input data based on the ABCD XML schema for collection data. Flat data has to be first written on specific Excel files that feature a Visual Basic macro generating the XML file. This made the adaptation of the template procedure very complex, any modification had to be implemented in three different systems (the XML parser in darwin, the stored procedure in XML and the macro in Visual Basic). The macro and XML parser were very complicated to test and debug. The complexity in time (the parsing of XML documents is time consuming, such as the conversion of the Excel data into XML) and space (XML files are much larger than flat CSV files) was also needlessly high. The usage of XML introduced bugs related to the syntax of the XML document that were harder to diagnose: the intermediate data structure handed by the PHP server had to be serialized on the hard drive and syntactically analyzed for each problem. Besides, the XML ABCD format was not exposed to a public web service on the Internet as the corresponding web service BioCASE works through a database connection: the files could not be reused for any other usage. XML could have been interesting to handle controlled vocabularies and external thesaurus, but this functionality is not present in DaRWIN. A flat tabular structure was converted into an hierarchical structure, before being converted again the other way round into the flat structure of DaRWIN data model. The Excel files were also complicated to handle, having more than 100 columns, the name of each of them being case sensitive, and had a constraining order, while the imported data for specimen most of time contain juste about 15 to 30 meaningful fields (label number, locality information scientific name).

Finally, the ABCD parser could work only on specimen data, and custom PHP parsers would have to be developed for any other type of data (people, localities etc...) . DaRWIN's team was also contacted in April 2019 by a team of scientists from the University of Rwanda for a JRS Biodiversity project . While they were interested in the possibility of integrating and cleaning external data into a reference collection database, this part of the system had to be simplified, to ease its documentation and usage for external users.

We decided to remove the XML part and parser and to replace it by a parser for CSV files that would follow the following concepts:

1. The data would be tab-delimited
2. The column name would follow a controlled vocabulary
3. Each column would be optional. The system could import specimens having already a collection number, or assign a new value (as numeric sequence) of this collection number is missing.
4. The column names would be case insensitive
5. Their order could be free

6. Geographical coordinates could be inserted in DMS (degree minutes seconds) while being converted in decimal degrees if they followed a consistent text pattern (eg "10°30' 45" W/E" or "N/S 14° 45')

These changes have been integrated gradually, first by keeping the XML existing parser and generating in-memory XML files inside of DaRWIN. This part has been finally removed and replaced by a PHP parser directly filling the staging table from the tab-delimited values. The import process was using four steps (creating of the Tab-delimited file, importing the data into the SQL *staging* tables, checking duplicate, and integrating the data in the normalized part of DaRWIN) instead of 6 (filling the template, generating the XML, parsing it, filling SQL *staging* tables, checking duplicate, and integrating the data in the normalized part of DaRWIN).

Finally, the jobs to import the data, that are asynchronous and background console operations, have been linked to the web interface of DaRWIN, while they were previously only available via Bash or DOS instructions (often provided by a SSH connection). This limited the importation tasks to IT-trained staff, while the new procedure is available for any user having the access rights to DaRWIN. However this introduced a moderate security risk, as the web interface has to execute shell commands. This risk can be mitigated by checking and controlling the type of parameters passed to the command as argument (limiting them as numeric values) .

This part of the work was surely the most complex and harder to test amongst the Dawin tasks within the framework of NaturalHeritage, as the completeness of imported data had to be cautiously checked, on more than 100 columns that could be combined in different ways. These checks could not be automated. The scalability (ie, ability to work on a great number of records) of the procedure had also to be verified, which implied huge amounts of data to be produced (which is sometimes almost as complex as developing the application) and lengthy test operations. We should actually have defined more rigorous and standardized test procedures. However, this kind of data is placed outside of the scope of unit testing which is easier to automate (which are more targeted and specific, but do not correspond to this scenario as they compare the behavior and backward-compatibility of new versions of a programme to a reference and stable behaviour). The import speed can be estimated by 2500 to 5000 records/hour for importing data from the source file and approximately the same duration for the check (detection of duplicates). Integration of the data in DaRWIN is faster (5000 to 10000 records/hour)

### **2.3 Additional importation templates and verification interfaces**

Once the importation workflow has been simplified and the XML part removed in design pattern, a design pattern that could be reused with other types of data was made available. This design pattern followed a three steps approach, each of them corresponding to asynchronous background operations:

1. Importation of tab-delimited data into the staging part
2. Iterative checks to detect and remove duplicated either by
  - Creating new values and doing batch updates or...
  - Choosing an existing value and doing batch update on the others records
3. Integrating the cleaned data into the normalized part of DaRWIN . This is also an iterative task that can be done on parts of the dataset, and that can be sprawled, interrupted and resumed on several sessions

The figure 2 describes the importation workflow for specimen data, that checks duplicates, disambiguate homonymes, and missing data, and correct then by single or batch updates on:

1. Peoples
  - Collectors
  - Identifiers (for taxonomic, or geological attributions)
  - Donators
2. Institutions
3. Taxonomic identification (if the taxa is missing it can be created)
4. Expedition
5. Sampling locations (using their station code as link, that becomes a mandatory field)

It is also possible to enforce or disable a unicity constraint on the main specimen code, while uploading the file.

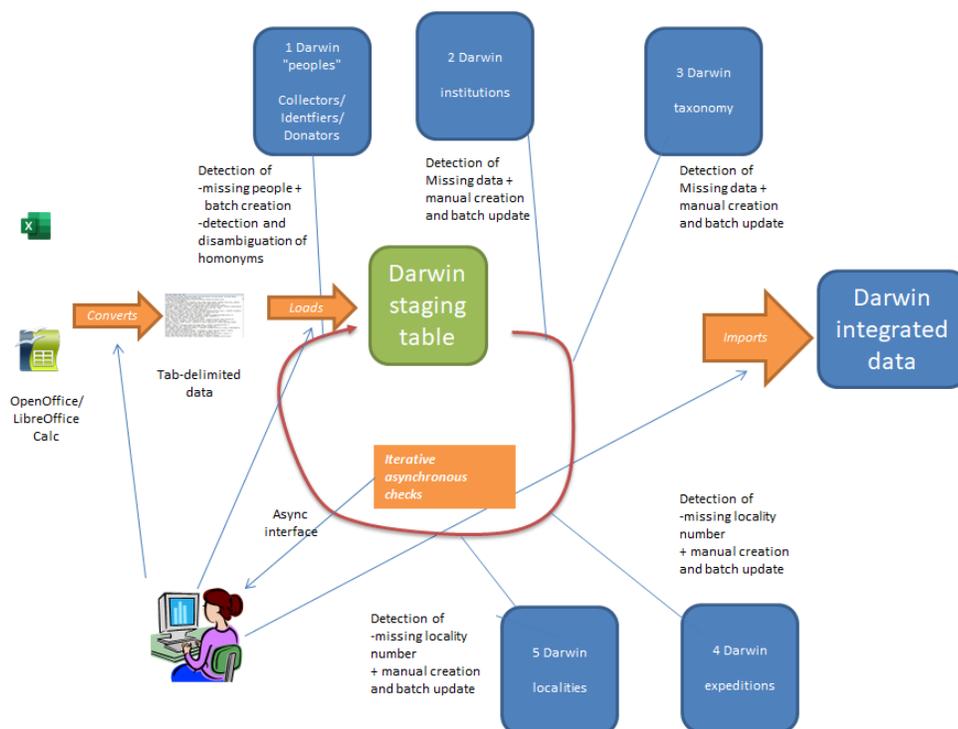


Figure 2. Importation workflow for specimen data

This template has been adapted to several other content types that can now be uploaded from tab-delimited files into DaRWIN and verified. The most important development effort didn't reside in the programming of the import logic, but in the development of visualization and data-management interface allowing users to access the staging tables and clean data from the Internet. A semaphore mechanism had to be implemented, to notify the users whether import tasks successfully ended or not, and to report errors (PHP and PostgreSQL exception are thrown to the web interface giving the status of an import to ease debug and correction of data, as they often give information about syntactical issues).

