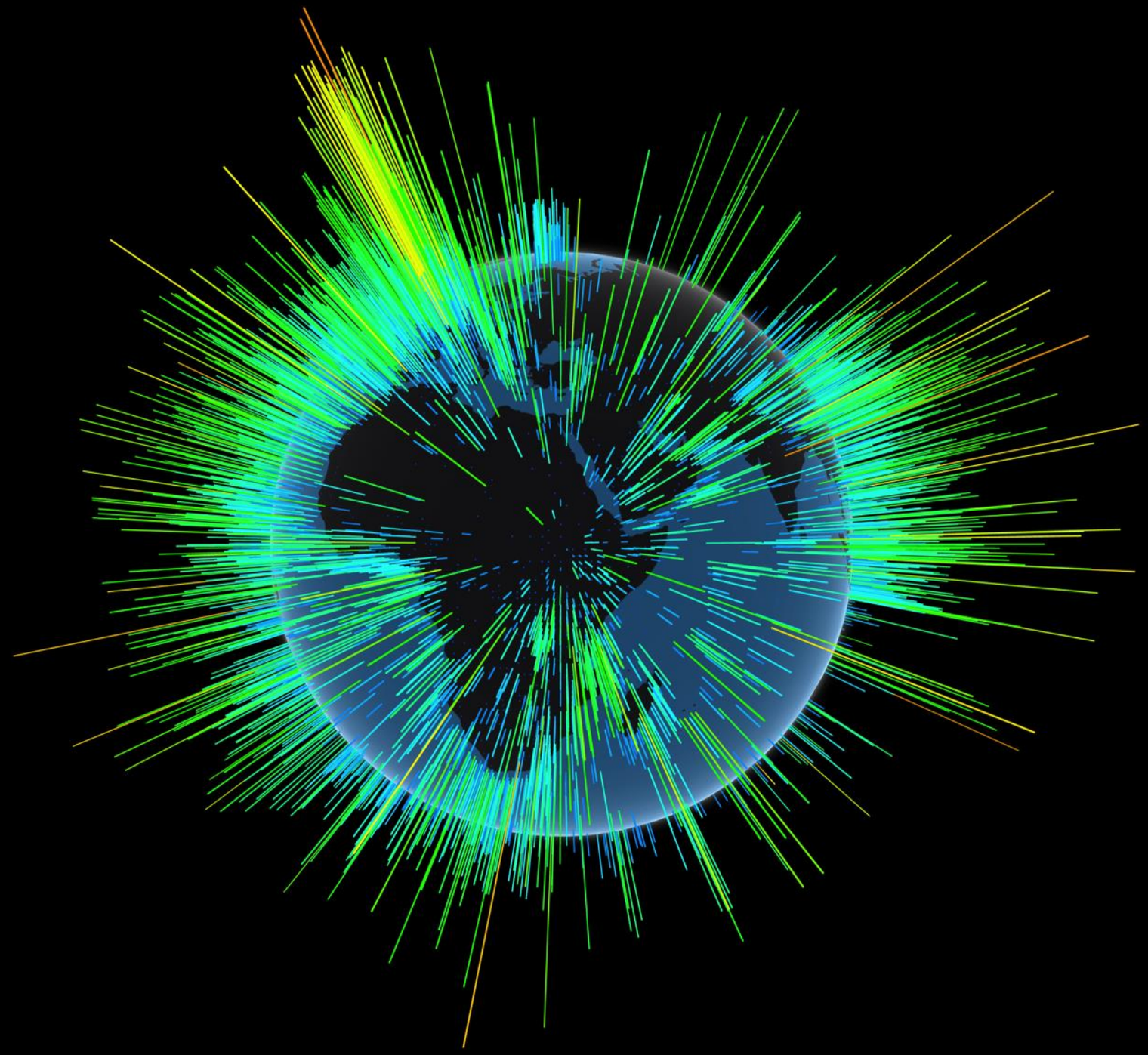
NATURAL HISTORY MUSEUM



The future of biodiversity informatics

Vince Smith, Natural History Museum, London

Empowering Biodiversity Research, Brussels, Belgium, May 21st 2015

Overview

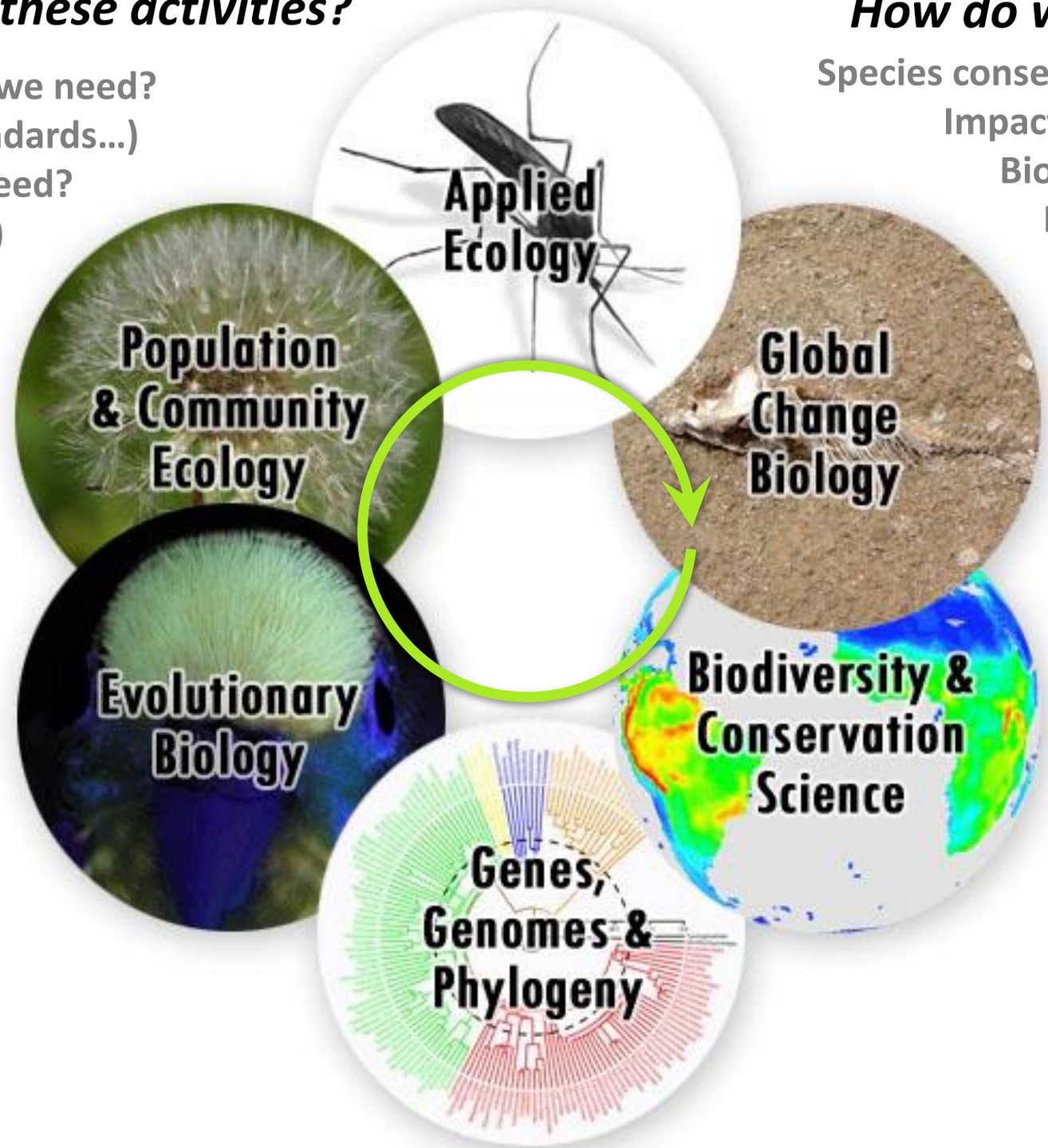
1. **Background – the biodiversity informatics domain**
   - The problem, (i.e. why are we here)
   - Toward an integrated view (strategy)
2. **Data mobilisation**
   - Mass digitisation
   - Crowdsourcing & digital volunteers
3. **Data synthesis**
   - Data aggregation & visualisation
   - Modelling
4. **Enabling technologies**
   - Computer vision
   - Remote sensing / field activities
5. **Adapting to the future**
   - Lessons learned (agility, flexibility, pace of change, risks)
   - Sustainability

1. Background

The problem – integrating biodiversity research

How to we join up these activities?

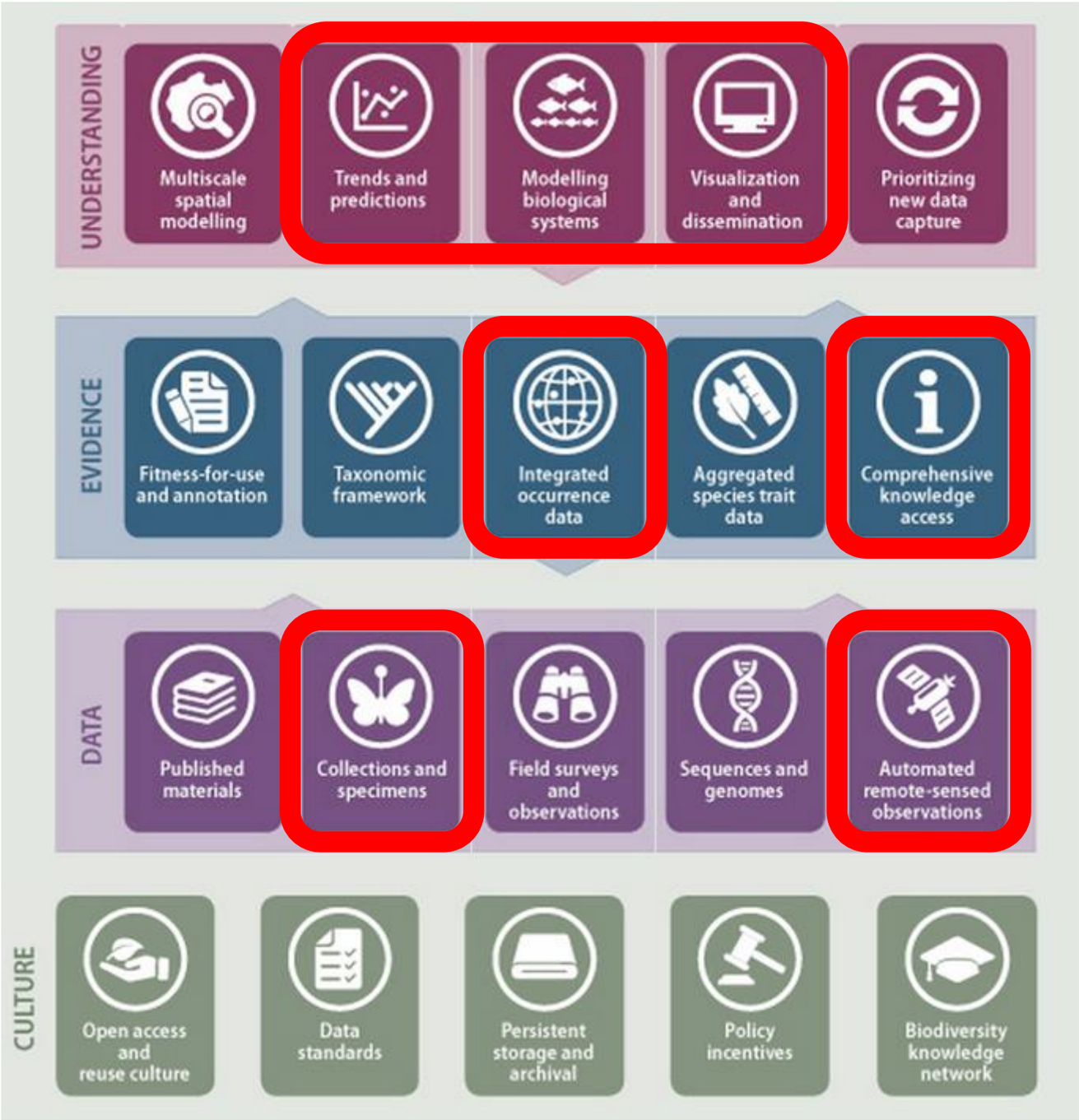
What infrastructures do we need?
(technologies, tools, standards…)
What processes do we need?
(Modelling, workflows…)
What data do we need?
(Genes, localities…)

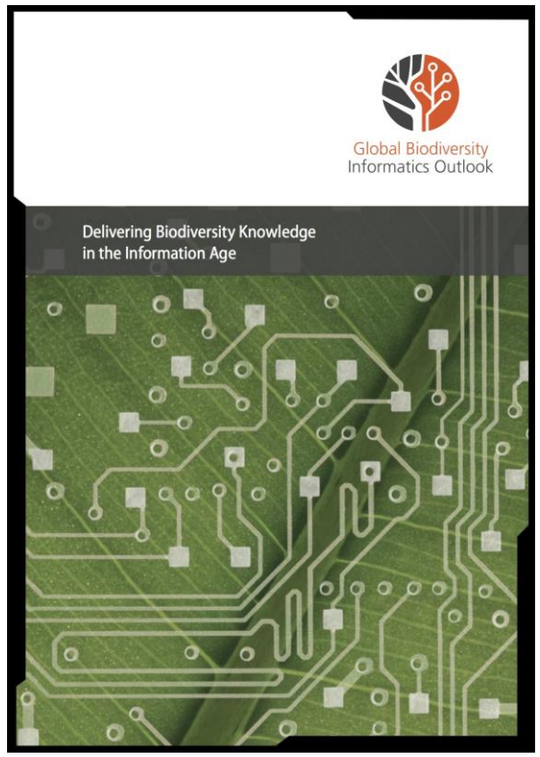


How do we use this as a tool?

