A Treasure Trove of Nature

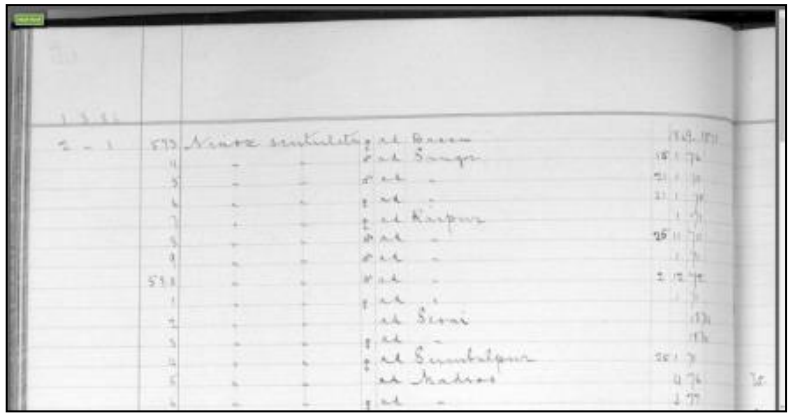
The advances and challenges of digitising natural history specimens

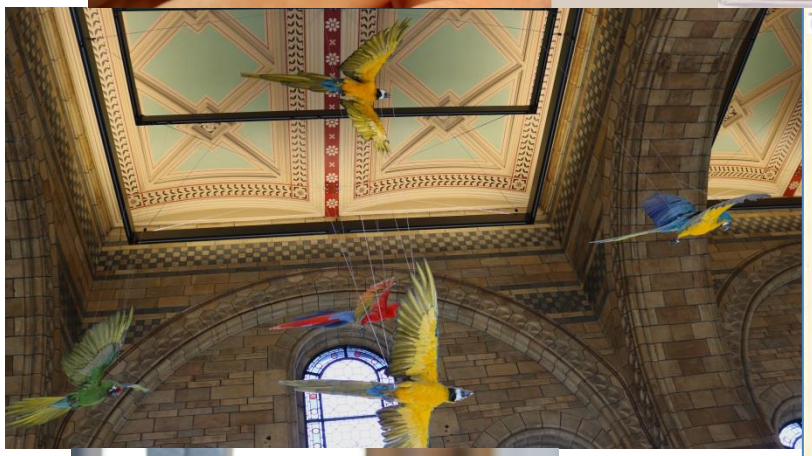
Steen Dupont and Laurence Livermore

16-05-2019 British Computing Society

NATURAL HISTORY MUSEUM

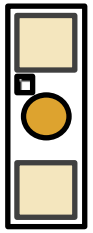

Nucella tetragona
(Sea snail)

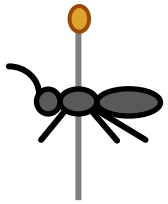




What is the composition of the collections?

Microscope slides (2.5M)





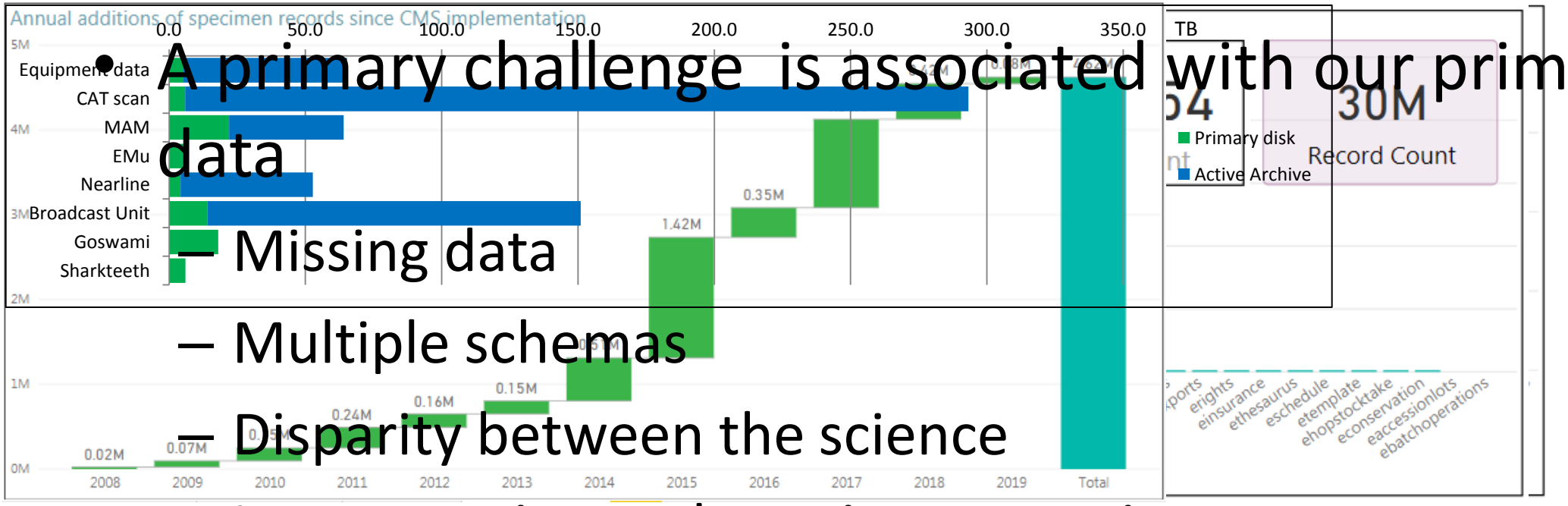
Pinned insects (~25M)

Herbarium sheets (2.8M)

Labelled segments account for >90% of our specimens by count!

How do we keep track of it all

- Good old index cards and catalogues
- Our collections management system
- A primary challenge is associated with our primary data
  - Missing data
  - Multiple schemas
  - Disparity between the science
  - Interpretation and non-interpretation



Summary

- We have lots of stuff, it’s all very different.
- Not much of it is represented in our data base
- Five years ago we started a digitisation programme to digitise everything.
- To tackle the variation of the collections we need industrial scale processes that are highly customisable
- There were some challenges along the way

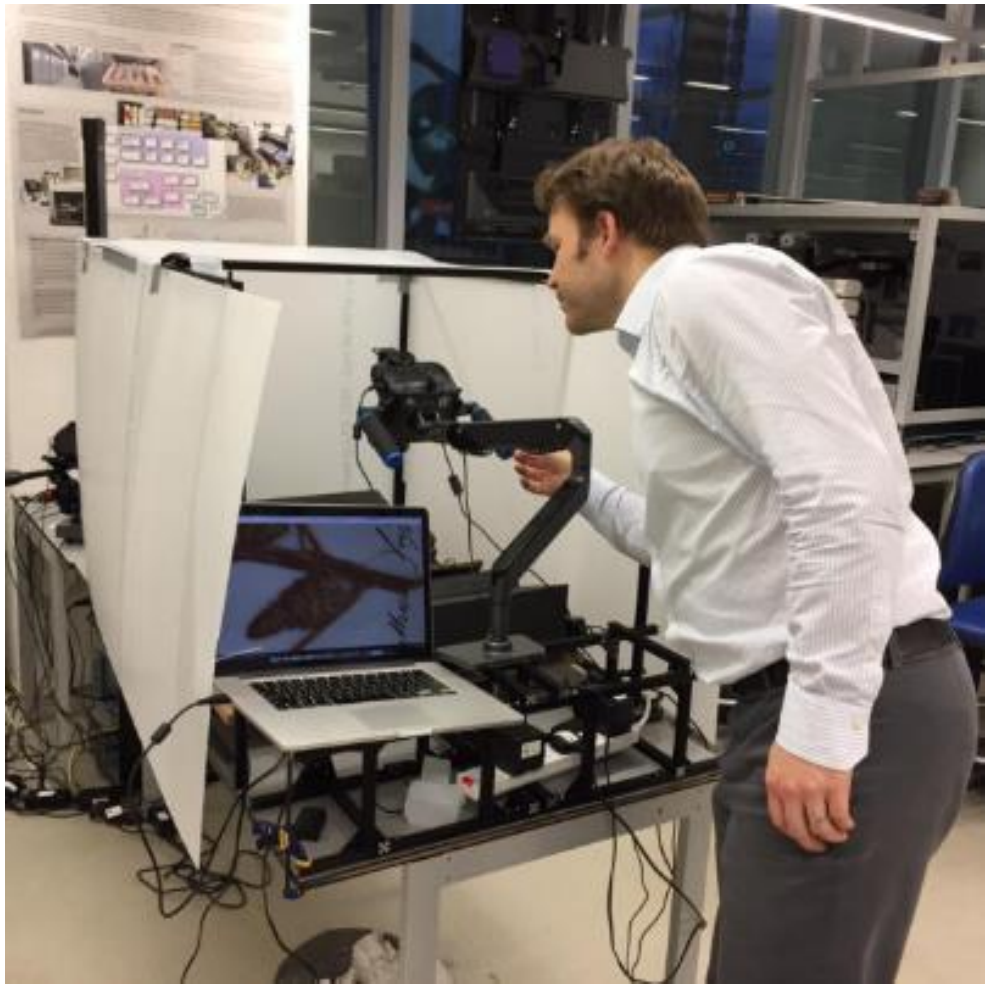
The Digital Collections Programme

(2014-2024, currently in phase 3)

- Embarking on an epic journey to digitise 80 million specimens
- Giving the global scientific community access to unrivalled historical, geographic and taxonomic specimen data
- Creating the foundation for a global initiative aimed at outlining and answering global biodiversity challenges.

https://www.nhm.ac.uk/our-science/our-work/digital-museum/digital-collections-programme.html

This is where we come in!



