USER MANUAL

TEMPLATE ABCDIMPORT2DARWIN: ENCODING DATA IN XML-FILE (ABCD SCHEMA STRUCTURE)

Version		Author(s)	Date	Description
1.0	ABCDImport2DaRWIN _Paleontology.xlsm = "CLASSICAL" version	MAdam	August 2014	Small adaptations and debugging
1.0 bis	ABCDImport2DaRWIN _Paleontology _taxonFullName.xlsm = "_taxonFullName" version	MAdam	December 2014	New version with simplified taxonomy

For information about ABCD - Access to Biological Collection Data: http://wiki.tdwg.org/ABCD
For information about EFG extension – Extension for Geosciences: http://www.geocase.eu/efg

TABLE OF CONTENTS

DECEMBER 2014 – UPDATES

After extensive tests made within the IT, it was shown that the “Create missing taxon” tool could lead to unforeseen errors. It was therefore decided that the creation of taxonomy had to be made in a previous step to the specimens import. A template for importing of taxonomy has therefore been prepared.

This template for taxonomy comes in an excel format. Each taxonomic level corresponds to a column. An export in xml format produces an xml file that can be imported directly into DaRWIN via an import interface for taxonomy, replacing the "Create missing taxon" button that is not available anymore.

For more information about this template, you can also read the user manual for the taxonomy import template.

From now on, 2 methods are available for specimens import…

Method 1 – Use of the revised template (_taxonFullName version) and of the template for taxonomy import

(preferred method)

The principle for this method follows the “DaRWIN good practice” that is to prepare your taxonomy previously to the import of specimens into DaRWIN.

How does it work ?

You have 2 templates at your disposal: the revised template for specimens “ABCDImport2DaRWIN_General_taxonFullName.xlsm” and a new template for importing taxonomy “TaxonomyImport.xlsm”.

The template for importing taxonomy should be used to create the missing taxonomy corresponding to the specimens you wish to import. More information about this template is available in the dedicated user manual.

In the new specimens template, only the “taxonFullName” field remains for taxonomy. You have to give the taxon name/identification with the exact same spelling than what is present in the “name” field of the taxonomic entry in DaRWIN. Remember that if you wish to give the full taxonomy in the template, you can add as many columns as you wish. At the moment of the export, a popup message will appear, warning you that unrecognized columns will not be present in your export file…

Besides, a tool was also developed to check if an entry already exists in the Taxonomy Catalogue in DaRWIN. It is the “Check_taxonFullName.xlsm » excel file. You need to enable macros and have an active internet connection to make it work properly. This tool will, through a custom ribbon and button, scan the DaRWIN database, to see if what is in the “taxonFullName” column already exists. You only have to copy-paste the column “taxonFullName” from the specimens template, to the “taxonFullName” column in the “taxonFullName” woorsheet of the excel file, and click on “CheckTaxon” in the custom ribbon. This way, you can check before your specimens import, if the taxon name you used in the “taxonFullName” columns will be recognized. If no, you can prepare a taxonomy template for importing taxonomy or create the missing entries directly through the DaRWIN interface.

Method 2 – Use of the template released in September 2014

The template released in September 2014 is still functional. You can continue to work with it. Only the use of the “taxonFullName” column is slightly different.

What is different ?

When you import a file, during the specimens check, you will be asked to select the corresponding entry in the Taxonomy Catalogue in DaRWIN (or to create the missing entry) for specimens which have unrecognized taxon names. Only one correction by taxon and by file is necessary: the name will be corrected automatically in each line where the taxon name is the same.

Warning ! Keep in ming that the column “taxonFullName” is used for checking if the taxon name is already present in DaRWIN.

> For identification up to lower taxonomical levels (genus or below), the column “taxonFullName” is used for checking if the taxon name is already present in DaRWIN. It is therefore strongly advise to use the exact same taxon name in this field “taxonFullName” than what is present in DaRWIN. If this column is empty at the moment of the export, it will be automatically produced by concatenation of what is in the “genus”, “species, « subspecies », « form_variety » and « author_year » columns. Risks of errors are therefore high…

> For identification to higher levels (above genus), the system check the name based on the lowest level referenced in the template.

When you fill many levels for taxonomy in the template, there is also a risk for unmatching if the name is spelled differently in DaRWIN or if the hierarchy is different in DaRWIN (one sublevel missing for example). We therefore advise you to specify only one or two higher levels corresponding exactly to what was encoded in DaRWIN and be careful with the name spelling. Example: only fill the “genus”, “species” and “taxonFullName” columns or, for an identification up to the genus level, only fill the direct parent, like the “familia” column.

