
**Indexing for Life**

# Instructions for the pilot projects (new GSDs)

Workpackages 2/3/5

Viktoras Didžiulis,

edited by Alastair Culham & i4Life team

2013-05-08

Capacities Programme of Framework 7: EC e-Infrastructure Programme – Virtual Research Communities - INFRA-2010-2

| | |
|---|---|
| Grant Agreement No: | 261555 |
| Project Co-ordinator: | Dr Alastair Culham |
| Project Homepage: | http://www.i4Life.eu |
| Duration of Project: | 36 months |
| Start Date: | November 2010 |
| End Date: | October 2013 |



# Table of Contents

What am I expected to do during the Pilot Project? ..... 3
What is the role of the pilot projects in the i4Life project? ..... 3
What are the milestones and deadlines? ..... 3
How much should I trust data in the piped datasets? ..... 3
How do I decide which dataset I need to download and where? ..... 4
How do I open the dataset in MS Excel? ..... 4
Can I use other software and what are general data import export instructions? ..... 5
What fields in the piped dataset are relevant to my project? ..... 5
How do I submit my database with updated taxonomic checklist? ..... 6
How do I make annotations in spreadsheets? ..... 6
I am creating a new GSD, how are my tasks different from tasks done by CoL GSD? ..... 8
How do I submit my database with updated taxonomic checklist? ..... 8
How do I submit my annotated spreadsheets? ..... 9
I still have questions whom should I contact? ..... 9

## What am I expected to do during the Pilot Project?

a) Download the dataset containing species scientific names for your group and their higher classification piped via the i4Life project tools from the Global Biodiversity Partners.
b) Check the scientific names provided in downloaded dataset against your knowledge and expertise.
c) Annotate names using the three defined fields for annotations to express your expert knowledge and opinion.
d) Submit the annotated spreadsheet back to i4Life buffer database.
e) Update your own database with additional names from the i4Life spreadsheets (where it is appropriate), and submit the new version of GSD checklist for updating the Catalogue of Life.

## What is the role of the pilot projects in the i4Life project?

a) New names piped from the Global Biodiversity Programmes have the potential to add missing names and taxa to global species databases to make them more complete.
b) Annotations provided by the Pilot Projects will be used by the Global Biodiversity Partners to tidy up their databases and related digital resources.
c) Global Species Databases updated with new i4Life names will be used to enrich the Catalogue of Life with new names and taxa, thus making it more complete.
d) In the long run Global Biodiversity Programmes will update their taxonomic backbones from the Catalogue of Life thus making their databases more complete.

## What are the milestones and deadlines?

**Table 1.** Relevant dates for the Pilot Projects.

| **Signature of consortium agreement** | **April/May 2013** |
|---|---|
| **Start of pilot projects** | April 2013 |
| **First experimental Name- and Taxon- batch Placement Agreements in place with GSD custodians** | April 2013 |
| **Progress report on pilot projects, led by MNHN (WP2 milestone M2.7a) (MS38)** | May 2013 |
| **Delivery of intermediate report (and possibly download) of first batch of incorporated/annotated names** | 15 June 2013 |
| **Exemplar placement data sets: First five placement data sets visible in the CoL import buffer (WP5, D5.1)** | August 2013 |
| **Final date for delivery of project results to i4Life** | 31st August 2013 |
| **Final report from the pilot project, led by MNHN (WP2 milestone M2.7b)** | October 2013 |

## How much should I trust data in the piped datasets?

Piped datasets contain species names with additional information like higher classification and data sources where available. As the data-sets are not produced by professional taxonomists but by people who use species names in their activities (monitoring, genetics, etc...), different kinds of errors or misconceptions are very likely. That is actually the very reason why we need your expertise to separate “good” names from “bad” ones by providing your taxonomical opinion in annotations.

Sometimes author names may contain broken characters caused by faulty conversion between UTF8 and other non-unicode character sets. We would appreciate if you could process these where possible, otherwise please annotate them as “Misspelled name” in ***gsd_comments_predefined*** field.

## How do I decide which dataset I need to download and where?

If you are already a data provider of the Catalogue of Life (i.e. a CoL GSD) then simply choose the data-set corresponding to the name of your species database. Optionally - have a look into the Unplaced Names data-set – a large list of “dirty” names for which our software tools were not able to identify any suitable expert to deal with.