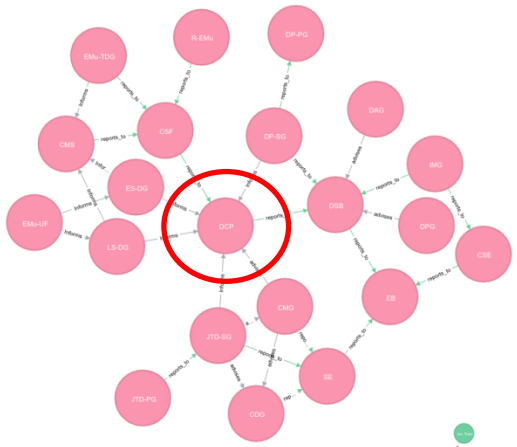
Hardware

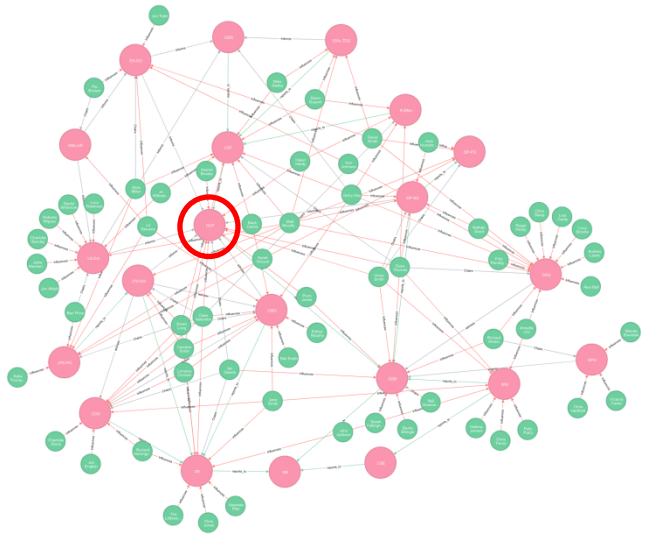
LAURENCE

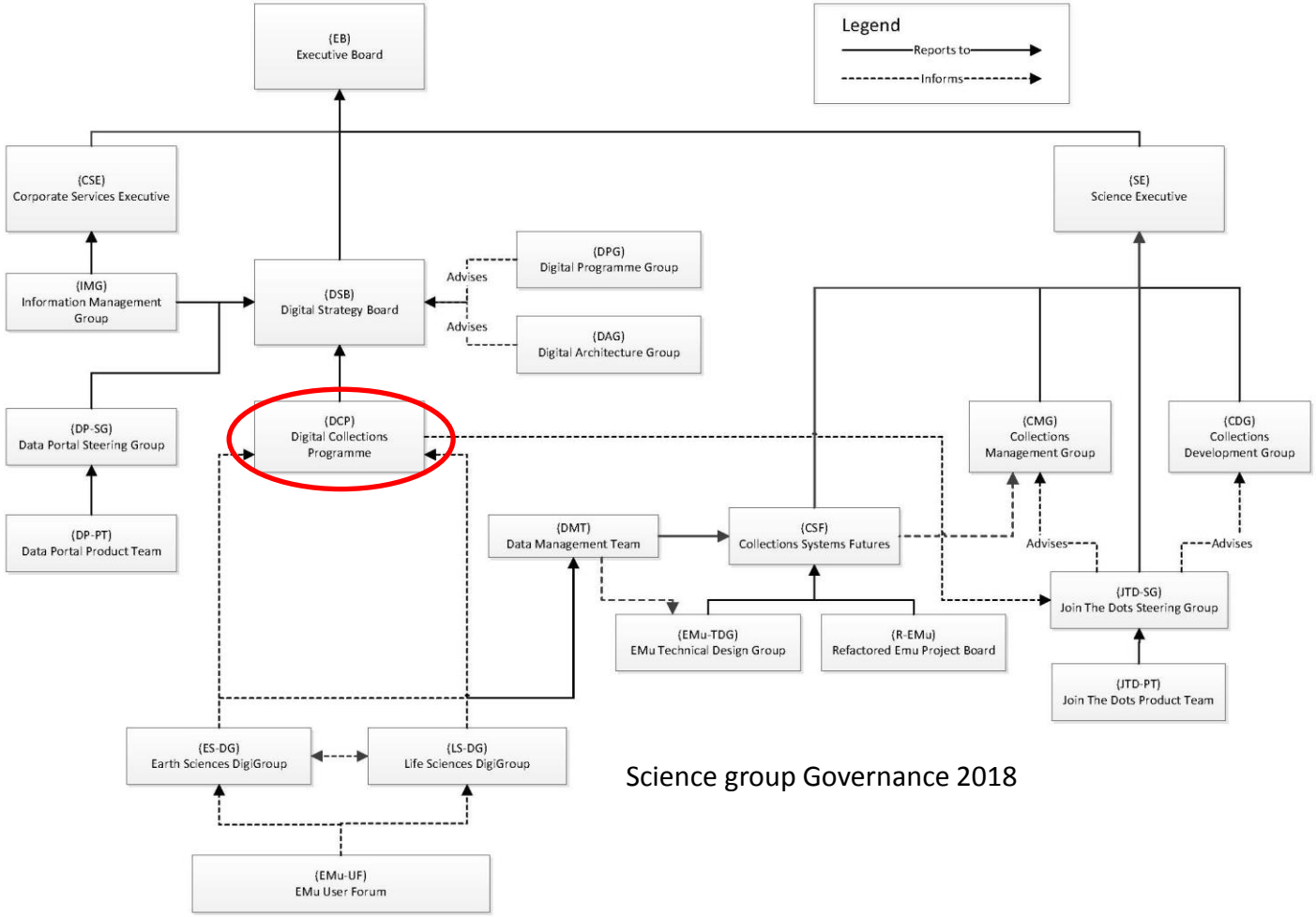
Software

Digitisation

And this is where we fit into the org chart







Science group Governance 2018

How do you start digitising 80 million objects?

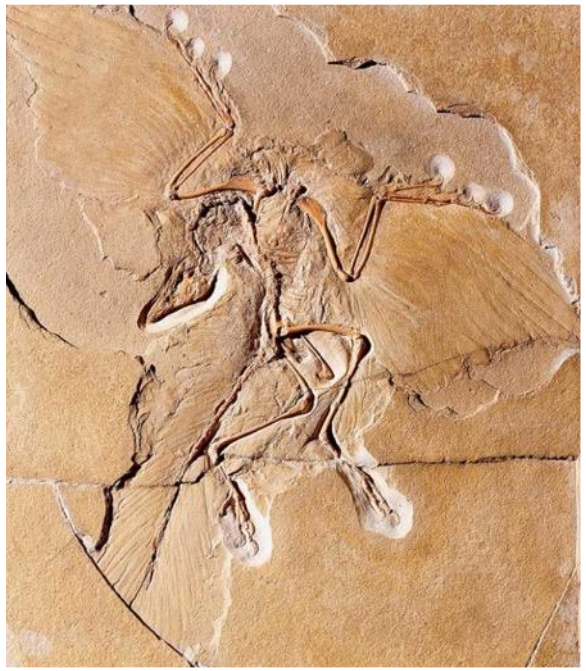
(Especially when you are not sure exactly what you have because nothing is digitised yet)

- Cultural change
- Developing processes (standards, policy > specimen audits)
- Practicality (cost, time, expertise)
- Prioritisation (research, curation, funding, public interest)

Why digitise the collections?

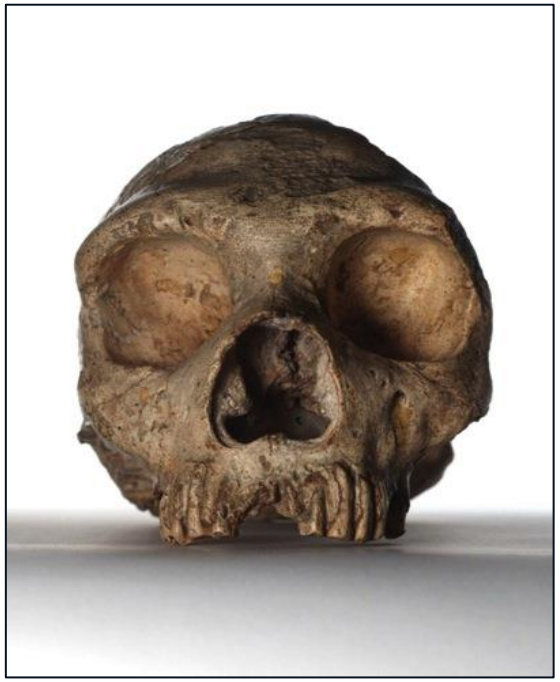
We have things that have changed the way we think and how we see the world

Missing link - *Archaeopteryx*



Darwin's finches

First *Neanderthal* skull



Why digitise the collections?

We have data that enables us to:

travel through time to visualise the impact of global changes

address spread and potential impact of diseases and their vectors

Ecography 40: 1152–1165, 2017
doi: 10.1111/ecog.02658

Subject Editor: Sarah Diamond. Editor-in-Chief: Miguel Araújo. Accepted 23 August 2016

The influence of life history traits on the phenological response of British butterflies to climate variability since the late-19th century

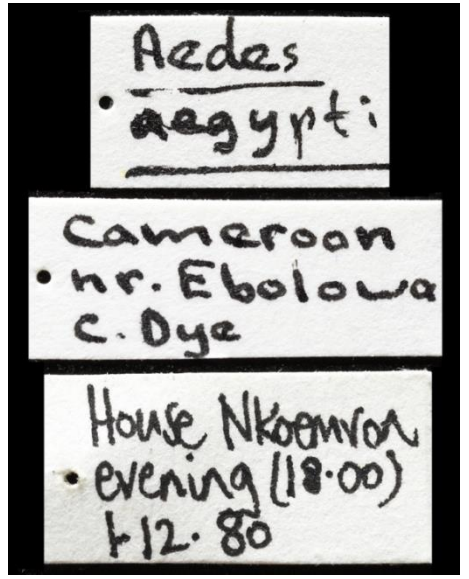
Stephen J. Brooks, Angela Self, Gary D. Powney, William D. Pearse, Malcolm Penn and Gordon L. J. Paterson

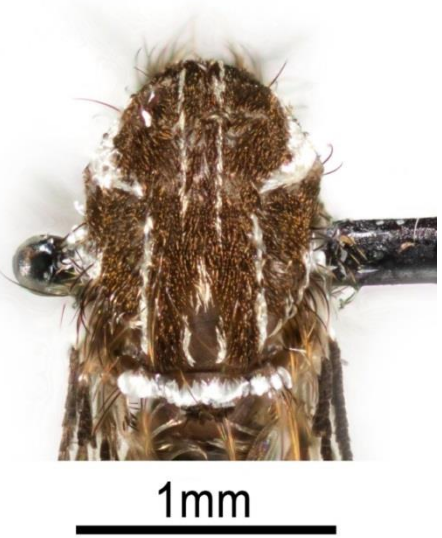
International Journal of Epidemiology, 2017, 1–10
doi: 10.1093/ije/dyw366
Original article

Original article

Spatial quantification of the world population potentially exposed to Zika virus

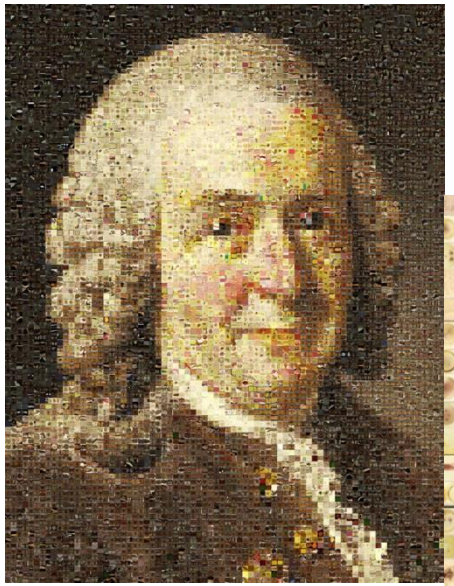
Alberto J. Alaniz,[1,2*] Antonella Bacigalupo[3] and Pedro E. Cattan[3]

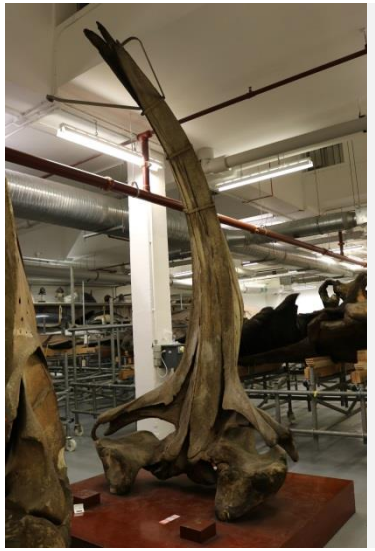




Why digitise the collections?

The data we create provides the underlying resource to make new innovative ways of presenting and interacting and engaging with our specimens






5 mm

What is a “typical” digitisation workflow?

(and what do we mean by digitisation?)