Comparison

	Method 1 ***ABCDImport2DaRWIN_General_taxonFullName.xlsm***	**Method 2** ***ABCDImport2DaRWIN_General.xlsm***
« taxonFullName » column	Only way to precise an identification; if empty, no identification… Can be a name for any taxonomical level.	Used to compare with the names in Taxonomy Catalogue in DaRWIN, for taxon names from genus (and levels below); if empty, automatically recomposed.
Higher levels and taxonomical hierarchy	No higher levels available in the xml ABCD exported file.	Higher levels, when specified, have to be spelled exactly the same way than in DaRWIN and follow the exact same hierarchy…
Conditions for immediate matching during specimens import	An entry in the Taxonomy Catalogue in DaRWIN has to exactly match with what is specified in the “taxonFullName” column of the specimens template. Warning! The matching is made for the character string specified in the “taxonFullName” column without special characters and lowercase (parenthesis, for example, are not taken into account for the matching).	For identification until genus level or below: > taxonFullName matching to a DaRWIN entry in the Taxonomy Catalogue, for the lower identification level that was specified in the specimens template > for each specified parents in the template, exact same spelling and hierarchy than in DaRWIN for the corresponding entry is obligatory > Advise: specify only the direct parent that exists in DaRWIN. For higher levels identification: >taxonFullName not useful > for each specified parents in the template, exact same spelling and hierarchy than in DaRWIN for the corresponding entry is obligatory > Advise: specify only the direct parent that exists in DaRWIN.
Additional tools	> “TaxonomyImport.xlsm” to prepare taxonomy before specimens import > “Check_taxonFullName.xlsm” to check if names in “taxonFullName” column are recognized	

USER INFORMATION

Template tested with Excel 2007.

1. Open the template document and save it with a custom name "[CUSTOMNAME].xlsm".

If a Security Warning appears, then Enable macros (click options… and select "Enable this content").

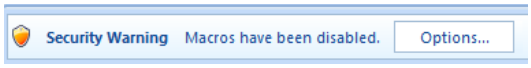


Figure 1 - Security Warning for macros

You can also access these options through the Excel Options > Trust Center ❶. Then, click on "Trust Center Settings" ❷. In the new window, in "Macro Settings" ❸, check "Enable all macros" ❹.

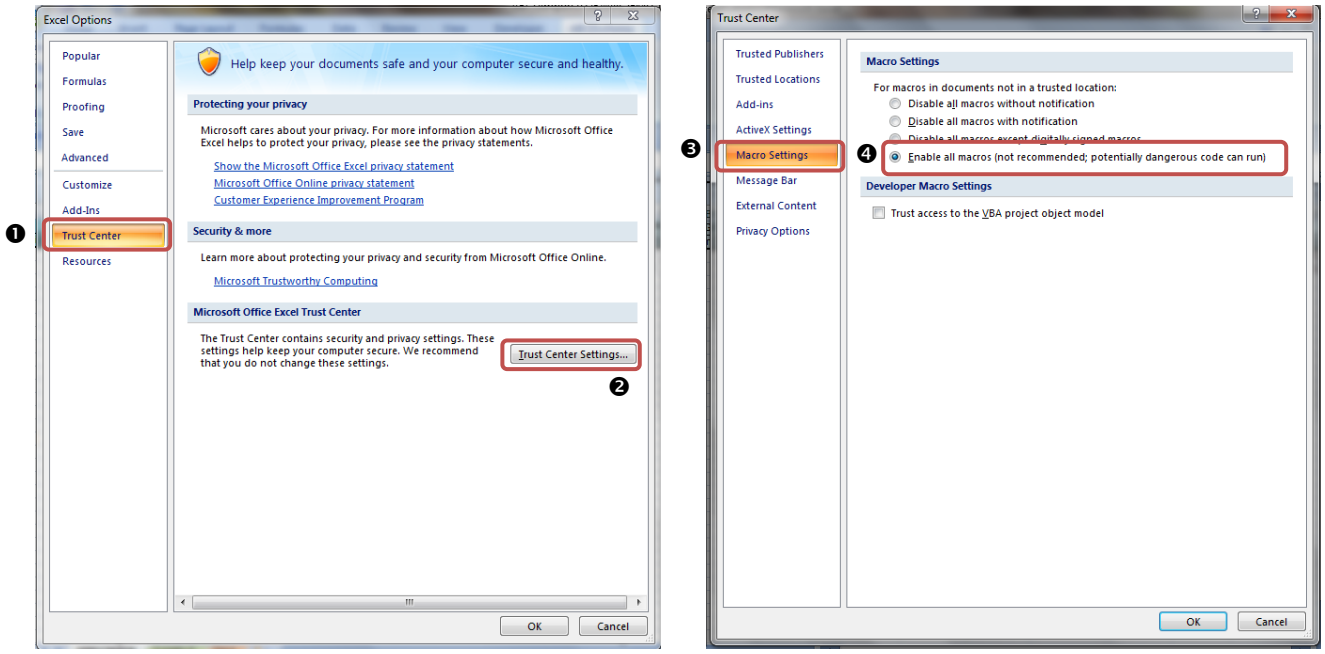


Figure 2 - Excel options: enable macros

2. Fill in the worksheets SPECIMEN

Information about the specimen is stored in a worksheet named SPECIMEN. Information about the taxonomy (phylum, class, order and family) is stored in a worksheet named TAXONOMY.

3. Before export, some checks can be made

For more details about the tools available for checking before export, see Technical Information. The *Quick Practical Guide* below gives you some practical information on how to use of the template. It gives you some minimal advices to complete the template properly. Nevertheless, we strongly advise you to read this user manual completely.