Species conservation & protected areas
Impacts of human development
Biodiversity & human health
Impacts of climate change
Food, farming & biofuels
Invasive alien species

A strategic view of biodiversity informatics



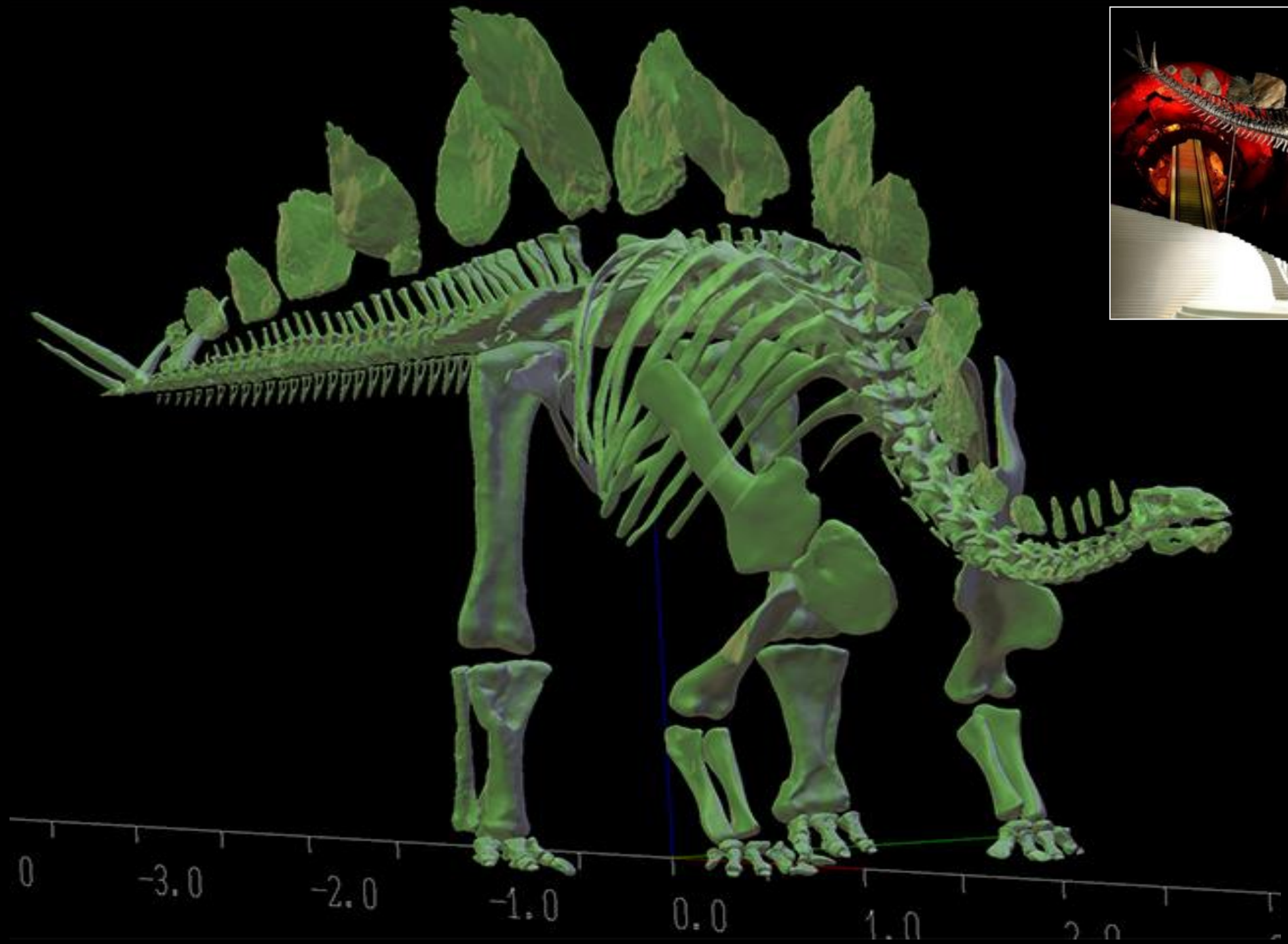


Global Biodiversity Information Outlook, 2013

"identifies the priority questions for biodiversity research, the tools needed to answer them and the steps to create those tools and deploy them."

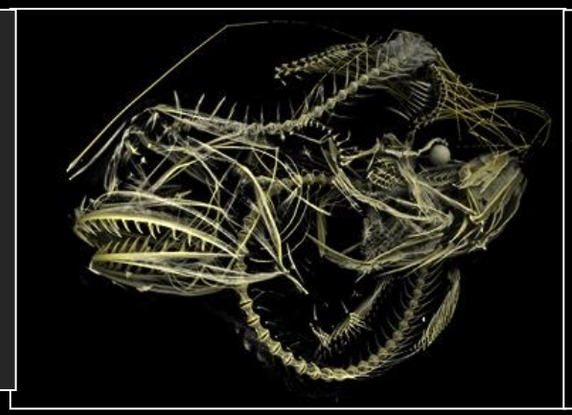
2. Data mobilisation

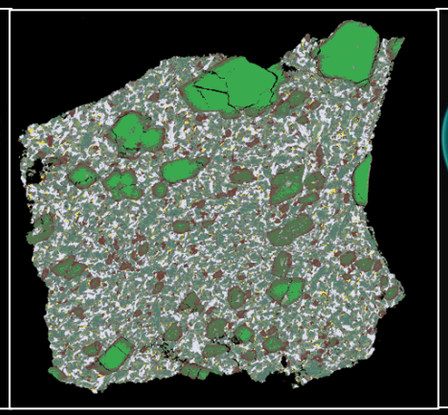
- Mass digitisation
- Crowdsourcing & volunteers

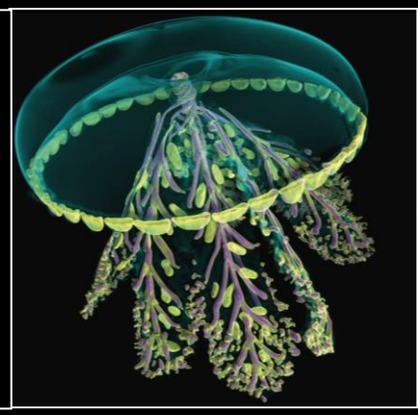


Digital surrogates of specimens:

- **3D models and printing**
- **CT imaging of internal structures**
- **Atomic, physical and chemical analysis**







An inordinate fondness for beetles:
400,000 species described so far, 100,000 types among the NHM's 10M beetle specimens

NHM collection

Collection area	No of objects	No of type specimens	Physical register	Digital data
Palaeontology	6,919,207	43,146	2,364,232	340,636
Mineralogy	423,563	615	425,000	402,727
Botany	5,863,000	172,750	127,200	645,222
Entomology	33,753,257	612,796	57,197	255,000
Zoology	27,501,350	325,000	1,986,000	1,160,216
Library & archives	5,460,000	-	-	-
TOTAL	**79,920,377**	**1,154,307**	**4,959,629**	**2,803,801**



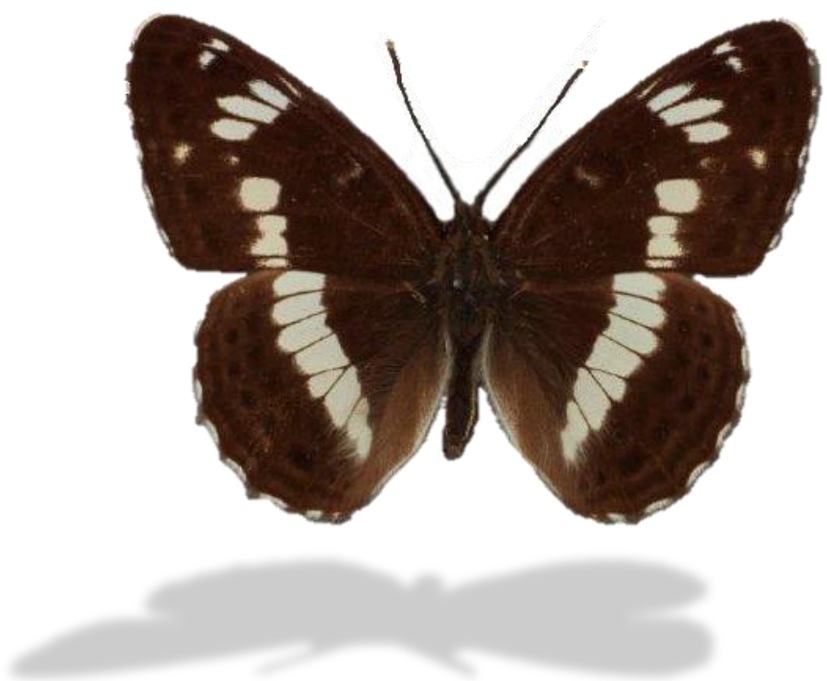

<3% of NHM specimens are digitised, & even fewer are 'computable'

NATURAL HISTORY MUSEUM

iCollections: 2013-2015

- A pilot for the Digital Collections Programme
- iCollections digitisation criteria:
  - Entire collection
  - No existing digitisation pipeline
    (pinned, slide & herbarium specimens)
  - High research potential
    (phenology, morphometrics, migration patterns, pest associations, automated species recognition)
  - Curation opportunity

- Data outputs
  - Image(s)
  - Metadata (what, when & where)
  - Georeference point localities



NATURAL HISTORY MUSEUM

NHM iCollections project:

- UK butterflies & moths
- 500k specimens
- 2 mins per specimen
- £1 per specimen

Large-scale digitisation:

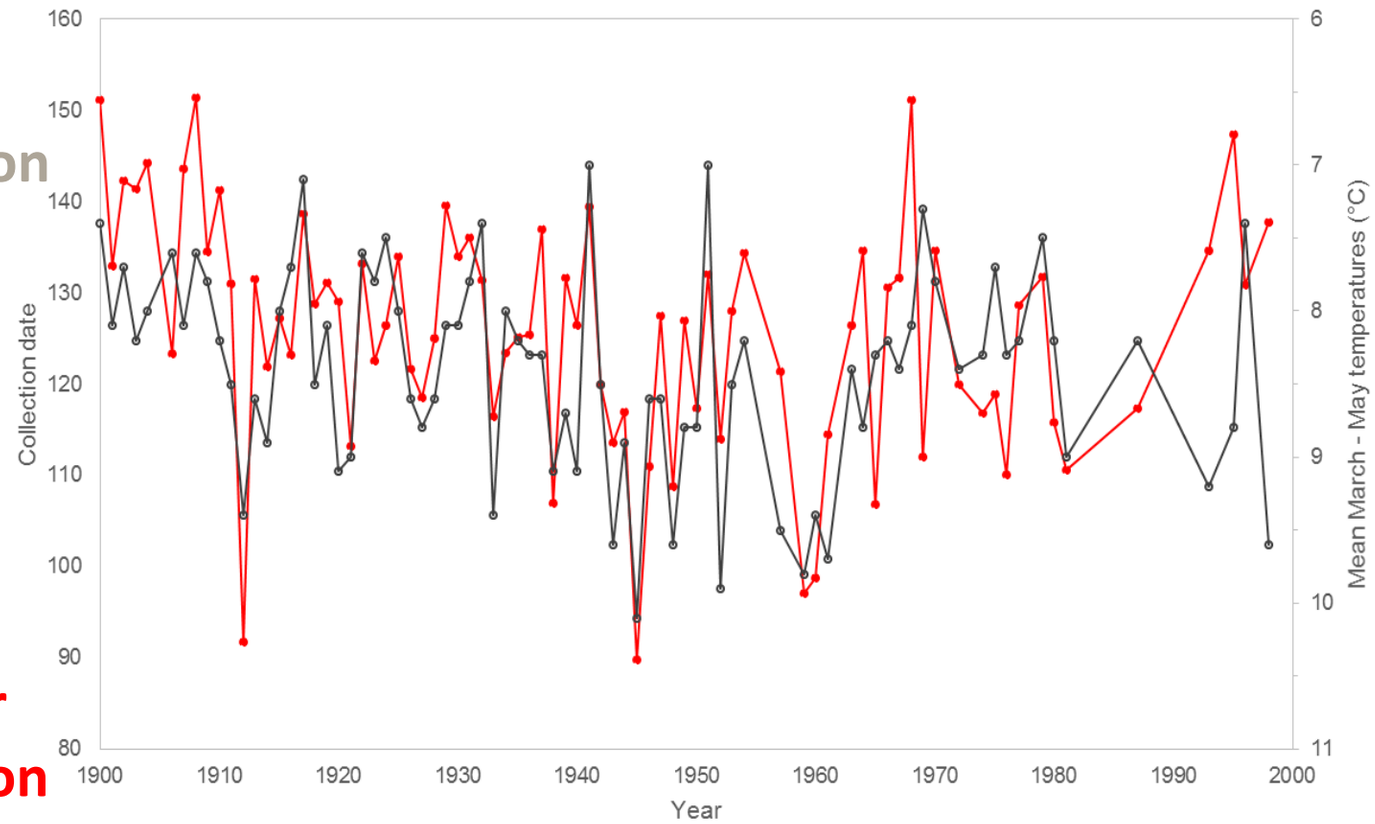
- High-throughput digitisation workflows
- Informatic pipelines
- Computer-assisted object recognition

iCollections digitisation process

iCollections research: long-term trends in phenology

Later collection

Earlier collection



Cooler springs

Warmer springs

Orange tip butterfly
(*Anthocharis cardamines*)