A “typical” digitised specimen

Higher Classification

Scientific name: *Ornithoptera victoriae regis* Rothschild, 1895
Family: Papilionidae

Location

Locality: Bougainville
Country: Solomon Islands
Continent: Oceania

Collection Event

Recorded by: A S Meek

Specimen

Catalogue number:	BMNH(E)102551
Preservative:	Dry - mounted
Individual count:	1
Sex:	Male
Life stage:	Adult
Barcode:	013602485

External Links

BHL Biodiversity Heritage Library

Catalogue of Life

Find more links on the GBIF View

Permanent URL: https://data.nhm.ac.uk/object/407b7063-f942-42f2-a107-885a82f8cc18/1557705600000

1. Collection

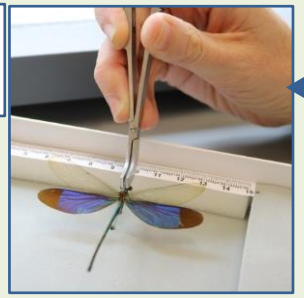
Locate specimens in the collection

Transport specimens to the imaging lab

2. Imaging Specimens

Return specimens to collection

Place specimen in template



Remove specimen & give it a unique identifier (barcode)

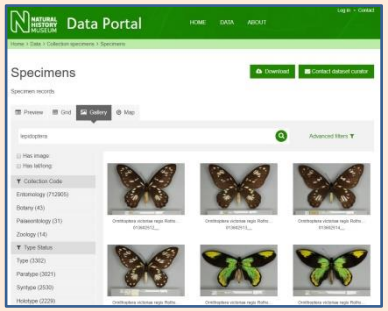
Capture image

3. Data Processing

Automated file renaming & processing

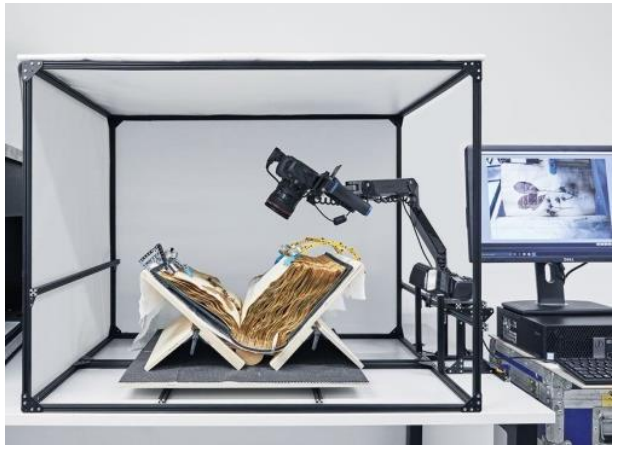
Import images into CMS

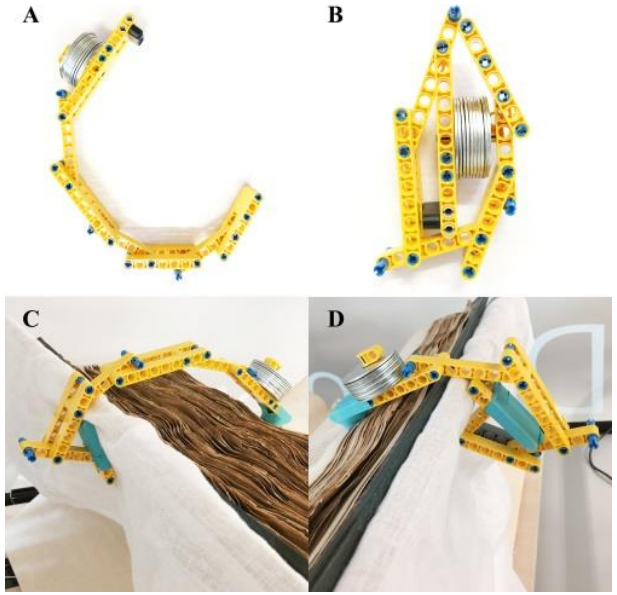
Release images & data online (NHM Data Portal)



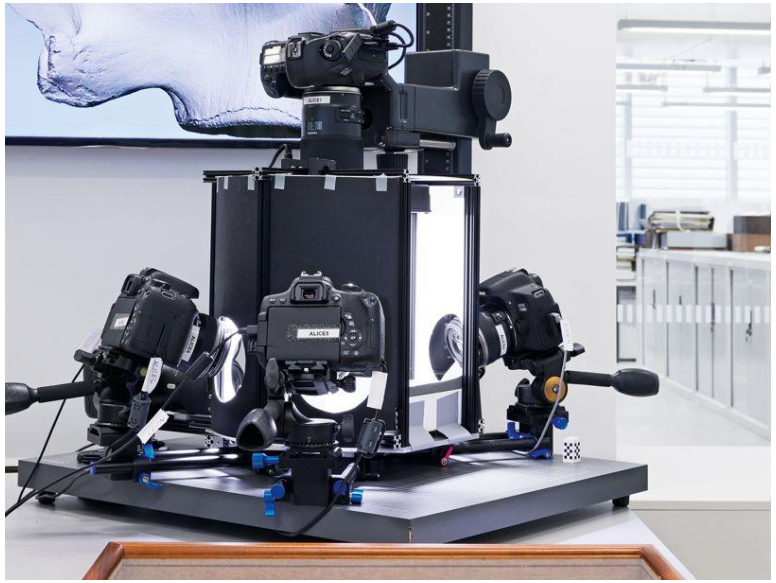
Innovations: Bridging the analogue-digital gap

