NHM Georeferencing & Mass Digitisation

MOBILISE Workshop

Warsaw, Poland
Feb 10-12 2020

Laurence Livermore

Natural History Museum, London

NHM Overview

- **Collections Management System:** EMu (Axiell ALM)
- **Main geographic schema:** Database of Global Administrative Areas (GADM)
- **Other geographic schemas:** World Geographical Scheme for Recording Plant Distributions (WGSRPD), various biogeographic regions
- **Transcription software:** Custom web application, Excel, CMS apps, direct into CMS
- **Georeferencing software:** Largely none (manual), Google Maps API (prior projects)
- **Georeferencing standards:** In-house, based on MaNIS/HerpNET/ORNIS
- **Data Portal:** Customised CKAN-based OS data portal, provides data to GBIF
- **Quality checks:** Manual integration of GBIF data but no “closed loop” / process

NHM Georeferencing Guidelines

Defines:

1. **Locality types** (10 named places, 5 offsets, 2 coordinates)
2. **Georeferencing procedure** (for determining decimal latitude and longitude)
3. **Determining extent** (in metres, also includes guidance on comments/assumptions/accuracy)

Does not cover footprint geometry/shapes

Public Data – Shared with GBIF

Category	Number of records	% of total collections
Data Portal – Public Specimens	4,525,711	5.6%
/w locality	2,355,085	2.9%
/w decimal lat & long	1,230,842	1.5%
+ georeference protocol	143,937	0.2%
+ geodeticDatum	15,157	<0.02%

georeferenceProtocol (Top 20)

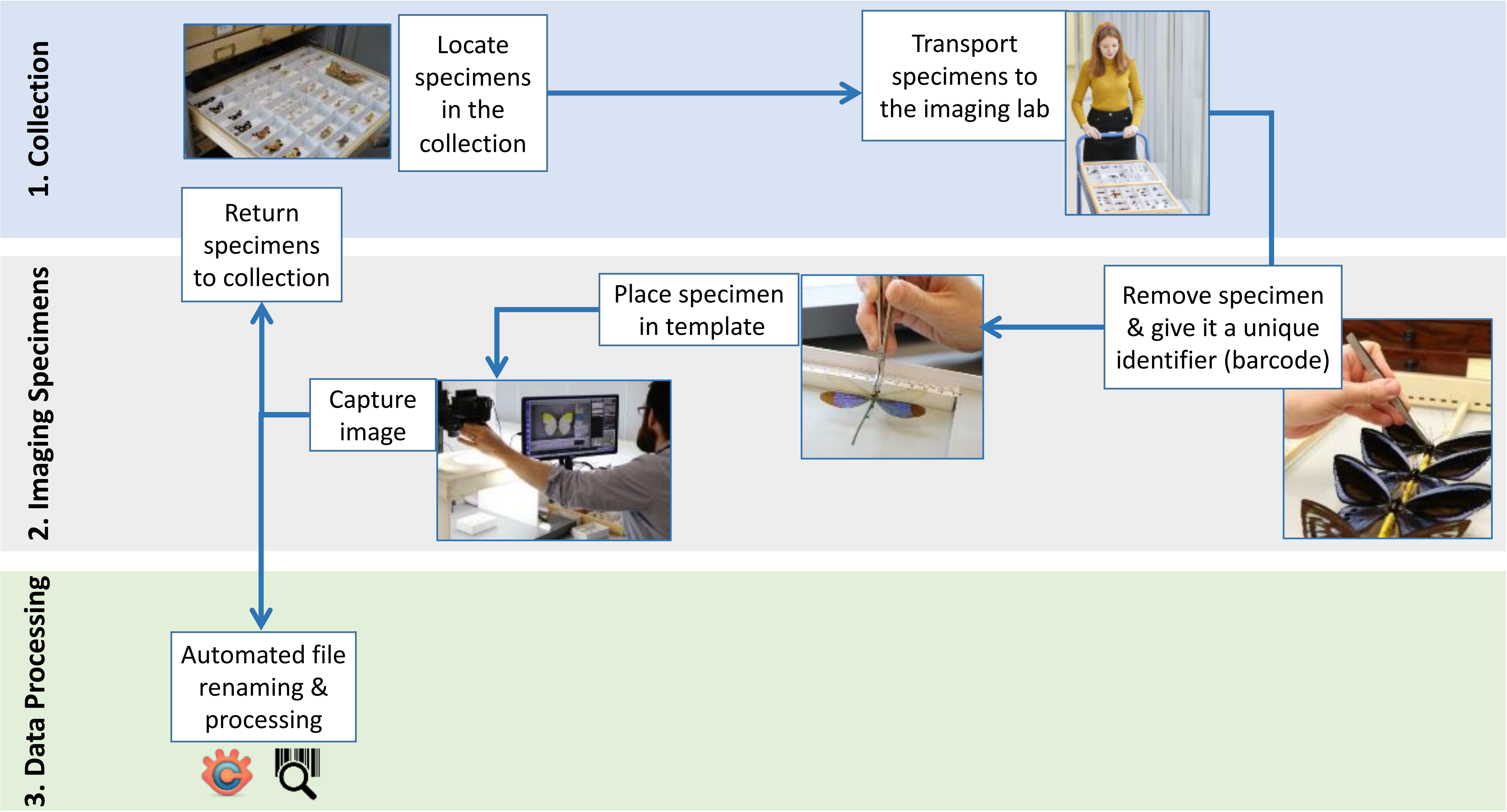
Google Earth	81256
iCollections protocol	23796
Area polygon calculation M. Penn	16668
Google Maps	7639
Google Map	6431
GPS	5756
Online Gazetteer	708
Microsoft MapPoint	520
via web	307
Gazetteer	250
Wikimapia	187
Bing Map	135
Specimens label	114
NE	95
PhD thesis of H.V. Hunt	80
Specimen label	62
Estimate	57
Grabagridref(Beds. NHS)	23
GBIF Best Practice	19
Grid ref.	15