### **3. INFRASTRUCTURE**

The adopted procedure simplifies the infrastructure needs as no specific server is requested by the new procedure. The complete process can be realized with Open Source technologies as the templates were developed for the proprietary Microsoft Office Excel but also for the Open Source LibreOffice suite. The main difference was in the programming which is in Visual Basic for the Excel macros and in Basic for the LibreOffice version.

## **4. RESULTS AND RECOMMENDATIONS**

### **4.1 The DaRWIN side changes**

#### **4.1.1 Taxonomy**

DaRWIN initial versions already featured a template mechanism to upload taxonomic hierarchies, but it was using a custom XML schema derived from ABCD (also requiring a Visual Basic macro) and had no validation interface allowing the user to check and validate data from the web interface.

Besides, the concept of "parallel" taxonomies (or taxonomical metadata) had been introduced in DaRWIN, allowing to publish different hierarchies for the same taxon and annotate their scientific accuracy. The initial importation mechanism used also one SQL transaction (either all data could be imported or none, error or taxonomic conflict cancelling the whole job).

The taxonomic template, and a substantial part of the database logic in the *staging* part of DaRWIN, needed therefore to be reworked. A template for a tab-delimited file has been defined, where users can provide a list of scientific names with upper ranks and authors within each row. The higher rank provided for each row is free (it can be the phylum, the order, the family or others... ) but has to be already created in DaRWIN,

which should build the complete descending taxonomic tree from the higher taxonomic level to the lower one.

For each attempt to create taxon, 4 types of operations can be detected:

1. The taxon is missing and could be successfully created in DaRWIN (his parent exists both in the file and the system)
2. The taxon is missing but couldn't be created (the parent in the file cannot be created in DaRWIN)
3. The taxon has another parent in DaRWIN, for the considered parallel taxonomy, at least one of its parent has another hierarchy (Upper level conflict with DaRWIN)
4. The taxon is present twice or several times in the file with different hierarchies, that contradict themselves (Upper level conflict within the file)

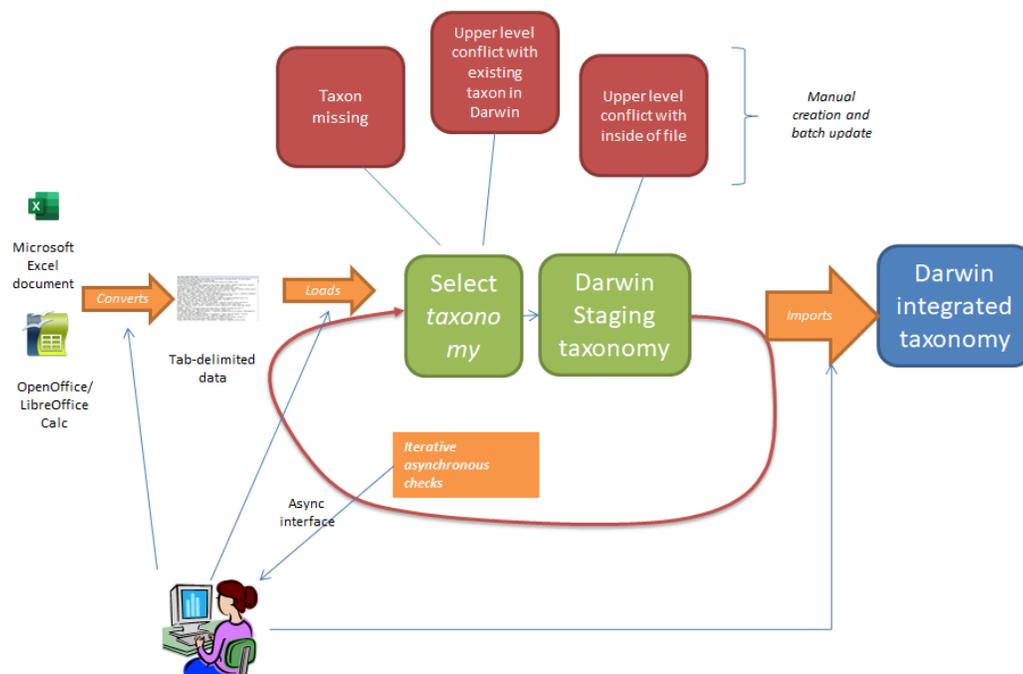


Figure 3. Importation workflow for taxonomic data

A validation interface has been developed (see figures 4 and 5). It features a pager with global statistics on the imported records, allowing the user to navigate through results. Each page displays 1000 rows.

- A simple color code (green rows for imported results and orange for errors, makes it more readable and allows users to rapidly identify issues).
- Each row contains a field where taxonomic hierarchies in DaRWIN and in the imported files are displayed as paths separated by “/”, allowing a rapid comparison.
- A button opens a modal window allowing you to manually create the taxon, and/or to correct the hierarchy of an existing taxon in DaRWIN.

- Buttons placed at the bottom of the web page allows to launch the check and integration again, and to change the parallel taxonomy that has been chosen in the import procedure. This allows users to handle problem without reimporting the tab-delimited file, and in several internet sessions

Count all : 2872  
Current page : 1 / 3  
Page : 1 / 3  
go

Message	Count
<a href="#">imported_taxon</a>	939
<a href="#">taxonomic_conflict</a>	42
<a href="#">taxonomic_hierarchy_already_exists</a>	12
<a href="#">taxon_to_be_created_without_suitable_parent</a>	7
<b>Total :</b>	1000

All data:

Message	Count
<a href="#">imported_taxon</a>	2665
<a href="#">taxonomic_conflict</a>	152
<a href="#">taxonomic_hierarchy_already_exists</a>	31
<a href="#">taxon_to_be_created_without_suitable_parent</a>	24
<b>Total :</b>	2872

[Download all](#) [Download unimported](#) [Recheck and reimport](#)

Figure 4. Validation interface for taxonomic import (pager and statistics)

Count all: 63  
Current page: 1 / 1  
Page: 1 / 1  
go

Message	Count
imported_taxon	42
taxonomic_hierarchy_already_exists	21
<b>Total:</b>	<b>63</b>

All data

Message	Count
imported_taxon	42
taxonomic_hierarchy_already_exists	21
<b>Total:</b>	<b>63</b>

Download all | Download unimported | Recheck and reimport

id	name	level	ref	name_cluster	imported	import_exception	compare_hierarchies	import
295388	Anthomyiidae	family	1		FALSE	taxonomic_hierarchy_already_exists	Staging hierarchy (Anthomyiidae (family))  Darwin hierarchy (Eucaryota (domain)/Animalia (kingdom)/Arthropoda von Siebold, 1843 (phylum)/Hexapoda Elmirle, 1816 (sub phylum)/Insecta Linnaeus, 1758 (class)/Pterygota Gegenbaur, 1879 (sub class)/Optera Linnaeus, 1758 (order)/Brachycera (sub order)/Muscomorpha Sharp, 1894 (infra order)/Anthomyiidae (family))	Counts taxon
295389	Macrochis	genus	1		TRUE	imported_taxon	Staging hierarchy (Anthomyiidae (family)/Macrochis (genus))  Darwin hierarchy /	
295391	Rhynchoceomys	genus	2		TRUE	imported_taxon	Staging hierarchy (Anthomyiidae (family)/Rhynchoceomys (genus))  Darwin hierarchy /	

Figure 5. Validation interface for taxonomic import (hierarchy viewer)

### 4.1.2 Localities

It has been decided to create a specific importation template for localities, in order to keep in line with the normalization of localities described above.