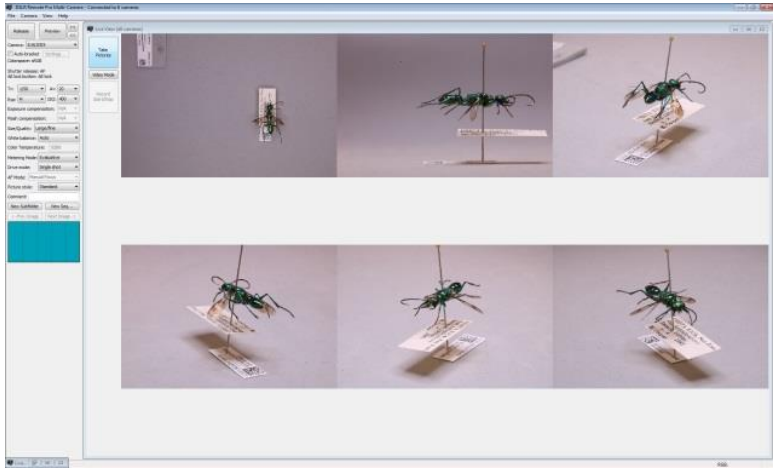
HerbIE



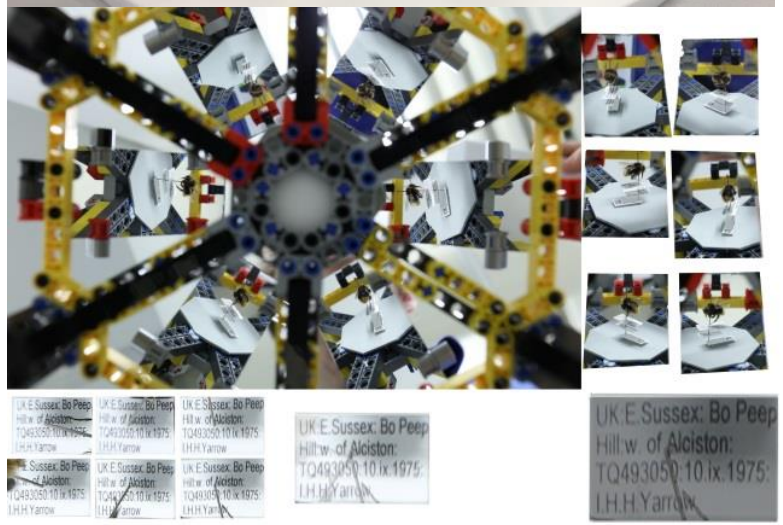


ALICE



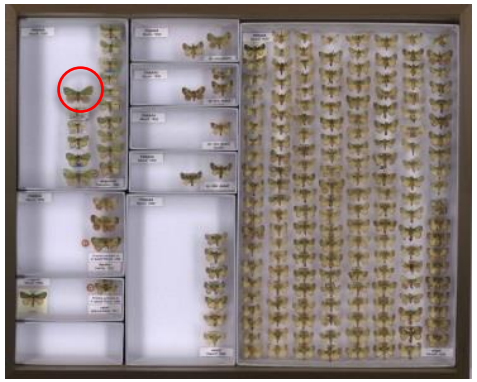


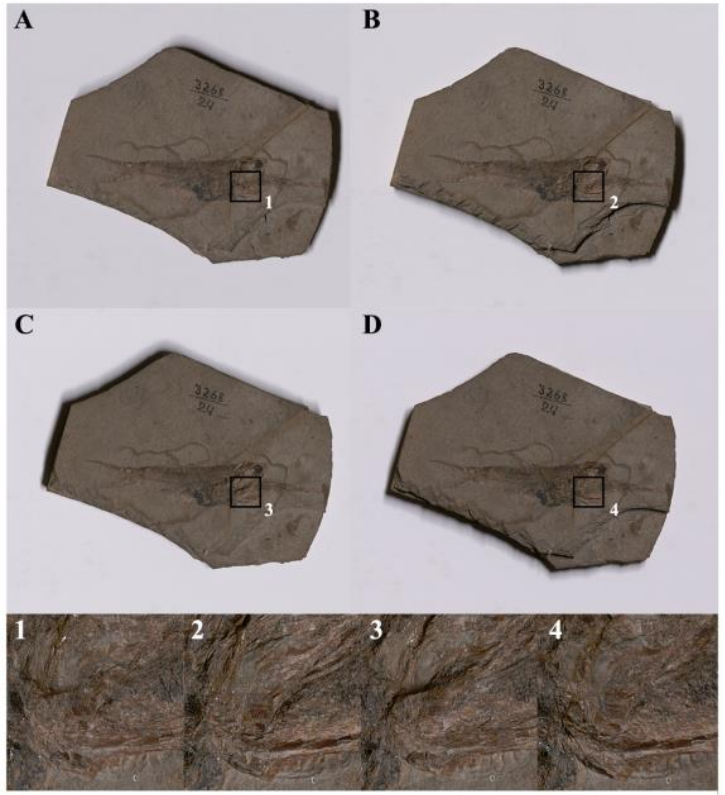
MALICE

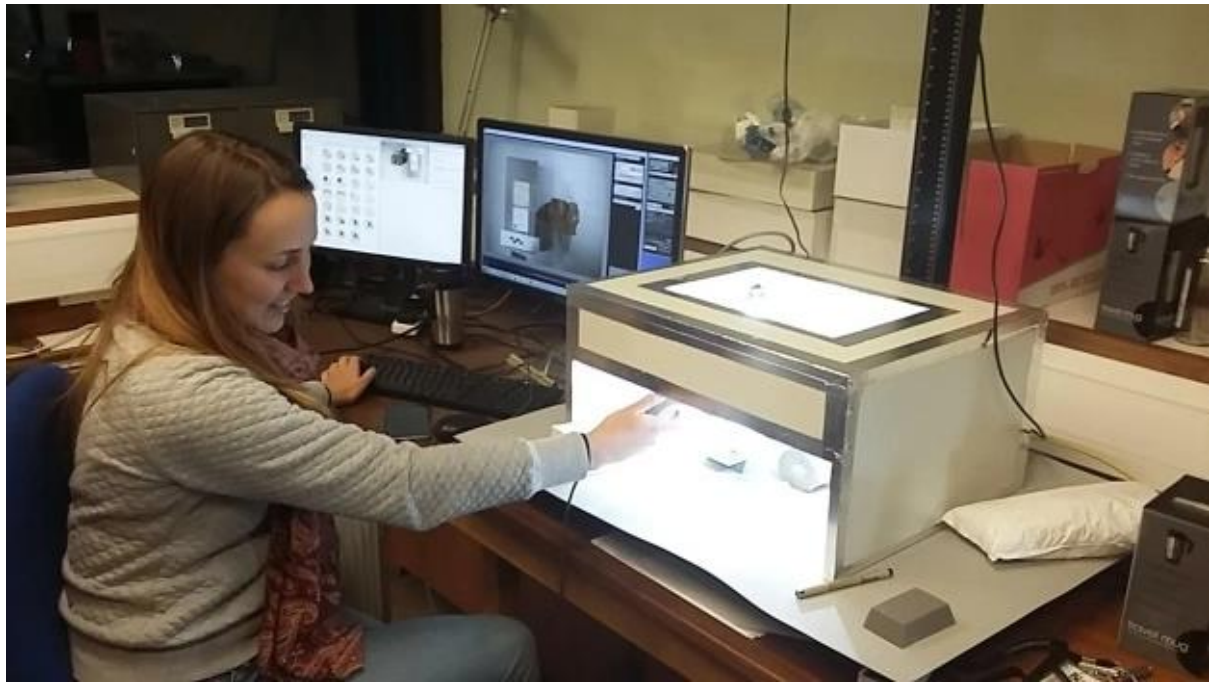


Innovations: Bridging the analogue-digital gap

Large format scanner

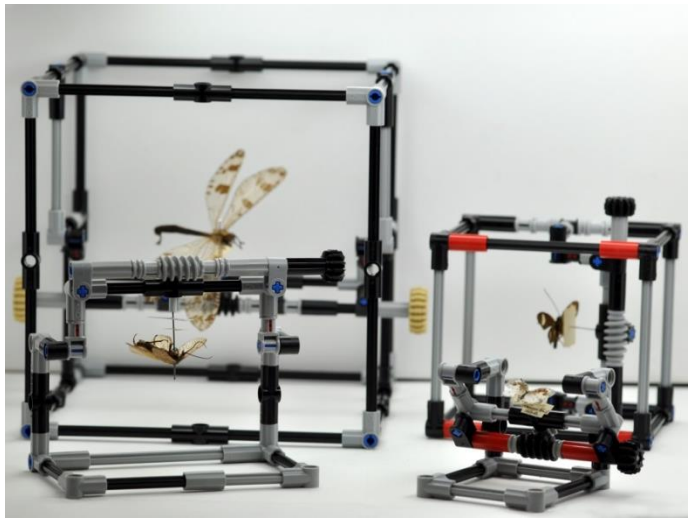


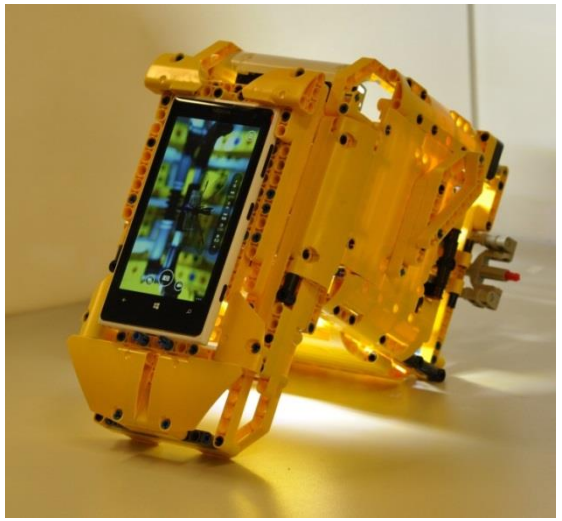


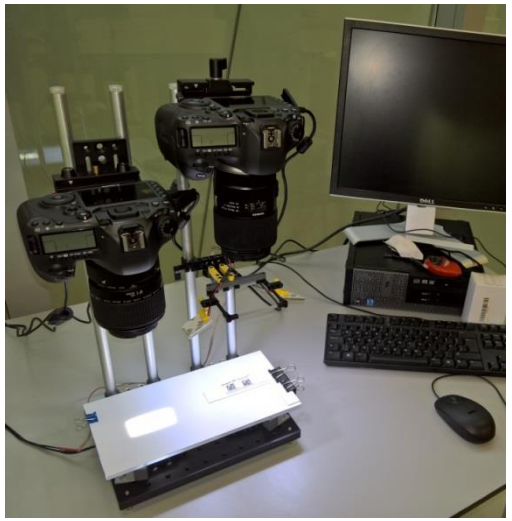


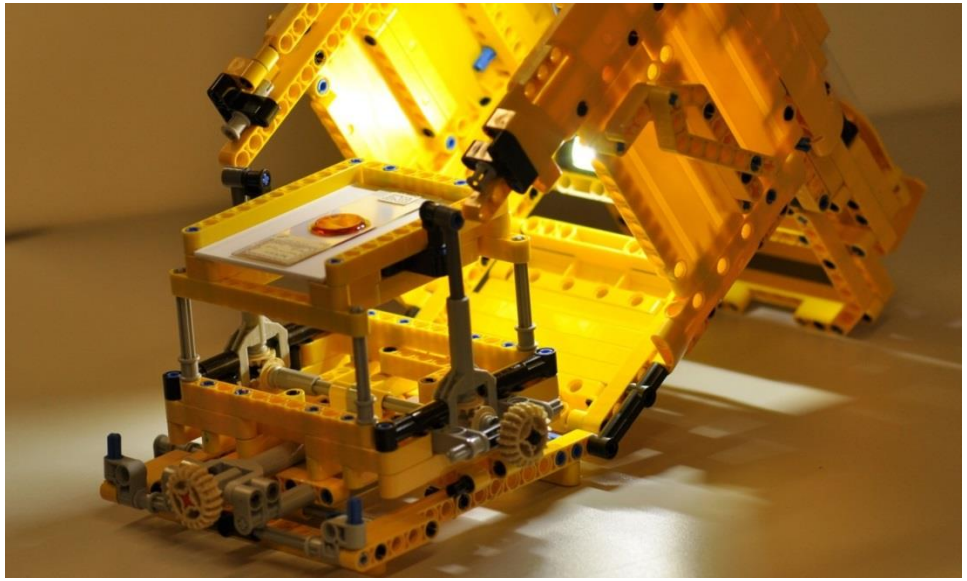
Imaging the Palaeontology collection

Innovations: LEGO and be Mobile







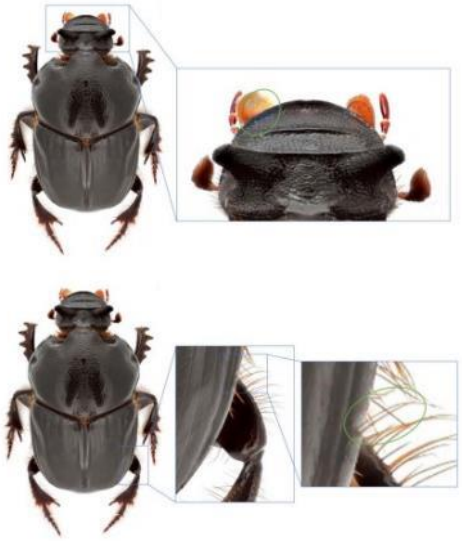


Innovations: Some data is hidden…

Enhancing insect specimen visualisations through combined focus stacking and multi-light acquisition

Christos Makris

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Master in Science
of
University College London.

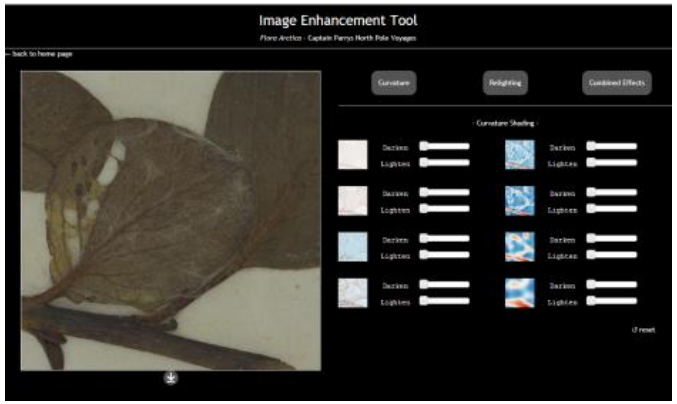


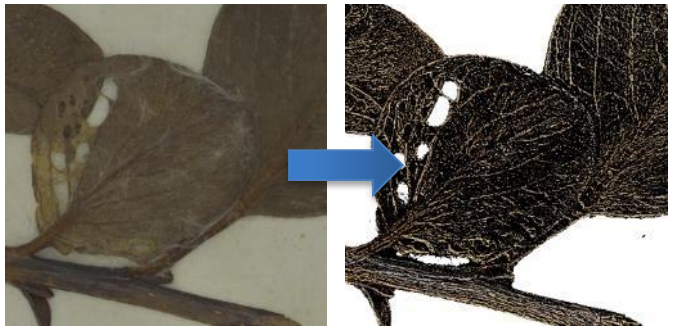
A System for Web-Based Interactive Enhancement of Multi-Light Photographs of Heritage Artifacts

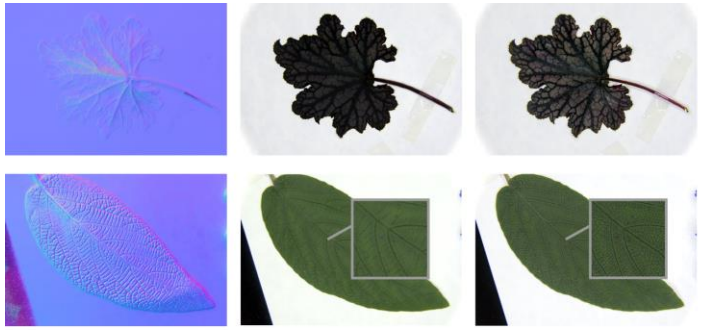
Student: Céline Dupuis

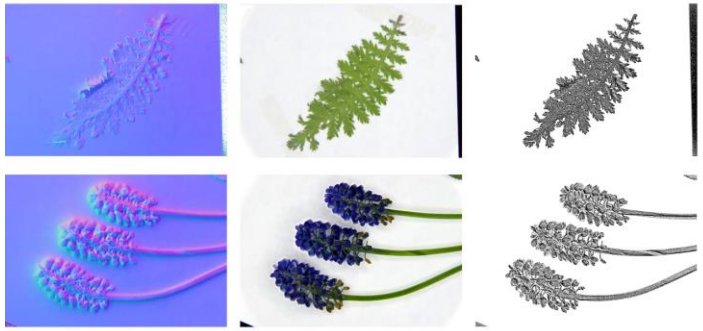
Supervisor: Prof. Tim Weyrich
External Supervisor: Dr. Steen Dupont

