THE ATLANTIC-WIDE RESEARCH PROGRAMME FOR BLUEFIN TUNA (GBYP Phase 12)

SHORT TERM CONTRACT FOR THE SUPPORT TO THE DEVELOPMENT OF THE ICCAT ELECTRONIC TAGS MANAGEMENT SYSTEM "ETAGS"

(ICCAT GBYP 11/2022)

Final Report

Hong Kong, 15 July 2023

Big Fish Intelligence Company Limited

Dr. Chi Hin Lam, contractor

This project is co-funded by the European Union

Background

Over the years ICCAT has released in the Atlantic Ocean and adjacent Seas many electronic tags on tuna, tuna-like and various shark species, to record information on the behaviour and migrations of those species. A large portion of those electronic tags are related to ICCAT research programmes (www.iccat.int/en/ResProgs.html) in which GBYP (about 59 internal archival tags and 446 satellite pup-up tags released since 2008) and AOTTP (about 430 internal archival tags and 169 satellite pop-up tags) are two major contributors. Overall, these tagging activities have behind a considerable investment. Yet, the associated information was never properly stored in a centralised relational database held in ICCAT, which would greatly improve its potential use in scientific studies, allowing for example the development (or improvement of existing ones) of more efficient and complex analytical tools.

Currently, the ICCAT electronic tagging information has various weaknesses:

a) Lacks a complete and efficient inventory (metadata: scientific programmes, deployment activities, tag event characteristics, raw data availability, manufacturer's raw binary files, resulting scientific work associated, etc.),

b) The raw binary files are spread across various laboratories and/or scientists (only a small portion is held in ICCAT, mostly associated to GBYP and AOTTP programmes),

c) The raw data files are archived using various structures and formats (in many cases reflecting the changes made manufacturers over the years),

d) The different models of electronic tags (internal archival, satellite pop-up) used over time and the output formats used by each manufacturer could be inconsistent in time (changes in software, data field policies, etc.)

The ICCAT existing and future electronic tagging information will be much more valuable to the ICCAT scientific community, if all the information is validated and stored in a centralised relational database, together with all the associated metadata. Therefore this "ETAGS" project is conceived to build the necessary software infrastructure to handle the enormous electronic data needs of ICCAT, and provide a means for future development and integration of its other data systems within the organization. This is achieved mainly through:

1. A simple, flexible flat file format "eTUFF" that serves as an intermediary exchange format to facilitate the consolidation of data products from various tag manufacturers
2. The electronic tag data in eTUFF format can then be uploaded to the a specialized database management system, Tagbase-server (github.com/tagbase/tagbase-server). Tagbase-server is a Flask (flask.palletsprojects.com/en/2.0.x) application which provides OpenAPI REST endpoints for ingestion of various files into the Tagbase SQL database (PostgreSQL engine).
3. Data can be managed through Tagbase-server, which allows connections to various visualization and analytical endpoints

Phase 1 – brief recap

The main activities of this phase covered the inventory (including metadata and raw binary files) of the electronic tagging of ICCAT, the support for the generation of the "eTUFF" file format through its associated tools associated (tags2etuff), and the work behind the Tagbase-server application (mainly the Postgres SQL database training: installation, schema learning, SQL tools, etc.). Our team has made major upgrades to the database management system and

customized the application for ICCAT. Metadata handling is also updated to match ICCAT's requirements on various data sharing and collection scenarios. A beta-release of Tagbase-server was delivered to the Secretariat on July 19, 2022 and it is been continually supported and revised with feedback from the end users. In summary, completed technological improvements on Tagbase-server in support of ICCAT's usage is listed as follows:

1. Modernized and upgraded components
2. Improved code quality
3. Review and eliminate vulnerabilities
4. Ensured maximum code traceability through Github issue tracking, release history & documentation
5. Code expansion developed in mind for long-term expansion/ capable deployment, e.g., Amazon Web Services & CloudFormation
6. Comprehensive logging
7. Real-time operational notifications

Given its complexity (combination of various technologies, large amounts of data to inventory, recover and treat, etc.) this project was envisioned to be completed over at least three phases. This report described the activities and work carried out during Phase 2 that took place between January 15 and July 15, 2023.