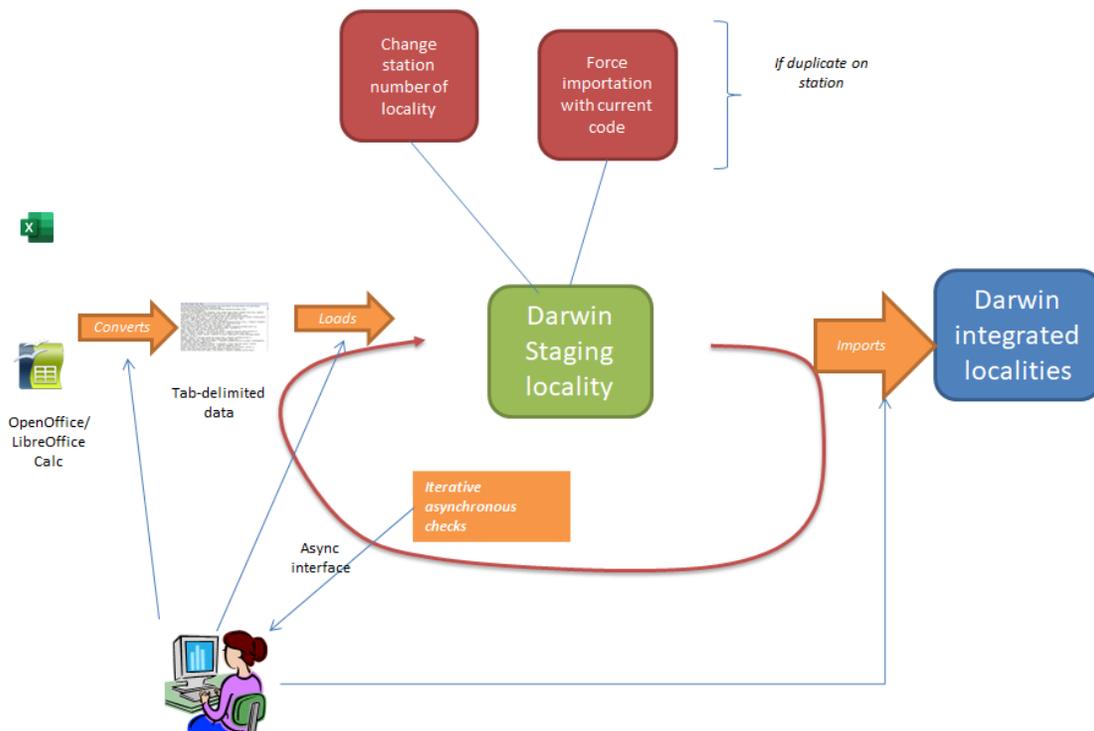


Figure 6. Importation workflow for locality data

### 4.1.3 Lithostratigraphy

A template has also been developed to import additional lithostratigraphic classifications, for fossil or mineralogical collections.

It has been decided to make the lithostratigraphic scale dynamically updatable, but to keep the existing chronostratigraphic frozen, as the chronostratigraphic scale is global and stable, while the lithostratigraphic scale is dependant from the location and less standardized.

The template columns are:

supergroup
group
formation
member
layer
sub_level_1
sub_level_2

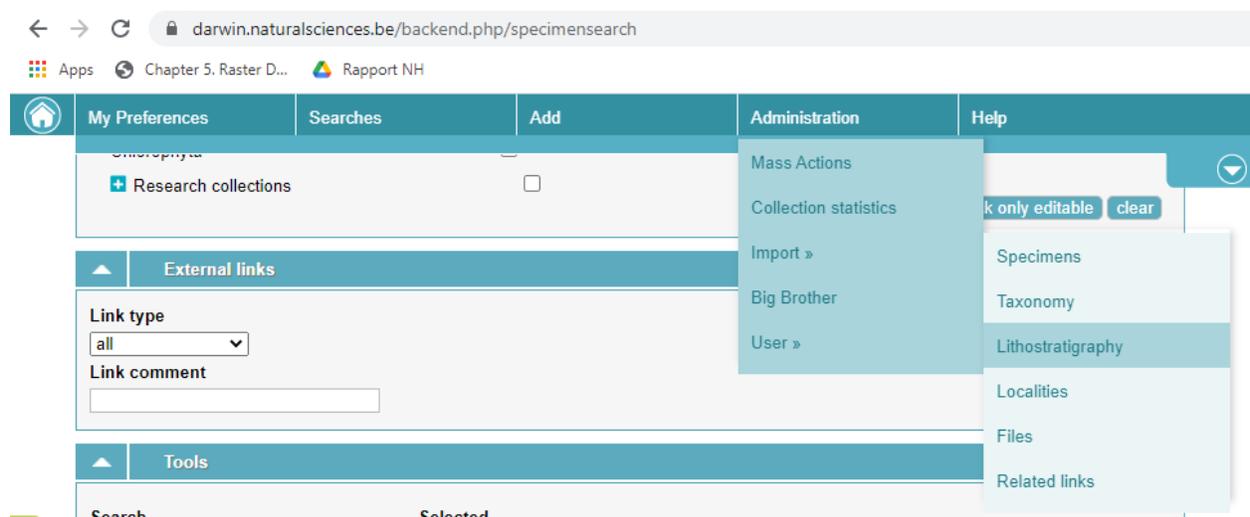


Figure 7. Access to the lithostratigraphy import

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	group	formation														
2	Kibarien moyen	Actuel														
3	Kibarien moyen	Anversien														
4	Kibarien moyen	Anversien														
5	Kibarien moyen	Anversien; ou Messinian (Bolderien) ID Age: 9														
6	Kibarien moyen	Bartonien (Asschien)														
7	Kibarien moyen	Bartonien (Wemmelien)														
8	Kibarien moyen	base miocène moyen														
9	Kibarien moyen	between lower and middle Eocene														
10	Kibarien moyen	Bolderien														
11	Kibarien moyen	Bruxellien														
12	Kibarien moyen	Bruxellian B1 (facies dit panisellen)														
13	Kibarien moyen	Bruxellian sup.														
14	Kibarien moyen	Bruxellien														
15	Kibarien moyen	Campinien q2														
16	Kibarien moyen	Coblentzien, Siegenien sup.														
17	Kibarien moyen	Continental Landenian														
18	Kibarien moyen	Continental Landenian (L2)														
19	Kibarien moyen	Continental Landenian L2														
20	Kibarien moyen	Couches de passage de l'Oligocène au Miocène														

Figure 8. Example of lithostratigraphic import

The template for specimens contains several fields to link specimens to geological classifications. These are:

**a. Paleontology / chronostratigraphy**

GeologicalEpoch
GeologicalAge
GeologicalAge3

**b. Lithostratigraphy**

lithostratigraphyGroup
lithostratigraphyFormation
lithostratigraphyMember
lithostratigraphyBed
lithostratigraphyInformalName

**c. Mineralogy (identification level)**

mineralogicalIdentification
mineralogicalIdentifier
mineralogicalIdentificationYear
mineralogicalIdentificationMonth
mineralogicalIdentificationDay

**4.1.4 Embedded multimedia files**

A template was also developed to embed multimedia files, that are stored as Media files in the filesystem of the DaRWIN server (hosting the PHP backend) and associated to specimen records.