Retrieved: February 2020 from NHM Data Portal (data.nhm.ac.uk)

Absent* DarwinCore Location Properties

1. locationAccordingTo
2. locationRemarks
3. **coordinateUncertaintyInMeters**
4. coordinatePrecision
5. pointRadiusSpatialFit
6. verbatimCoordinates
7. verbatimCoordinateSystem
8. verbatimSRS
9. **footprintWKT**
10. **footprintSRS**
11. **footprintSpatialFit**
12. georeferencedBy
13. georeferencedDate
14. georeferenceProtocol
15. georeferenceSources
16. georeferenceVerificationStatus
17. georeferenceRemarks

*Not present in NHM Data Portal, some fields are present in our CMS but not well-used, standardised or included in protocols. Some absent in CMS.



1. Collection

Locate specimens in the collection

Transport specimens to the imaging lab

2. Imaging Specimens

Remove specimen & give it a unique identifier (barcode)

Place specimen in template

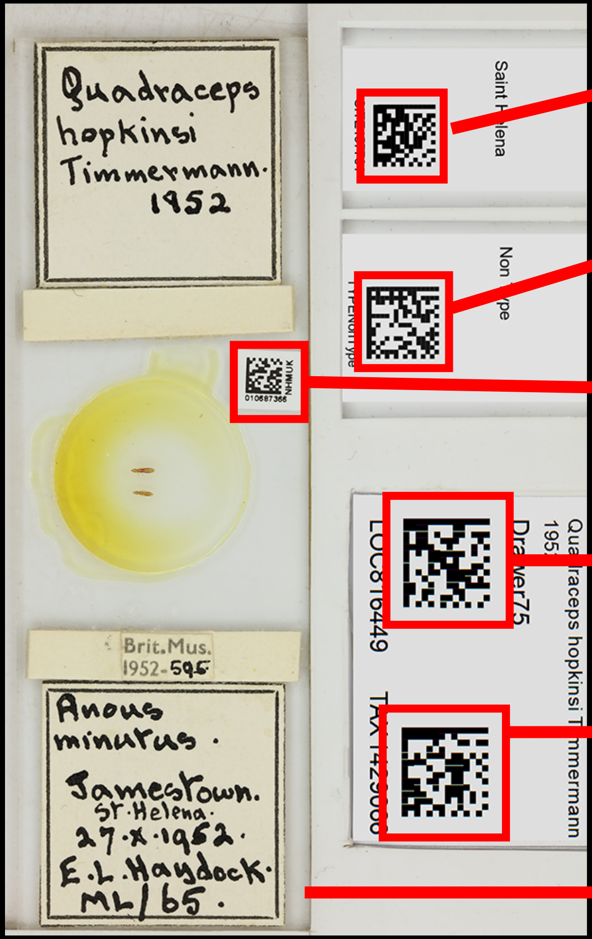
Capture image

Return specimens to collection

3. Data Processing

Automated file renaming & processing

Allan E, Livermore L, Price B, Shchedrina O, Smith V (2019) A Novel Automated Mass Digitisation Workflow for Natural History Microscope Slides. Biodiversity Data Journal 7: e32342. https://doi.org/10.3897/BDJ.7.e32342