4. Click on "Export2ABCD" to export the data

When running the Export2ABCD code, an XML-file following the ABCD schema structure is produced and this XML-file can be saved in a folder defined by the user. During the export, Excel is unavailable (the worksheet could disappear or turn into blue during the process). This may take several minutes, depending on the number of lines and the quantity of information.

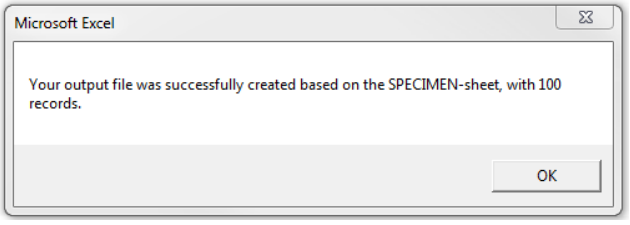


Figure 3 - Summary of your export

QUICK PRATICAL GUIDE

Template structure

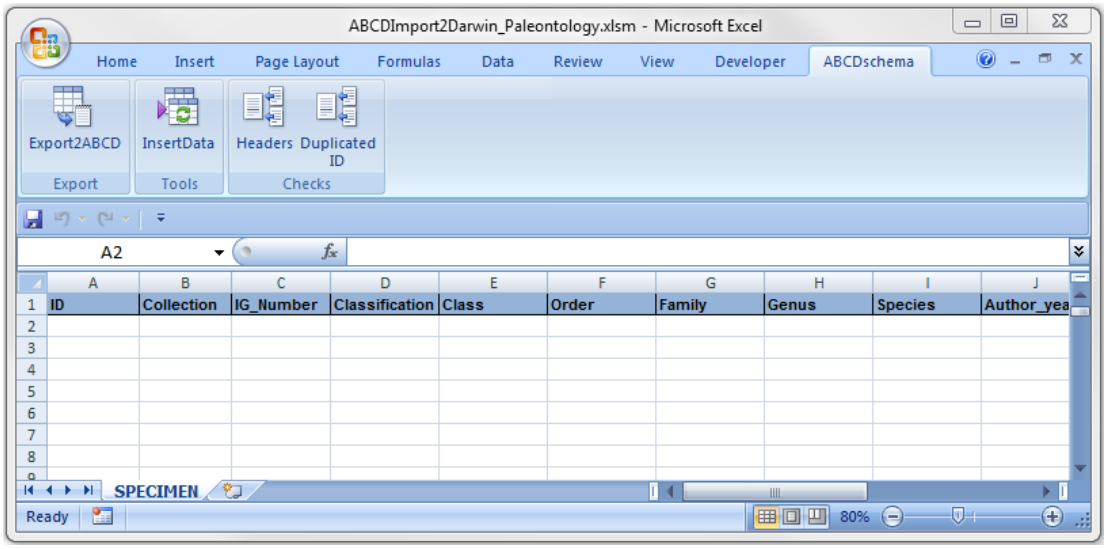


Figure 4 - Template structure

Verify the presence of:

- **The worksheet named 'SPECIMEN', containing information about your records**
- **A title for each column**, written **in the first row** and using the exact same name and spelling as in the pre-established list of supported fields available in the appendix of this document. If this condition is not fulfilled, the information will not be exported to the XML ABCD formatted file. You can add as many columns as you wish, for internal purpose but the information they contain will not be exported to the XML ABCD formatted file.
- **A column for IDs**, named 'ID'
  An ID is not required, but remember that links between specimens and hosts or other kind of units (e.g., part of specimen) are established thanks to it. Using the voucher/accession number attributed in the museum collections as the ID will allow to retrieve information of each linked object within DaRWIN. If you use your own IDs, this link will not be possible…

Values

No values are required. Nevertheless, remember that this tool helps you to import data into DaRWIN, as collection management system. Incomplete information is therefore of limited interest and relevance.

Only use special characters (=, +, /, *, &, #, $, etc.) in Excel cells **if strictly necessary**, neither as first character nor in subsequent characters. Excel is a data analysis software, therefore it will try to interpret these characters and this may raise errors during export to the XML ABCD formatted file.

Note: Since the use of "." (dot) in specimen ID is frequent, several tests were made to see if an error could occur. It appears from these tests that the use of a dot surrounded by letters is not a problem. Example: INV.2367. The collection manager can define a default prefix in the collection settings. Only the numeric part should therefore be specified as specimen ID, the prefix being automatically added during the import in the collection.

If you don't have information for a cell, leave it blank. This should limit the presence of uninformative values in your exported XML ABCD formatted file or possibly the number of errors to correct during the import into DaRWIN.

For some columns, the program expects **specific formats or predefined values** given in the list of supported fields available in the appendix of this document. If this format is not respected, the value cannot be taken into account or replaced by default values and you could end up with errors or unexpected values in your exported XML ABCD formatted file.

Checks

You can check the presence of duplicated IDs and the correct structure of the template (name of SPECIMEN-sheet and column titles), by using the **buttons in the groups "Tools" and "Checks" in the custom "ABCDschema" menu**.

Use filter (select the heading row, click on "Filter" in the "Data" menu) to check your values. You can **see whether the expected values or formats were used**.