These files are available in the backend part of DaRWIN which is password protected. This may be relevant for data that may not be publicly disclosed to (.e.g. Material for future paper).

This import consists of two files:

- A Zip file, which is uploaded and unzipped on the server, that contains the files
- A tab-delimited file, that has to be called *meta.txt*, which describes the compressed files

The fields of *meta.txt* are:

Field name	Field Description	Mandatory
UnitID	The main specimen code of the associated specimen	x (or UUID)
filename	the name of the file in the associated ZIP	x
title	The title (caption) of the file	
description	Free text description of the file content	
sub_type	Sub-type of the file	
mime_type	mime type of the file (to help the client to choose the appropriate viewer or player)	x
technical_parameters	the technical parameter of the file (pixel resolution, sampling rate etc...)	
internet_protocol	the internet protocol of "external_uri"	

field_observations	field parameters of the object (e.g. water temperature, salinity etc...)	
external_uri	link to an external resource describing or completing the data	
uuid	the UUID of the associated specimen in DaRWIN	x (if no UnitID provided)

#### 4.1.5 Links to remote multimedia files

Finally, a fifth template was created to associate remote multimedia documents available on the Web to existing DaRWIN, as links that are batch-created. This can be images, sounds, description of the specimen in an on-line publication, related specimens in other databases, links to DNA sequences in GenBank...

DaRWIN also features an IIIF client (Mirador) which is synchronized with this template. It can be minked to the "virtualcol" platform which itself gets the UUID identifiers from DaRWIN. This template allows exchanging data between the two systems.

Field from this template are the followings:

Field name	Description	Values(ex)
<b>UnitID</b>	The main collection code of the specimen in DaRWIN	INV.2090
<b>UUID</b>	The uniform unique identifier of the specimen in DaRWIN (if no UnitID provided)	89f7383b-87c2-47a5-946c-1032bef0ae73
<b>URL</b>	The URL of the resource to link	
<b>Type</b>	The type of link (abbreviate)	DNA; IIIF ; CITES; Nagoya
<b>Comment</b>	Link description (searchable in DaRWIN)	

## 4.2 Input Templates

### 4.2.1 Need for an import tool

The DaRWIn web interface is very complete and allows you to enter all data needed concerning specimens. It's very useful when 1 or 2 specimens are to be encoded but it's time consuming if there are a lot of specimens. There is also a need to be online and connected to the DaRWIn server to enter data.

For data coming from outside, there is also no common template: data may come from text files, excel, databases in various formats in data structure and data format. So a tool was needed to work offline (at home or in field work) and to import lots of specimens in one step.

As spreadsheets are known by nearly everyone working with data, this kind of file has been chosen. DaRWIn can also use csv files to import lots of data and a first version of an import template has already been done in excel.

- The new template will use the possibility to create forms above the spreadsheet to more easily enter data that are sometimes spread in many sheets of a workbook.
- Export to csv is easy and automatically done by buttons in the spreadsheet and the generated files can be used to be imported in DaRWIn without other treatment.
- A link to a tool to check taxonomy has also been added in the spreadsheet.

The advantages of this tool and of the forms are an easy way to fill in data, and a view of all the data in one screen, so data can be checked easily for completeness and integrity.

Two versions of the template have been developed:

- a first one in Microsoft Excel and VisualBasic for users using the desktop version of Microsoft Excel on Windows and Mac. This template is not working with Office 365 online or with libreOffice or previous versions of Excel.
- a second template in Calc and basic for users using the Open Source LibreOffice suite on Windows, Mac and Linux OS.

The template offers the possibility to import data in DaRWIn both at RMCA and RBINS. The template may be filled as a simple spreadsheet, by filling in each sheet one after the other but as there is a risk to write data at the wrong place if we choose the wrong line on a sheet, it's better to use the forms that gather all fields of the same line in simple forms.

### 4.2.2 RMCA Excel template

#### a. Sheets

Data are splitted into several sheets:

Code, location, DNA, ecology, taxonomy, counts\_storage and acquisition.

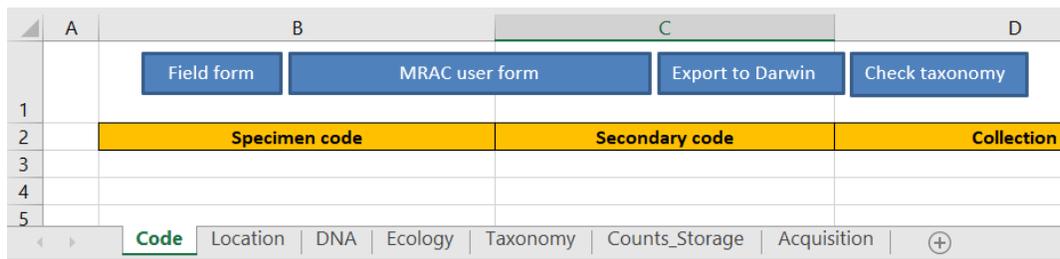


Figure 9. Different sheets of the template

#### Code

	Specimen code	Secondary code	Collection	Entered by	Description - Notes
1	SP19-005				
2	SP19-005				

#### DNA

A	B	C	D	E	F	G	H	I
Code		Specimen info		Fin-clip info				Notes
Specimen code	Location Code	Tag number	DNA box	Tube number	Horizontal position	Vertical position	DNA notes	
1	SP19-005	Loc-001	19004	23	25		2	6 form wing
2	SP19-005	Loc-001	19004	23	25		2	6 form wing

#### Ecology

A	B	C	D	E	F	G	H	I
Code		Ecology parameters						
Specimen code	Location Code	Water temperature	Hour(HH:MM)	pH	mV	Conductivity (µS/cm)	O2 dissolved (%)	
1	SP19-005	Loc-001	20	0.524305556	7	45	52	10
2	SP19-005	Loc-001	20	12:35	7	45	52	10

J	K	L	M	N
				Notes
O2 dissolved (mg/l)	hPa	Air temperature	Relative humidity	Ecology notes
6	1020	21	82	very warm
6	1020	21	82	very warm

#### Acquisition

A	B	C	D	E	F	G	H	I
Code		Acquisition information			Acquisition dates			Notes
Specimen code	Location Code	Type	From	Day	Month	Year	Acquisition notes	
1	SP19-005	Loc-001	Donation	Mr X	25	12	2012 cadeau	
2	SP19-005	Loc-001	Donation	Mr X	25	12	2012 cadeau	

Figure 10. Columns of sheets Code, DNA, Ecology, Acquisition

Location

Code		Location names and description					
Specimen code	Location Code	Continent	Country	State-province	Municipality	Exact site	
1	SP19-005	Loc-001	Africa	Madagascar	Anta	Antanarivo	near Antanarivo
2	SP19-005	Loc-001	Africa	Madagascar	Anta	Antanarivo	near Antanarivo

DMS coordinates										Decimal coord.
Degrees N/S	Minutes N/S	Seconds N/S	N/S	Degrees E/W	Minutes E/W	Seconds E/W	E/W	Latitude	Longitude	
12	25		5 S	15	45		12 E			
12	25		5 S	15	45		12 E			