- **Many species emerging earlier and earlier each year**
- **Initial collection date & temperature highly correlated**
- **A unique marker on phenological response before recent climate change**
- **Longer time perspective than most observational records (BMS post-1976)**
- **Museum data available for rare or hard to record species**

NATURAL HISTORY MUSEUM

Brooks, Self, Toloni & Sparks, 2014, *Int. J. Biometeorol.*
DOI 10.1007/s00484-013-0780-6

NHM Digital Collections Programme

"To collate, organise and make available to global scientific and public audiences one of the world's most important natural history collections, delivering:

- an online specimen- / lot-level data base for all holding

- core meta-data and / or images for key parts of the collection, and

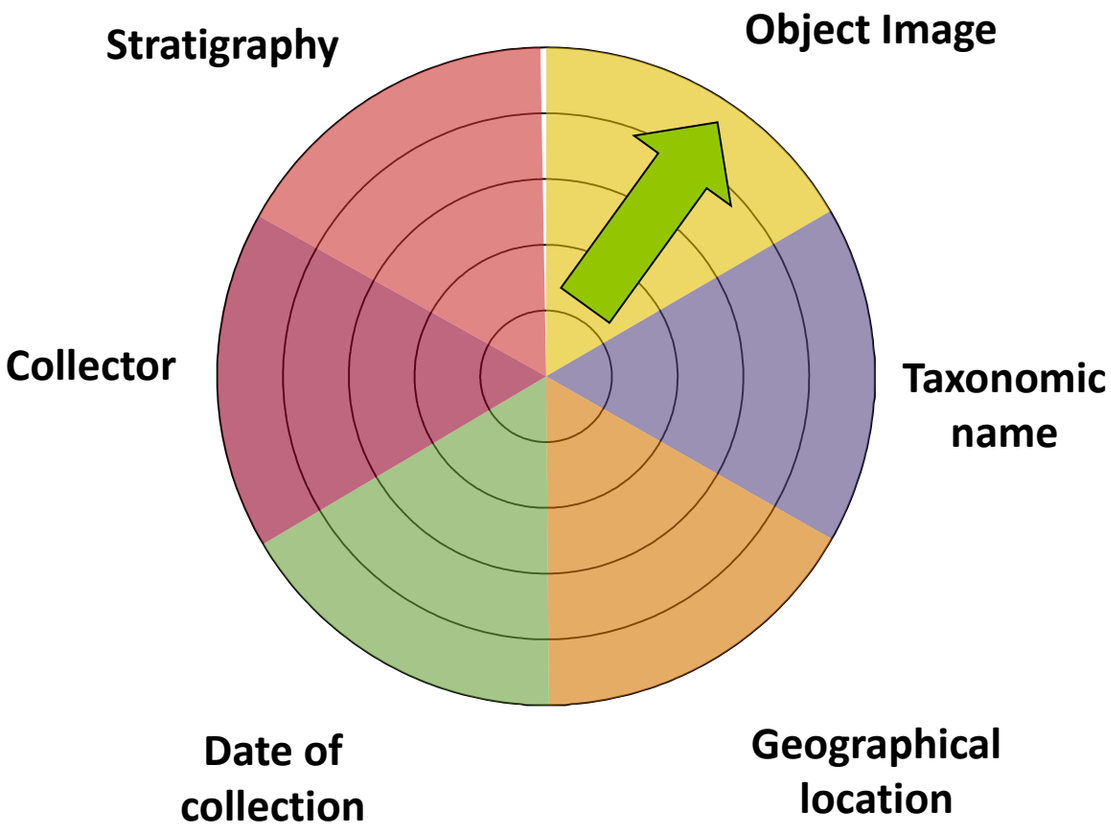
- flexible informatics and visualisation tools"

Target = 20 million specimens digitised in 5 years

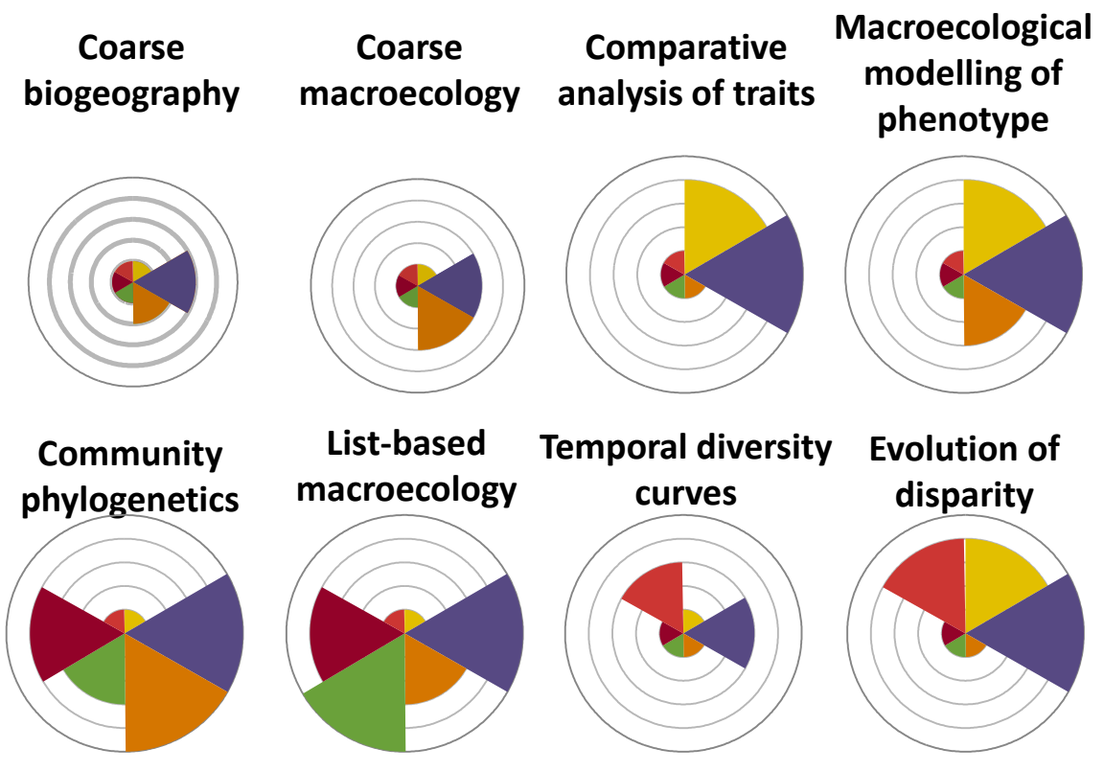
	2 year	5 year	10 year
POLICY & PROTOCOL	Defined data policy and standards	Policies embedded in NHM operating practises	Leaders of process in the digital curatorial world
DATA CAPTURE	Prioritised digitisation Workflows piloted	Portfolio of mass digitisation output projects	Some major collections digitised
PEOPLE & SKILLS	Task force formed and operating	Best-practice processes integrated into training	Digital curation as a core part of our practice
INFRASTRUCTURE	Refined collections database, tools & hardware	Future collections database implemented	Broad connections to other large digital collections
STAKEHOLDERS & GOVERNANCE	Key user communities engaged	Peer to peer development	Proactive engagement of emerging audiences
PARTNERSHIPS	Partners involved in pilot projects	Fully funded digitisation portfolio	Major international coalitions
RESEARCH	Research-orientated projects & initiatives	Collaborative research material published	Major contributions to grand challenges
ACCESS	Live NHM Data Portal	Tools , visualisations & analytics	Integrated global network of users

Question-oriented prioritisation

Core meta-data categories:



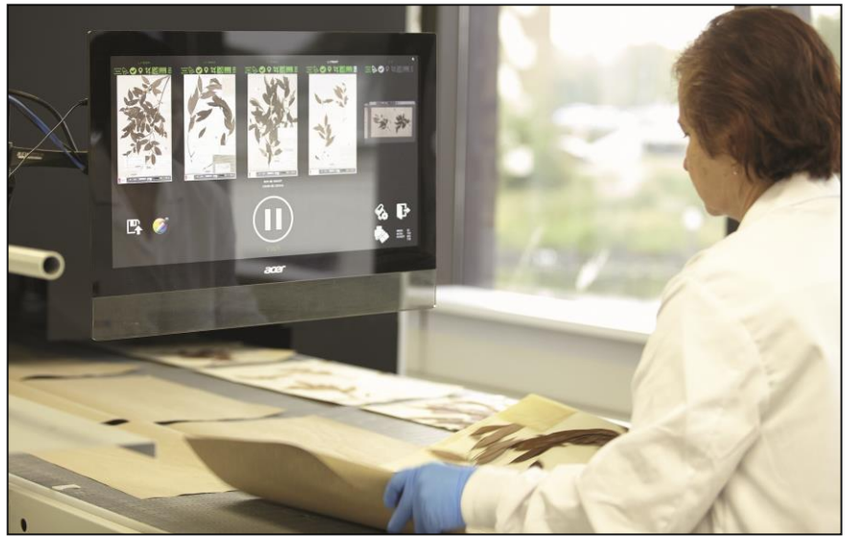
What meta-data are required for different types of study?



Digital Collections Programme: pilot projects

Pilot 1 – Herbarium sheets

- Trial partnership with Kew & Picturae (cf Naturalis)
- Conveyor-based imaging equipment
- Outsourcing of label transcription
- C.70,000 sheets
- January-April 2015

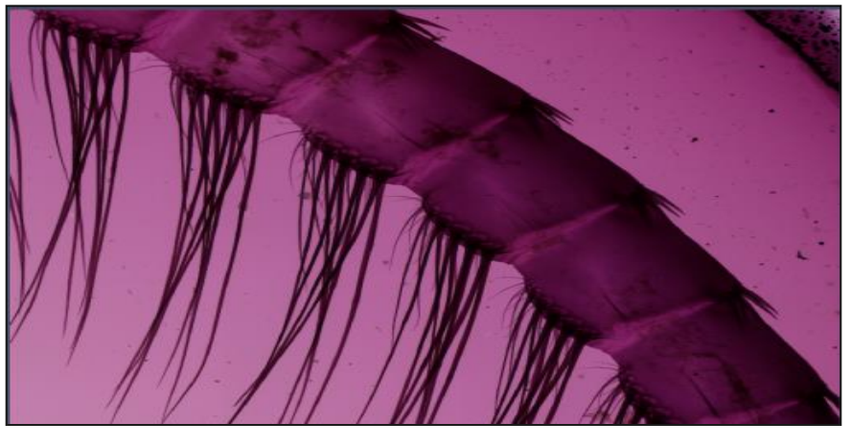


Pilot 2 – Palaeontology

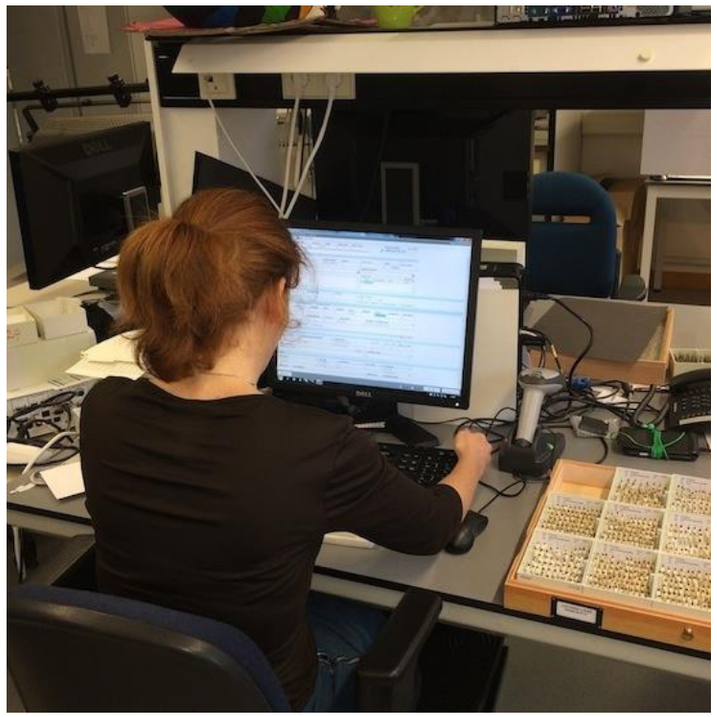
- British Fossils (Mesozoic vertebrates)
- Standardised photography
- Capturing taxonomic & stratigraphic metadata
- February 2015 – March 2016

Pilot 3 – Microscopic slides

- High-throughput slide scanner
- Satscan to capture label information
- Call for pilot project ideas
- April 2015 onwards