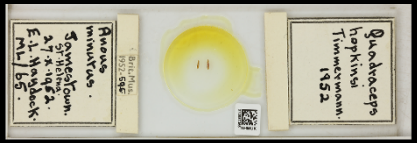
Locality: SITE157761 *(Saint Helena)*

Type: TYPENonType *(Non-type)*

Specimen ID: 010687385

Storage Location: LOC816449 *(Drawer 75)*

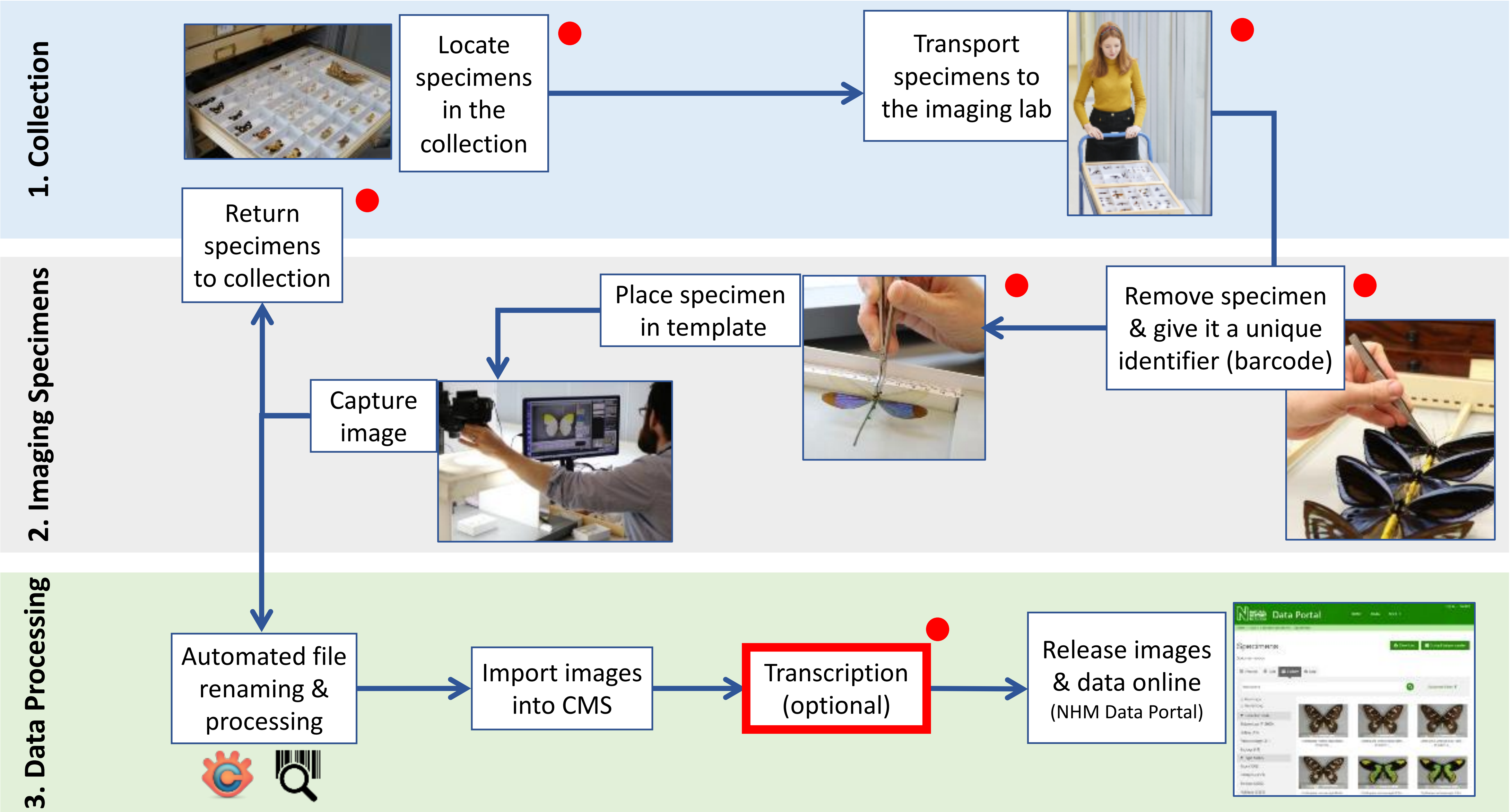
Taxonomy: TAX1429066 *(Quadraceps hopkinski)*



Processed and imported into institutional systems (CMS, public portal)

We would use more but we've hit the limits of our software…

Allan E, Dupont S, Hardy H, Livermore L, Price B, Smith V (2019) High-Throughput Digitisation of Natural History Specimens. Biodiversity Information Science and Standards 3: e37337. https://doi.org/10.3897/biss.3.37337



1. Collection

Locate specimens in the collection

Transport specimens to the imaging lab

2. Imaging Specimens

Remove specimen & give it a unique identifier (barcode)

Place specimen in template

Capture image

Return specimens to collection

3. Data Processing

Automated file renaming & processing

Import images into CMS

Transcription (optional)

Release images & data online (NHM Data Portal)

Allan E, Livermore L, Price B, Shchedrina O, Smith V (2019) A Novel Automated Mass Digitisation Workflow for Natural History Microscope Slides. Biodiversity Data Journal 7: e32342. https://doi.org/10.3897/BDJ.7.e32342

Digital Collections Programme T2 2016-2019

- In three years we digitised **~245,000 specimens across 11 projects**
- All projects photographed labels with locality information
- **Eight projects had partial transcription** ~71,000 specimens (29%)
- **Three had full (interpreted) transcription** ~11,000 specimens (4%)
- **One had complete georeferencing** 6,760 (2.8%) due to project/grant research requirements

Priority is mass digitisation – looking for effective, scalable methods to capture locality data and to georeferenced it

