SHORT TERM SCIENTIFIC MISSION (STSM)
SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: COST Action 16224
STSM title: Scoping a database for the European raptor specimen bank
STSM start and end date: 25/11/2019 to 20/12/2019
Grantee name: Konstantinos Vlachopoulos
Grantee home institution: University of Thessaly
Host name: Dr. Rene Dekker, Dr. Paola Movalli
Host institution: Naturalis Biodiversity Center

PURPOSE OF THE STSM

(max.200 words)

The aim of the STSM was to contribute to the ERBFacility's Research Objective relating to the development of a distributed European Raptor Specimen Bank (ERSpeB) for contaminant monitoring by scoping out a database for the ERSpeB. The distributed ERSpeB will improve collaboration among the Collections Arena (natural history museums, environmental specimen banks and other collections) and the Analysis Arena (analytical labs, ecotoxicological research institutes) by providing an online database of raptor carcasses held by collections in freezers, allowing for provision of tissue samples for contaminant analyses.

More in depth it is expected that the proposed ERSpeB database will mine data in

Funded by the Horizon 2020 Framework Programme of the European Union

real-time, attach any field contextual data or data from other online databases that refer to species distribution (e.g. GBIF) and most importantly to be able to interpolate with LIFE apex database, IPCheM and DiSSCo.

This mission had the following objectives:

1) To determine the database specifications: In particular to investigate the requirements against the availability of specimens and related information in collections.
2) To outline the database design with a view to ensuring data efficiency, accessibility and interoperability with other related databases (notably Norman, IPCheM, GBIF).

DESCRIPTION OF WORK CARRIED OUT DURING THE STSM

(max.500 words)

1. Determination of database specifications

In order to investigate the range of data necessary to include in the database, both from the perspective of the collections, and from the perspective of the analytical labs, a questionnaire was distributed to these two focal groups (Collections Arena, Analysis Arena). The questionnaire was made as simple as possible in order to allow efficiency of completion and to aid consistency, with most questions limited to yes/no answers, and only two open-ended questions. The questions were slightly modified depending on the targeted Arena. An invitation to complete the questionnaire was issued by email. The questionnaire was distributed to 73 natural history museums and 23 analytical labs involved in the ERBFacility network. Through this questionnaire we assessed the willingness of collections to share their data through the proposed ERSpeB specimen database and determined the data fields required to meet the needs of both Arenas. An attempt was made to assess the temporal boundaries of the database through one of the open questions. The questionnaire responses were analysed and visualised graphically using the R language (R. 3.6.2) for programming and statistical computing and the package “ggplot2”.

2. Database design

To develop a proposed database design, meetings were held within Naturalis with experts in museum curation, database development and ecotoxicology. Specifically, key issues concerning the database development from a collection perspective were discussed with Dr. Rene Dekker and Steven van der Mije, and from the analysis perspective with Dr. Paola Movalli. An overall ERBFacility perspective was provided by Guy Duke. Database design and key issues relating to interoperability were discussed with Dr. Koureas Dimitris and Dr. Sharif Islam of the European Research Infrastructure DiSSCo which is coordinated out of Naturalis. Database flowcharts were drawn to summarize the outcome of these discussions (see output).

DESCRIPTION OF THE MAIN RESULTS OBTAINED

(max. 500 words)

Overall, 20 natural history museums, 1 environmental specimen bank and 12 analytical labs responded to the questionnaire, from 20 European countries. Despite the small sample size of respondents, responses were well distributed across European providing a good indication of both available and required data.

Requirements of ecotoxicologists/analytical labs and proposed fields by users-curators

The questionnaire survey resulted in two key outcomes. The first is an overview of the data recorded on raptor specimens by collections. The second is a consensus on the data field required by the Analysis Arena in order to identify specimens of possible interest for contaminant monitoring. It is clear that the data fields most useful for the Analysis Arena (n=12) in this respect, are: species scientific name, collection registration number, date of death, geographical area/location, sex, age class, cause of death. These data fields are in line with the data that most of collections record when receiving and storing a fresh raptor carcass.

Type of data storage and registration

According to the responses from collections (n=21), it is clear that not all collections currently record the relevant data electronically, with some using paper records. This will be a constraint to uploading data to the ERSpeB database or to mining this data.

Data sharing and policy

The most frequent answer to the question “Are you willing to provide your freezer data to an online ERSpeB specimen database” was “Not sure yet” (n=10) and “Yes” (n=8). Only three respondents indicated that they would not share data with an ERSpeB database. The variance in responses may relate to differing collection policies, with some collections being less ready to share data.

Temporal boundaries

Responses on the issue of the temporal boundaries for the database were inconclusive, possibly due to ambiguity in the question. Four respondents suggested the database should include data on frozen specimens dating from before the year 2000, three suggested only including specimens collection in the year 2000 or later, and two suggested it depends on the aim of the monitoring. A couple of respondents suggested much earlier start dates, perhaps because the question did not sufficiently focus on more recent frozen specimens.

Database frequency update

One of the major concerns during this STSM was the type of database update system and how often the data should be updated. We believe that most of respondents (from both the Collection and Analysis Arenas) would value a fully real-time database but the questionnaire design does not allow us to make any conclusions in this respect.

Database design

Database flowcharts were drawn to better understand and document the information flow among the different process stages (e.g. specimen registration, accession, sampling and sample provision/loan). The flowcharts were drawn using the draw.io app.

A full technical report is in preparation.

FUTURE COLLABORATIONS (if applicable)

(max. 500 words)

This STSM proposes a database system that will be interoperable with individual collection databases and collect all the required information for specimens that are stored in collections freezers. The intention is that the database user will obtain a real time picture of what specimens are available in participating collections. In order to achieve this, next steps include:

- The data of the raptor specimen that are stored in collections' freezers should be digitized and stored in a digital format.
- Maintaining real-time data on raptor samples in freezers across collections in Europe requires discipline by collections to keep their individual databases up to date – both in terms of new specimens entering freezers, and specimens being removed from freezers.

Moreover, in order to allow data sharing and exchange of the ERBFacility database with the different individual collection databases systems, a unique "key" field is required that will identify each specimen and follow it persistently. Even if a specimen leaves a collection freezer and becomes tissue samples, the specimen record should not completely disappear from the specimen database. This need can be served by a unique persistent identifier (UID). A unique identifier (UID) is a numeric or alphanumeric string that is associated with a single entity within a given system. With the UID it is possible to address a specimen (database entity), so that it can be accessed and tracked even if it is loaned to another collection or transformed to samples for analysis.

DiSSCo project (https://www.dissco.eu/) is a distributed system of scientific collections and is testing the use of UID and how specimens that are hosted in natural history museum (NHM) collections can be linked together with all other relevant data, e.g. on species, genomes, phenotypes, geography, geology and the environment, in ways that drive novel, integrative research (e.g. data on the distribution of living species held by the Global Biodiversity Information Facility (GBIF) and date on the genetic sequence information held by GenBank and Ibol etc.). The ERBfacility database offers a possible pilot for DiSSCo although the following challenges will need to be overcome:

- Getting collection curators to record raptor carcasses information on arrival and attach a unique identifier (UID) before the specimen is stored in a freezer.
- Any specimen-level data that are stored in the ERSpeB database must be interoperable with related sample and contaminant-level data, such as that held in the Norman and IPCheM databases.

The three Arenas participating in ERBFacility need to be informed about the proposed database in order to allow interoperability with their database plans.