One template = one xml file = import in one collection in DaRWIN

It is not possible to import only a part of records from one xml ABCD formatted file in one collection and the remaining records from the same file in another collection. Once the xml file is created, each record it contains can be imported in one unique collection.
If some specimen are not yet published or should not be visible for everybody, they should be stored in another template and imported in a private collection. You can always transfer these specimens to another collection once they are published. You can for example create a collection and add a public sub collection and a private one, that can be grouped later.

Number of "exportable" rows

The template should not contain more than 3000 rows. If you wish to import more records than 3000 in one collection DaRWIN, you should split this dataset into different templates of 3000 rows.

TEMPLATE STRUCTURE

The template for encoding consists of the worksheet SPECIMEN, containing data with regard to the specimen (collecting, identification, etc.) where the “ID” column is required (even if the cells are left blank)

This template was designed to minimize the requirements when encoding, which supposes that your data is clean. This allows for a large range of data that can be encoded, but also implies that the values encoded in the Excel cells will be exported as such. In some cases, for example in fields containing date, character strings or alphanumerical data could raise errors during the import.

During the export, the macro will look for which information is stored in the template thanks to the title row (located in the first row of the worksheet). Only data stored in columns where the title was recognized will be exported to the XML ABCD formatted file. Errors will occur if the titles are not well spelled and/or not in the first row of the worksheets. The order of columns doesn’t matter. You can add as many columns as you wish, keeping in mind that they will not be recognized and thus the information they contain will not be exported to the XML ABCD formatted file.

For more information about the format restrictions and correct title spelling for each field, see the list of supported fields available in the appendix of this document.

A tool was added to verify the correct mapping of columns in the template before export. It will tell you which columns are not recognized and if the SPECIMEN-sheet is well found. If the required IDs column (‘specimenID’) is missing the export will be stopped and an error will be raised.

A warning message will also pop up if a column is not recognized, telling you which headers are concerned. You can decide to go on with the export, by clicking “Yes”: the program export your data without taking unrecognized columns into account. You can abort the export, by clicking on “Cancel” in case of misspelling…

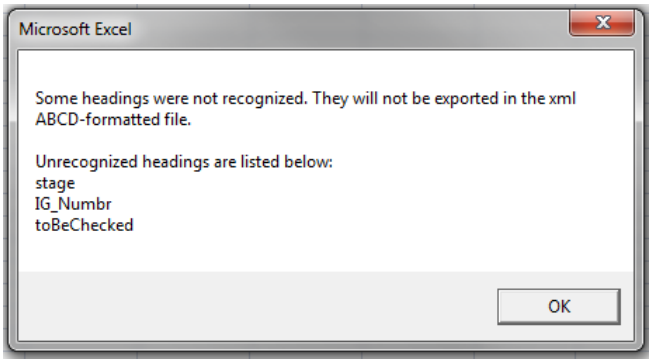


Figure 5 - Result of columns mapping

In this example (Figure 5):

- toBeChecked could be additional information, for internal purpose
- IG_Numbr is obviously misspelled and should be corrected to IG_Number
- stage is available for the export, but the correct name for this field is epoch

Except columns for IDs, you can decide to keep only the relevant columns for the data you want to store in the template. The presence or absence of columns is completely customizable. For example, if you never mention the ocean or the sea where you collected your specimens, these columns do not have to be present in your template.

ADDITIONAL INFORMATION TO FILL THE TEMPLATE

Some requirements or limitations for filling the template are necessary in order to concur with the ABCD schema or with the DaRWIN structure. They are listed in the following paragraphs.

1. Introduction

You don't have to complete each cell. No values are required. Nevertheless, remember that you use this template to import your data into DaRWIN, as collection management system. Incomplete information is therefore of limited interest and relevance, for you and for any other scientist.

If you don't have any information for a cell, leave it blank. For example, when you don't know each level for taxonomy, do not add a dash or a question mark, leave the cell empty. The macro doesn't make the difference between real values and a dash, a question mark, "NA", etc. Consequently, it will export the exact character string that it reads in the cells and you will end up with this uninformative values in your exported XML ABCD formatted or possibly with more errors to correct during the import into DaRWIN.

Only use special characters (=, +, /, *, &, #, $, etc.) in Excel cells **if strictly necessary**, neither as first character nor in subsequent characters. Indeed, Excel is a data analysis software. It will therefore try to interpret these characters and this may raise errors during export to the XML ABCD formatted file.

Nevertheless, as the use of a dot (".") in specimen ID is frequent, several tests were made to see if an error could occur. It appears from these tests that the use of a dot surrounded by letters is not a problem. Example: INV.2367. Remember that it will be stored as such in the "code" field of the Codes Widget, and not split into a prefix and a code. If you wish to add a prefix for one collection, the curator of the collection should edit the collection to explicitly give a default prefix to the code.

For some columns, the program expects specific formats or predefined values given in the list of supported fields available in the appendix of this document. If this format is not respected, the value cannot be taken into account or can be replaced by default values and you could end up with errors or unexpected values in your exported XML ABCD formatted file.

Using the fields for other kinds of information than what is expected will result in errors being raised or may lead to irrelevant information in the database.