The original datasets from the Global Biodiversity Partners used for the call are available for download at http://www.catalogueoflife.org/piping/webservice/gsd/. Use the username and the password given to you to authenticate and get access to the service.

In addition to the above there are updated datasets provided by the piping process after the call with more names available at http://www.catalogueoflife.org/piping_new/webservice/gsd/ .

If you wish you can choose more names than you agreed in the original contract. Or you may choose the exact number of names as agreed but use additional newer datasets. If you choose more names there may be some extra funding options available. Please consult with Thierry Bourgoin (bourgoin@mnhn.fr) to get more information on funding.

## How do I open the dataset in MS Excel?

a) Download MS Excel template from **http://www.i4life.eu/pilot-projects/.**
b) Click on the "Data" tab (5th from the left in the top menu).
c) On the left-hand side of the tab find "Get External Data" section and click "From Text" menu item (icon).
d) “Import Text File” window will open. Into the "File name" field of the new window copy and paste a GSD link that is provided by the webservice (http://www.catalogueoflife.org/piping/webservice/gsd/) of the piping tools for your database. For example http://www.catalogueoflife.org/piping/webservice/gsd/WoRMS_Asteroidea, or pick any other.
e) Click "Import" button and, when it asks for a username/password, enter them into corresponding fields and click OK.
f) Text Import Wizard window will appear, check “Delimited” checkbox and click “Next” button, then leave all settings as they are (“Tab” checked under Delimiters, double quote chosen as Text qualifier) and click “Finish”.
g) Leave all the default settings if it asks to select the top-left cell (i.e. =$A$2, please make sure that it points to row 2 of the column A), hit OK and wait a moment until data finishes loading.
h) Try to scroll down - the header should remain visible as it is "frozen".
i) Save all into a tab separated spreadsheet when finished (Save as, choose “Text (Tab delimited)(*.txt)” in the “Save as type” drop-down menu, name it *annotated.txt*, click “Save”).

Use of the template is optional; however it provides a quick way to annotate records by reducing amount of typing and checks for errors in annotation fields. For large datasets you may have a more efficient procedure in place. In this case skip downloading the template and load dataset directly into MS Excel by following these steps:

a) Go to File/Open and paste URL of your GSD into "File name" field, click Open button;
b) "Connect to www.catalogueoflife.org" dialogue window will appear asking you to provide a username and a password;
c) Text Import Wizard window will appear, check "Delimited" checkbox and click "Next" button, then leave all settings as they are ("Tab" checked under Delimiters, double quote chosen as Text qualifier) and click "Finish".
d) Annotate using your procedure then save file (Save as annotated.txt) on your computer once finished.

## Can I use other software and what are general data import export instructions?

All the datasets are tab delimited text files and can be imported into any spreadsheet program (e.g. Open Office Calc, or other) or RDBMS (e.g. MS Access, MySQL, SQLite or other). When finished annotating in these programs, please export them back into a tab delimited text file (use utf-8 character encoding) preserving order of fields as it is in the original dataset and observing annotation rules as described further in "How do I make annotations in spreadsheets?".

## What fields in the piped dataset are relevant to my project?

The dataset contains 32 fields (columns). The main fields that you as a GSD are expected to check are: ***genus***, ***specificEpithet***, ***scientificNameAuthorship***, ***infraspecificEpithet*** and ***scientificName***. Names of the 5 fields are self-explanatory, e.g. ***genus*** field is for storing genus name, and so on. Field ***scientificName*** stores full species scientific name. It provides all the same information as fields ***genus***, ***specificEpithet*** and ***infraspecificEpithet*** but without ***scientificNameAuthorship***. The ***scientificName*** field should only be checked when ***genus*** and ***specificEpithet*** fields are empty or NULL (e.g. technical faults of data processing).

Please ignore ***in_col*** field, because you are required to assess ALL names in the spreadsheet. Please take as much as possible of the "good" names into your database.