GPS		Collecting information				
GPS Weight Points	Day/Night_catch	Altitude(m)	Collectors	Collecting method	Expedition project	
none	Day	1200	Merlijn	Apstein net	Merlijn et al	
none	Day	1200	Merlijn	Apstein net	Merlijn et al	

Collecting dates						Notes
Start day	Start month	Start year	End day	End month	End year	Locality notes
10	11	1955	11	11	1955	loc notes
10	11	1955	11	11	1955	loc notes

Figure 11. Columns of sheet Location

Taxonomy

Code		Taxonomy									
Specimen code	Location Code	Temp. species field name	Kingdom	Phylum	Class	Order	Family	Genus	Species	Subspecies	Author and year
1	SP19-005	Loc-001	temp taxon	Animalia			Sphingidae	Nephele	densoi		(Kefenstein, 1870)
2	SP19-005	Loc-001	temp taxon	Animalia			Sphingidae	Nephele	densoi		(Kefenstein, 1870)

Identification		Type	Notes		
Identifier	Day	Month	Year	Type	Taxonomy notes
Jimh	10	12	1956	Specimen,Lectotype	no notes for taxo
Jimh	10	12	1956	Specimen,Lectotype	no notes for taxo

Counts\_storage

Code		Specimen info		Relationship					
Specimen code	Location Code	Sex	Life stage	Type	Parasite species	Parasite code	Host species	Host code	
1	SP19-005	Loc-001	female	chrysalis	Host	baobab	par001		
2	SP19-005	Loc-001	female	chrysalis	Host	baobab	par001		

Amounts				Notes	Storage location							
Amount males	Amount females	Amount juveniles	Total number	Notes for amounts	Institution	Building		Floor	Room	Lane	Column	Shelf
1	1		2	2 not sure	RMCA RMCA RMCA	Palais de l'Afrique	Palais de l'Afriq	2 2 3	23 24	2 3 4	3 4 5	5 6 7
1	1		2	2 not sure	RMCA RMCA RMCA	Palais de l'Afrique	Palais de l'Afriq	2 2 3	23 24	2 3 4	3 4 5	5 6 7

Status		Specimen part	Container ID	Container type	Container medium	Subcontainer ID	Subcontainer type	Subcontainer medium	Storage notes
Dry---good state---	Fresh---incomplete body	head	719 719 720	Cabinet Cabinet Jar	alcohol	199 200	Box Jar	unknown alcohol	
Dry---good state---	Fresh---incomplete body	head	719 719 720	Cabinet Cabinet Jar	alcohol	199 200	Box Jar	unknown alcohol	

Figure 12. Columns of sheets Taxonomy and Counts\_storage

Two additional sheets are hidden for the user and contain more technical info:

- The first hidden sheet contains predefined lists that can be completed if necessary and that are used in the form combo boxes and lists:

Figure 13. Technical sheet Lists

The second sheet contains info about the mapping with Darwin and Virtual Collections:

	A	B	C	D	E	F
1	Name excel	Field name	Sheet	order	Name darwin	Name Virtual collection
2	Specimen code	TB_Fieldcode	Code	2	UnitID	Code
3	Secondary code	TB_Sec_code	Code	3	additionalID	
4	Collection	TB_Collection	Code	4	collection	Collection in Institution
5	Entered by	TB_Label_createdby	Code	5	label_created_by	
6	General notes	TB_GeneralNotes	Code	6	Notes	Description
7	Full code	TB_FullCode_Joc	Location	2		
8	Location Code	TB_SamplingCode	Location	3	samplingcode	
9	Continent	CB_continents	Location	4	continent	
10	Country	CB_Countries	Location	5	country	Country
11	State-province	TB_Province	Location	6	Province	
12	Municipality	TB_Locality	Location	7	Municipality	
13	Exact site	TB_ExactSite	Location	8	exact_site	Location details
14	Degrees N/S	TB_Deg_NS	Location	9	LatitudeDMSDegrees	
15	Minutes N/S	TB_Min_NS	Location	10	LatitudeDMSMinutes	
16	Seconds N/S	TB_Sec_NS	Location	11	LatitudeDMSSeconds	
17	N/S	CB_N	Location	12	LatitudeDMS_N_S	
18	Degrees E/W	TB_Deg_EW	Location	13	LongitudeDMSDegrees	
19	Minutes E/W	TB_Min_EW	Location	14	LongitudeDMSMinutes	
20	Seconds E/W	TB_Sec_EW	Location	15	LongitudeDMSSeconds	
21	E/W	CB_E	Location	16	LongitudeDMS_W_E	
22	Latitude	TB_Latitude	Location	17	LatitudeDecimal	Coordinates
23	Longitude	TB_Longitude	Location	18	LongitudeDecimal	

Figure 14. Technical sheet Column\_matching

**b. Buttons**

On the first sheet, on top of the sheet, are displayed 4 blue buttons:

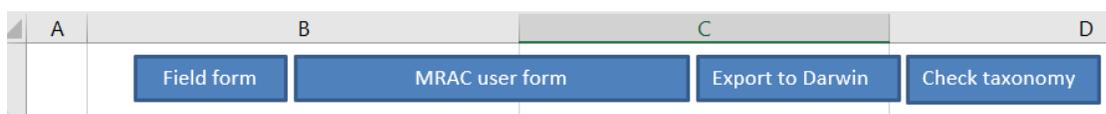


Figure 15. Buttons on first sheet

The 2 first buttons, “Field form” and “MRAC user form” are used to call 2 different forms. The first one contains only a limited set of fields and is intended to be used more in the field whereas the second one contains all the fields corresponding to every column of the 7 sheets. These forms will be described in paragraph 4.2.3. The next button calls the export features for DarWIN.



If you want to create a new line in the sheets, click on “New record”. It will clear the form and the new data will be saved on a new line. If you want to copy an existing line, go to that line/record with the navigation buttons and click on Duplicate record: it will create a new line with the same data and you will have to change only the necessary data as the code.

The button “Clean content” empties all fields.

## c.2 MRAC user form

Use of this second form is similar to the first one. Only the content is different because it contains all the fields.

Because the number of fields is more important and to keep readability of the form, it has been divided in 2 tabs, “General info” and “Secondary info”. General tab contains data about codes, taxonomy, sampling information. Taxonomy is much more detailed and complete taxonomy can be entered as well as the type, author and other info.

Figure 17. First screen of MRAC user form

“Secondary info” tab contains data about ecology, specimen parts, counts, relationship and acquisition. Acquisition is also new in regard to the field form and allows to mention the origin of a specimen other than a collect in the field.

Specimen parts are much more detailed: a container and subcontainer may be defined, with an ID, a type, a medium. Place of the container can be precisely given, as well as state of the specimen part. DNA being considered as a part, DNA info can be given in that section.

A new record has to be created for each part.

The screenshot shows the 'Specimen data entry' form, second screen. It features several sections with various input fields and dropdown menus:

- Ecology:** Includes sub-sections for 'Water' (Temperature, pH, mV, Conductivity, Dissolved O2) and 'Air' (Temperature, RH, P. Atm.).
- Specimen parts:** Includes sub-sections for 'container', 'Subcontainer', 'Storage' (Institution, Floor, Column, Building, Room, Shelf, Lane), and 'DNA' (Tag number, DNA box, Tube number, Horiz. position, Vert. position, DNA notes).
- Counts:** Includes fields for Total, Sex, Stage, and Juv.
- Relationship:** Includes a dropdown for 'This specimen' and a field for 'of'.
- Acquisition:** Includes fields for Type, Date, From, and Notes.