2. ID number

a. Column names for IDs

For each row, an ID should be present. This ID should be unique and correspond to the voucher ID/accession ID attributed to the object in the museum collections. Nevertheless, for specimen that you don't own, you could possibly not know the ID. You can therefore use a custom ID, to allow cross-referencing between the information about the specimen and the samples.

Following the definition of the ABCD concept "UnitID", *"The unit ID should provide the key by which a specimen or specimen component can be identified. Preferrably, the unit ID should be stable in the database, so that it also can be used to find the same record again (e.g. for data exchange purposes)."* [1]

The identification number of an object is used to distinguish the object from other objects in the collection or department. It should be unique within the collection. This uniqueness also allows to build relationships between objects.

Within an institution, these unique identification numbers can be completed by an additional identifier, as the acronym of the institution and the collection or department to which it belongs. In this way, uniqueness is extended within the institution and even beyond the institution.

[1] http://wiki.tdwg.org/twiki/bin/view/ABCD/AbcdConcept0140

One ***suggested*** format to construct your identification numbers is the following, ***but you can use the format that suits you the best***: [YYYY]_[CollectionOrDatasetCode]_[SubGroup]_[Iterative_nb]. The year should consist of 4 digits. The collection or dataset code may include an acronym representing the expedition and/or the institutional registration number. The subgroup may be the name or an acronym of the order/family concerned. The iterative number is a unique number in the collection/the department or the subgroup. Example: 2013_RBINS23134_AVES_01034.

Remember that an unique identifier (ID) will be attributed to each encoded specimen in the DaRWIN database at the moment the specimen is created. This ID is guaranteed to be unique and stable among the whole DaRWIN database. This database ID is not to be confused with the specimen ID that is used within departments and that is not guaranteed to be unique, even if more convenient to use for scientists and curators. The specimen ID is imported as a code, with the category "main", in the DaRWIN database. If the collection curator defined a default prefix and/or suffix for the collection, you just need to write the numeric part of the code in the template, and the prefix and/or suffix will be automatically added during import. Remember that the whole content of the specimen ID will be present in the field "Code" of the "Codes" widget in DaRWIN. Example: INV.2367 in the specimenID field will be stored in the "code" field in the "Codes" widget and not split into a prefix and a code. To have such a subdivision, the prefix "INV" has to be defined as default for the collection by the curator, and sole 2367 should be present in the specimenID field.

b. Duplicated IDs

Duplicates IDs are not allowed in the "ID" column. You can check the presence of duplicates for these by using the corresponding tool available in the ABCDschema menu. The reason behind that is if an association is set on one unit thanks to its ID, duplicated IDs could lead to irrelevant associations in DaRWIN (association with several records that have the same ID).

3. Insert Data Tool

This tool is available in the "ABCDschema" menu and allows the insertion of data in SPECIMEN-sheet from another excel file. This excel file must contain the same columns as in the template in the exact same order (see below). Moreover, the sheet containing data to copy must be named "ExportABCD". A query can be created in the access database to produce such a file (see appendix for more details).

ExportABCD-sheet and SPECIMEN-sheet structure for "classical" version of the template:

ID	Collection	IG_Number	Classification	Class	Order	Family	Genus	Species	Author_year	Epoch	Age	Age_bis	Country	Locality	Comment

ExportABCD-sheet and SPECIMEN-sheet structure for " taxonFullName" version of the template:

ID	Collection	IG_Number	TaxonFullName	Epoch	Age	Age_bis	Country	Locality	Comment

4. Pre-export checks

You can also check the presence of correctly named worksheets, the mapping of column names and the presence of duplicated IDs before exporting the data, to allow a correction. These checks are available by clicking the buttons "1. Headers" and "2. Duplicated ID" in the group "Checks" of the custom "ABCDschema" menu. A window will appear, listing the potential problems. It is highly recommended to run these checks before trying the export.

Finally, you can use the filter (select the heading row, click on "Filter" in the "Data" menu) to check your values. This way, you can see if you use only the expected values or formats in the different columns. For example, you can check if the same names have the same spelling or if fields where only specific values are supported do not contain erroneous values, etc.

In the example below (Figure 6), instead of "Zoological" or "Botanical", that are the two supported values for the "Classification" column, the values "Zoolgical" was written by mistake. Such values could not be present in the exported XML ABCD formatted file, or they could raise errors. Thanks to the filter, you can display only rows with these values, and correct them.