In addition there are 5 fields (***order***, ***class***, ***phylum***, ***kingdom*** and ***higherClassification***) that provide higher taxonomic ranks for the names - these are only as good as the data source used by the project partner and in some cases may be empty (or NULL) where no data has been presented. After your taxonomic scrutiny it is up to you to decide under what rank and status the names should be placed in your database. We used these fields (where available) to identify which species names go to what GSD accounts. Field ***higherClassification*** provides a full taxonomic hierarchy as used by a data provider.

Further are the fields that can provide additional information to assist in placing a name into your database. Any of them can be empty or NULL where no data has been submitted by the Global Biodiversity Programmes. Field ***taxonomicStatus*** indicates whether a name is an accepted name or a

synonym. Fields ***taxonRank*** and ***verbatimTaxonRank*** indicates whether it is a name of a species (sp.) or an infraspecies (subsp. or var.). Field ***acceptedNameUsageID*** points to the ***taxonID*** field value for an accepted name of a species.

When linking synonyms to accepted names please take into account that species names should NOT be linked between different providers (i.e. value in the ***provider*** field for all linked names should be the same – if it is "1" for an accepted name then it should be linked only to synonyms having "1" in the ***provider*** field). Field ***parentNameUsageID*** is used for subspecies only and points to ***taxonID*** field value of a parent species name, please also ensure that the ***provider*** values of all the names are the same.

Field ***namePublishedIn*** may contain a literature source and field ***source*** may provide a link (URL) to more information on the name. Field ***taxonRemarks*** may contain additional remarks from the Global Biodiversity Programmes, however in some cases that additional information may be a reference too.

## How do I submit my database with updated taxonomic checklist?

We expect GSDs to send us a new version of database with added i4Life names and taxa through routine CoL update process using Annual Checklist Exchange Format (see http://www.catalogueoflife.org/colwebsite/content/contributing for details).

## How do I make annotations in spreadsheets?

We are expecting you to provide your opinion on validity of scientific species names for all names in the spreadsheet despite whether they are present in the Catalogue of Life or not. Please, take a note, your new GSD will replace incomplete dataset in the Catalogue of Life and "old" taxa and names previously supplied by regional databases will disappear in the taxonomic sector of your responsibility.

There are three fields for annotations in the datasets (field gsd_status, gsd_comments_predefined, gsd_comments).

(1) In the field **gsd_status** please write either **Rejected** for "bad" names you don't want to place into your database or **Placed** for names which you are taking in your database for delivery in the Catalogue of Life. It is understandable that sometimes decision on placement or rejection of a name may be quite complicated and not straightforward at all. For example there may be a spelling error in the name which otherwise would be new and suitable for placement – should you reject or place it? Therefore there are two more fields to assist you in making decisions (and annotations).

(2) Field **gsd_comments_predefined** can only have 10 categories of annotations. List of standard predefined categories is given in column "Category" in Table 2.

(3) Field **gsd_comments** is reserved for your comments which may include scientific names or free texts as described in column "Comment" of the Table 2.

**Table 2.** A list of standard annotation categories for unvetted names.*

| ***Category***<br><br>*(it should be indicated in the* ***gsd_comments_predefined*** *field)* | **Description** | **Comment**<br><br>*(what to do next)* |
|---|---|---|
| **Incomplete name** | All sorts of Latin names, with incomplete or abbreviated genus name, species or infraspecific epithets, names without authorstring (e.g. *S. aguabonita Jordan, Salmo a. Jordan, Salmo aguabonita*). | If, by the chance, you know complete name, please, input it to your GSD and give it in the spreadsheet field ***gsd_comments*** without any additional text. |
| **Chresonym** | Scientific name with not validated authorship, which refers to published usage of the name rather than to the true author of the name. Please visit http://en.wikipedia.org/wiki/Chresonym for more details.<br><br>Example:<br><br>***Actinopus crassipes* (Keyserling, 1891)**<br><br>*Pachyloscelis crassipes* Keyserling, 1891: 3, pl. 1, f. 1<br><br>*Actinopus crassipes* Strand, 1916b: 81 - chresonym<br><br>*Actinopus crassipes* Mello-Leitão, 1923a: 18, f. 128 - chresonym<br><br>*Actinopus crassipes* Bücherl, 1957: 384, f. 5 - chresonym<br><br>*Actinopus crassipes* Schiapelli & Gerschman, 1962b: 72, pl. II, f. 3 - chresonym<br><br>*Actinopus crassipes* Lucas & Bücherl, 1965: 89, f. 1-18 - chresonym | If, by the chance, you know original author of the name, please, put full binomial/trinomial with correct authorship to your GSD and give it in the spreadsheet field ***gsd_comments*** without any additional text. |
| **Name with unresolved nomenclatural status** | All sorts of not validly published names, or names with the type of unknown location, etc. – Latin names which you are not taking in your taxonomic checklist. | These names should stay in the spreadsheet and do not go in your GSD. |
| **Hybrid formula** | You may find hybrid formulas as "*Dianthus caryophyllus × D. plumarius*" in the spreadsheet. | Keep these hybrid formula names in the spreadsheet only, but add both names of parents in your GSD. If hybrid has binomial as "*Dianthus × allwoodii*", please add it to your database, trying to avoid cross symbol in species epithet field. |