The form also includes a navigation bar at the bottom with buttons for 'New record', 'Save record', 'Clean content', 'Duplicate record', and a 'Go to record n°' field.

Figure 18. Second screen of MRAC user form

As for Field form, various combo boxes and lists are already filled with data to facilitate the work of the user.

The screenshot shows the 'Specimen data entry' form, focusing on lists and dropdown menus:

- Specimen info:** Includes fields for Specimen code, Secondary code, Entered by, and Collection.
- Taxonomy:** Includes fields for Temporary field species name, Author, Kingdom, Family, Genus, Species, Subspecies, Order, Identified by, Date, Det. St., and Type. A blue arrow points to the 'Type' dropdown menu, which contains: Specimen, Allotype, Cotype, Epitype.
- Sampling information:** Includes fields for Code, Date, Continent, City, Coordinates, Latitude, Longitude, Collectors, Expedition, and Location notes. A blue arrow points to the 'Exact site\*' dropdown menu, which contains: Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon. Another blue arrow points to the 'Collecting methods' list, which contains: Agassiz trawl, Amphipod Trap, Anchor, Angling, Apstein net, Argos buoy, Artificial substrate frame, Aspirator, Baited traps, balance à crabes, Beam trawl.

The form also includes a navigation bar at the bottom with buttons for 'New record', 'Save record', 'Clean content', 'Duplicate record', and a 'Go to record n°' field.

Figure 19. Examples of lists in the form

To make the navigation easier if there are a lot of records, a field "Go to record n°" has been added at the bottom of the form.

“Search” tab contains some fields to do a search in data. Search can be done on the most important fields of each section. It’s only a help to quickly find back one or more lines. If there are results, you can navigate through the results only with the navigation buttons. Click on “Reset” to go back to the whole set of data.

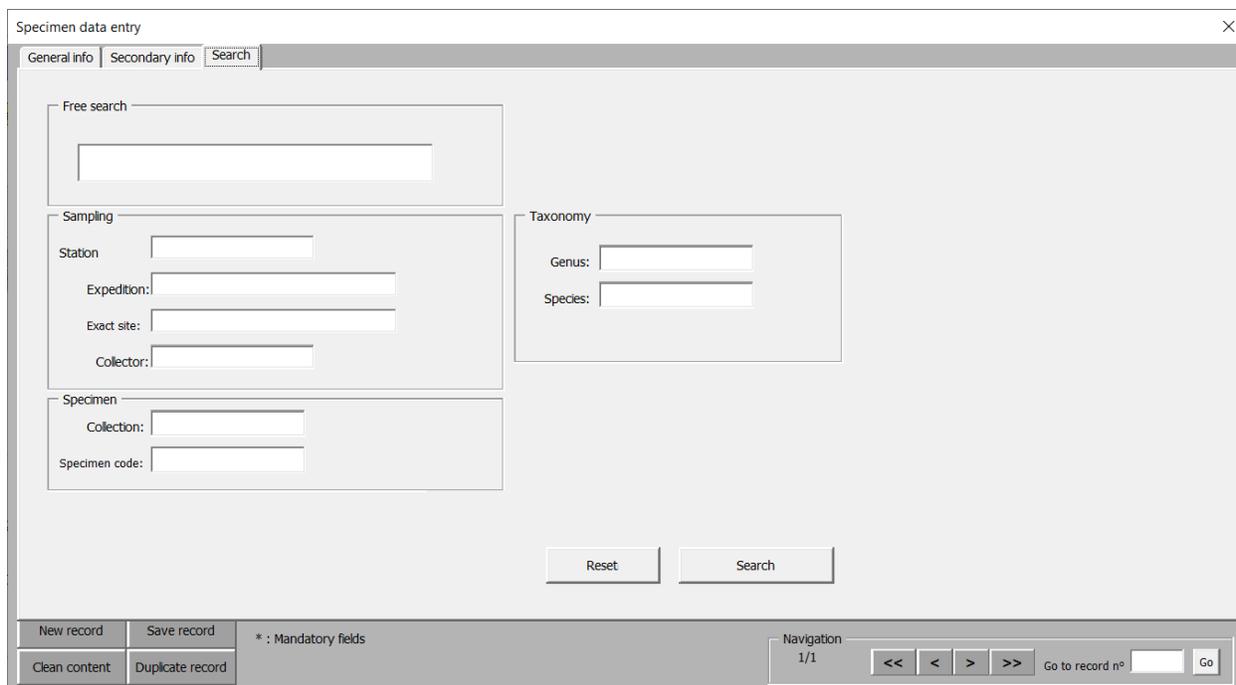


Figure 20. Search tab

#### d. Export

Data may be exported to DaRWIN.

Three csv files are generated: 1=taxonomy, 2=locations and 3=specimens. These 3 files are imported in DaRWIN in 3 consecutive steps. These files are generated automatically by clicking on the button “Export to Darwin” of sheet 1(Code). A popup window will ask you where to save the files.

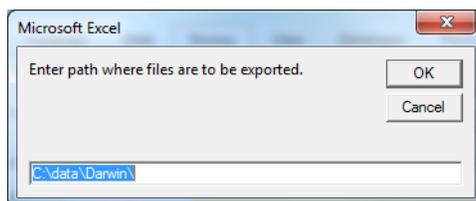


Figure 21. Export popup window

d.1 Taxonomy check

The last button on the first sheet is “Check taxonomy”.

It launches a browser to display a web service allowing users to check the taxonomy of the csv file “taxonomy” exported with the button “Export to DarWIN”. Taxonomy is checked against GBIF, IUCN, WoRMS, Fishbase:

**Welcome to the Natural Heritage taxonomy checker**

Mail :

Select Tab-delimited to upload:  Darwin\_im...\_taxo.txt

Has header row :

Column index of the name field (first = 1) :

Column index of the kingdom field (first = 1) [optional] :

DARWIN (RBINS):

GBIF:

GBIF (Vernacular names):

IUCN:

WORMS:

Figure 22. Taxonomy check interface

finishedPAGE = 1 NB\_PAGES = 1

Match type	Count
EXACT_OTHER_AUTHOR	2
MISSPELLING	2

[1 Last](#)

class	genus	order	family	phylum	gbif_id	kingdom	species	gbif_url	gbif_rank	gbif_class	gbif_genus	gbif_order	gbif_author	gbif_family	gbif_phylum
	Nephele		Sphingidae		5124095	Animalia	densoi	http://api.gbif.org/v1/species/match?verbose=true&name=Nephele+densoi+	species	Insecta	Nephele	Lepidoptera	Keferstein, 1870	Sphingidae	Arthropoda
	Hippotion		Sphingidae		1862368	Animalia	geryon	http://api.gbif.org/v1/species/match?verbose=true&name=Hippotion+geryon+	species	Insecta	Hippotion	Lepidoptera	Boisduval, 1875	Sphingidae	Arthropoda
	Hippotion		Sphingidae		1862368	Animalia	gerion	http://api.gbif.org/v1/species/match?verbose=true&name=Hippotion+gerion+	species	Insecta	Hippotion	Lepidoptera	Boisduval, 1875	Sphingidae	Arthropoda

Figure 23. Taxonomy check results

### 4.2.3 LibreOffice template

One of the challenges of the Natural Heritage project is to promote the use of Open Source solutions. The Royal Belgian Institute of Natural Sciences decided to evaluate a free open source tool as the MS-Excel template is highly dependent on the MS Office version and request an A5 licence as the Office 365 online is not compatible with the Visual basic macros.

