Biodiversity Volunteer Portal

Preparing BVP export data for import into EMu

Overview

This document outlines the process of taking transcribed label data from the Biodiversity Volunteer Portal and putting it into a form for importing into EMu, the Australian Museum database.

John Tann
Australian Museum

December 2013

Contents

Background

Australian Museum natural history specimens from Entomology, Malacology and Marine Invertebrate collections are being digitised using volunteers in a streamlined process. Volunteers in the DigiVol lab work with one batch at a time, and photograph each specimen with its label. At the end of each batch, a short record is created for EMu, and the photographs are uploaded to the Biodiversity Volunteer Portal as an expedition. Online, other volunteers working from home transcribe information on the labels into pre-defined fields. Transcription is followed by a dedicated validation process. As expeditions are completed, label information is made accessible publicly as interim occurrence data through the Atlas of Living Australia. Completed label data is downloaded from the BVP, passed through a series of semi-automatic quality steps, and loaded into EMu.

BVP data preparation

Download

BVP data is downloaded as a series of CSV files – one for each expedition. Each CSV file holds label data for an entire expedition, generally several hundred to a few thousand specimens each with up to 40 fields of information. CSV files are converted to spreadsheets.

Fields summary

Subject	Field
Transcription	transcriber, validator, comments, notes, web link
Administration	catalogue number, collection code
Date	start date, end date
Collector	up to 4 collectors
Location	country, state, precise location, verbatim latitude, verbatim longitude, derived latitude, derived longitude, uncertainty, derived locality, altitude, depth, site number
Method	sampling protocol
Event	field numbers
Taxonomy	scientific name, author, identifier, type status

BVP label data is downloaded as a CSV file with about 40 fields of information

Spreadsheet preparation

Extra columns are added to the spreadsheet

- Ocean – used by Malacology
- District – used especially for islands
- Township
- CEC – Collection Event Code
- SiteNumber – site numbers

Cleaning

Label data is messy. Abbreviations are inconsistent and plentiful. Spelling is imperfect. Specimen labels are commonly handwritten and transcription errors are not unusual.

The dataset of transcribed label data is passed through a series of manual and semi-automatic processes to align and clean the data. A copy of the original data is kept for reference.

Data cleaning processes

Subject	Cleaning process
Dates	• Outliers are detected and fixed
Collectors	• Multiple collectors are separated, eg *M.S. and B.J. Moulds* becomes two people: 1) *M.S. Moulds* and 2) *B.J. Moulds*. • Only the first four collectors are used for any single event. • Non-collector names are removed.
Locations	• Descriptions are made consistent and more readable, eg '*2.5kms (about 1.5 mi) nth of t-off*' will be changed to '*2.5 km N of Turnoff*' • Site Numbers and Collection Event Codes are discovered and assigned • Lord Howe Island and other 'stand-alone' islands are assigned as Districts • Antarctica is re-assigned to Australian Antarctic Territory • Country and State are assigned where absent • Historical names of countries are replaced – eg British New Guinea, New Hebrides, Deutsch-Ostafrika • For PNG, USA and other countries, States and Districts are assigned where known
Methods	• Sampling protocols are assigned a preferred method. For example, MV light and M.V. Lamp are probably the same thing
General	• Abbreviations are made consistent, spelling is corrected

Note: Changes to descriptions are made for matching purposes. Original data is retained.

Reference picklists

An attempt is made to match information on each label against existing records in EMu. If on a label the name of a person matches a Party, or a location matches a Site, or a combination of Date, Party, Site and Collection Method matches an Event, the corresponding EMu IRN is assigned. Each collection is treated separately; for instance, malacology records are checked against only Malacology Parties, Sites and Events.

For each Department in EMu, relevant data from the Parties, Sites and Collection Events modules are extracted, cleaned and converted into three picklists. The preparation of picklists involves both manual and semi-automatic processes, and should be carried out when significant numbers of Parties, Sites or Events recorded in EMu change.

Quality checks

Semi-automated methods are used to highlight potential errors such as an incorrect or missing date, collector or site. A visual check is carried out on the entire dataset.

Importing into EMu

The spreadsheet of cleaned, referenced and checked BVP data is loaded into EMu as a bulk import. Further spot checks are carried out.

Schedule of effort

The following table gives a breakdown of effort for preparing a spreadsheet of transcribed and validated labels ready to be imported into EMu.

Note that each expedition is different. The output of each expedition produces a varying quality of transcribed material, and each department has its own peculiar data requirements. Some batches will match a significant number of existing parties, sites and events, while others won't. Some batches require more manual effort for checking and correcting and will take much longer to process.

Example: Stiletto flies expedition

Statistics

Event	no. of records	% total	matching	% match
Dates	582	98%	496 dates	85%
Collectors	572	97%	778/781 party IRNs	99%
Locations	506	85%	457 site IRNs	90%
Methods	99	17%	99 methods	100%
Events using BVP			249 events	42%
Events using scripts			277 events	47%
Total	**592**	**100%**	**327 events**	**55%**

Preparation for EMu

Table of effort to prepare BVP export data of stiletto flies for import into EMu.

Activity	time	manual or automatic
Check, correct and add country, state, islands	0.5 hour	manual
Check, correct, and extract from verbatim transcription: location, methods, lat-long, elevation	3 hours	manual
Convert, correct and check missing dates	0.5 hour	semi-automatic
Unpack and check people	1 hour	manual
Run scripts to clean people and match parties. Repair and iterate	1 hour	semi-automatic
Run scripts to clean places and match sites. Repair and iterate	1 hour	semi-automatic
Run scripts to match events.		semi-automatic
Check and repair obvious outliers	2 hours	manual
Re-run scripts, prepare for EMu, write-up.	1 hour	semi-automatic
Total time	**10 hours**	

BVP data preparation process chart