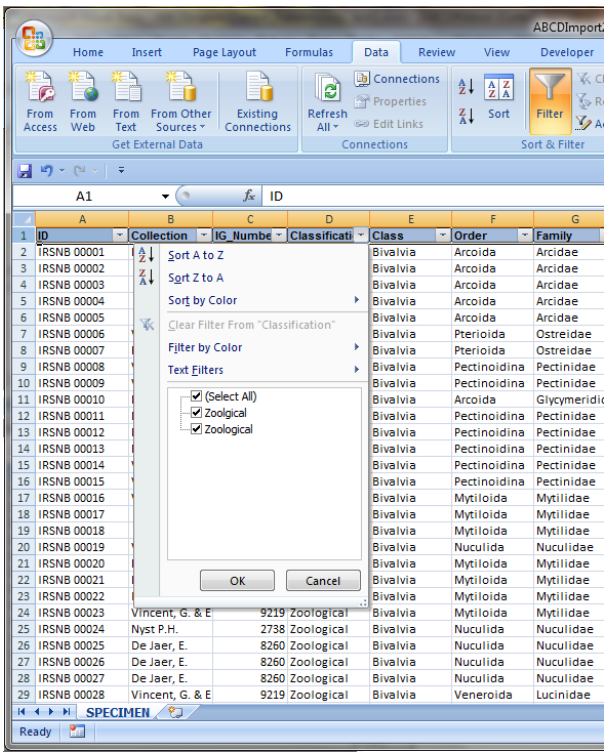


Figure 6 - Filter data in Excel

TECHNICAL INFORMATION – ABCDSCHEMA TAB

1. Commands

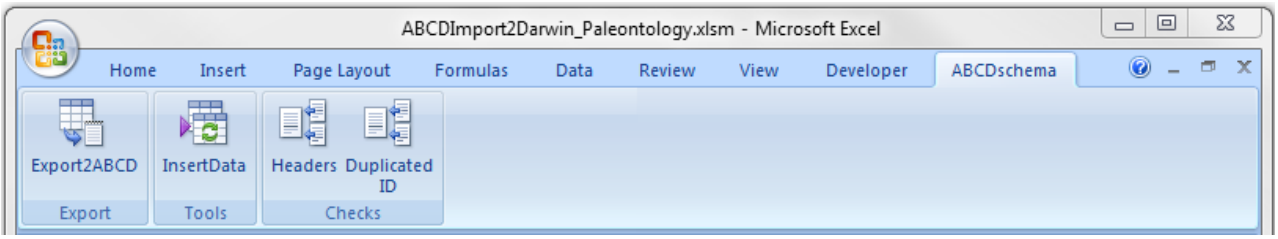


Figure 7 - ABCDschema Menu

a. Export group

- *Export2ABCD* : creates XML file that matches the ABCD schema with the data contained in the Excel file.

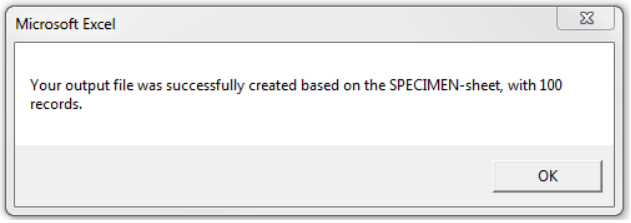


Figure 8 - Result of the export to XML ABCD formatted file

b. Tools group

- *Insert Data*: allows you to copy data from another workbook with the exact same structure as the template (same order of columns).

c. Checks group

- *1. Headers* : checks if each column title is recognized and if the SPECIMEN-sheet and the 'ID' column are found.
- *2. Duplicated IDs* : checks if no duplicated IDs are present.

2. Output

a. Name and extension

[USERDEFINED_NAME].XML

The export will create a file with extension .xml. You can choose the name and the folder where you wish to store this XML ABCD formatted file.

b. Structure

Globally, units are encoded step by step, looping within the SPECIMEN-sheet. One unit is created for each row.

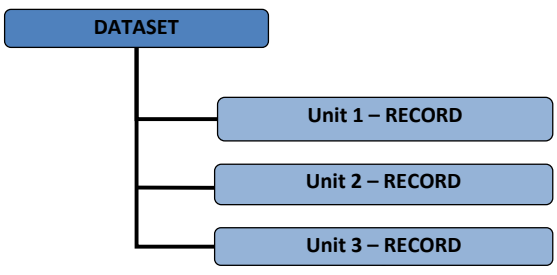


Figure 9 - Export XML file structure

GLOSSARY

XML and ABCD XSD schema

XML stands for eXtensible Markup Language and is a markup language much like HTML.

XML was created to structure, store, and transport information. Documents are therefore encoded in a format that is both human-readable and machine-readable. It is mainly used as intermediate format between two computers or softwares.

Information is structured through tags. These tags surround information by creating one element. Different elements can be fitted together, to create a hierarchical structure. A tag begins with "<" and ends with ">". Each element has a start-tag *<tag>* and end-tag *</tag>*. Empty tags takes the form *<tag />*.

XML Schema can be used as template for structuring information in your XML file. It rigorously defines the structure of your document. ABCD (Access to Biological Collection Data) is a predefined format to store biodiversity collections, developed by TDWG (Taxonomic Database Working Group). Several extension were developed:

- Extension for Geosciences (EFG)
- Extension for DNA data (ABCDDNA)
- Extension for herbarium collections (HISPID)

Visual Basic For Applications and macros

Excel has a language called VBA (Visual Basic for Applications). This language enables to program excel to automate several tasks. A macro is nothing but a set of instructions you give Excel in the VBA language.

The code for exporting your data filled in the template into an XML ABCD formatted file was prepared in a macro using VBA.