The choice was made to use LibreOffice with the use of basic as macro language. The option to enter media files data is also not yet developed in this version.

#### a. Sheets

The use of the LibreOffice template is the same as the excel template.

Data are spread into 5 sheets and a form allows us to fill all the sheets together, for one record. The number of sheets is smaller because the data are organized differently.

Here are the columns of the 5 sheets:

##### Taxonomy

Taxonomy												
Code	Temp. species field name	Family	Author	Genus	Author	Species	Author	Subspecies	Author			
test011	Felis à points noirs	Felidae	Gray	Thriachomys	Trouessart, 1880	Thriachomys aperoides	(Lund, 1839)					
test021	Big ape	Pongidae	Elliot	Pan	Oken, 1816	Pan troglodytes	(Blumenbach, 1775)					
test031												
test041				Felis	Linnaeus, 1758							
test051												

Relationship												
Identifier	Day	Month	Year	Type	Taxonomy notes	Type	Parasite species	Parasite code	Host species	Host code		
Yoyen Emmanuel	13	03	1955	Cotype	perhaps new species	Host	Camexlectularius s	rmca1023s				
Yoyen Emmanuel	10	12	1953	Allotype	perhaps new species	Parasite			ganjs lupus	rmca001		
						Parasite			ganjs lupus	rmca001		
	10	12	1953	Allotype-Holotype	perhaps new species	Parasite			ganjs lupus	rmca001		

##### Specimen

Specimen info																				
Specimen code	Secondary code	Collection	IG	Entered by	Media URL	General notes	Type	From	Day	Month	Year	Notes	Sex	Life stage	Males	Females	Juveniles	Total	Notes	
test011	B-023	coll1s	IG1s	ij1ts	file1s	No general notes	Donation	Achille	08		1726	Given by director	hermaphrodite	subadult	2	14	3	20	2 males	
test021	C-019	coll2	IG2	ij2	file2, file3	notes	Acquisition	Agena	10	12	2015	3 months after death of owner	Female	juvenile	1	1	1	1	Juvenile probably female	
test031																				
test041																				

##### Location

Location names and description																	
Specimen code	Station number	Continent	Country	Original country name	State/province	Municipality	Ocean	Sea	Exact site	Degrees N/S	Minutes N/S	Seconds N/S	Degrees E/W	Minutes E/W	Seconds E/W	E/W	
test011	lccode611	Asia	Bahrain	old Bahrain	Miratus	Riffa	Southern Ocean	Bellinghousen Sea	center of country	1	1	1,280	S	1	1	1,500	W
test021	lccode621	Europe	Italy		Roma	Roma city	Atlantic ocean	Gulf of Guinea	center of country	12	4	50	S	5	53	47	E
test031	lccode631	Africa	Democratic Republic of Congo	belge		matadi	Atlantic ocean		near road to Kinshasa								
test041	lccode641	Europe	Belgium			brussels			test								

DMS coordinates																		
Decimal coord.	Latitude	Longitude	Origin of coordinates	Web service	Web service ID	Web service URL	GPS Weight Points	Line or area in WKT	Coordinates accuracy(m)	Original coordinates	From	To	Accuracy	From	To	Accuracy	Height	Accuracy
	-5.82566	13.46090	Historical-Label	OpenStreetMap Nominatim	4882355592	https://nominatim.org	no points1	Point(4 2,5 311)	1,25	org in DMS1	100	120	5	10	12	1,23	44	0,2
	50.84656	4.35170					no points2	Point(4 2,5 4)	100	org in DMS2	102	104	2	21	27	1	77	3

Figure 24.Columns of sheets Taxonomy, Specimen, Location

Sampling

Collecting Info												Collecting time				Notes			
Code	Collectors	Collecting method	Expedition project	Start day	Start month	Start year	Start hour	End day	End month	End year	End hour	Locality notes							
test011	Abbas	Agstein net,Armfall substrate frame	Project Angola	11	1956	12:13				1957	14:16	locality near the sea							
test021	Aaron	Agstein net,Aspirator	Italy 2020	01	1956				12	1956									
test031	Abbas	Arges buoy,Aspirator	Congo 2020	01	2020	10:15	20	03	2020	12:00		Attention, coordinates retrieved from OpenStreetMap based on city and country!							
test041	Abbas	Agstein net,Aspirator		01	2015				12	1956									

Ecology parameters												
Hour of measure(HH:MM)	Water temperature	pH	Salinity	Tot. asp. solids (ug/l)	Streamflow	uvV	Conductivity (µS/cm)	O2 dissolved (%)	O2 dissolved (ug/l)	Air pressure (hPa)	Air temperature	Relative humidity
12:45	21	6	0.00	7.00	1200	1200	120	45	10	1020	35	85.00
12:00												85.00
15:45												85.00
12:00												85.00

Biomes					
Ecology notes	Biogeog. realms Terr. and freshwater	Biogeographic realms Marine	Biome Terrestrial	Biome freshwater	Biome Marine
found in forest NA: Nearctic	Eastern Indo-Pacific	Tundra	Temperate floodplain rivers and wetlands	Littoral/intertidal zone	
found in forest NA: Nearctic	Eastern Indo-Pacific	Tundra	Temperate floodplain rivers and wetlands	Littoral/intertidal zone	
found in forest NA: Nearctic	Eastern Indo-Pacific	Tundra	Temperate floodplain rivers and wetlands	Littoral/intertidal zone	
found in forest NA: Nearctic	Eastern Indo-Pacific	Tundra	Temperate floodplain rivers and wetlands	Littoral/intertidal zone	

Storage

Storage location												Status		Specimen part		Container ID
Code	Institution	Building	Floor	Room	Lane	Column	Shelf	Status		Specimen part		Container ID				
1	test011	RBINS RBINS RBINS RBINS RBINS RBINS	Alle des dinos alle6 Main building	1 2 5 14	1 2 6 1	1 2 4 1	1 2 6 1	1 2 8 1	fresh--good state--Loan	fresh--good state--	head body leg arm parts	Container01 1 1 1				
2	test021	RMCA RMCA	CAPA1 Palais de l'Afrique3	1 2	1 2	1 2	1 2	1 2	fresh--good state--Loan	fresh--good state--	head body					
3	test031	RBINS RBINS RBINS RBINS RBINS	Alle des dinos alle6 Main building	1 2 5 1	1 2 6 1	1 2 4 1	1 2 6 1	1 2 8 1	fresh--good state--Loan	fresh--good state--	head body leg arm					
4	test041	RMCA														

Status - Container - Type of medium		Subcontainer ID		Subcontainer type		Subcontainer medium		Storage notes		DNA		Vertical position		Notes
Container type	Container medium	Subcontainer ID	Subcontainer type	Subcontainer medium	Storage notes	Tag number	DNA box	Tube number	Horizontal position	Vertical position	Notes			
Microscopic slide Alcohol flask jar	alcohol algar alcohol	SubID12 SubID22 None	Microscopic slide Alcohol flask jar	alcohol algar alcohol	storage in boxes imported	tag011	box11	tube11	21	41	no details1			
Microscopic slide Alcohol flask	alcohol algar	SubID12 SubID22	Microscopic slide Alcohol flask	alcohol algar	storage in boxes imported	tag011	box11	tube11	21	41	no details1			
Microscopic slide Alcohol flask jar	alcohol algar alcohol	SubID12 SubID22 None												

Figure 25. Columns of sheets Sampling and Storage

b. Buttons

As for the excel template, it's much easier to enter data via the form. There is only one form here (no field form). So there are 3 buttons on the first sheet:

A	B
	Code
Input form	Specimen code
	1 test011
	2 test021
Export to Darwin files	3 test031
	4 test041
Check taxonomy	5 test051
R	

Figure 26. Buttons on the first sheet.