Many organisations are now digitising at scale

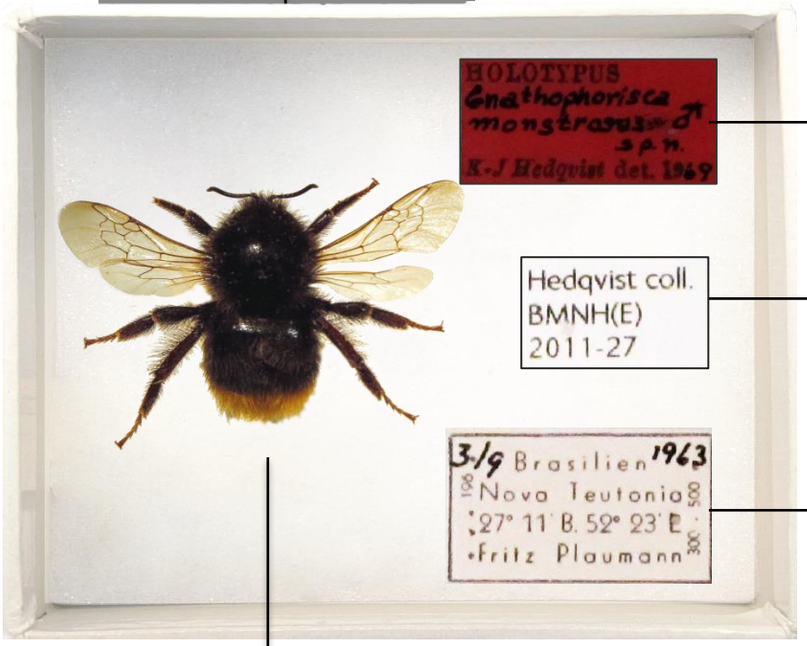


e.g Naturalis: 33k Specimens per day, 3 shifts (6am-10pm), herbarium complete in 1.5 years
€1.29 Euros per specimen image (if outsourced), transcription at similar cost

Digital Collections Programme: informatics challenge areas

Digital exhibitions
(Virtual tours)

Automating mass digitisation
(e.g. Wellcome Digital Library)



Transcription
(e.g. Zooniverse)

OCR
(e.g. reCAPTCHA)

Georeferencing
(Online mapping tools)

Computer vision / image recognition
(e.g. Image Search)

Data mining & linkage
(e.g. Biodiversity Heritage Library)

Data Cleaning
(e.g. OpenRefine)

Data Visualisation
(e.g. CartoDB)

Big data research
(e.g. PREDICTS)

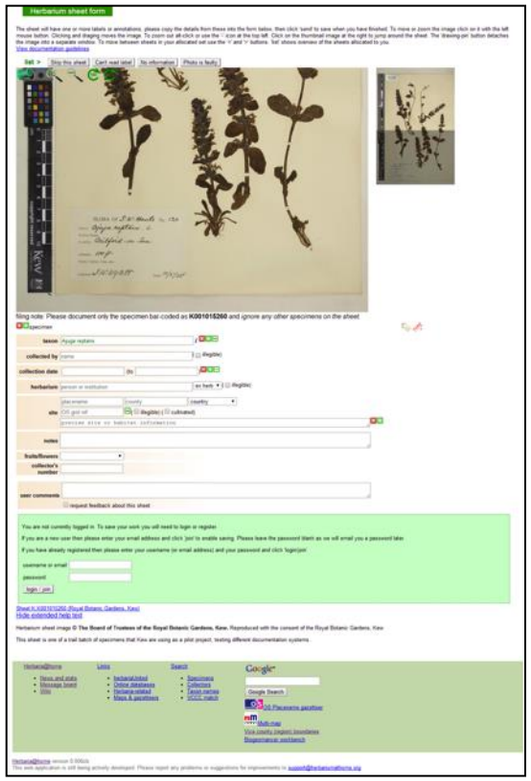
Search
(e.g. WikiData)

NHM in an App
(cloud computing & storage)

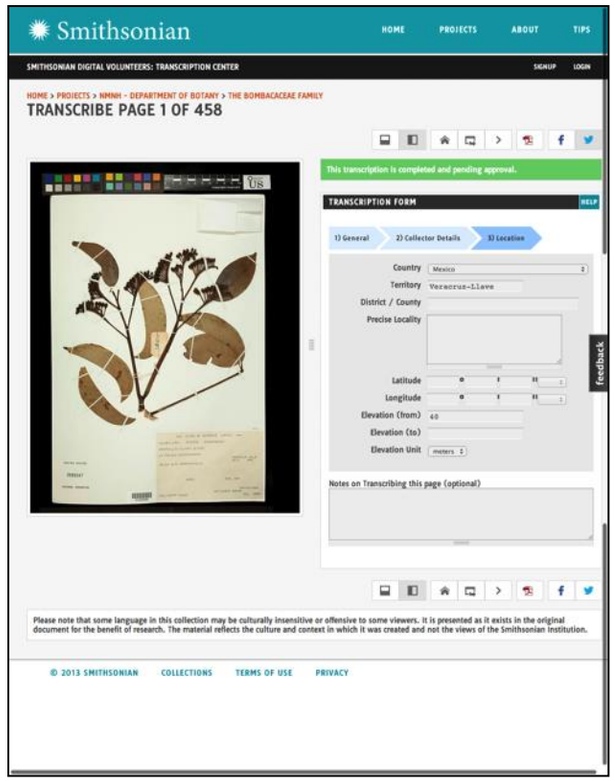
Crowdsourcing

NATURAL HISTORY MUSEUM

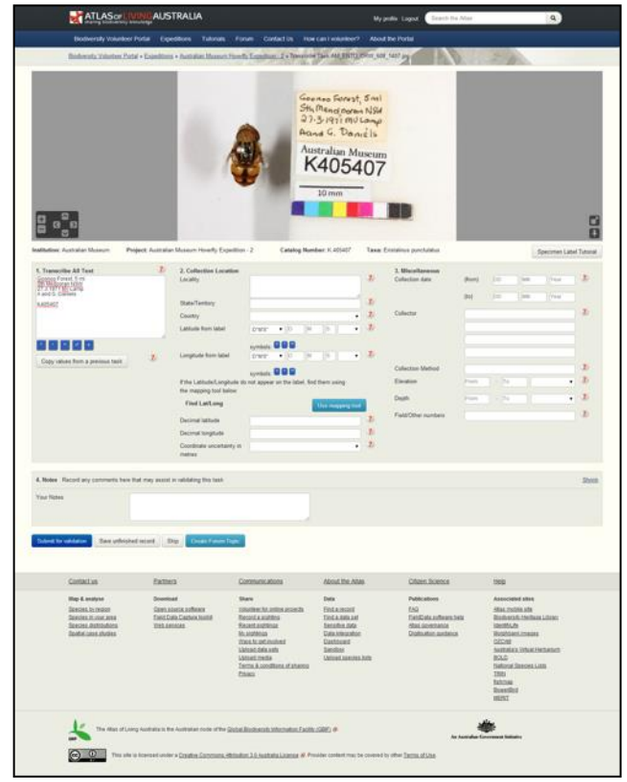
Crowdsourcing platforms



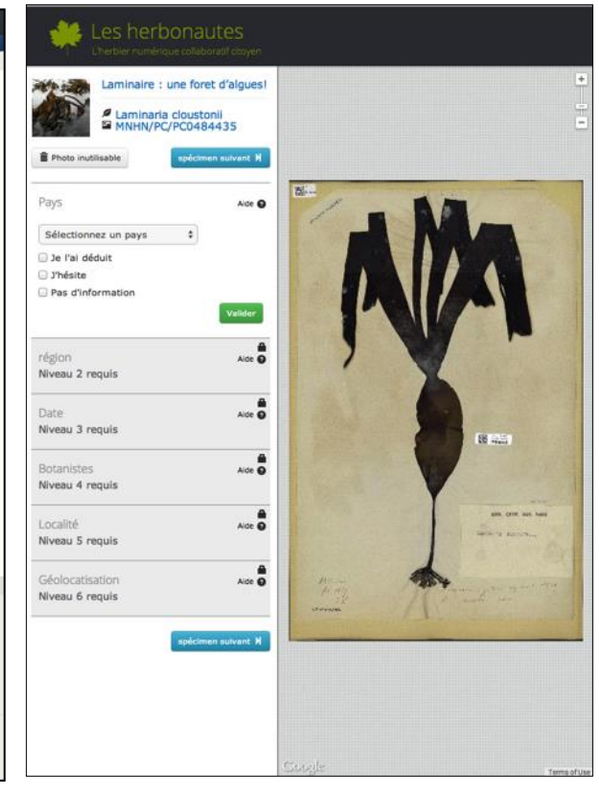
Herbarium @Home
http://herbariaunited.org/atHome/



Smithsonian Transcription Center
https://transcription.si.edu/



Atlas of Living Australia
http://volunteer.ala.org.au/



Les Herbonautes
http://lesherbonautes.mnhn.fr/

Notes from Nature
http://www.notesfromnature

Crowdcrafting
http://crowdcrafting.org

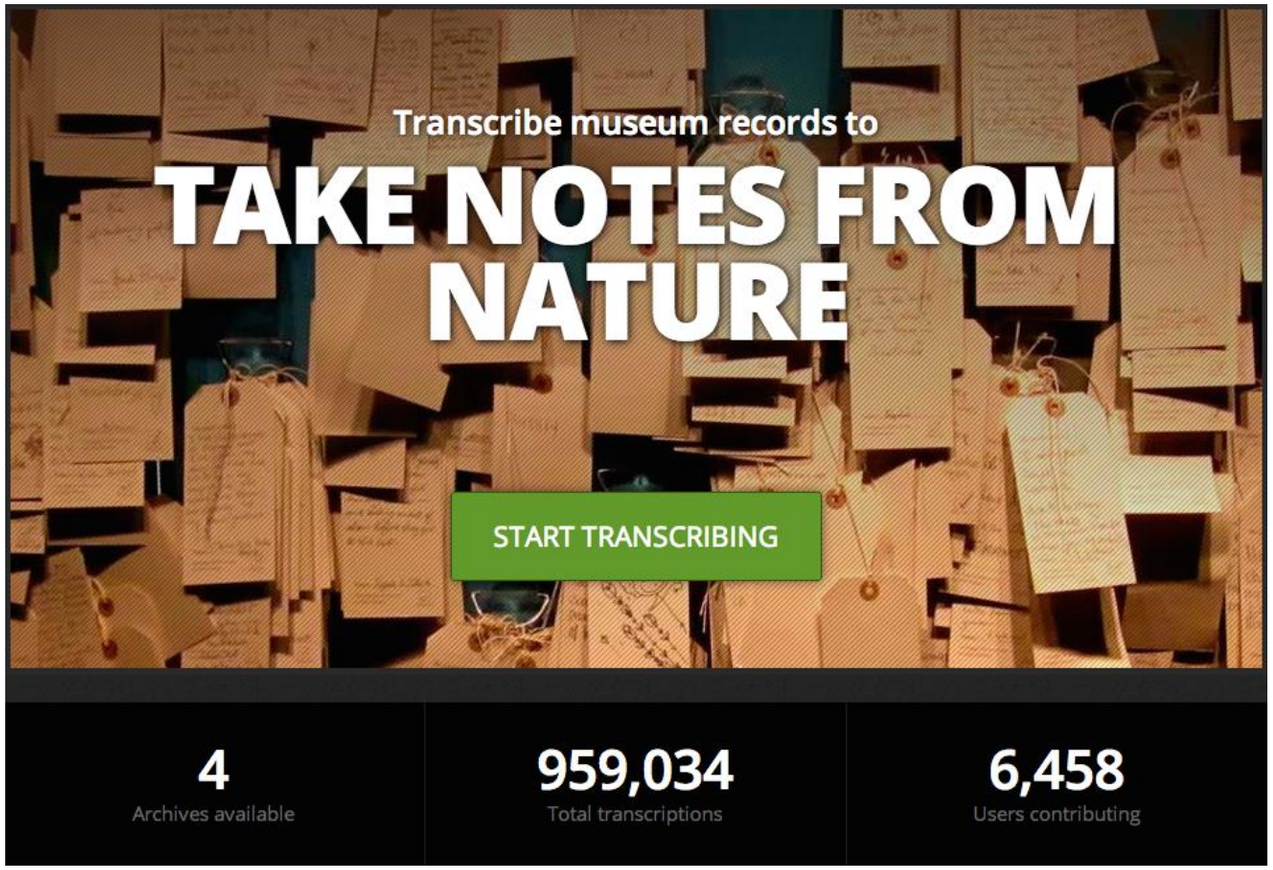
NATURAL HISTORY MUSEUM

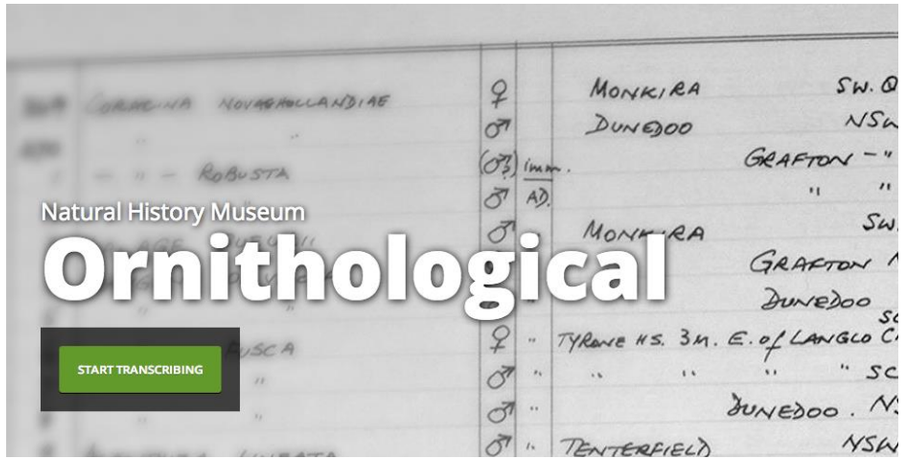
Notes from Nature (Zooniverse)

