## CAB2: A step towards Biodiversity data enrichment (Topic 12)

*Aim/problem/goal*

Natural history specimens may be sampled for sequencing, and these specimens/sequences published or cited in literature and deposited in repositories like ENA. Links between these types of data are rarely explicit, so that it is not straightforward to connect a specimen to a sequence or literature. The goal of this project was to (re-)establish these links, making use of Computer Vision models and ad hoc text mining scripts to extract more data from the specimens.

*Method*

We made use of Computer Vision (CV) to identify indications of sequencing on specimen images. We then retrieved the corresponding sequences (in ENA), gene annotations and references in literature (in TreatmentBank and ENA flat files), either by matching identifiers or by mining through common properties such as taxonomic names and their corresponding identifiers. We also tried to leverage results from the GBIF clustering algorithms, which cover ENA sequences and published specimen data.

*Results*

A relatively small set of specimens was identified as (probably) having been sequenced (467 out of 3,184 specimens processed). There was considerable complementarity to previous results, but again only a fraction could be unambiguously matched to ENA sequences. Poor identifier propagation is a fundamental blocker, but we also had great difficulty in taxonomic matching between specimen and sequencing data, in particular through the ENA API. The GBIF clustering method was very conservative for this sort of matching and yielded almost no results, given the large variability and inconsistency in how identifiers are provided to the two infrastructures. The connection to literature showed similar issues with different representations of identifiers or even their total absence, with the added complication of trying to identify the material citations from text, tables or supplementary material in the first place.

*Conclusion*

Linking between these different data types is currently very difficult and labour-intensive. Scientists should be strongly encouraged to make use of persistent identifiers to maintain links in all sources and infrastructures to support and promote this. Computer Vision worked well, but showed scaling issues of costs. Sequencing labels are often a mix of sparse handwritten and typed text, for which free algorithms currently do not yield satisfactory results. In addition, the Computer Vision approach is likely to only cover specific cases and it's still challenging to connect the specimens to the sequences, as ENA identifiers are rarely used on the specimens themselves. Large-scale clustering approaches, such as performed by GBIF, could yield more results. Taxonomic interfacing between the different data sources should be improved. We made use of Wikidata as a broker, but these data are not always up-to-date and can suffer from taxon rank discrepancies. Taxonomic resolution options through the ENA API were very limited, so we had to resort to mining through data dumps instead.