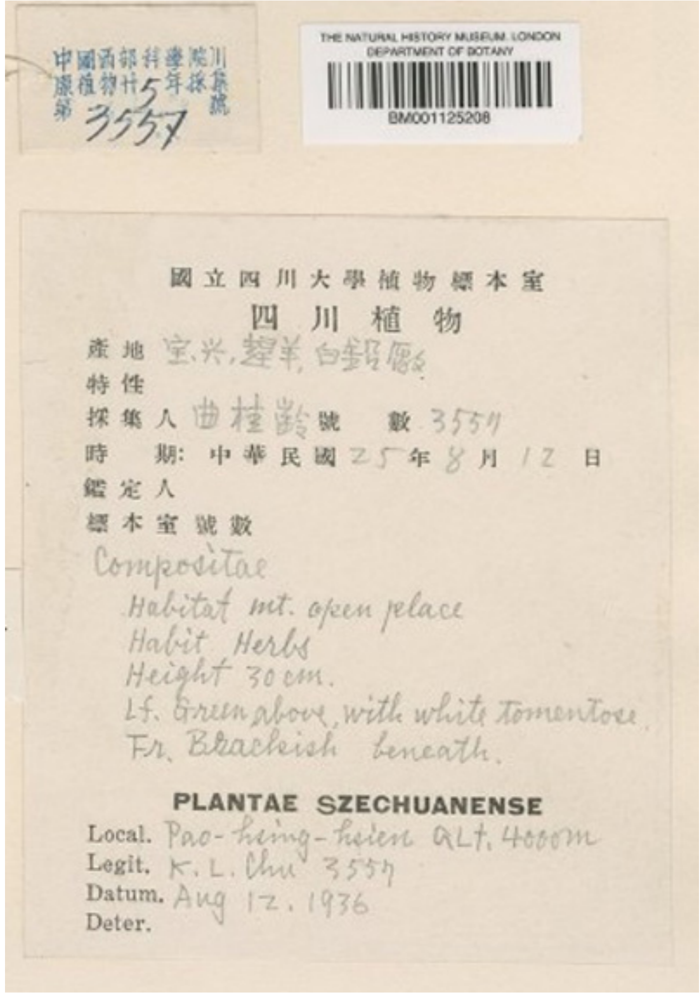
Project 1 – UK Bumblebees

- 6,760 UK bumblebee (Bombus sp.) specimens
- 52 person days
  - 32 days imaging (Pinned standard, barcode workflow)
  - 11 days transcription (Excel, 644/day, n=6,236), verbatim + atomised
  - 9 days georeferencing (80 sites/day, **total unique georeferenced sites = 716**)
  - Google and Ordnance Survey – followed internal procedure
- **Note:** specimens not publicly available – currently under research embargo

Project 2 – Chinese Botanical Types

- 3,736 Chinese botanical type specimens (JSTOR GPI)
- Transcription split into 3 stages:
  - Grouped by collector (from existing data)
  - Checked/verified all data (5 Romanization systems)
  - Georeferenced data
- ~60 person days (~62 specimens/day)
- Suppl. data: collectors lists; spelling vars., data sources.

NB: In addition to knowledge of collections and taxonomy, georeferencer had following language skills: English C2, Chinese (written) B1, French A2, Latin A2, [Hungarian Native] – they also encountered Russian and Japanese localities.



中國西部科學院川康植物廿5年採集第3557號

THE NATURAL HISTORY MUSEUM, LONDON
DEPARTMENT OF BOTANY
BM001125208

國立四川大學植物標本室
四川植物

產地 宝兴,趕羊,白鋁廠
特性
採集人 曲桂齡 號數 3557
時期: 中華民國 25 年 8 月 12 日
鑑定人
標本室號數

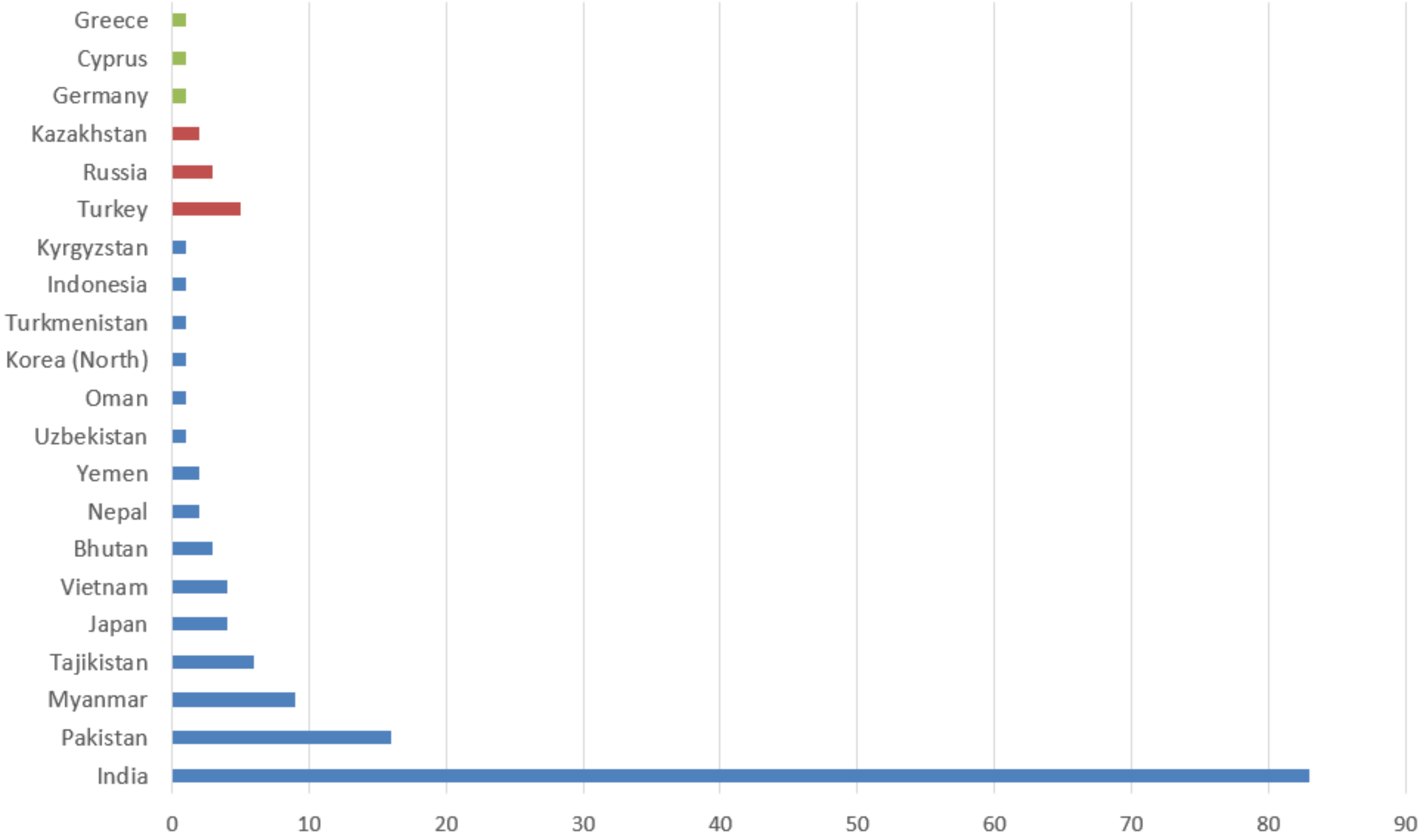
Compositae
Habitat mt. open pelace
Habit Herbs
Height 30 cm.
Lf. Green above, with white tomentose
Fr. Brackish beneath.