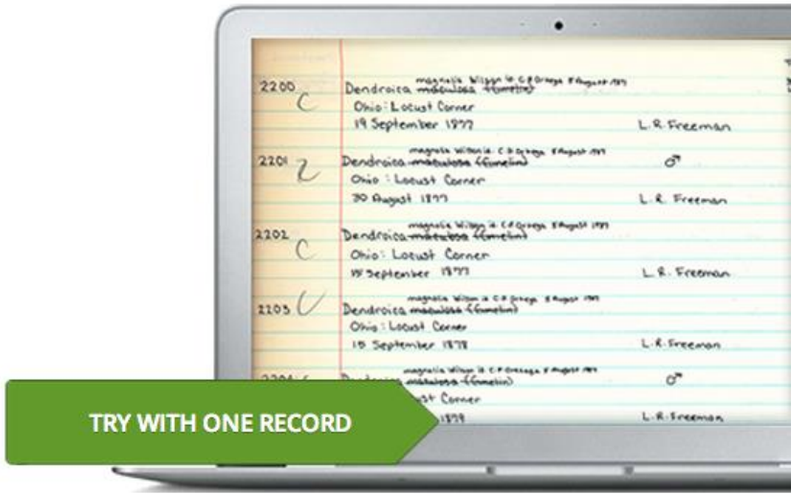
NHM Ornithology Registers (1837 – 1990)

Progress

- Total Images: 4,563
- Records: c. 230,000
- Complete Images: 2,957
- 329,454 transcriptions
- Circa 1/3rd by one person



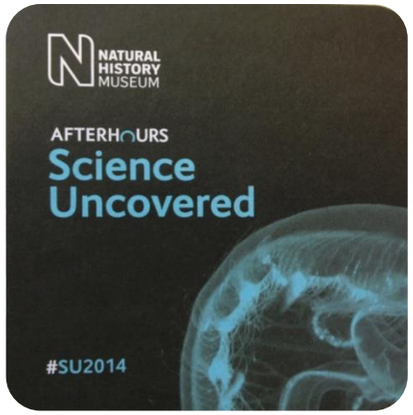


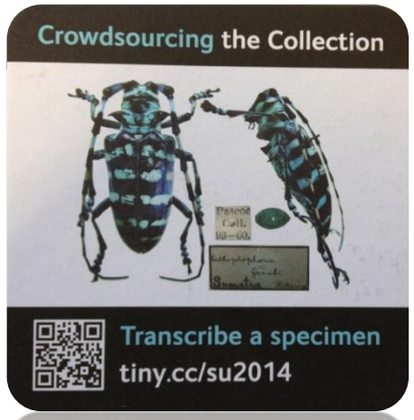


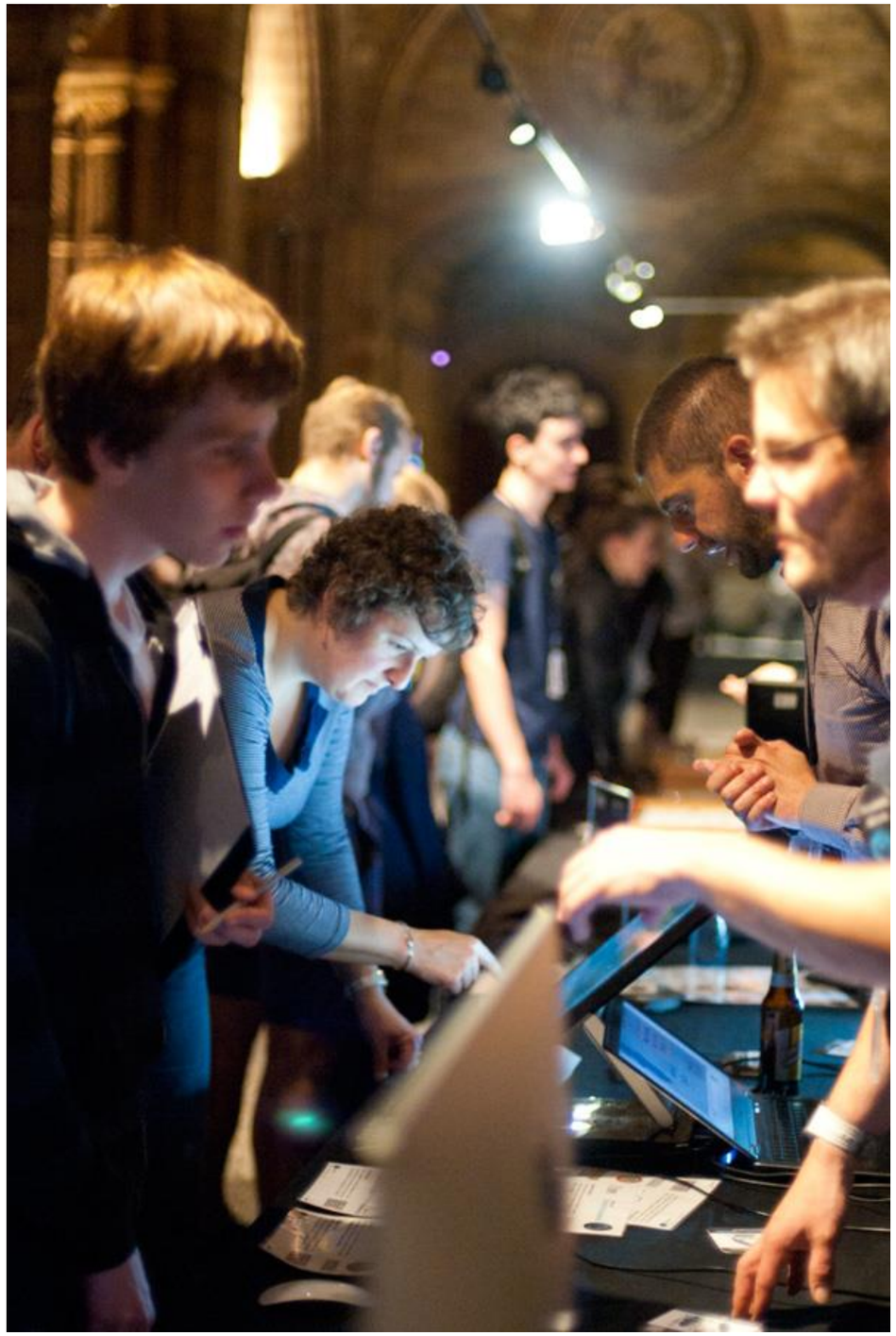
NATURAL HISTORY MUSEUM

Science uncovered 2014: “Crowdsourcing the Collection”

- Demonstrate digitisation process
- Engage the public in transcription
  - Digitise; web publish images; public transcription; data publication
- Dedicated mobile website



