A Data Mobilization Project in a Regional Herbarium

Biodiversity Data Mobilization Workshop Use Case 1 - Exercise Solution

Exercise 1: Suggested Solution

STAKEHOLDERS AND ROLES

People and institutions	Affiliations	Stakeholder	Roles
The University of White Plains	Institution	Institution	--
The Plant Biology Department	Institution	Institution	--
Professor of Plant Systematics	Institution;	Collections; Research;	Conservator; Curator; Collection manager; Researcher; Database Administrator
The Departmental Admin [Missing ROLE]	Institution	Institution	Human Resources; Finance
Faculty Students [Missing ROLE]	Institution	Research	Researcher
Two retired botanists who regularly volunteer	Institution; External	Research; Collections;	Volunteer
University's central web-team	Institution	Web Team	Designer; Content Developer; Web Developer
The Head of the Plant Biology Department	Institution; Project	Project Management; Research	Principal Investigator; Project Manager; Researcher
Hosted servers	External	Technology	Network Administrator

GOALS & TASKS - Simplified solution using Before-During-After stages

Stage	Task	Role
Before	Specimen/Object staging	Conservator; Curator; Collection Manager
Before	Curation	Conservator; Curator; Collections Manager
Before	Hiring Staff	Human Resources; Finance

Before	Equipment Purchasing	Human Resources; Finance
Before	Specimen/Object Staging	Volunteer
Before	Curation	Volunteer
Before	Database Modifications	Designer; Web Developer; Database Administrator
Before	Provisioning Storage	Network Administrator
During	Curation	Conservator; Curator; Collections Manager
During	Data Review	Curator; Collections Manager; Researcher
During	Quality Checking	Curator; Collections Manager; Researcher
During	Data Cleaning	Researcher
During	Image Capture	Volunteer
During	Transcription	Volunteer
During	Publishing	Designer; Web Developer; Database Administrator
During	Provisioning Storage	Designer; Web Developer; Database Administrator
After	Re-publishing	Designer; Web Developer; Database Administrator

GOALS & TASKS - Advanced solution using PMBoK stages

Stage	Task	Role
Initiating	Specimen Object Staging;	Volunteers/Intern; Collection Manager
Initiating	Provisioning Storage;	Network Administrator
Initiating	Equipment Purchasing;	Departmental Admin [MISSING ROLE]
Initiating	Hiring Staff;	Departmental Admin [MISSING ROLE]
Initiating	Equipment Setup	Helpdesk [ADDITIONAL RESOURCE NEEDED]
Initiating	Permissions [MISSING TASK]	Legal [MISSING ROLE]; Registrar [ADDITIONAL RESOURCE NEEDED]
Planning	Curation;	Curator; Collections Manager
Planning	Data Standards;	Data Specialist [ADDITIONAL RESOURCE NEEDED]
Planning	Database Modifications	Database Administrator; Programmer
Planning	Provision IPT [MISSING TASK]	Web Developer; Programmer
Monitoring	Quality checking;	Quality Checker [ADDITIONAL RESOURCE NEEDED]
Monitoring	Data Cleaning;	Transcribers [ADDITIONAL RESOURCE NEEDED]
Monitoring	Curation	Curator, Collections Manager
Executing	Publishing;	Principal Investigator; Database Administrator; Collection Manager
Executing	Transcription;	Transcribers [ADDITIONAL RESOURCE NEEDED]; Students
Executing	Data Cleaning;	Transcribers [ADDITIONAL RESOURCE NEEDED]
Executing	Image Capture	Imagers [ADDITIONAL RESOURCE NEEDED]
Closing	Re-publishing;	Principal Investigator; Database Administrator; Collection Manager
Closing	Specimen/Object Return	Volunteers/Intern; Collection Manager

Exercise 2: Biodiversity Data Capture

See Excel spreadsheet

Exercise 3: Biodiversity Data Quality and Standardization

3a. data cleaning

Description of the suggested data cleaning procedures				
Column	**Hint**	**Error**	**Type**	**Subtype**
Country col.	Filter on column then visual check	Country not documented in two cells	Technical	completeness
YE	Filter on column then visual check	2088 is out of the bounds (1 record)	Technical	bounds
countryCode	Filter on column then visual check	Code should include only letters, some cells have coordinates (3 records)	Technical	data type
MO	Filter on column then visual check	In some cells the month (OCT) is not in number format (10) (3 records)	Technical	Data format
taxonRank	Filter on column then visual check for blanks	One record with missing rank	Technical	completeness
Elevation	Filter on column then visual check	9000 is not a valid elevation (1 record)	Technical	bounds
phylum	Filter on column then visual check	Perez S is not a phylum (1 record)	Technical	data type
coordinate Uncertainty	Filter on column then visual check	Uncertainty should be in meters not degrees (1 record)	Technical	Data format
Link to the result spreadsheet				