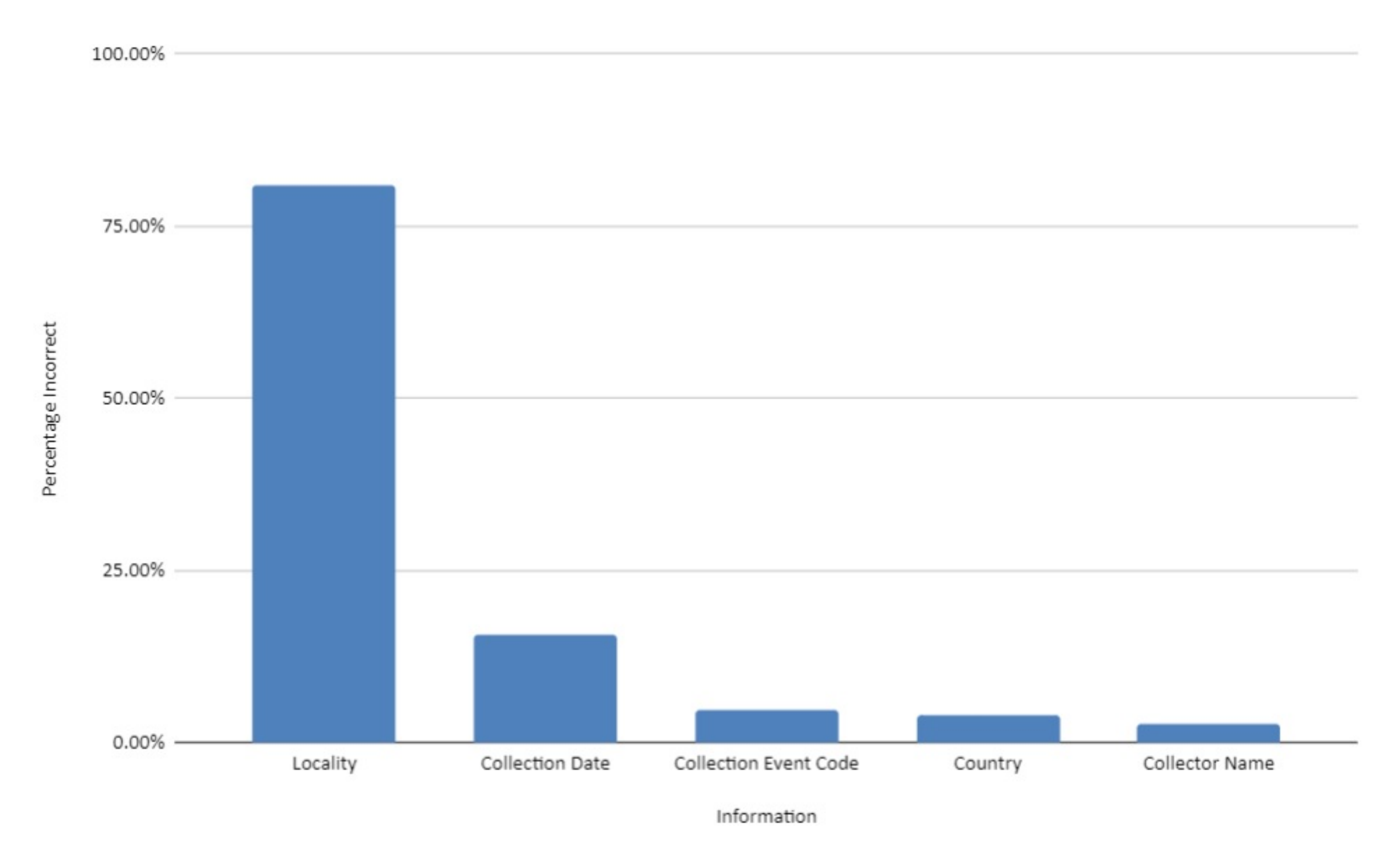
PLANTAE SZECHUANENSE

Local. Pao-hsing-hsien ALt. 4000m
Legit. K. L. Chu 3557
Datum. Aug 12. 1936
Deter.