MSc ICT Innovation
September 2017







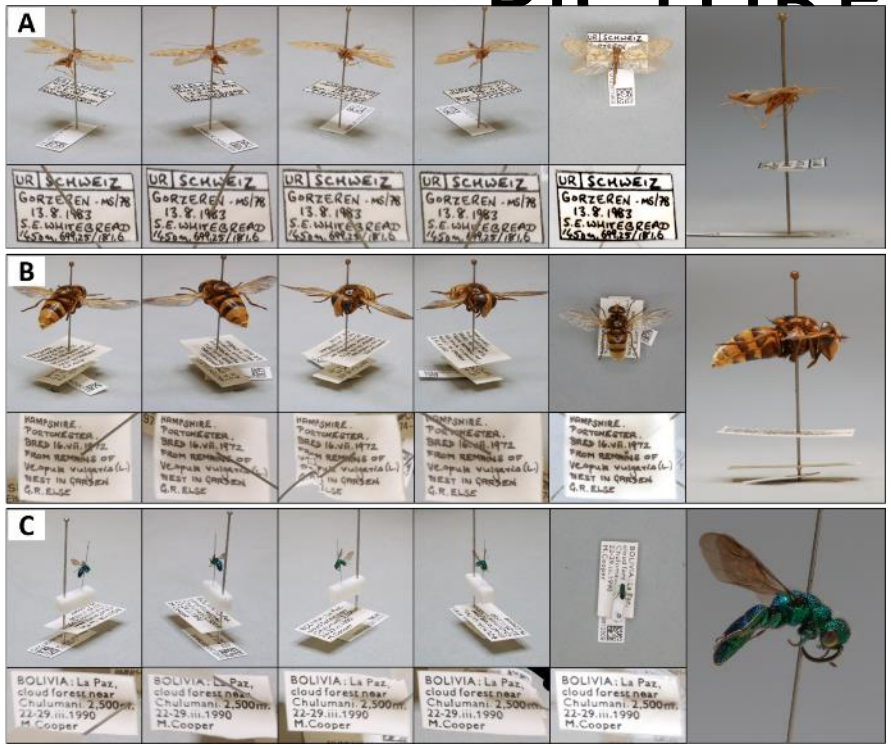


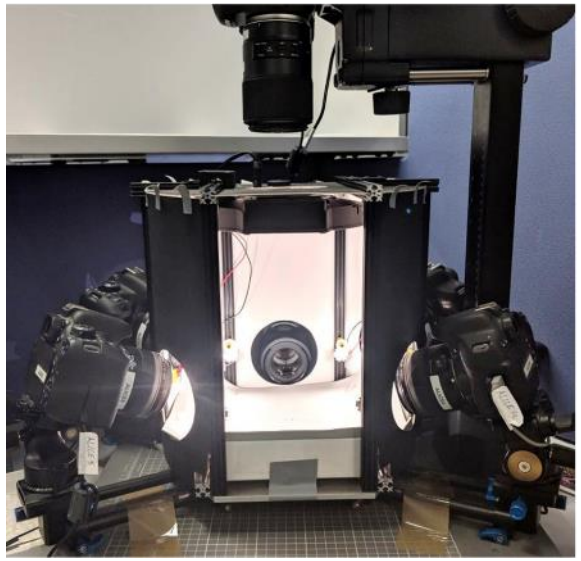
Handover

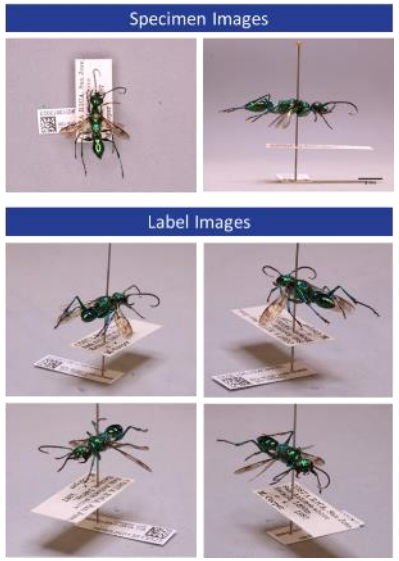
Data Challenge: LEGACY Data Needs Reconstructing

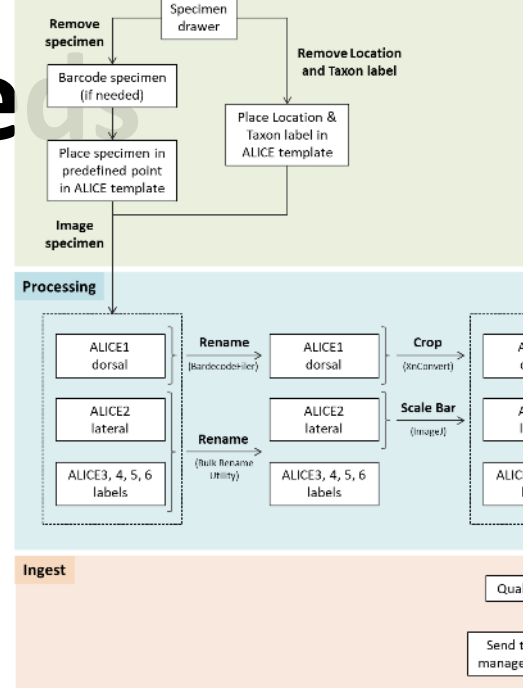
Data Challenge: Some Data Needs Reconstructing

THE ALICE CHALLENGE WITH PICTURES – 2 slides?





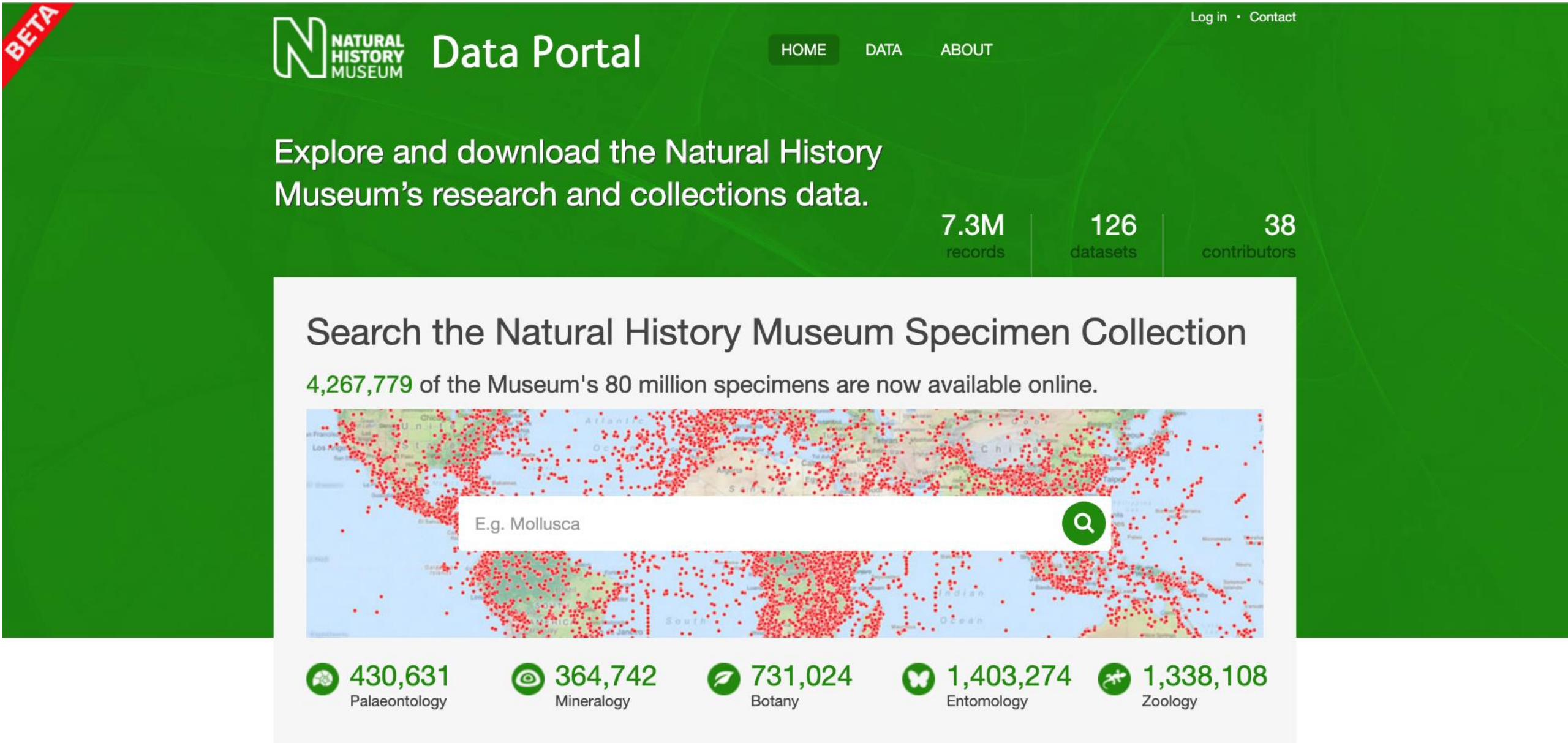




https://doi.org/10.312

Data Challenge: How do we share data?

Data Portal



https://data.nhm.ac.uk/

Back to resource | Download | Contact record curator

Normal view | Darwin Core view | GBIF view

2018.9.4.12-13

GBIF quality indicators: NO ERRORS

Classification

Scientific name:	*Salmo trutta* Linnaeus, 1758
Author:	Linnaeus, 1758
Class:	Actinopterygii
Order:	Salmoniformes
Family:	Salmonidae
Genus:	*Salmo*
Species:	*trutta*
Higher classification:	Actinopterygii; Salmoniformes; Salmonidae

Location

State province:	Wales
Country:	United Kingdom
Continent:	Europe
Water body:	Bugeilyn
Higher geography:	Europe; United Kingdom; Wales; Powys

Data Challenge: How do we measure impact?

Data Portal Stats

Data Portal

Refreshed: 31-03-2019

14/15 | 15/16 | 16/17 | 17/18 | 18/19

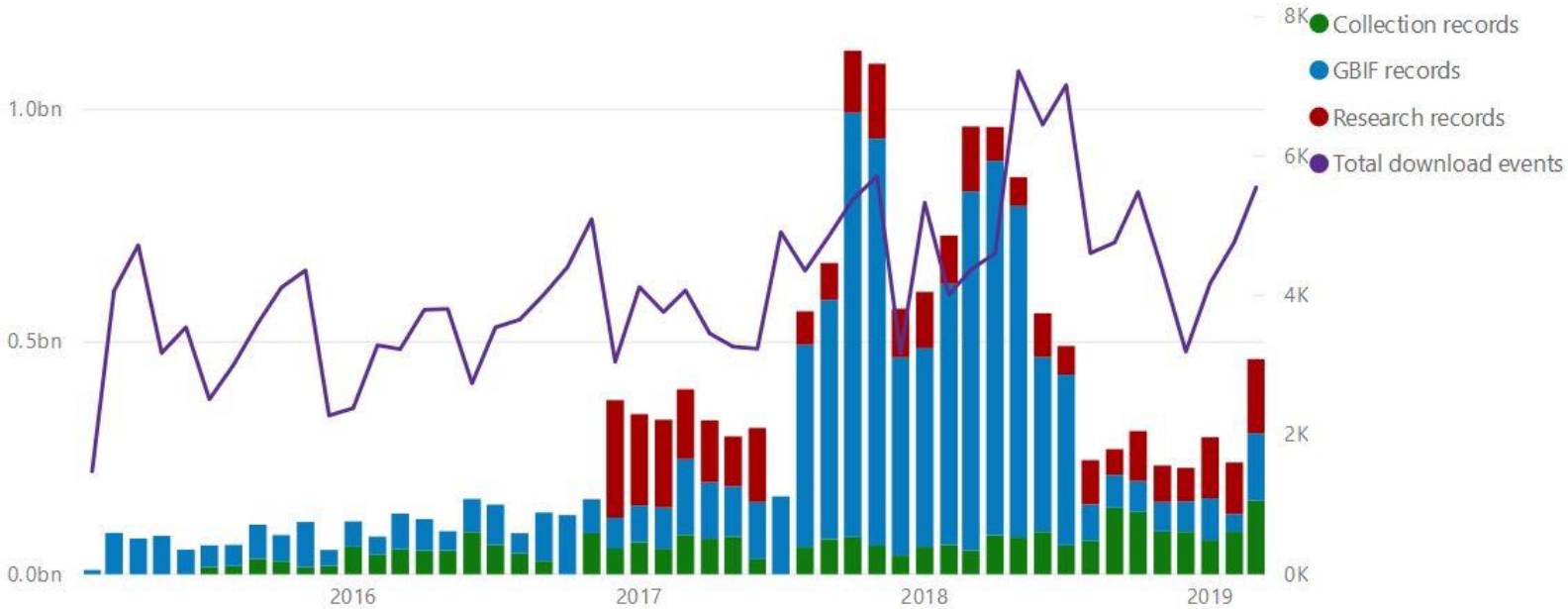
Between February 2015 and March 2019, 16.2bn of the Museum's records were downloaded during 206.02K separate download events.