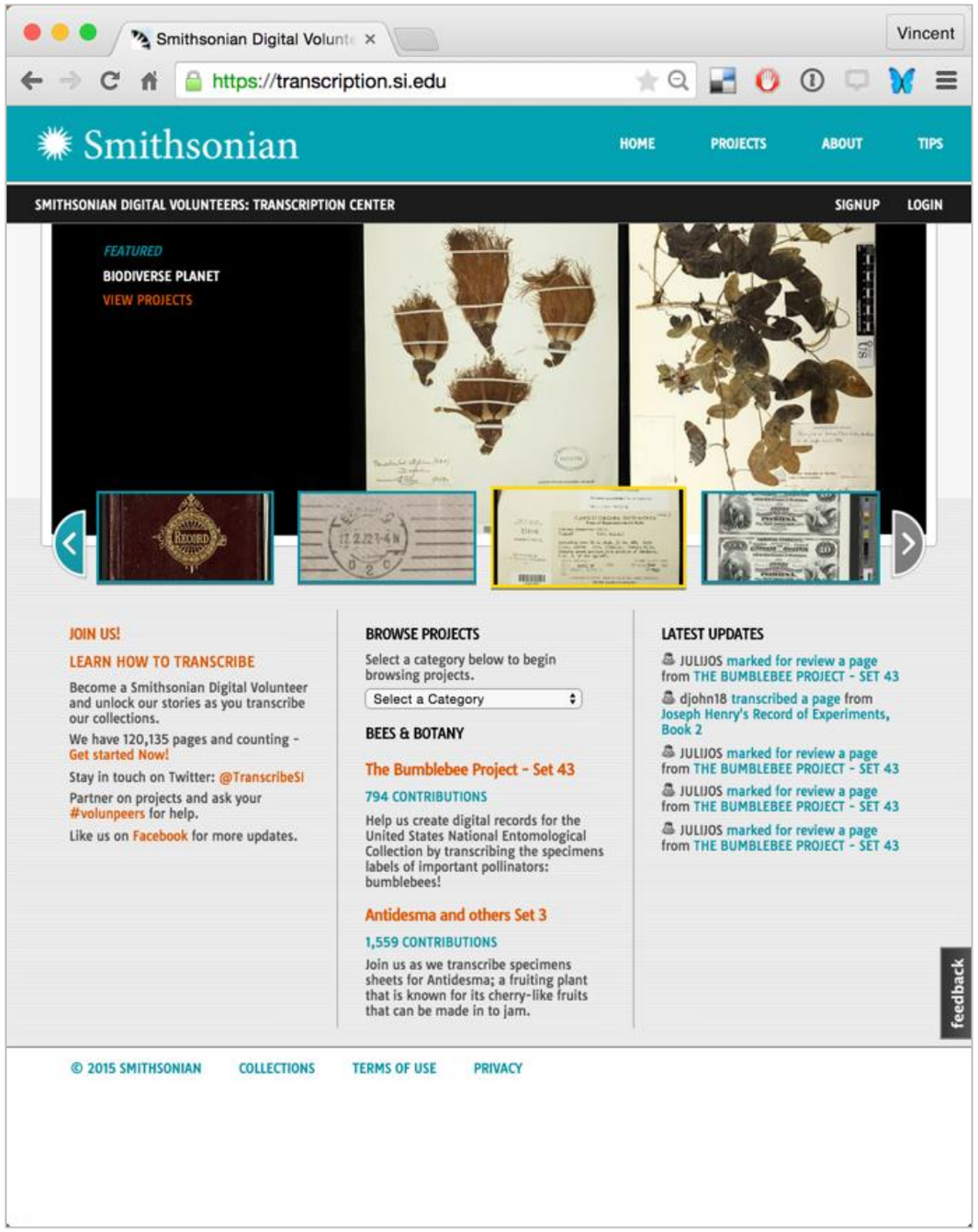




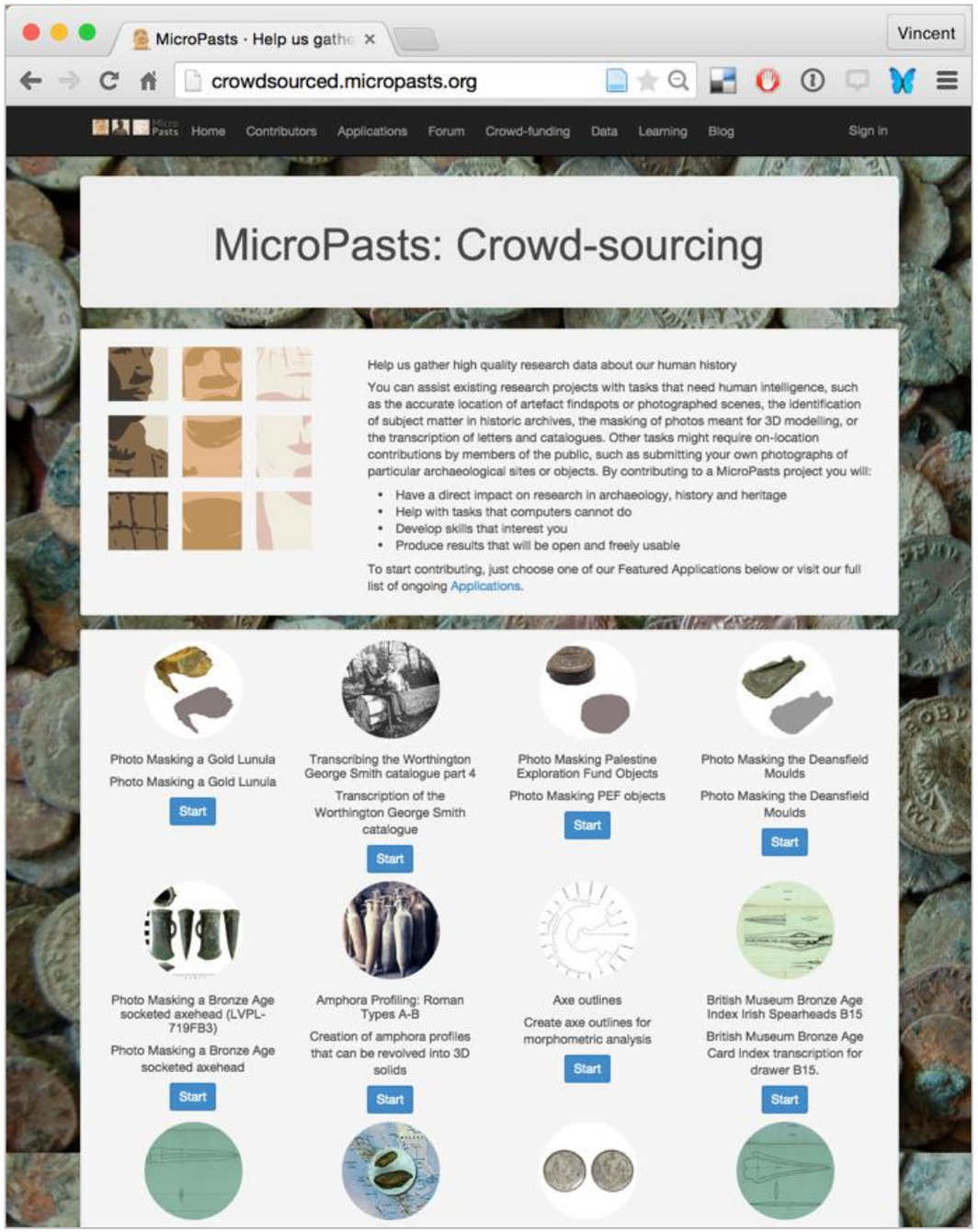
NATURAL HISTORY MUSEUM

DEMO

Platforms currently in technical review



Smithsonian Transcription Centre



PyBossa

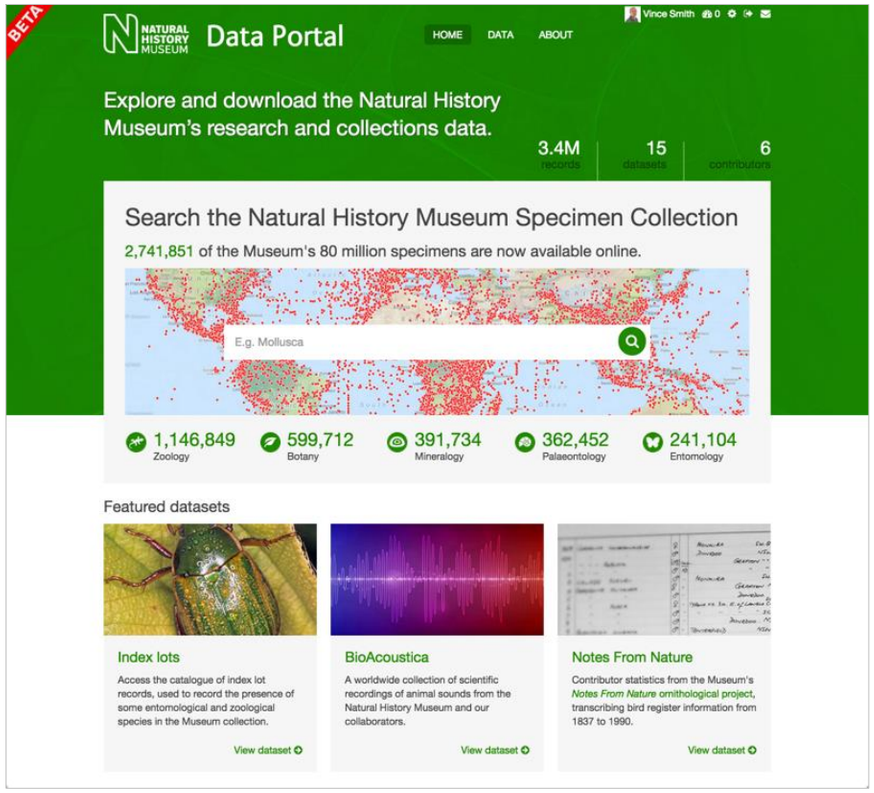
NATURAL HISTORY MUSEUM

3. Data Synthesis

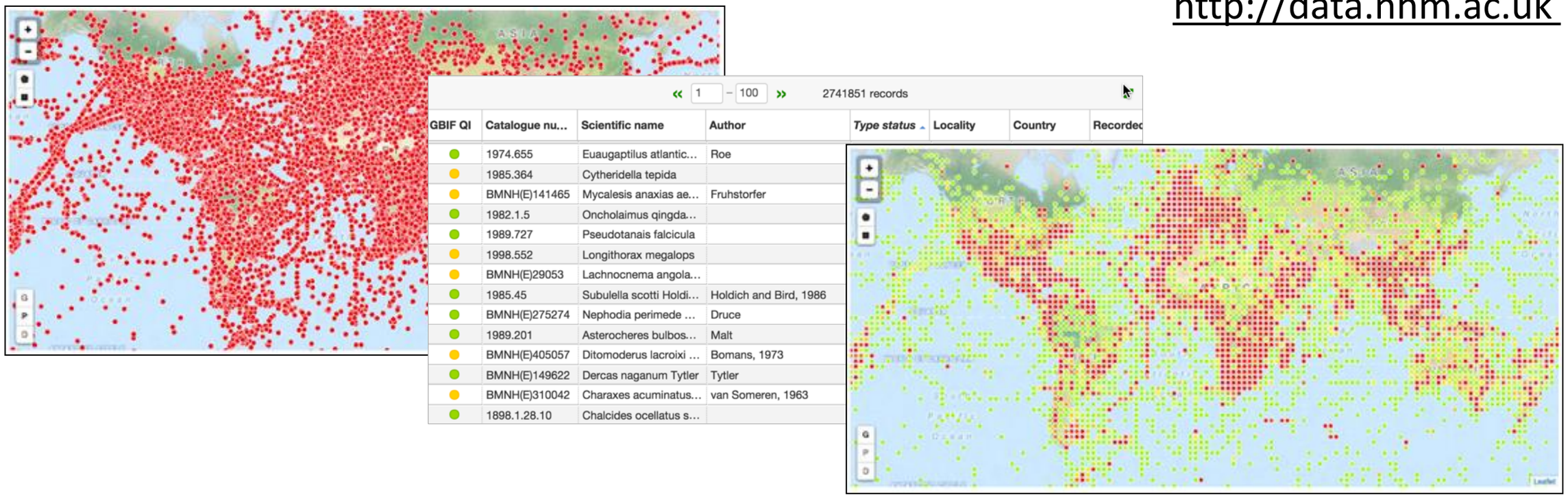
- Data aggregation & visualisation
- Modelling

The NHM data portal

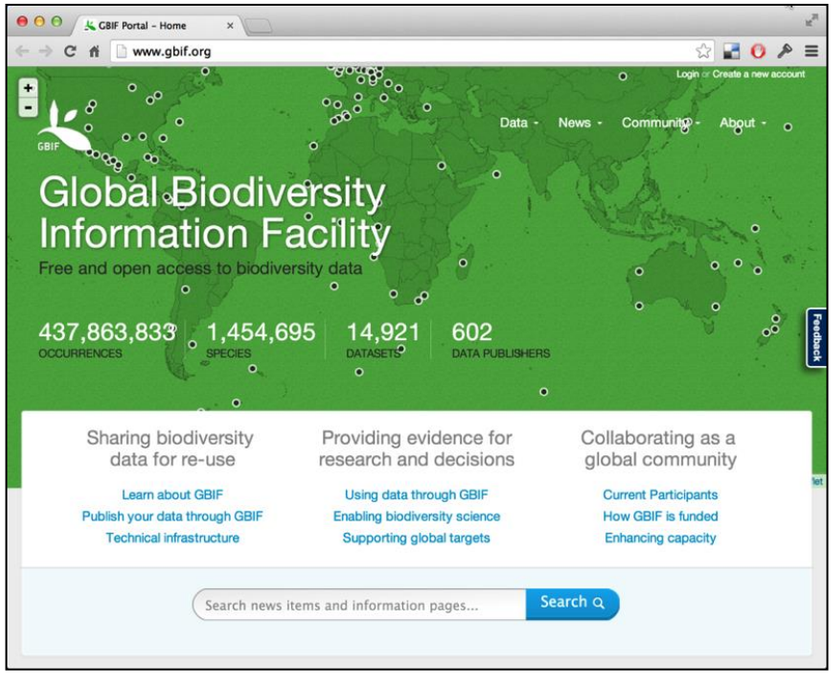
- A platform for deposition and discovery of NHM collections & research data
- Promote innovation & collaboration through easy access & reuse (website, API & download)
- Integrates with our collection management system
- Handles heterogeneous datasets of NHM scientists
- Stable, citable (DataCite) identifiers on datasets & GUIDs on records to measure impact
- Technically sustainable & scalable
- Open data policy (CC-Zero, CC-BY)



http://data.nhm.ac.uk



External services supporting data quality indicators



Via GBIF API

GBIF QI	Catalogue nu...	Scientific name	Author	Type status
	RT Lowe 2000 ...	Bromus diandrus Roth	Roth	
	BM001147086	Frullania microphylla (...	(Gottsche) Pearson	
		Daphnusa ocellaris W...	Walker, 1856	
	BM000559415	Capsicum annuum (D...	(Dunal) Heiser & Pickersgill	
	1949.1.19.30	Crossaster papposus...	(Linnaeus, 1767)	
	PM P 43052 (2)	Orbitolina birmanica ...	Sahni, 1937	
	BM000798867	Chasalia kolly (K.Sch...	(K.Schum.) Hepper	Isotype
	BM000003217a	Salix arctica Pall.	Pall.	
	Carlos Types - ...	Polystichum viviparu...	Fée	Isotype
	1998.3.12.1-50	Neolepidapedon sp.		
	50021	Solanum galapagens...	S.C.Darwin & Peralta	
	PM OS 16045			
	BMNH(E)70713	Mellicta athalia		
	1974.1.25.142	Rhodeus suigensis M...	Mori, 1914	