Specimen BM001125208 - Example of a label written in Chinese and English with more information in Chinese than in English.

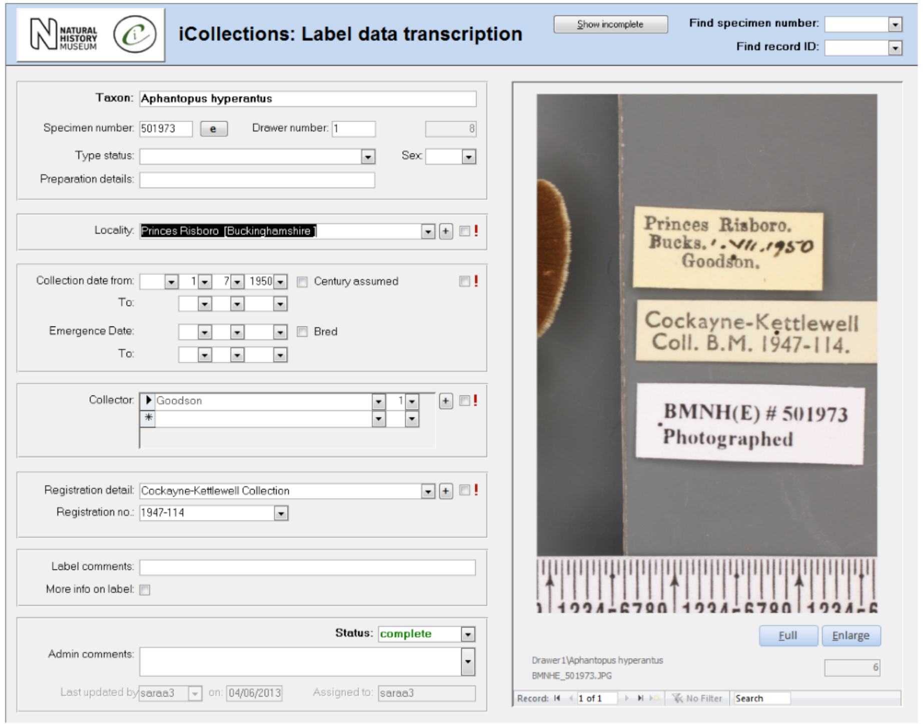
Lohonya K, Livermore L, Penn MG (In Review) Georeferencing the Natural History Museum's Chinese type collection: of plateaus, pagodas and plants


Greece
Cyprus
Germany
Kazakhstan
Russia
Turkey
Kyrgyzstan
Indonesia
Turkmenistan
Korea (North)
Oman
Uzbekistan
Yemen
Nepal
Bhutan
Vietnam
Japan
Tajikistan
Myanmar
Pakistan
India
0
10
20
30
40
50
60
70
80
90


100.00%
75.00%
50.00%
25.00%
0.00%
Percentage Incorrect
Locality
Collection Date
Collection Event Code
Country
Collector Name
Information

Project 3 – iCollections (2013-2016)

- 380,686 UK butterfly & moth specimens
- Transcription split into 4 stages:
  1. Transcription
  2. Taxon name mapping
  3. Site processing
  4. Georeferencing
- Protocol not part of any records!



iCollections: Label data transcription

Find specimen number:
Find record ID:

Taxon: Aphantopus hyperantus
Specimen number: 501973 Drawer number: 1 8
Type status: Sex:
Preparation details:

Locality: Princes Risboro [Buckinghamshire] □

Collection date from: 1 7 1950 □ Century assumed □
To:
Emergence Date: □ Bred
To:

Collector: Goodson 1 □

Registration detail: Cockayne-Kettlewell Collection □
Registration no.: 1947-114

Label comments:
More info on label: □

Status: complete
Admin comments:
Last updated by saraa3 on: 04/06/2013 Assigned to: saraa3

Drawer1\Aphantopus hyperantus
BMNHE_501973.JPG
6

Record: 1 of 1

“Semi-automated georeferencing functions based on Google Maps and various georeferencing software tools, such as Biogeomancer, allow at least 10% of the total sites variants to be georeferenced quickly and accurately.”

https://bdj.pensoft.net/article/19893/

Challenges

- **Specimen-by-specimen approach to georeferencing inefficient:** Collections largely organised by taxon - a few with varying degrees of geographical suborganisation

- **CMS/Data Portal issues**
  - Poor implementation of DwC Location class in CMS (non-standard, missing fields)
  - Confounded “site” model in CMS
  - Limited standards/checks in CMS outside of GADM hierarchy
  - Limited integration of georeferencing tools in CMS
  - Lack of CMS field mapping to Data Portal DwC-A

- No automated improvements or process for fixing of GBIF-flagged geo issues

Summary of Tasks

Task 8.1

Landscape evaluation

- Evaluate platform based approaches inc. tool/service registries
- Identify sources of data
- Identify & create training/reference datasets
- Assess potential to use EUDAT, EGI, EOSC, AWS etc
- Reuse & integrate reports from previous DiSSCo-related work