A first one to open the form, a second to export files to DaRWIN and the last one to check taxonomy.

c. Forms

Use of the LibreOffice form is the same as for the excel template. Navigation buttons allow you to go from one line/record to another and 4 buttons allow you to save, duplicate, clean and add a record (these buttons are displayed in this form as icons).

As in the excel template, data are displayed on several screens (5 screens here in place of 2 in excel). To go from one screen to the other, click on the big buttons on the right. A special button "Search" shows a screen where you can search data.

First screen is Taxonomy. It contains info about taxonomy, type, identification, interspecies relations.

The screenshot shows the first screen of the LibreOffice form. It features several sections:
 

- Taxonomy:** Includes fields for 'Temporary taxon name' (Felis à points noirs), 'Family' (Echimyidae), 'Genus' (Thrichomys), and 'Species' (Thrichomys apereoides). It also has a 'Type' dropdown menu with options like Cotypte, Epitype, and Holotype, and a 'Notes' field containing 'perhaps new speciesss'.
- Identification:** Includes 'Identified by' (Vreven Emmanuel) and 'Date' (13/03/1955).
- Interspecies relations:** Includes 'This species is' (Host) and 'of' (Cimex lectularius s).

 On the right side, there is a navigation sidebar with buttons for 'Taxonomy', 'Specimen', 'Location', 'Sampling', and 'Storage'. Below the sidebar are icons for save, add, duplicate, and delete, and a 'Search' button. Blue arrows point to these elements with labels: 'Navigation buttons', 'Buttons to change section', 'Buttons for actions on record: save, New, Duplicate, Clean form', and 'Button to search'.

Figure 27. First screen of the LibreOffice form. Buttons.

Second screen is Specimen and contains info about codes, acquisition, sex, stage, counts.

The screenshot shows the second screen of the template, titled 'Specimen info'. It contains:
 

- Specimen info:** Fields for 'Collection' (coll1s), 'Entered by' (jims), 'Media URL' (file1s), 'I.G.' (IG1s), 'Main code' (test011a), and 'Sec. code' (B-023). A 'General notes' field contains 'No general notes'.
- Acquisition:** Fields for 'Type' (Donation), 'Date' (08/1726), and 'From' (Achille). A 'Notes' field contains 'Given by director'.
- Sex - Stage - Counts:** Fields for 'Sex' (hermaphrodite) and 'Stage' (subadult). A 'Counts' table shows: Total (20), Males (2), Females (14), and Juveniles (3). A 'Notes' field contains '2 males'.

 The right sidebar and bottom navigation buttons are identical to the first screen.

Figure 28. Second screen of the template, Specimen

Third screen is Location and contains info about exact geographic place, coordinates, altitude, depth and more technical data.

Figure 29. Third screen of the template, Location

A special function has been added in this LibreOffice version: it's possible to get the coordinates of a place, based on the country and the municipality. It works with a webservice as OpenStreetMap. When you have filled in the country and municipality, click on the button "Get coordinates". If it can find coordinates, they are written in decimal latitude and longitude and the origin of the coordinates is written below: calculated from OpenStreetMap.

Figure 30. Coordinates fields of the form.

A fourth screen is the sampling. It contains all data about how and when specimens were collected and also local ecological info, as well as larger ecological info in Biogeography.

Figure 31. Fourth screen of the template, Sampling

The last screen is Storage. As for the excel, many parts may be added by clicking on the “Add a part”.

Figure 32. Fifth screen of the template, Storage

There is a special screen to do a search in data. Search can be done on the most important fields of each section. It's only a help to quickly find back one or more lines. If

there are results, you can navigate through the results only with the navigation buttons. Click on “Reset filters” to go back to the whole set of data.

Figure 33. Search screen of the template

#### d. Export

Button “Export to Darwin files” on first sheet has the same function as in the excel, to export files that have to be imported in DaRWIN

#### e. Taxonomy check

Button “Check taxonomy” has the same function as in the excel template.

### 4.3 Integration of previous databases and Import of data in DaRWIN

The import procedure is a very important tool for the specimen data:

The total of records at the RMCA DaRWIN is 695.704 records, corresponding to 1.912.700 specimens. The import procedure allowed us to import 259.883 records which represent 37,3 % of the total of the database.

The total of records at the RBINS DaRWIN is 696.446 records, corresponding to 4.360.000 specimens of which 73.800 were imported with the excel template (10,6 %).

### **4.3.1 Mapping of the RBINS MISTA database**

Data from the RBINS MISTA database (polar missions in the Antarctic), originally in Microsoft Access format have been mapped to the DaRWIN importation templates and would be ready to be imported into DaRWIN after check from the scientists. This database was chosen as a case study for the Natural Heritage project as it was a very complex database including specimens from different institutions.

In contrast to other collections, MISTA has taken more time because data were in a very complex database and it has been difficult to extract data correctly. MISTA also influenced the original development of DaRWIN, for example by putting dates in the geographical data because data of MISTA contain a majority of specimens caught in sea (Antarctica) and during cruises. A catch can begin at a point A at day J and end at a point B at day J+3, which is not the case with terrestrial specimens caught at a precise place. Stations and expeditions were also as important in the original database as specimens and because of this, it influenced the development of an import in 3 steps: taxonomy, localities and specimens. This allows to import first a list of stations based on cruises.

The data are now imported in a working study in DaRWIN and will be integrated in the main collections after a final checking of the data.

A total of 5.831 records of MISTA corresponding to 132.190 specimens were imported. This is the equivalent of 1 FTE of manual encoder during 1 Year.

### **4.3.2 The RBINS Geology Collection**

The collections of Geology use Microsoft Access as a database. A previous attempt of import in RBINS DaRWIN was made in 2017 but the import was cancelled thanks to import errors in the validation of sampling locations.

It was now possible to import the data again using the 3 steps import procedure.

A total of 40.069 records are now available in the DaRWIN Collection Management System.

### **4.3.3 The RBINS Paleontology Collection**

The collections of Paleontology also use Microsoft Access as a database.

Data was exported as XLS files and templates were prepared for further import in DaRWIN. The process is not yet completed as we need first to control the chronostratigraphy reference system existing in the main database.

More than 40.000 type specimens will be imported with this procedure in 2021.

### **4.3.4 RMCA zoology**

Vertebrate data from the DataPerfect/Drosera system have been imported by using the template in 2019 and 2020.

This concerns the following collections:

- 21.216 records of Reptilia
- 16.069 records of Amphibians
- 16.000 records of Ornithology
- 451 records of ichthyology

Data from the Invertebrate collections were also imported:

- 1.276 records of trichoptera
- 28.936 records of Acari
- 185 records of Ephemeroptera
- 11.830 records of Crustacea
- 9.220 records of Myriapoda
- 1.541 records of Vermes
- 67.747 records of Coleoptera
- 1.837 records of Echinodermata
- 275 records of snails

#### **4.3.5 Mapping of RMCA wood biology data**

Data from the wood biology department have been successfully imported into DaRWIN in 2020, using the tab-delimited templates presented above. These data (83300 records) were originally conserved in offline Excel format.

Authors: Jean-Marc Hegers, Franck Theeten, Marielle Adam & Patrick Semal