Collection records made up 80% of the total.

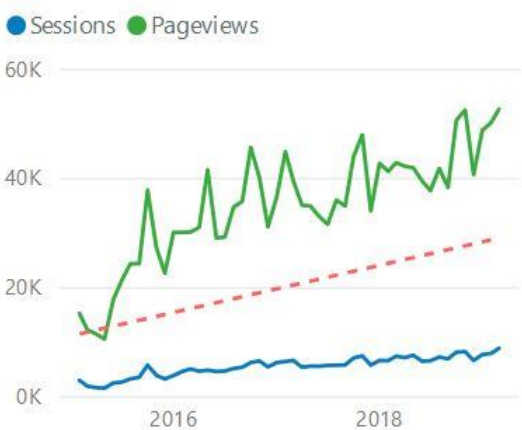
37% were downloaded from the Data Portal and 63% were via GBIF.

1.75M Data Portal pages were viewed during 281.58K sessions. 74.81% of sessions originated outside of the UK.

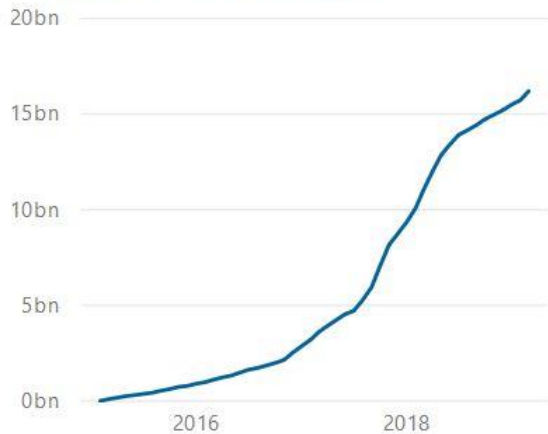
Record downloads



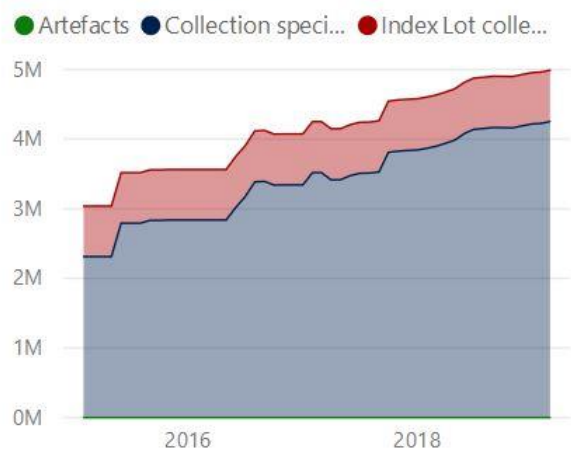
Sessions and pageviews



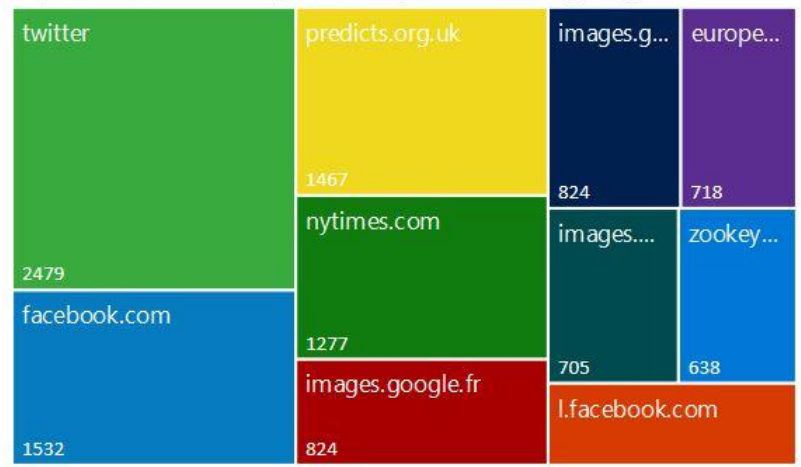
Download events (cumulative)



Count of collection records



Top referral sources by session (ex. search engines)



Data Challenge: How do we measure impact?

Research Impact (Papers)

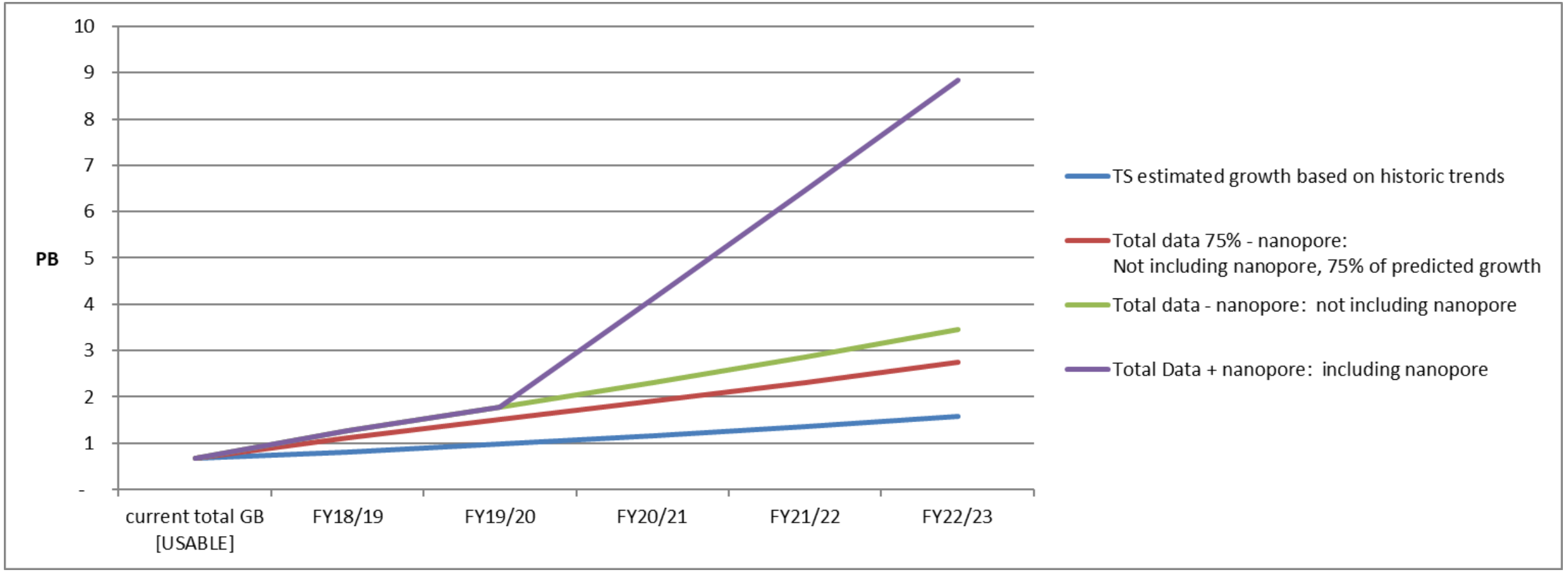
Data Challenge: How do we annotate?

Community use, Wikidata?
Control, access issues?

What other challenges do we have?

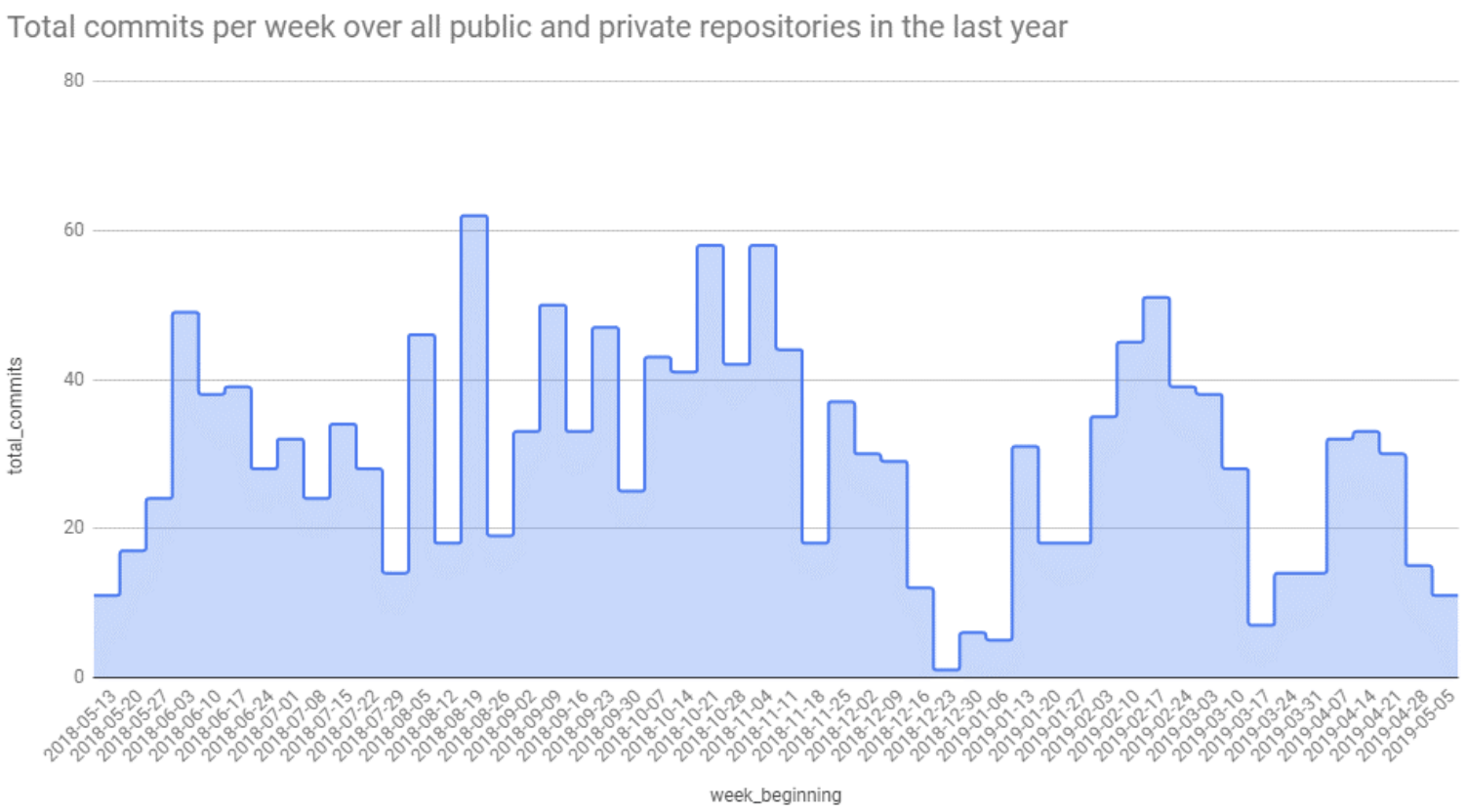
- Legacy data, legacy standards, and legacy practices
- Parts of parts of parts
- Missing data (did you know there are no comprehensive digital lists of most of the natural world – even just the names of species!?)
- Data integrity
- Scalability
- Data validation
- Sharing and enhancing