D8.1 Report (NHM)

Task 8.2

Development of tools, services & workflows

- *Services include:*
- Optical character recognition data capture and analysis
- Data mining and linkage
- Image analysis
- Georeferencing
- Human interaction and crowdsourcing

- Mature services will be containerized and incorporated into SDR workflow & tool registry

D8.2 Demonstrator (NHM)

Task 8.3

Development of SDR cloud platform

- Platform will run workflows in SDR registry
- Controlled user access (ELIXIR authentication & authorisation)
- Create detailed standardized provenance metadata of derivative datasets
- Derivative datasets and outputs packaged as Research Objectives (ROs)
- Ledger log of all ROs & specimen UIDs

D8.3 Demonstrator (UNIMAN)

Task 8.4

Data delivery & exploitation

- User acceptance testing (against DiSSCo user stories)
- Training and workshops to promote use
- Delivery of data to collections management systems
- Exploitation of data by third parties

D8.4 Report (Naturalis)

External

Specimen Data Refinery Workflows

External

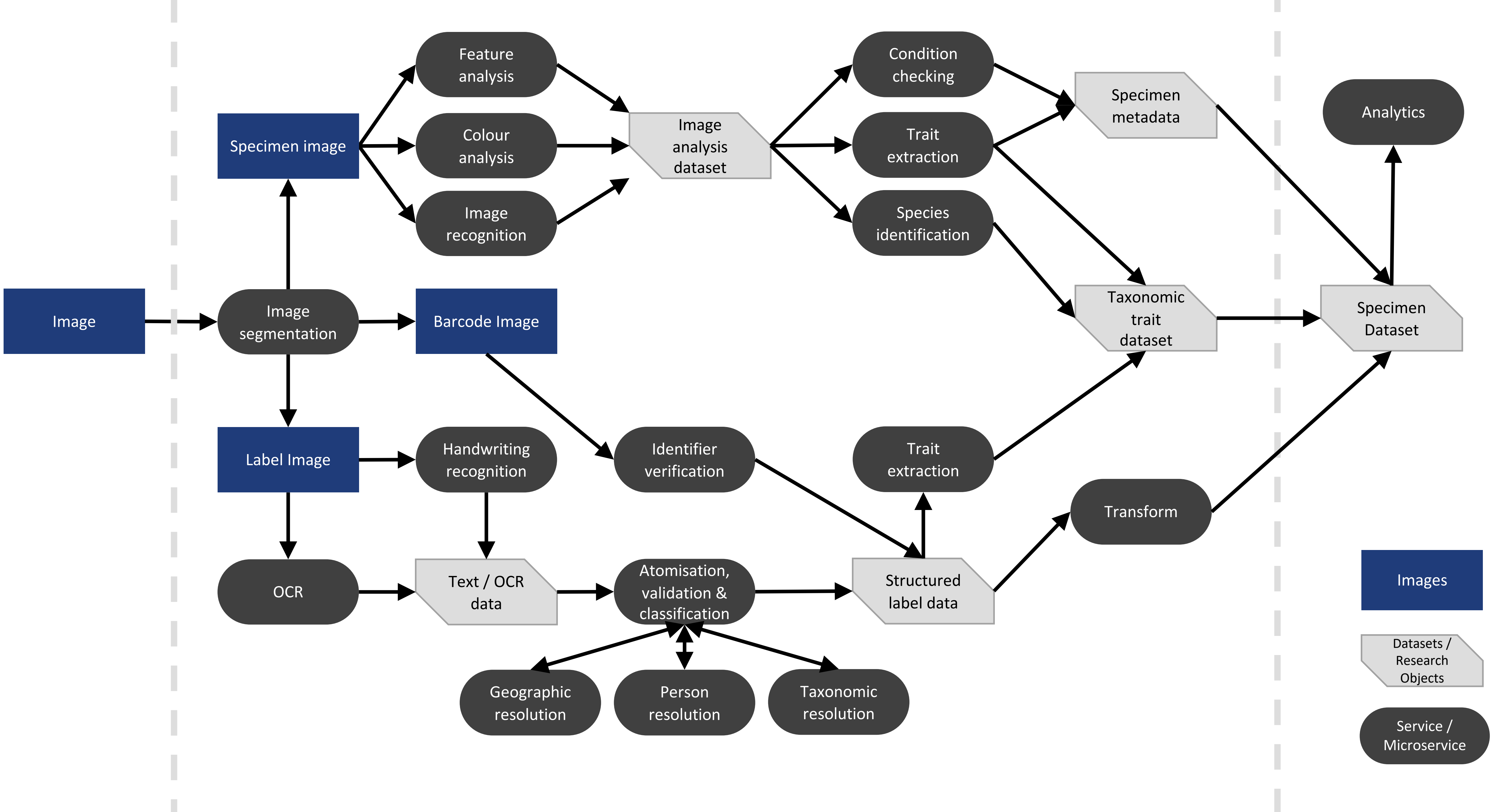
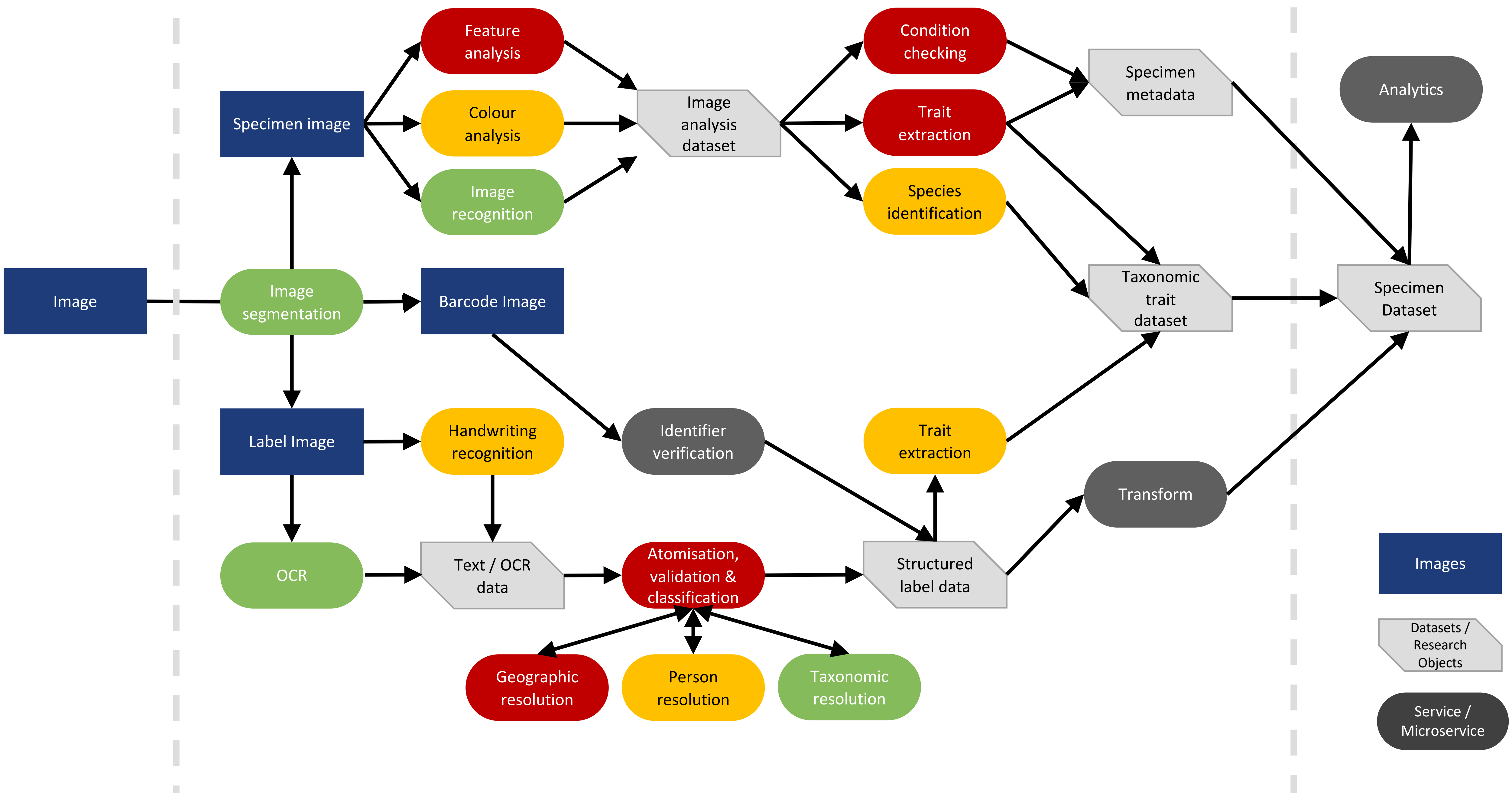