Major errors

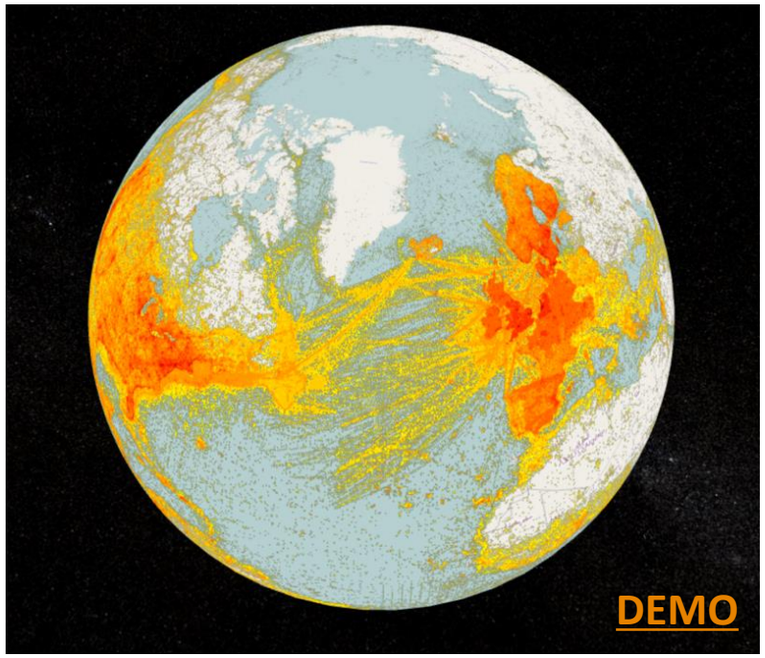
Minor errors

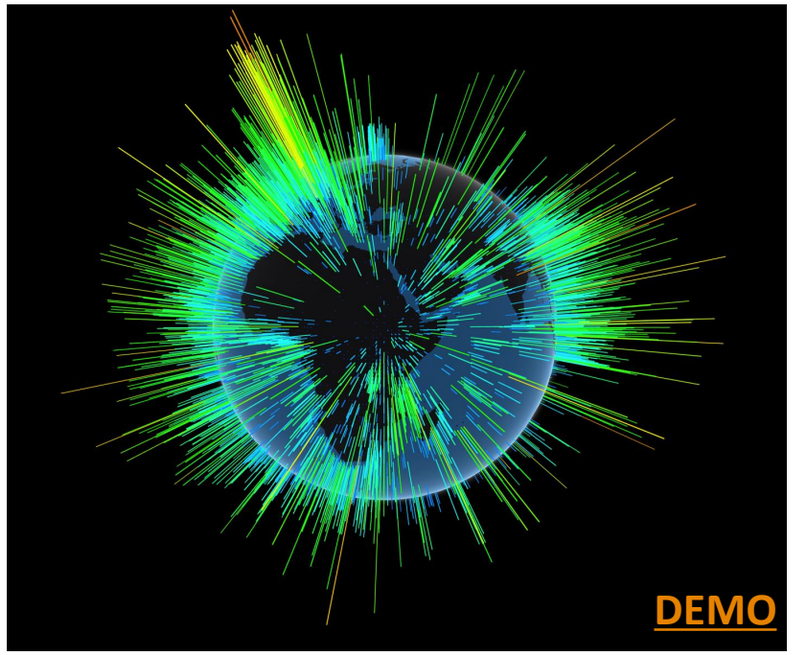
No errors

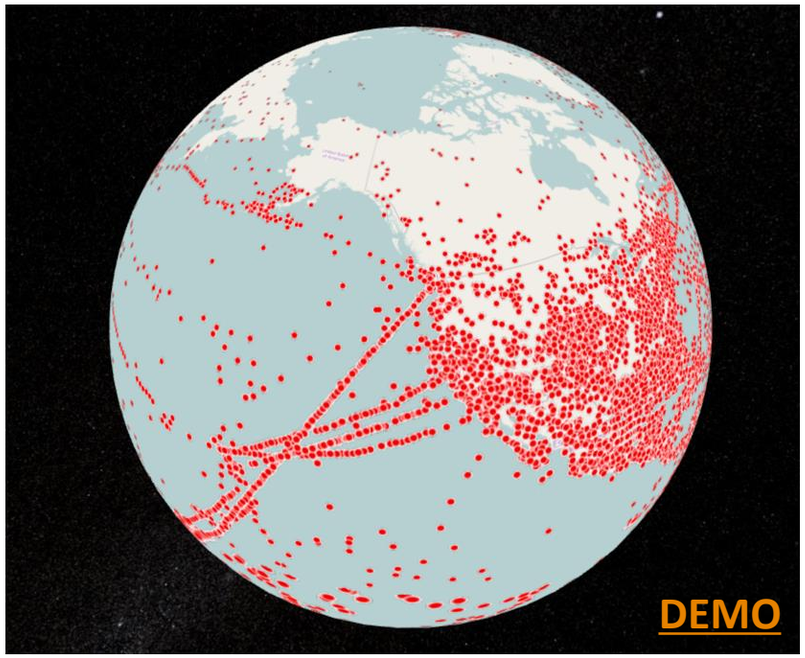
NATURAL HISTORY MUSEUM

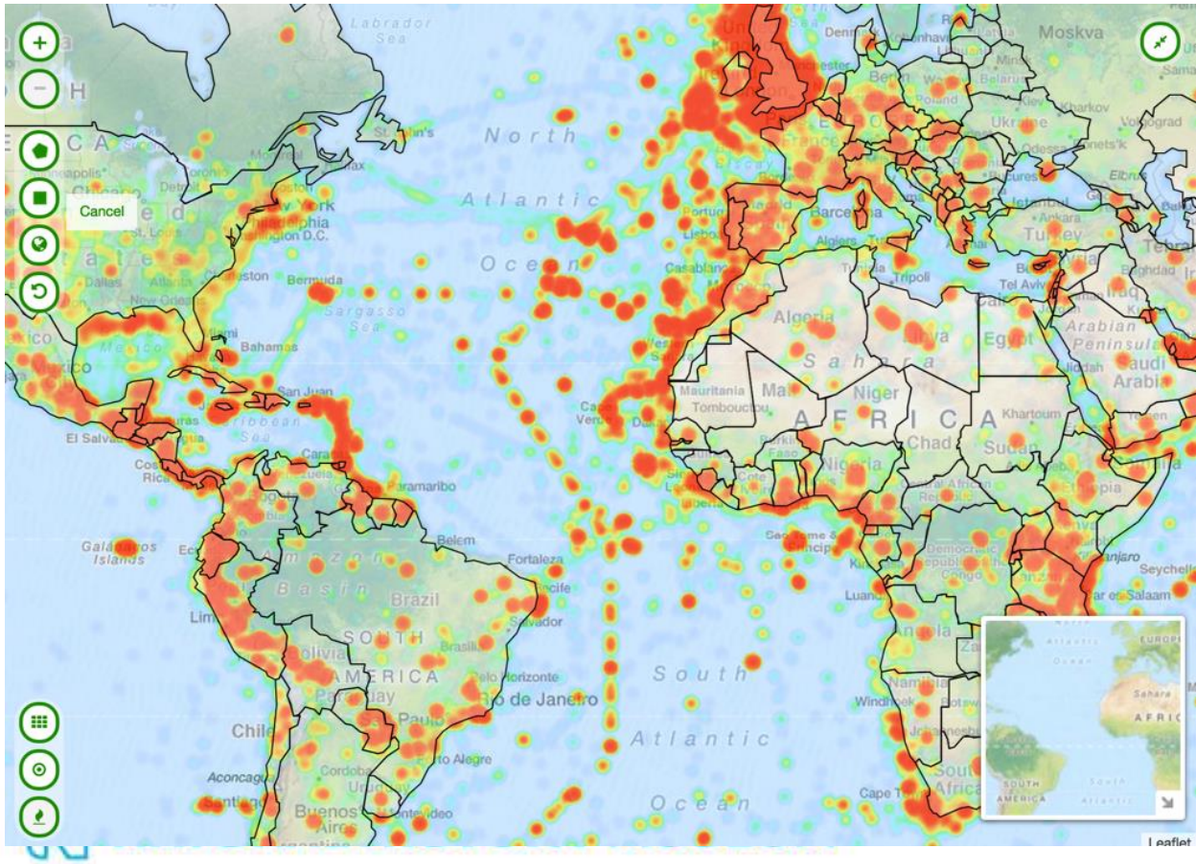
nb. similar services offered by CRIA for Brazilian data

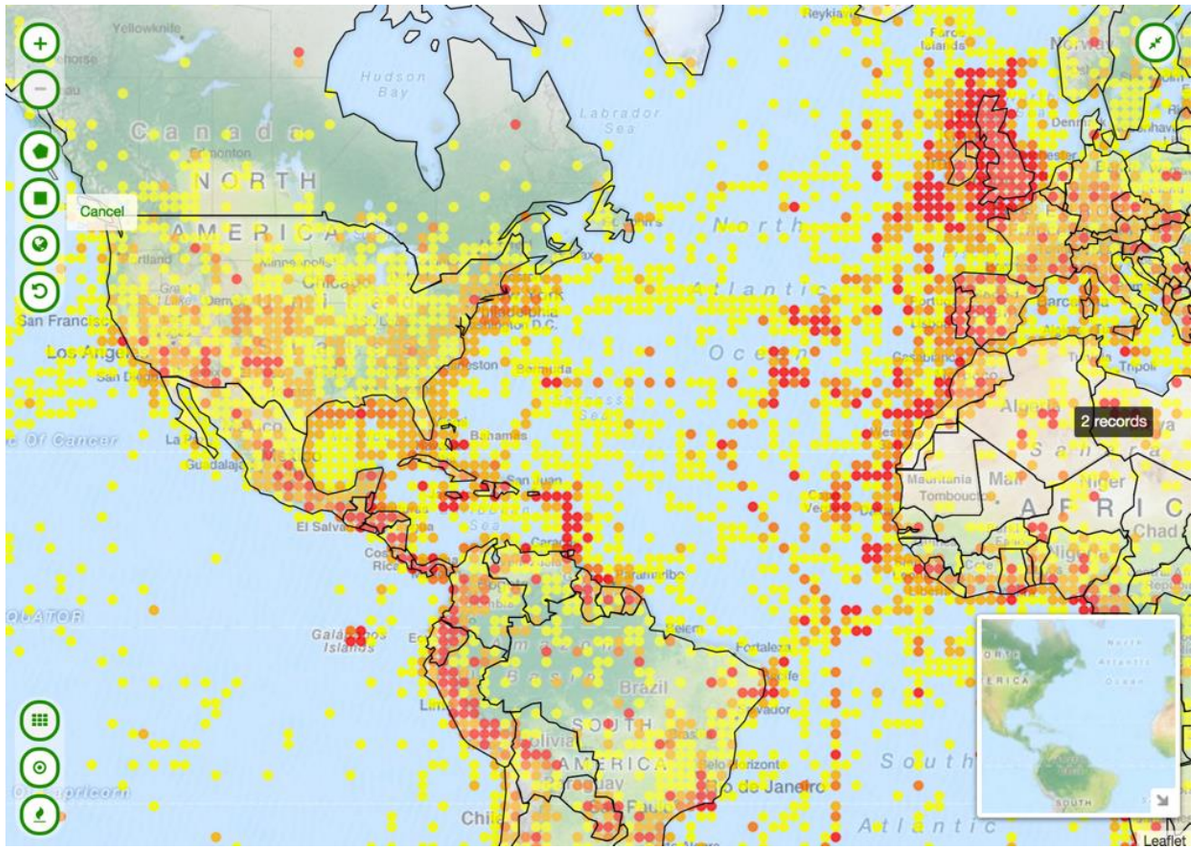
Data visualisations (embedded & via API)



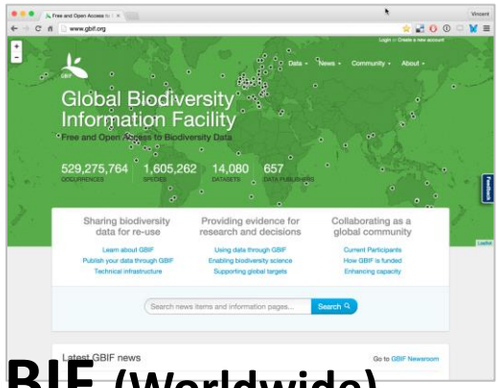




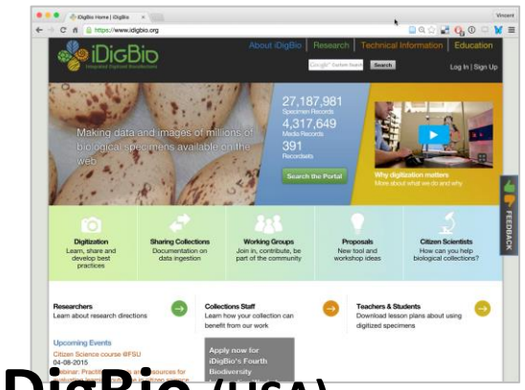




Consolidation & sharing of biodiversity data portals

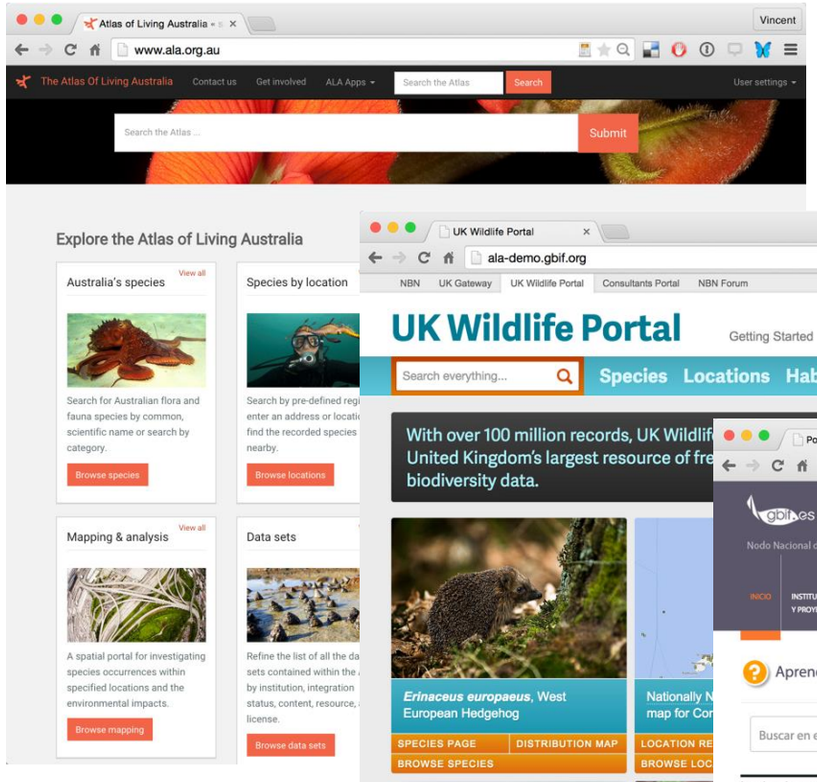


GBIF (Worldwide)



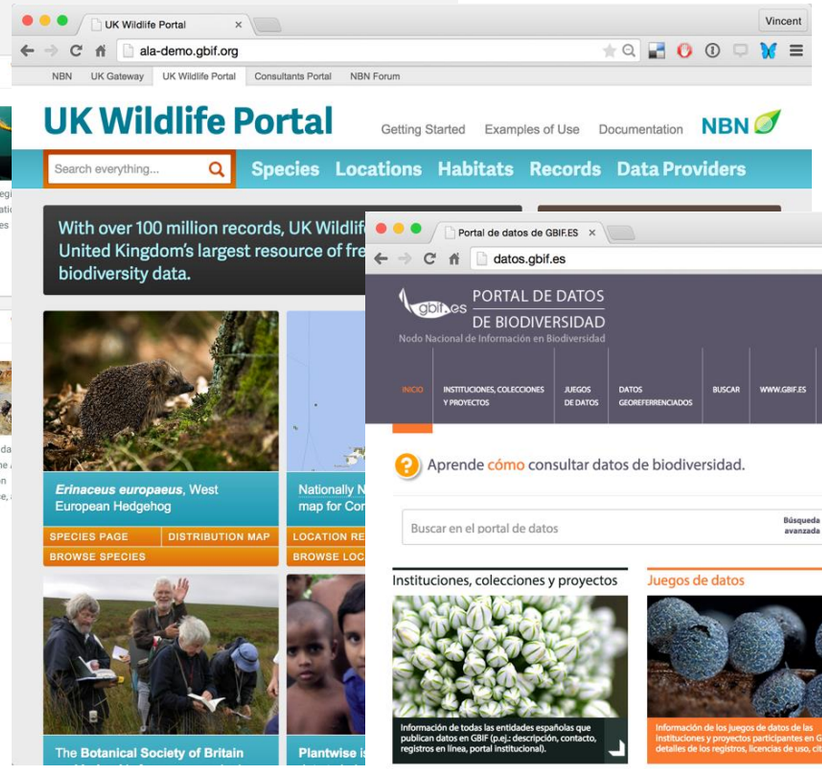
iDigBio (USA)

CRIA (Brazil)

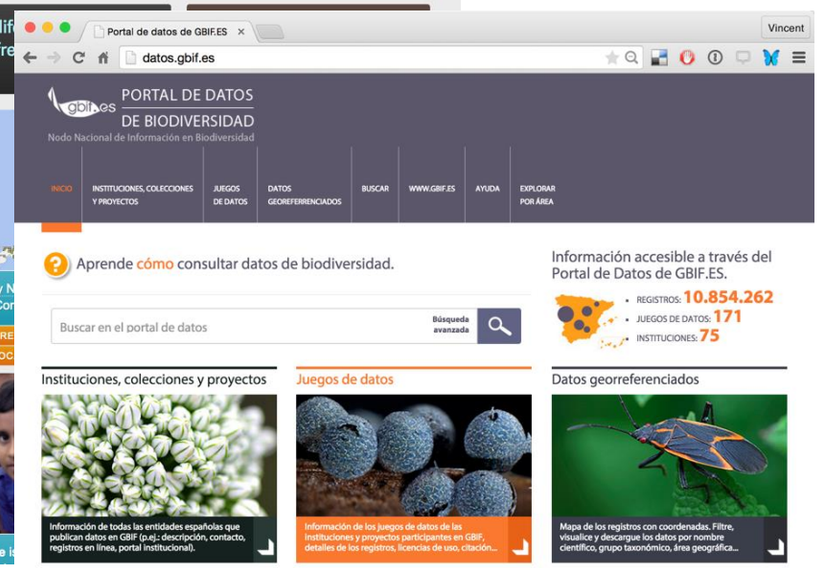


Atlas of Living Australia

Australia



U.K.