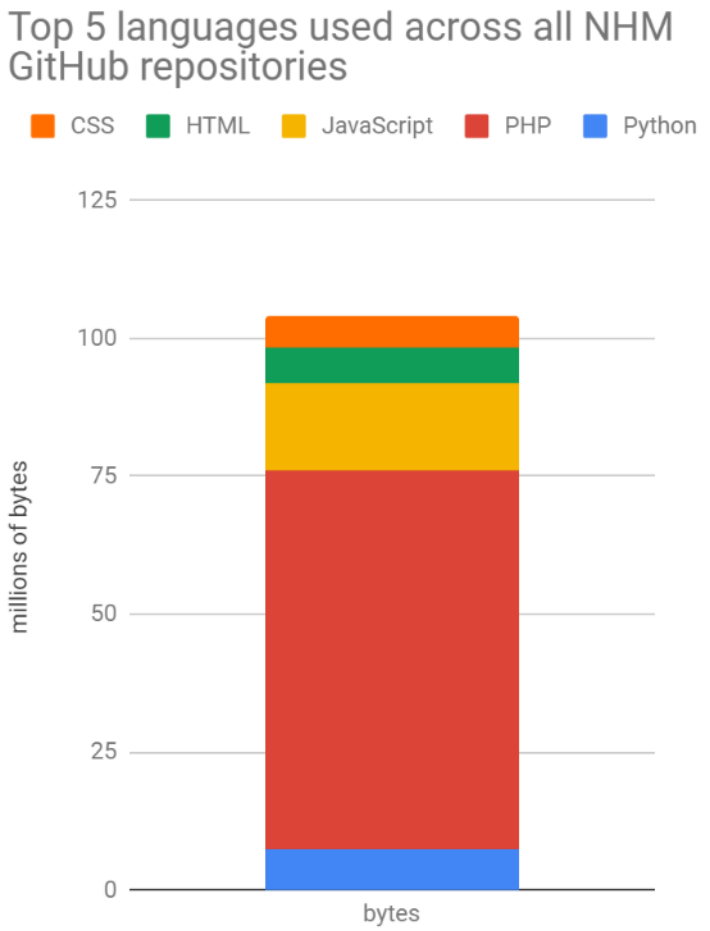
What is our data footprint?



How much do we make ourselves?

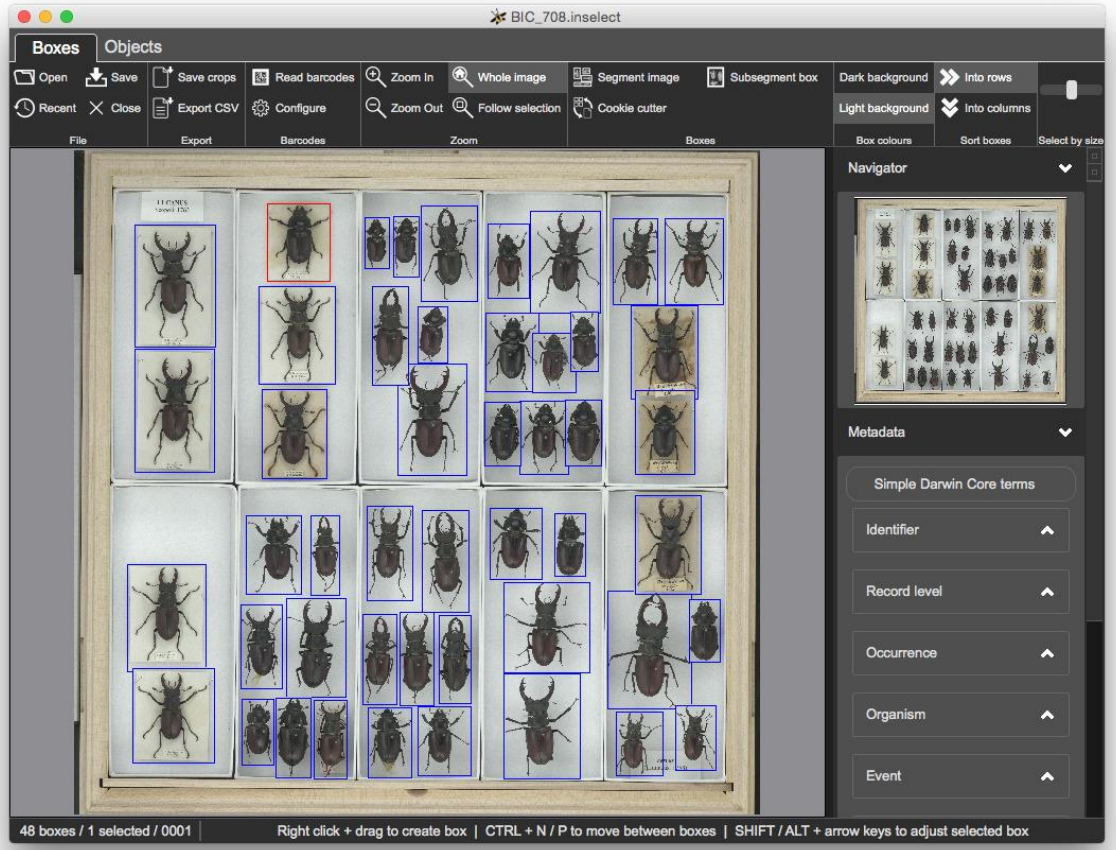
Segway into software





Inselect

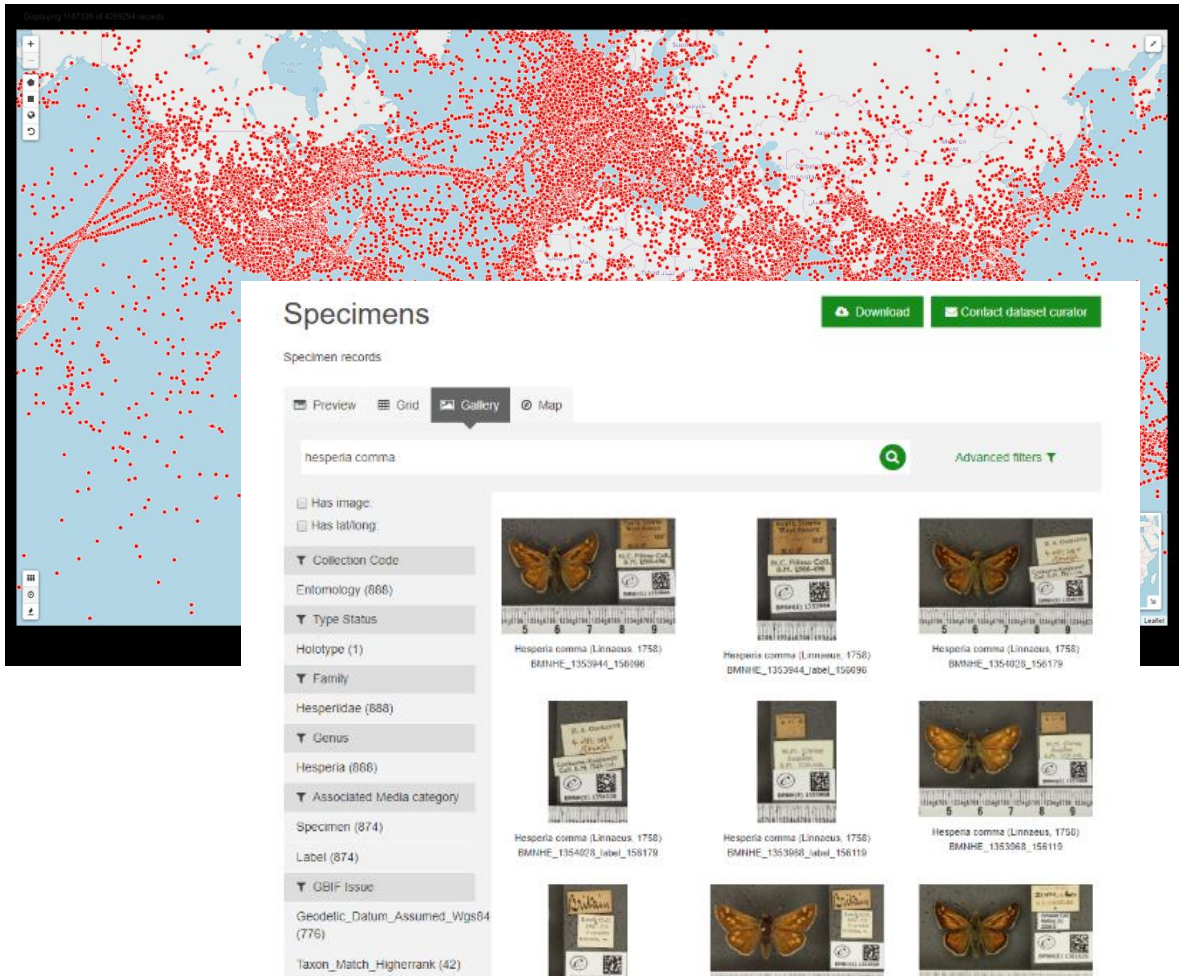
- Originally developed to assist with whole drawer imaging of pinned insects but can be used for any bulk annotation of multi-specimen images
- Automated/assisted placement of bounding boxes
- Automatic barcode reading and capture
- Crops out specimen-level images,
- Capturing metadata such as catalogue numbers, location within the collection, and possibly information on labels and
- Associating metadata with the cropped images
- Allows users to write YAML metadata templates



https://naturalhistorymuseum.github.io/inselect/

Data Portal

- Primary access point for users who wish to search and download the Museum's scientific data
- 4+ million specimens available
- 100+ datasets from 30+ contributors
- For every visitor using our physical collections, 10+ visitors download data from our digital collections
- Written in Python and is built on CKAN
- Supports RDF, rich API, plans for more LOD!