Image
Image segmentation
Specimen image
Barcode Image
Label Image
Feature analysis
Colour analysis
Image recognition
Image analysis dataset
Condition checking
Trait extraction
Species identification
Specimen metadata
Taxonomic trait dataset
Analytics
Specimen Dataset
Handwriting recognition
Identifier verification
OCR
Text / OCR data
Atomisation, validation & classification
Structured label data
Trait extraction
Transform
Geographic resolution
Person resolution
Taxonomic resolution

Images
Datasets / Research Objects
Service / Microservice

Specimen Data Refinery Workflows



External

External

Image

Image segmentation

Specimen image

Barcode Image

Label Image

Feature analysis

Colour analysis

Image recognition

Image analysis dataset

Condition checking

Trait extraction

Species identification

Specimen metadata

Taxonomic trait dataset

Specimen Dataset

Analytics

Handwriting recognition

OCR

Text / OCR data

Identifier verification

Atomisation, validation & classification

Geographic resolution

Person resolution

Taxonomic resolution

Structured label data

Trait extraction

Transform

Images

Datasets / Research Objects

Service / Microservice

Acknowledgements

Stephanie Walton – SDR RAG slide/analysis

NHM Digitisation Team (Louise Allan, Robyn Crowther, Elizabeth Devenish, Phaedra Kokkini, Krisztina Lohonya, Nicola Lowndes, Olha Shchedrina, Peter Wing), **NHM Data Management Team** (Malcolm Penn, Charlotte Barclay, Hillery Warner), **iCollections Team**, **Curatorial Teams**.