Data Processing Steps for Great Lakes Restoration Initiative – Pollinator Task Force Bee Records Produced by the USGS-USFWS Interagency Native Bee Lab

There are \_\_\_\_ main components to sharing the bee occurrences records associated with the USFWS GLRI project on the Global Biodiversity Information Facility platform.

This document will provide details for each step in the data workflow.

## Step 1: Accessing data from the USGS/USFWS Interagency Native Bee Lab (BIML)

The occurrences records are provided by BIML. Each specimen is entered into the SQL Server database at BIML using an established workflow. The BIML database captures occurrence and sampling event details using fields that do not follow the Darwin Core standard. The structure of the database requires that project-identifying information be stored in a field titled “email”. When specimens from the GLRI project are entered into the database, the value “USFWS\_GLRI” is entered into the email field.

BIML also utilizes a table titled “Project\_Identifiers\_Table” (PIT) to capture project-level information associated with the specimens being identified. The fields in the PIT are Darwin Core formatted and provide relevant record-level fields (see data dictionary ###). This table is not part of the SQL Server database. Prior to starting the data processing workflow, this table will be exported to a file location that can be accessed by the data manager and BIML team.

Once specimen identifications are finalized, the data are ready and ready to be shared with the GLRI team and published to GBIF. The data are exported from the SQL server database in the form of a dollar sign ($) delimited flat file. This flat file is created by a data workflow that provides some initial QA/QC. Prior to starting the data processing workflow, the flat file is manually uploaded to a file location that can be accessed by the data manager and BIML team.

## Step 2: Cleaning the data and mapping to Darwin Core standard

There are two R scripts that are required to prepare the data for upload into the GBIF Integrated Publishing Toolkit (IPT). The scripts can be accessed in the GitHub Repo folder titled “scripts”. The first script, titled “filter\_and\_join\_tables\_BIML\_all.R” is used to:

1. Remove all records that do not have confirmed species-level identifications present,
2. Reassign the value in the email field to “BIML” if the value does not match any of the other datasetID values in the PIT table,
3. Join the details of the PIT table to the records (1:N) based on matching values in the email field of the flat file and the datasetID field in the PIT table, and
4. Export the new .csv file, titled “species\_project\_joined\_new.csv” for use in the second R script.

Once the outputted file is uploaded to the GitHub Repo in the “data” folder, the second R script, titled “crosswalk.R” can be ran. This script is used to:

1. Fixe values in the “name” field,
2. Remove records that do not have valid species-level identifications,
3. Adjust other field values to be compliant with the Darwin Core Standard,
4. Maps all fields to the Darwin Core Event Core, Occurrence Extension, and Extended Measurement of Fact standards, and
5. Outputs three .csv files for each set of records belonging to unique projects based on the datasetID value.

The GLRI dataset is exported in a set of three .csv files. The first contains the records that will be mapped to the Event Core in the GBIF IPT. The second contains the records that will be mapped to the Occurrence Extension in the GBIF IPT. The third contains the records that will be mapped to the Extended Measurement of Fact in the GBIF IPT. These three files will be used to complete the next step.

## Step 3. Creating the GBIF Project and mapping the data in the Integrated Publishing Toolkit (IPT)

The occurrence records are published to GBIF via the Integrated Publishing Toolkit (IPT). Within the IPT, the GLRI Bee project profile is used to:

1. Map the .csv files to each Darwin Core Standard,
2. Produce the Project-level metadata EML file for the Darwin Core Archive,
3. Set visibility and publication specifications,
4. Register the records to GBIF, and
5. Assign Resource Manager permissions individuals associated with the project.

Resource manager permissions are required to publish data, update project metadata, and adjust visibility and publication settings.