| | | |
|---|---|---|
| **Misspelled name** | All kind of published orthographic variants or typos, both in paper and digital media. | Please input correct name to your GSD and give it in the spreadsheet field ***gsd_comments*** without any additional text. |
| **Fossil name** | Latin name given to fossilised specimen. Names of fossils are out of subject area of the Catalogue of Life. | Keep these names in the spreadsheet only. If you decide to include fossils in your GSD, please, clearly indicate that they are fossil names in ***gsd_comments***. |
| **Unidentified specimen** | All sorts of names used for unidentified specimens/samples (e.g. *Salmon sp.1*, *Gagea aff. lutea* - aff. for "affinis", cf. for "confer") | If, by the chance, you are able to identify specimen/sample and provide correct scientific name, please, input this name to your GSD and give it in the spreadsheet field ***gsd_comments*** without any additional text. |
| **Non scientific name** | You may find strange names delivered by our partners, which are not Latin (e.g. *Lathyrus with white flowers from Argentina*) | These names should stay in the spreadsheet and do not go in your GSD. |
| **Name from other taxon** | A name which does not belong to your GSD sector and was placed in the list by mistake | If, by the chance, you know correct taxon to which this name belongs, please, give it in the spreadsheet field ***gsd_comments*** without any additional text. |
| **Other** | | Please, add your additional comment in the spreadsheet field ***gsd_comments*** as free text. |

*****from "Roadmap for i4Life pilot projects" (http://www.i4life.eu/i4lifewebsite/wp-content/uploads/2012/12/Roadmap_for_i4Life_pilot_projects.pdf)

## I am creating a new GSD, how are my tasks different from tasks done by CoL GSD?

a) You will be expected to provide a new taxonomic checklist in the CoL which will replace incomplete data from regional databases ( ITIS, NZIB, China, etc.).
b) You will be expected to annotate all names in your i4Life spreadsheet.

## How do I submit my database with updated taxonomic checklist?

We expect GSDs to send us a new version of their database with added i4Life names and taxa through routine CoL update process using Annual Checklist Exchange Format (see http://www.catalogueoflife.org/colwebsite/content/contributing for details). As a new provider, you

are recommended to contact CoL executive editor (Yuri Roskov, y.roskov@reading.ac.uk) for advice on how to proceed.

## How do I submit my annotated spreadsheets?

We expect GSDs either to put the completed files named *annotated.txt* (can be zipped into *annotated.zip*) for us to download from your server and send us URL, or to upload to "i4life_annotated_names" directory in your CoL contributor account at https://www.catalogueoflife.org/accounts. The "i4life_annotated_names" directory in GSD accounts will be created as the pilot projects are initiated. Please note that we use a self-signed SSL certificate for encryption purposes only and browsers will complain that the certificate is not valid for identification, please ignore all the warnings and proceed to the accounts.

## I still have questions whom should I contact?

Regarding technical side of the project please write to the systems manager Viktoras Didziulis (v.didziulis@reading.ac.uk) and developer Kwok Yin Cheung (k.y.cheung@reading.ac.uk). Any questions related to project management or funding should go to Alastair Culham (a.culham@reading.ac.uk) and Magda Sitko (m.h.sitko@reading.ac.uk). Taxonomical questions and questions related to the content of the Catalogue of Life should be directed to Yuri Roskov (y.roskov@reading.ac.uk) and Thomas Kunze (t.kunze@reading.ac.uk).