STRUCTURING DATA INTO THE DARWIN CORE STANDARD

CESP Project: Strengthening Zimbabwe’s GBIF node through collaboration with GBIF Spain



# INTRODUCTION

Through this use case, we will convert an original database into the Darwin Core standard by mapping the original/verbatim fields with the DwC terms.

Our main objectives are:

* To become familiar with the Darwin Core standard terms.
* To find the equivalence between the source database and Darwin Core terms using templates.
* To discuss potential differences between mappings and know the best practices carried out in the community.

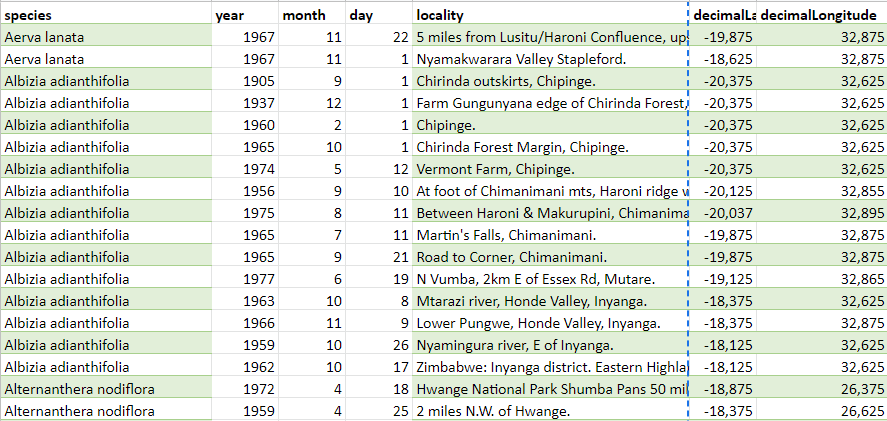
# YOU WILL NEED

* The exercise data [Trees of food importance in biodiversity hotspots of Zimbabwe.](https://drive.google.com/open?id=1bVdD38rpTx5fE_L-oaqPOumidgVT1D3w)
* [Simplified Darwin Core spreadsheet template](https://drive.google.com/open?id=10k-oj00U-aHpuMSyYaPH5OUBMdHa5ecM). It is a very simplified version with very few terms.
* Darwin Core quick reference guide <http://rs.tdwg.org/dwc/terms/index.htm>

INSTRUCTIONS

**Step 1. Become familiar with the original data and with the Darwin Core standard**

1. Study the original data provided *Trees of food importance in biodiversity hotspots of Zimbabwe*. Become familiar with it, identify the names of the fields and their content.



1. Look for the direct equivalence between the fields in your table and the Darwin Core terms. You can do this by adding a row below the header and moving the original information to the template. You can use the standard page to see the definitions of the terms in case they are not so intuitive.

**Step 2. Completion of mandatory fields**

1. You will notice that not all mandatory fields (highlighted in blue) in the template have a corresponding attribute with the original table. Read the definitions of these elements and complete as much data as possible.

|  |
| --- |
| [**basisOfRecord:**](https://dwc.tdwg.org/terms/#dwc:basisOfRecord) The specific nature of the data record. Recommended best practice is to use the standard label of one of the Darwin Core classes.   * PreservedSpecimen * FossilSpecimen * LivingSpecimen * HumanObservation * MachineObservation * Sample   [**institutionCode**](https://dwc.tdwg.org/terms/#dwc:institutionCode)**:** he name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record.  Examples: “Univalle”, “UQ”, “UCN”  [**collectionCode**](https://dwc.tdwg.org/terms/#dwc:collectionCode)**:** The name, acronym, coden, or initialism identifying the collection or data set from which the record was derived.  Examples: “HUV”, “PPH” (Proyecto páramos y humedales),  [**catalogNumber**](https://dwc.tdwg.org/terms/#dwc:catalogNumber)**:** An identifier (preferably unique) for the record within the data set or collection.  Example: “UV215”, “001”, “46-2300MI2008AV0021”  [**occurrenceID**](https://dwc.tdwg.org/terms/#dwc:occurrenceID)**:** An identifier for the Occurrence (as opposed to a particular digital record of the occurrence). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the occurrenceID globally unique.  [institutionCode]:[colectionCode]:[catalogNumber]  Example: “UCN:MH-ORNIT:46-2300MI2008AV0021”  **To build occurrenceID the elements institutionCode, collectioncode and catalogNumber must be documented. To perform the construction of this identifier in bulk in the data set you can use a combination function in Excel such as ‘CONCATENATE’. Located in the first cell for the first record in the OccurrenceID field type the following function:**  **=CONCATENAR(C2;":";D2;":";E2)\***  **(C2, D2 and E2 are the sample cells in which the values for institutionCode, collectioncode and catalogNumber are found. Select the ones that apply to your case.**  **Once "Enter" you will have your data successfully combined. Drag this function to run on all records. Once the identifiers are assigned, select them in their entirety, copy and paste over the same cells as “values only” format to ensure that the cells save the data as text and not as a formula.** |
| [**scientificName**](https://dwc.tdwg.org/terms/#dwc:scientificName)**:** The full scientific name, with authorship and date information if known. When forming part of an Identification, this should be the name in lowest level taxonomic rank that can be determined. This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term.  Ejemplos: "Roptrocerus typographi (Györfi, 1952)" [Genus + Specific Epithet +Scientific Name Authorship], "Vespertilionidae" [Famy], "Manis" [Genus], "Quercus agrifolia" [Genus + Specific Epithet]. |

**Step 3. Indirect mappings**

1. In cases where a direct mapping cannot be performed, indicate what actions you would perform on the fields to be able to do a correct mapping (e.g., split a field content, concatenating fields, etc.).
2. Make a list of other possible fields present in the template that should be present to share the data (may include some derivatives from the previous point). Bear in mind those values that can be inferred from other fields content or the metadata.

Examples:

* Country
* CountryCode
* EventDate
* VerbatimDecimalLatitude
* VerbatimDecimalLongitude
* VerbatimCoordinates

COMPLEMENTARY INFORMATION

- **Darwin Core quick reference guide:** <http://rs.tdwg.org/dwc/terms/index.htm>

- **Darwin Core Questions & Answers Site:** <https://github.com/tdwg/dwc-qa>

- **Wieczorek et al. 2012. Darwin Core:** An Evolving Community-Developed Biodiversity Data Standard. Plos One, [**https://doi.org/10.1371/journal.pone.0029715**](https://doi.org/10.1371/journal.pone.0029715)