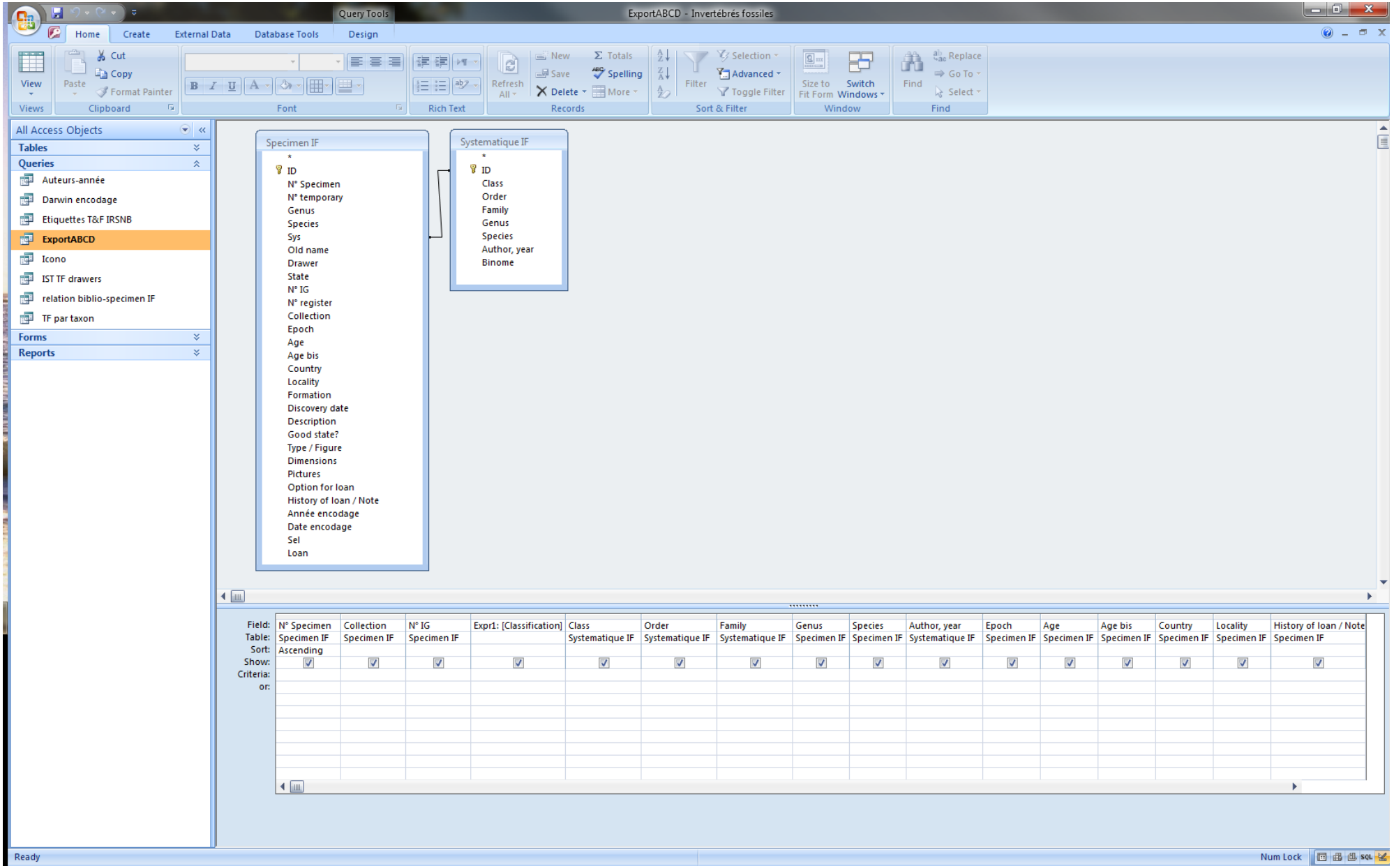
APPENDIX

1. General list of supported fields with expected format, description and example(s)

In the SPECIMEN-sheet, the following fields are supported corresponding to the information that can be retrieved in DaRWIN after import. Warning : Classification, Class, Order, Family, Genus, Species, Author_year are not available in the "_taxonFullName" version of the template

Field	Description	Example
ID	Unique identifier of specimen if exists. Proposed format: [YYYY]_[CollectionOrDatasetCode]_[SubGroup]_[Iterative_nb]. The year should consist of 4 digits. The collection or dataset code may include an acronym representing the expedition and/or the institutional registration number. The subgroup may be the name or an acronym of the order/family concerned. The iterative number is a unique number in the collection or the subgroup.	IRSNB_2738_BIV_00006
Collection	Name or code for the project, expedition, etc. for as complementary information for the collection name, choosen in DaRWIN at the moment of the import	Vincent, G. & E
IG_Number	Institutional number given to each new group of items acquired by the institution and recorded in the collection registers	9219
Classification	Is it zoological or botanical?	Zoological
Class	Class	Bivalvia
Order	Order	Pterioida
Family	Family	Ostreidae
Genus	Genus	Ostrea
Species	Species	wemmelensis
Author_year	Author, year or (Author, year) for the species	Glibert, 1936
Epoch	Epoch/Age (International chronostratigraphic nomenclature)	Eocene
Age	Age/Stage (International chronostratigraphic nomenclature)	Bartonian
Age_bis	Age/Stage (Local chronostratigraphic nomenclature)	Bartonien
Country	Country (administrative name)	Belgium
Locality	Locality (administrative name) or urban administrative division	Neder-Over-Heembeek
Comment	Additional information/remark about the specimen	valve fissurée et présence de fragments détachés

