Memo: Taxonomic Backbone in CaBOL

Stefan Otto (15/11/2020)

**Situation**: In the CaBOL Database we need to link every sample/CaBOL ID to a species name in a hierarchical system based on a checklist of accepted species. Since such a checklist does not yet exist in quality (accepted species) nor quantity (up-to-date number of species) for the Caucasus Ecoregion, we need to compile this checklist in CaBOL ourselves and agree on the species, which are to be included in this list as well as on the phylogeny linking these species to genera, families, orders etc. The result will be the taxonomic backbone of the CaBOL Database.

Task 1: Compiling a Checklist and Phylogeny for CaBOL

The basis for the Taxonomic Backbone is the Georgian Biodiversity Database (<http://biodiversity-georgia.net/> Giorgi Chaladze et al.). Giorgi has sent me the database in September 2020 ([link](https://www.dropbox.com/s/i5qst69nd1yqbks/species.sql?dl=0)). It consists of a large table (26687 entries), in which taxonomic groups/species are linked to higher taxa in a child-parent manner: Taxon -> Parent.

I tried to normalize this table by creating separate tables for species, genera, families, orders and linking taxa automatically to the respective taxon one level up. Problem here was: in the table not all taxa are linked to their direct Parent one level up, e.g. species -> genus. For example, Testudo graeca is directly linked to Testudinidae instead. This can be seen on the GBD website and makes sense there but not for our work. Also, sometimes taxa could not be normalized in the database tables automatically because some names had added empty spaces (like ‘Stefan\_’ instead of ‘Stefan’), preventing automatic matches.

The normalized tables therefore include a significant number of mismatches like family names in the table “genera” etc., which need to be resolved manually. Also, maybe I just did not manage to normalize the tables and Chris can do this easily. Nevertheless, it could be a good first exercise for the taxonomic coordinators to clean this up in cooperation with Chris in Bonn setting up the software tools for doing this!

The resulting very raw database can be downloaded here as .sql: [Link](https://www.dropbox.com/s/2wwplyexg59jfw8/TaxonomicBackbone_RawDatabase.sql?dl=0). The combined normalized table from this database as Excel file: [Link](https://www.dropbox.com/s/6fnytl2c7kyei5w/TaxonomicBackbone_RawDatabase.xlsx?dl=0).

**Proposed Strategy**: We pick out the species names (=all taxa with at least one space in the middle) from the Georgian Biodiversity Database and link them to an existing higher taxonomy.

Example of the resulting normalized table (combination of tables, ) with mismatches:



Task 2: Implement a Database structure for storing this Phylogeny in the CaBOL Database

The CaBOL Database will be implemented and maintained at ZFMK (Christoph “Chris” Erhardt) and will be moved to ISU after servers and staff have been established there (timeline goal?).

The CaBOL Taxonomic database will be the DiversityTaxonnames module of DiversityWorkbench. Here taxonomies and checklists can be maintained and enriched with further information on habitats, IUCN status, distribution, etc.



*Abramis brama* (Linnaeus 1758) in the database  **DiversityTaxonNames\_Vertebrata** from the German checklists of fishes. The core information about *A. brama* includes the nomenclatural code used for the name, together with the author and year from the first desciption. Additional information consist of the the taxonomix hierarchy (=linkage to taxon names in the same database), common names (here: „Brasse“, the german name for  *A. brama*), a reference to a published list, as well as informations about the status of the species in the german red list together with the reference. Synonyms, Images, Geographic information as well as futher references and usages in specimens lists can also be added.

**Database Specifications:**

* Easy linking to existing databases (GBOL, BOLD, …) and data access for CaBOL staff (unrestricted) and via the web interface/Wordpress at ggbc.eu (restricted access for the public)
* Easy maintainance and updating of the phylogeny by taxonomic coordinators, e.g. adding/deleting species, genera and adding others, attributing existing species to another genus … manually fixing errors, mismatches etc.

Task 3 – Backbone Maintainance: Attributing certain taxonomic groups to taxonomic coordinators as fields of responsibility during the duration of the CaBOL project

Taxonomic coordinators are responsible for updating the phylogeny and species status in the database and in case of questions and updates every species/taxon has a taxonomic coordinator who can be asked/held responsible for maintainance.

**Proposed Strategy**:

1. In general: Taxonomic coordinators manage all taxa within the database within their field of responsibility, either themselves in cases they have taxonomic expertise in this group or by actively looking for and finding taxonomic experts outside CaBOL who update these taxonomic groups.

**Fields of Responsibility**:

* **Plants:** Kosta Kereselidze (ISU)
* **Lower Invertebrates, Molluscs, Crustacea**: Levan Mumladze (ISU)
* **Arachnida:** Stefan Otto (ISU)
* **Insects:** Giorgi Japoshvili (AgrUni), Stefan Otto (ISU), (Coleoptera) Mark Kalashyan (SCZHE), Nils Hein (ZFMK)
* **Macroinvertebrates (benthic):** Marine Dallakyan (SCZHE) & Vardan Asatryan (SCZHE)
* **Vertebrates:** Levan Mumladze (ISU)
* **Vertebrates:** (fish) Bardukh Gabrielyan (SCZHE)
* **Vertebrates:** (Amphibians/Reptiles/Molluscs/Fungi) Marine Arakelyan (YSU)
* **Vertebrates:** (Mammals) Astghik Ghazaryan (YSU)
* **Parasites:** (Ticks, Mosquitoes, Birds, Lichens) Sargis Aghayan (YSU)

**Strategy for the development of the CaBOL Taxonomic Backbone:**

* + - Firstly, the structure of the Georgian Biodiversity data is updated and links between levels are fixed (species -> genus).
		- Species are taken (genus + species epithet) and linked to an existing taxonomic hierarchy.
		- Taxon lists are updated by taxonomic coordinators and external experts (firstly based on the fauna and flora in Georgia). Taxonomic coordinators collect and review sources of the Armenian and Georgian fauna and flora, which is to be added bit by bit.
		- Armenian and Georgian colleagues add **Caucasian** species and possible taxonomic conflicts are resolved in a friendly and constructive manner. “Every (accepted) species name counts”
		- Resulting Taxonomic Backbone is released and used
		- New species names coming from the lists of collected specimens are added by the respective taxonomic coordinators in a timely manner.