Scratchpads

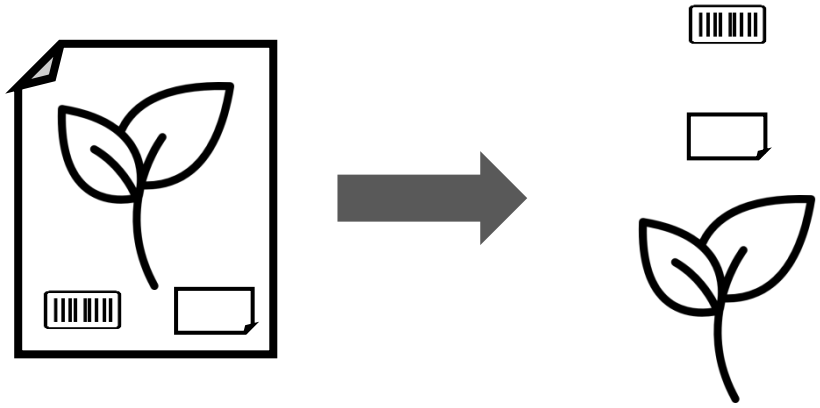
- Ask Ben for some stats? Might scrap

New project: Specimen Data Refinery

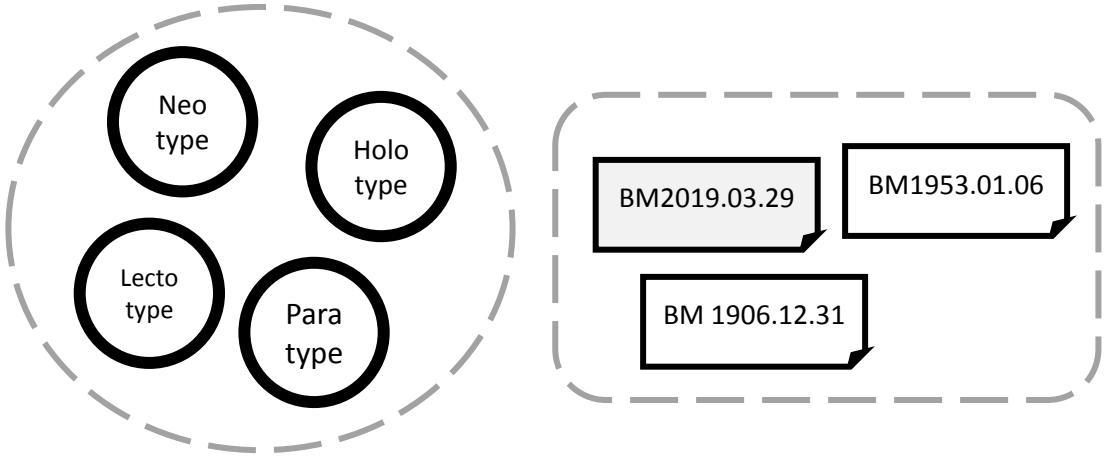
Goal:

"Develop a platform that integrates **artificial intelligence** and human-in-the-loop approaches to **extract**, **enhance** and **annotate data** from digital images and records at scale."

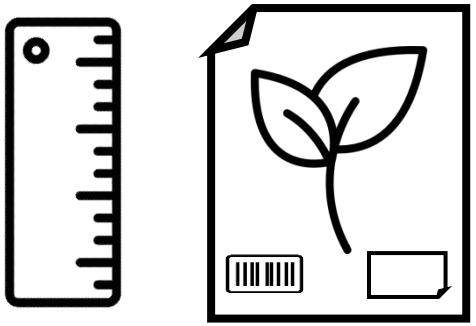
Allow curators & researchers to create and run repeatable and citable workflows resulting in datasets with rich self-descriptive metadata based on GUIDs and persistent identifiers



Segment and crop parts of images

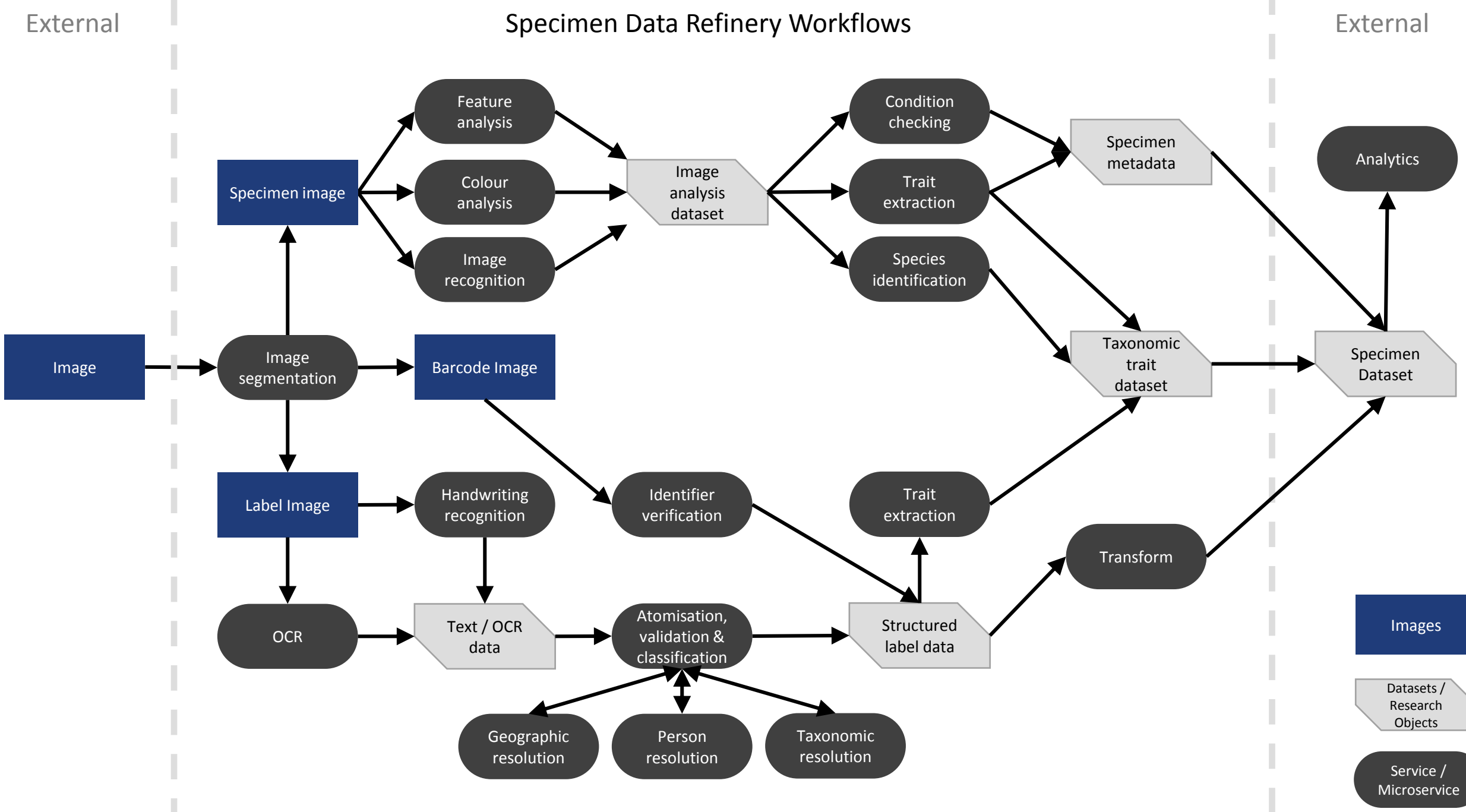


Group similar specimens and labels
(based on size, shape, colour, landmarks)



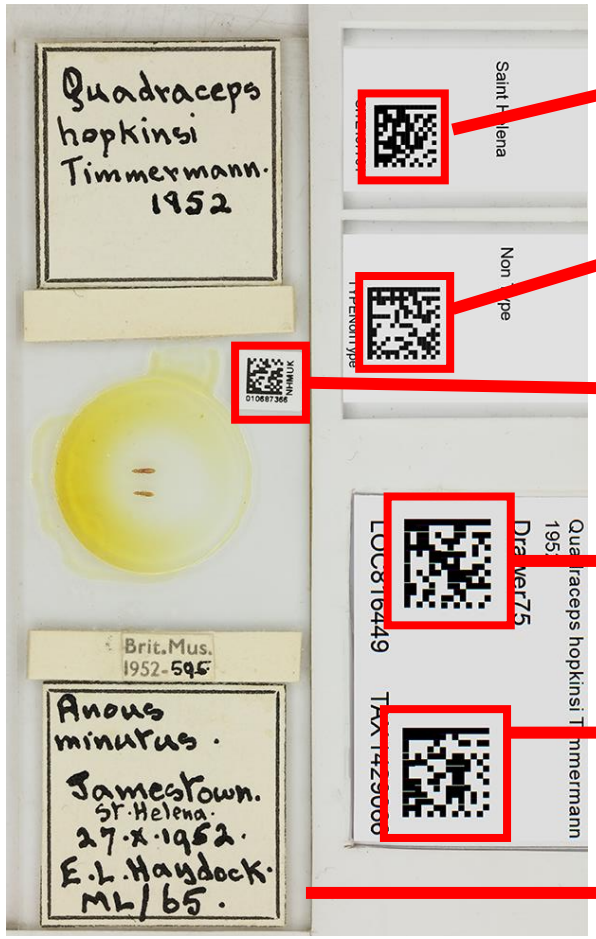
Measure specimens and labels

Georeference text



Original diagram by Matt Woodburn – Thanks!

Existing Example (jury-rigged)



Locality: SITE157761 *(Saint Helena)*

Type: TYPENonType *(Non-type)*

Specimen ID: 01687366

Storage Location: LOC816449 *(Drawer 75)*

Taxonomy: TAX1429066 *(Quadraceps hopkinski)*

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

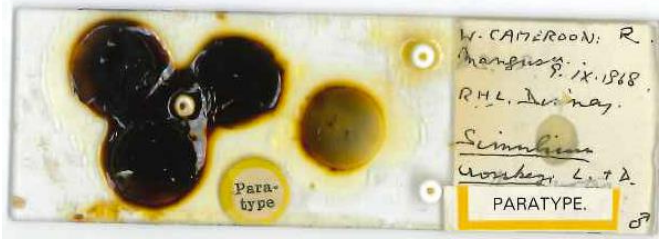
Publication in review: Allan et al (2019). A Novel Automated Mass Digitisation Workflow for Natural History Microscope Slides. Biodiversity Data Journal

Potential Applications

Easier

Harder

- Condition checking of specimens
  (e.g. gum chloral/phenol balsam discoloration, verdigris, pyrite oxidation)



Microscope slide with gum chloral discoloration

- Natural language descriptions of specimens
  (e.g. for public, curators, researchers)

This is a Matchsafe. We acquired it in 1980. It is a part of the Product Design and Decorative Arts department. Its dimensions are Overall: 6.4 cm (2 1/2 in.)

- Taxonomic trait extraction
  (e.g. phenology, morphology, biological relationships)

Slide image courtesy of Zoe Jay Adams (NHM). NL example: Matchsafe; Overall: 6.4 cm (2 1/2 in.); Gift of Stephen W. Brener and Carol B. Brener; 1980-14-911 (http://cprhw.tt/o/2CQGZ/)

Opportunities?

- Data Cleaning
- Community Data annotation
- Automation
- Robotics?

Acknowledgements

Thank you to:

Helen Hardy, Vince Smith, Ian Golding, Algirdas Pakštas, Paul Ward, Matt Woodburn, Dave Smith, Hillery Warner, James Ayre, Charlotte Barclay, Sarah vincent, Ben Price, jen Pullar, Louise Allan, Robyn Crowther, Lizzy Devenish, Phaedra Kokkini, Laurence Livermore, Krisztina Lohonya, Nicola Lowndes, Olha Shchedrina, Peter Wing, Steve Suttle and Glen Moore.

For facilitating and providing material for this talk

and thank you to all of you for listening