3b. other data management tools

Description of the suggested data cleaning procedures				
Column	**Hint**	**Error, solution and lessons learned**	**Type**	**Tools used**
Full Name	Are the names valid?	**Correction for the Eriocaulaceae family** Unmatched names: Paepalanthus alpestri Pepalanthus Erioculon Paepalanthus karsten	Nomenclatural	Global Names Resolver
lat/lon	Are all coordinates consistent and in decimal format?	**Correction for the Eriocaulaceae family** No, 6 records have coordinates are in degrees minutes seconds	Format	Canadensys coordinate conversion
lat/lon	Are all occurrences taking place in Guatemala?	**Correction for the Eriocaulaceae family** No, 2 are outside Guatemala Mali - Bay of Bengal	Geographic	InfoXY or Google Maps
eventDate		Creation of the column eventDate and conversion in the proper format.	Format	Excel tools + Canadensys date format conversion
Link to the result spreadsheet				

Exercise 4: Biodiversity Data Publishing

Description of the suggested data publishing procedure

I logged in to the IPT with my credentials, then opened the ***Management tab***, where I used the form at the bottom to create a new resource. I selected *Occurrence* for the ***Type*** and clicked on the ***Create*** button without selecting a file.

In the ***Source Data*** section of the ***Overview page*** I chose the cleaned occurrence data set (*12b IPT Data Publishing - Poales Dataset (ready for publishing).csv*) and uploaded it by clicking on the ***Add*** button. I verified that the source data format understood by the IPT was correct by reviewing the data (by clicking on the eye icon). Then, I clicked on the ***Save*** button to initiate the upload. The resulting page showed that the file had been uploaded and contained 760 rows and 32 columns, as expected.

In the ***Darwin Core Mappings*** section of the ***Overview page*** I selected ***Darwin Core Occurrence*** from the dropdown list, then clicked on the ***Add*** button. In the new page, I selected the one possible data source and clicked on the ***Save*** button. This opened the ***Mapping Source Data*** page. I noticed that 21 of the fields in the source data were auto-mapped, and that there was a warning that the

basisOfRecord is required. I scrolled to the bottom of the page to see which fields were not auto-mapped and found ***YE, MO, DA, Country col., lat, lon, Locality col., Elevation, Full name, Authorship*** and ***Other name***. Knowing that the ***lat*** and ***long*** were cleaned and formatted to decimal degrees, I mapped ***lat*** to ***decimalLatitude*** and ***long*** to ***decimalLongitude***. Looking back at my source file, I remembered that the fields ***YE***, ***MO*** and ***DA*** stood for ***year***, ***month*** and ***day*** respectively so I mapped them to the adequate Darwin Core terms on the IPT. ***Country col.*** and ***Locality col.*** were mapped to the ***country*** and ***locality*** Darwin Core fields. ***Elevation*** was mapped to both ***minimumElevationInMeters*** and ***maximumElevationInMeters***, ***Full Name*** to ***scientificName***, ***Authorship*** to ***scientificNameAuthorship*** and ***Other Name*** to ***vernacularName***. Then clicked on the ***Save*** button and reviewed the list of values that showed up under these fields to confirm that they were as expected. They were!

I still had the warning that ***basisOfRecord*** is required, so I mapped that value to the constant "***PreservedSpecimen***" (as my source data came from herbarium sheets) and clicked on the ***Save*** button. There was no longer a warning about ***basisOfRecord***. So I clicked on the ***Save*** button, then clicked on the ***Resource Title link*** at the top of the page to return to the ***Overview page*** where it showed in the ***Darwin Core Mappings*** section that the ***Darwin Core Occurrence*** mapping included 34 terms (the 32 original source fields (including ***elevation*** that was mapped twice), and ***basisOfRecord***).

Next I clicked on the ***Edit*** button to enter the first of many metadata entry forms - ***Basic Metadata***. On the ***Basic Metadata*** page I chose ***Specimen*** as the Subtype, as this Occurrence dataset came from a list of herbarium specimens. I composed the rest of the metadata based on the information provided in the use case document, and I chose a licence adequate for GBIF data publishing. When finished, I returned to the ***Manage Resources*** page by clicking on the ***Resource Title*** link at the top of any of the Metadata entry pages.

I chose to set the ***Visibility*** of the resource to ***Public*** even before publishing, because I remembered that if I publish for the first time with the ***Visibility still private***, I would have to publish again after making the resource ***Public*** in order for it to actually appear in the IPTs Home page.

Then I clicked on the ***Publish*** button in the ***Published Versions*** section and the resulting ***Publishing status*** confirmed that all went well.

I looked at the ***Published version*** section of the ***Overview page*** and saw that so far I had published version 1.0. I then clicked on the ***Home page*** and confirmed that my new Occurrence dataset was on the list with 760 published records. *I am happy!*

Link to the result published dataset

Example from a published Occurrences dataset on the GBIF France IPT:
http://ipt.gbif.fr/resource?r=confluences_cirripedes

Use case concept and content developed by Alberto González-Talaván, Néstor Beltrán, Nicolas Noé and Sharon Grant.