2. Query in Access DB



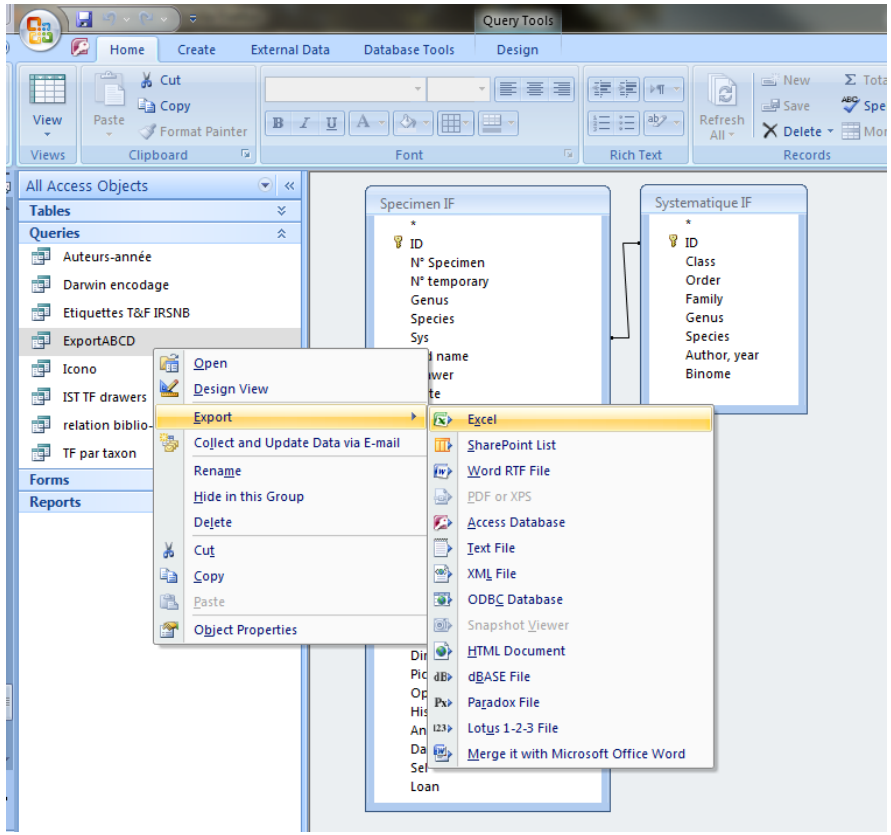
A query can be created in the DB access, that allows the export in excel of the data to import into DaRWIN.

This query should be named "ExportABCD". The order of the columns has to follow the same order than the template, as displayed below. A column classification is created after N° IG. Fill in the criteria needed for your export.

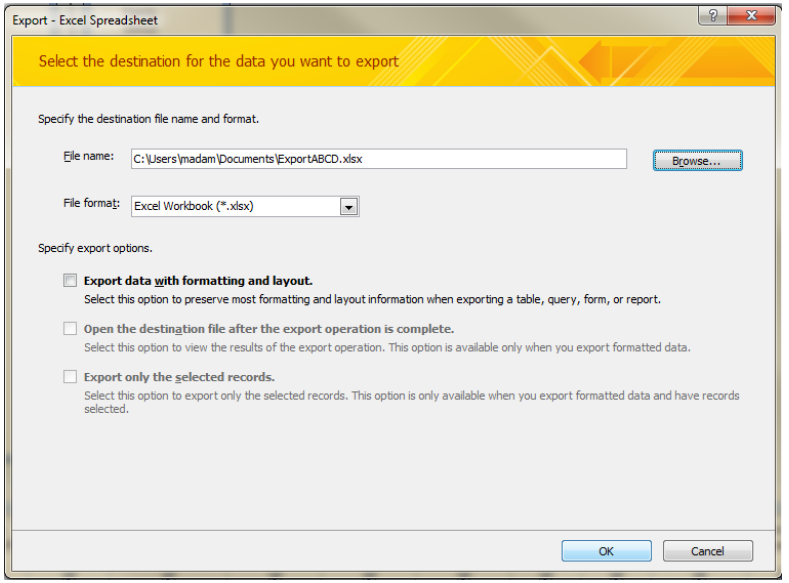
Field:	N° Specimen	Collection	N° IG	Expr1: [Classification]	Class	Order	Family	Genus	Species	Author, year	Epoch	Age	Age bis	Country	Locality	History of loan / Note
Table:	Specimen IF	Specimen IF	Specimen IF		Systematique IF	Systematique IF	Systematique IF	Specimen IF	Specimen IF	Systematique IF	Specimen IF	Specimen IF	Specimen IF	Specimen IF	Specimen IF	Specimen IF
Sort:	Ascending															
Show:	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Criteria:																
or:																

On right-click on the ABCDexport query, select export, then excel (❶). You will be asked to choose the repository where you wish to save the export file (❷). Then, you will have to precise what classification is concerned (ie Zoological or Botanical) (❸).

❶



❷



❸