Phase 2 - Progress to date

Please refer to Appendix 1 for links to file repository for Phase 2 activities. These links will remain active for the Secretariat to obtain the files and replicate/ relocate as it sees fit.

A. Full implementation of multiple-tracks support and improvements in database operations

The original data model assumed a 1-to-1 relationship between a particular tag deployment and its associated geo-positioned track. However, due to the limitations and updates to light-based geolocation algorithms, it is common to have multiple track solutions that either represents different processing options or improvements on estimation over time and/ computation capabilities. We provided additional coding to generate an eTUFF file consisting of a tag's metadata and its improved track solution, for example from Wildlife Computers' GPE3 and Kalman-filter algorithms (Trackit or Ukfsst). This ability allows the evolution of a dataset, for example, an initial track was first generated by manufacturer algorithms and then subsequently improved by running more sophisticated third-party geolocation methods.

In the development of supporting multiple track solutions, we identified an important need of the Secretariat that it must be able to identify duplicate files and data contents, such that redundant information is not incorporated into the database. Repeated datasets could significantly inflate the database, create confusion and cause data access problems. The capability to conduct "duplicate detection" is beyond our initial scope of work and is technically complex. However, we decided to tackle this challenge and updated the data model to enable "track changes" through multiple "checksums" or file content signatures and "fingerprint" every incoming eTUFF file into Tagbase-server. These checksums do not rely on any file names, and are created based on each individual file and the content it contains.

In brief, four checksums were created for each incoming eTUFF file: a) an overall file checksum which allows us to track on a "file" basis if a previously ingested eTUFF file is uploaded to Tagbase-server, and b) three file content checksums that track changes in the metadata, tag measurement data, and track data. If any of these 3 content checksums is different, we will

ingest the new information in the corresponding content. We have constructed all the necessary code and infrastructure for file level and content level “track changes” in Tagbase-server, however, we will need to conduct more load testing, e.g., import hundreds of eTUFF files to find out if additional improvements are required. During this development, we also identified a “concurrency problem” that is a result of our capability to parallel process multiple incoming eTUFF files that is sent to Tagbase-server via a zipped file. We engineered checks to ensure no duplicate files are ingested in a parallel processing operation. Again, these complex operations were implemented in anticipation for future scenarios where large amount of eTUFF files are sent to the Tagbase-server for incorporation. We have now completed a stable and flexible data model that will allow the Secretariat to house large volume of data and grow with new types of electronic tag output files. The entity-relational model is represented below:

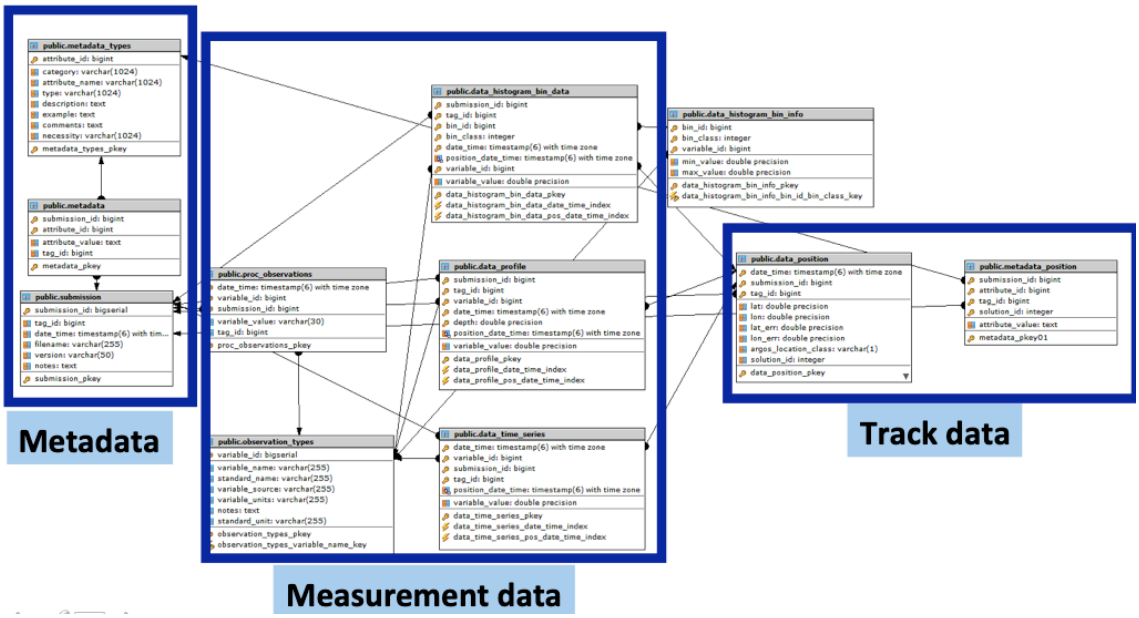

Figure 1 Entity-Relational model of Tagbase-server

Another enhancement is the full incorporation of the PostGIS extensions to Tagbase-server. PostGIS provides additional GIS functionalities and its additional was done to fulfil the request made by the Secretariat. Lastly, we also constructed stored procedures that allow the database administrators to quickly delete existing data from the database. This provides a cleaner and efficient way to remove unwanted data by the Secretariat.

B. Extended ETUFF support of /tag files importation from Lotek Wireless and Microwave Telemetry

We completed the support of generation of eTUFF files from outputs of Lotek Wireless software, TagTalk and LAT Viewer Studio (Fig. 2). This work involved significant communication and exchanges with Lotek Wireless technical support staff, as to our surprises, even Lotek staff has no clear ideas on their file outputs. This is further complicated by the lack of up-to-date or reliable documentation in any of their manuals.

Export from tal talk?

GMT Time Correction = 7200
Lg00 (Basic Log) Log:

Rec #	Date	Time	ExtTemp [C]	Pressure [dBars]	LightIntensity	IntTemp [C]	WetDryState	WetDryChange	C_TooDimFlag	Valid Flag
0	02/01/2017	02:00:00	20.56	-0.13	88	21.6	1	1	0	1
1	02/01/2017	02:00:15	20.56	-0.13	88	21.6	1	1	0	1
2	02/01/2017	02:00:30	20.56	-1.13	88	21.6	1	1	0	1

Export from Lat Viewer Studio

TimeS	ExtTemp	Pressure	LightInter	IntTemp	WetDryState	WetDryChange	C_TooDimFlag
09/03/2018 00:00	30.54	0	142	30.52	1	1	0
09/03/2018 00:00	30.56	0	143	30.52	1	1	0
09/03/2018 00:00	30.54	0.5	143	30.52	1	1	0

Export from postgres

TimeS	ExtTemp	Pressure	LightIntensity	IntTemp	WetDryState	WetDryChange	C_TooDimFlag
560131200	25.26000023	0	97	25.13999939	1	1	0
560131215	25.26000023	0	97	25.13999939	1	1	0
560131230	25.26000023	0	97	25.13999939	1	1	0

Figure 2 Variations of file headers from Lotek Wireless software tools

However, our team has worked extensively to cover outputs from multiple versions of the Lotek software suites. We cautioned that more testing would be required given the aforementioned internal confusion at the manufacturer level. A visual summary of the completed workflow for Lotek is illustrated below:

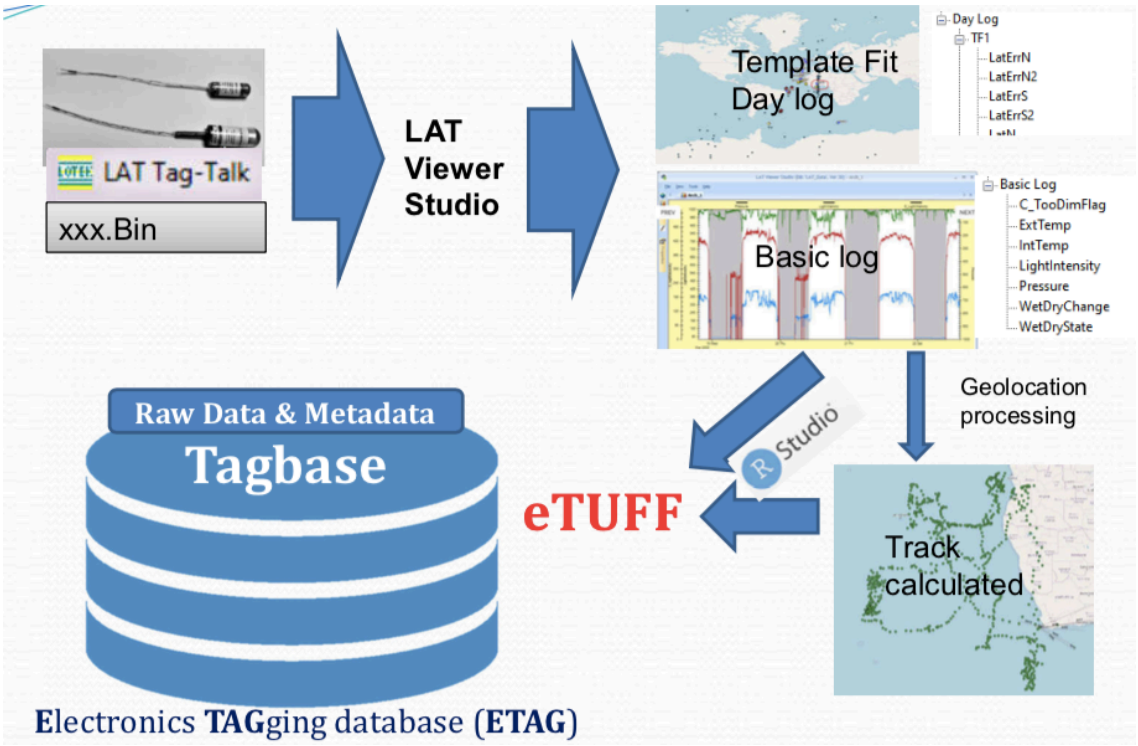

Figure 3 Example ingestion workflow of Lotek Wireless tag files into Tagbase-server

Secondly, we also completed the eTUFF generation for Microwave Telemetry (MT) Excel-based file outputs. A visual summary of the completed workflow for MT, as well as Wildlife Computers, is illustrated below:

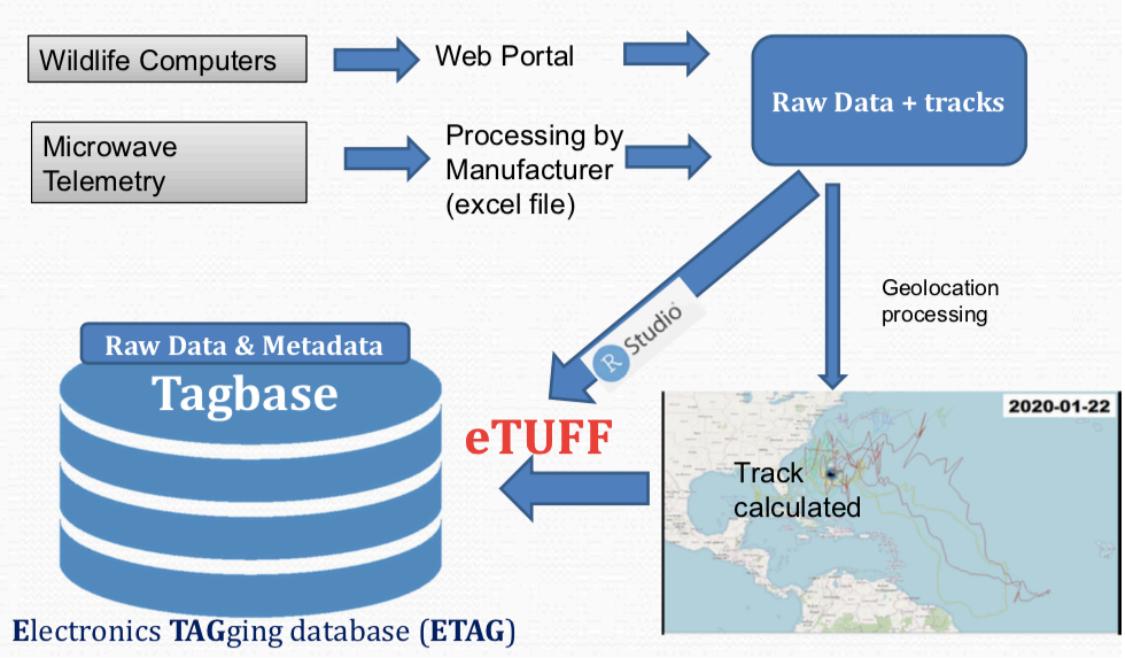

Figure 4 Example ingestion workflow of Microwave Telemetry and Wildlife Computers tag files into Tagbase-server

C. Continued support of eTUFF uploads to ETAGS database

In order to support the use of Tagbase-server in ICCAT, our team has been actively and promptly responding to eTUFF import issues and technical questions from the Secretariat throughout the Phase 2 period. Our response time is usually within half an hour, and troubleshooting and bug fixes were often completed within two or three days. Towards the end of this phase, web-based conferencing meetings were held on a weekly basis to go through issues and conduct walkthroughs. Our team remains firmly committed to assist the Secretariat to import more eTUFF files and grow the database as quickly as possible.

D. Streamlined data ingestion work

Since the end of Phase 1, we are already providing the feature to import of an eTUFF file over the web via command-line prompts and the REST API. However, after multiple consultations with the Secretariat, it was identified that batch import of eTUFF files was highly desirable. The consensus between the Secretariat and our team was to utilize existing software tools (mostly Windows-based) that ICCAT routinely used to do batch import. We therefore developed infrastructure to support simple “drag-and-drop” file transfer operations with the MobaXterm software. Multiple eTUFF files can be transferred through such a transfer protocol, and alternatively via a single or multiple zip files containing many eTUFF files which have much smaller file sizes and can be transferred more efficiently across the web.

Continuity tasks

A number of developmental areas are initiated in Phase 2 and will continue into Phase 3, and progress on each of them is briefly highlighted.

1. Support of legacy track outputs generated by CLS Track-and-Loc services

Only during the last month of the current phase, we received the notification that GBYP has a large number of tracks that were processed by CLS’s Track-and-Loc proprietary service and are

in a specific file format. In light of timing and the volume, this task is ticketed and will be handled after the completion of Phase 2.

2. Customization of components based on new data access patterns

We have developed a few modes of file ingestion pipelines in Phases 1 and 2, providing support to file transfer from FTP, HTTPS, bulk upload via zip file etc., given ICCAT's various number of potential data contributors from around the world. A follow-up step in Phase 3 is to identify preferred other important data access patterns and streamline an ingestion routine that is best set up for the Secretariat. The same customizations will be applied to other tools we have built for system notifications and logging. The goal will be to make a responsive system that helps the Secretariat to acquire and import data into Tagbase-server.

Acknowledgements

This work has been carried out under the ICCAT Atlantic-Wide Research Programme for Bluefin Tuna (GBYP), which is funded by the European Union, several ICCAT CPCs, the ICCAT Secretariat, and other entities (see https://www.iccat.int/gbyp/en/overview.asp). The content of this document does not necessarily reflect ICCAT's point of view or that of any of the other sponsors, who carry no responsibility. In addition, it does not indicate the Commission's future policy in this area.

Appendix 1. File repository for Phase 2

A. The scripts to create the eTUFF files from the different tagging brands

https://drive.google.com/drive/folders/1lI56jC8GbH6QhubYl04kmHK6XHmzzvlD?usp=drive_link

and the associated directory structure

https://drive.google.com/drive/folders/1rLhI34gOCet4U-Q7QuKq3j_WF-4Qj7Xe?usp=drive_link

B. Selected eTUFF file examples

https://drive.google.com/drive/folders/1_sQLk8NZd5tcITkACu8B1Eg09dO0hX8f?usp=drive_link

C. An example of database import scripts to import the eTUFF into the repository.

https://github.com/tagbase/tagbase-server/wiki/Ingestion-and-Access-Patterns

D. A backup of the eTag database.

https://github.com/tagbase/tagbase-server/wiki/Installation