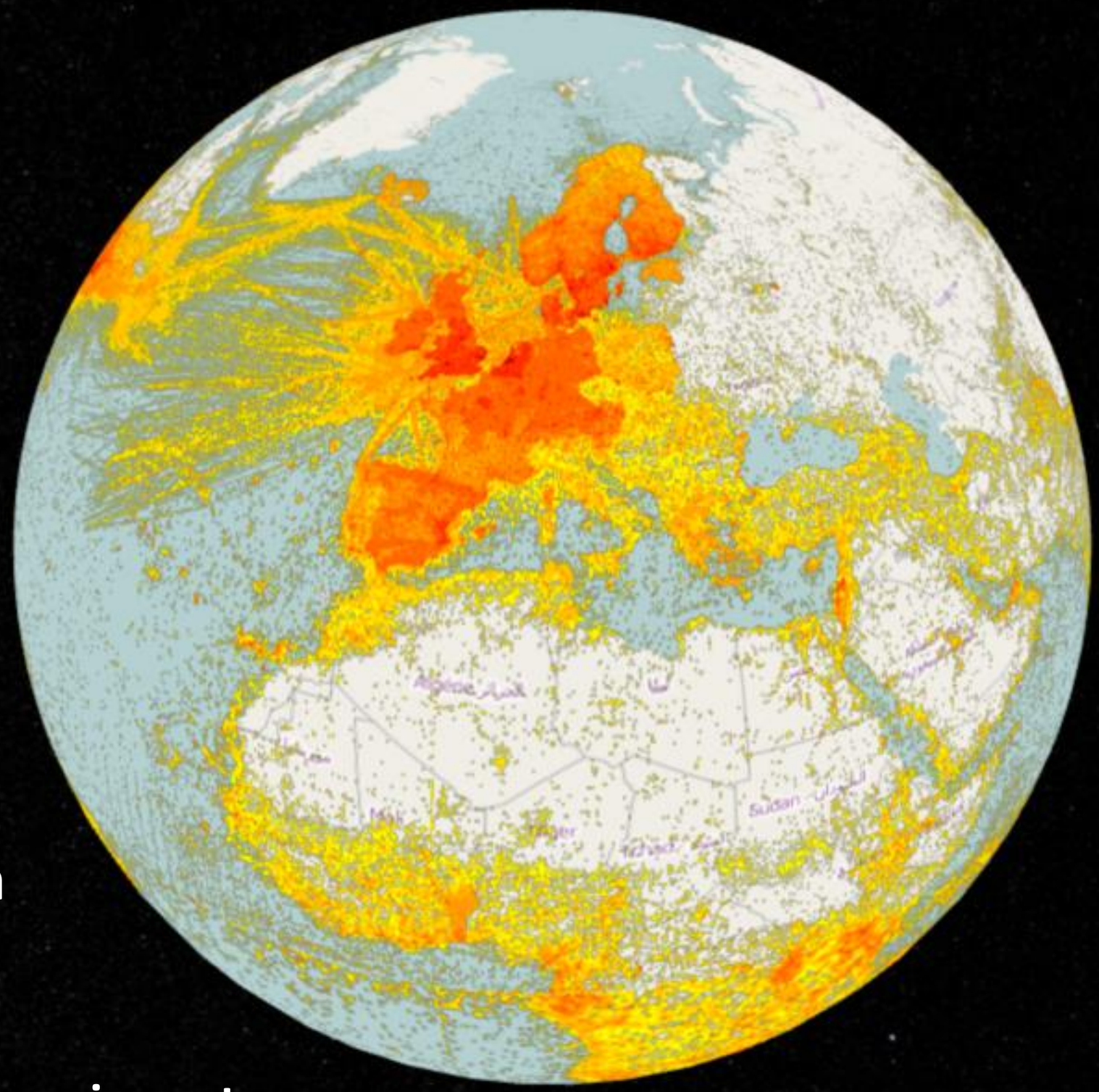


Spain

Growing reliance on a network of dependable data services

Big data research opportunities

Use NH collections data to explore changes on biodiversity over space & time



1.5-3 BILLION SPECIMENS

1.9 million species

300 years of collection

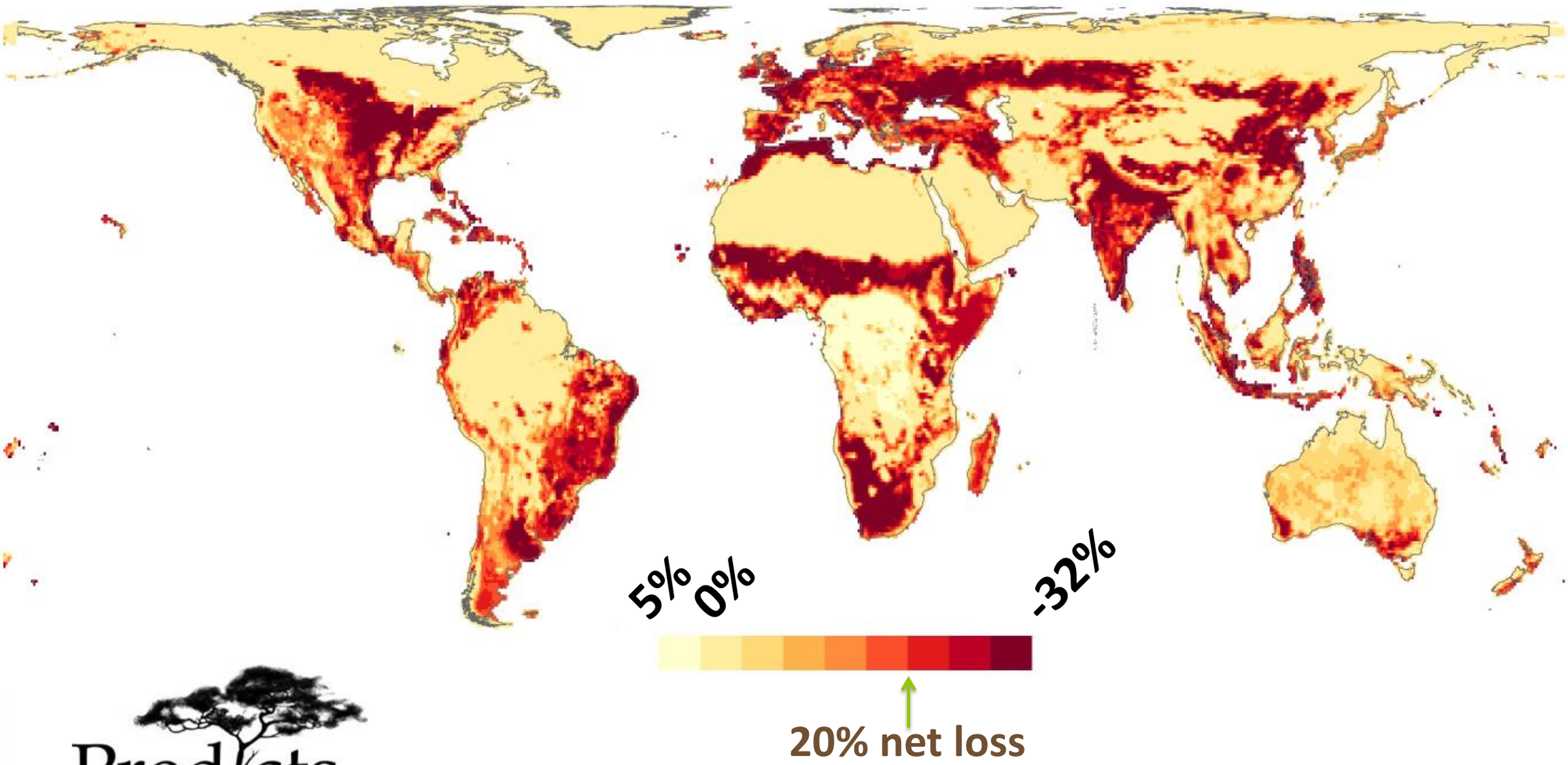
<u>Goals</u>

Quantify human impacts

Predict how to mitigate human impacts

Collection data analysis to model species richness

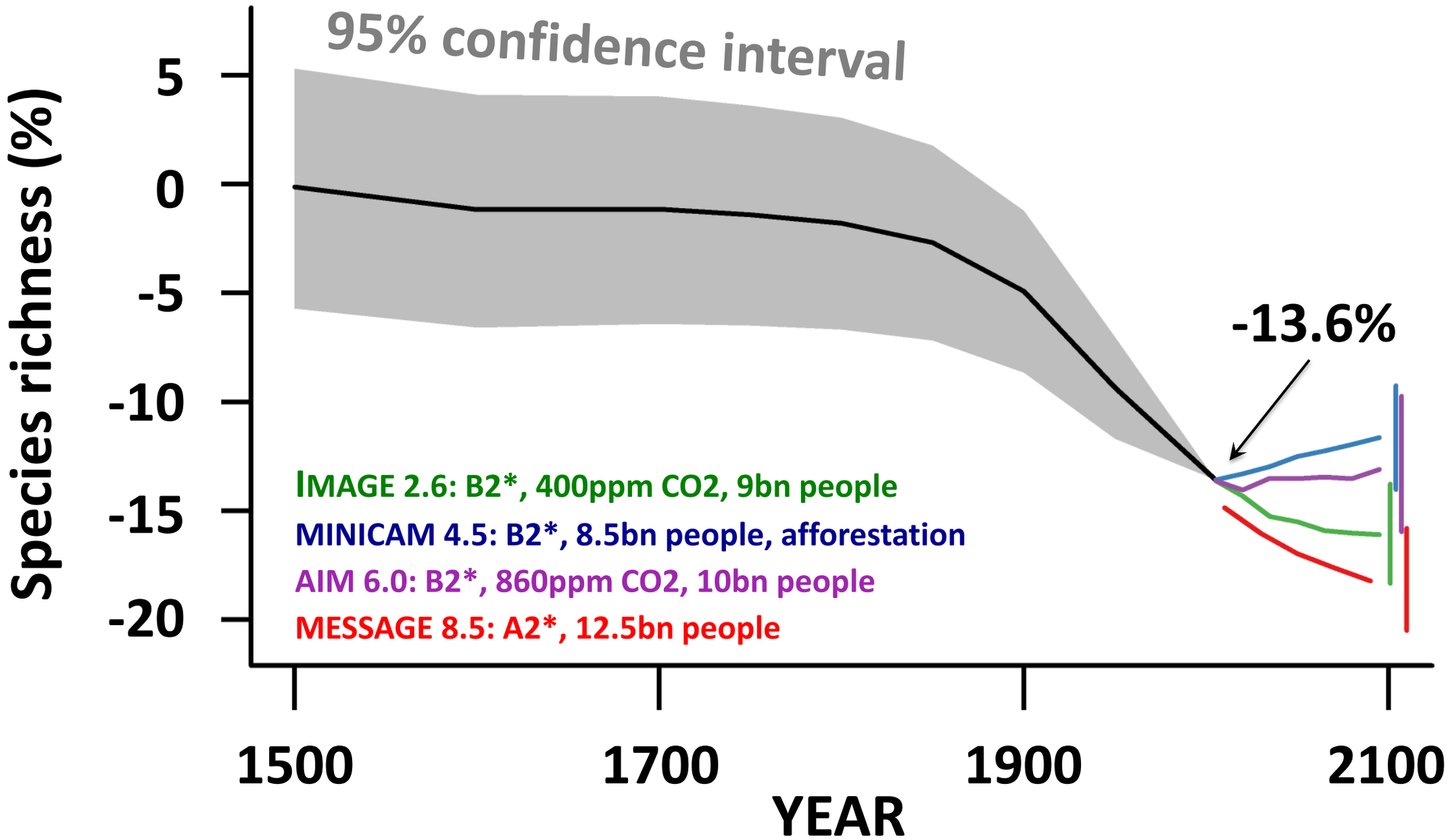
Species richness lost by 2005



Predicts
www.predicts.org.uk

Newbold, Hudson et al, Nature doi:10.1038/nature14324

Predicting effects on biodiversity under climate change models

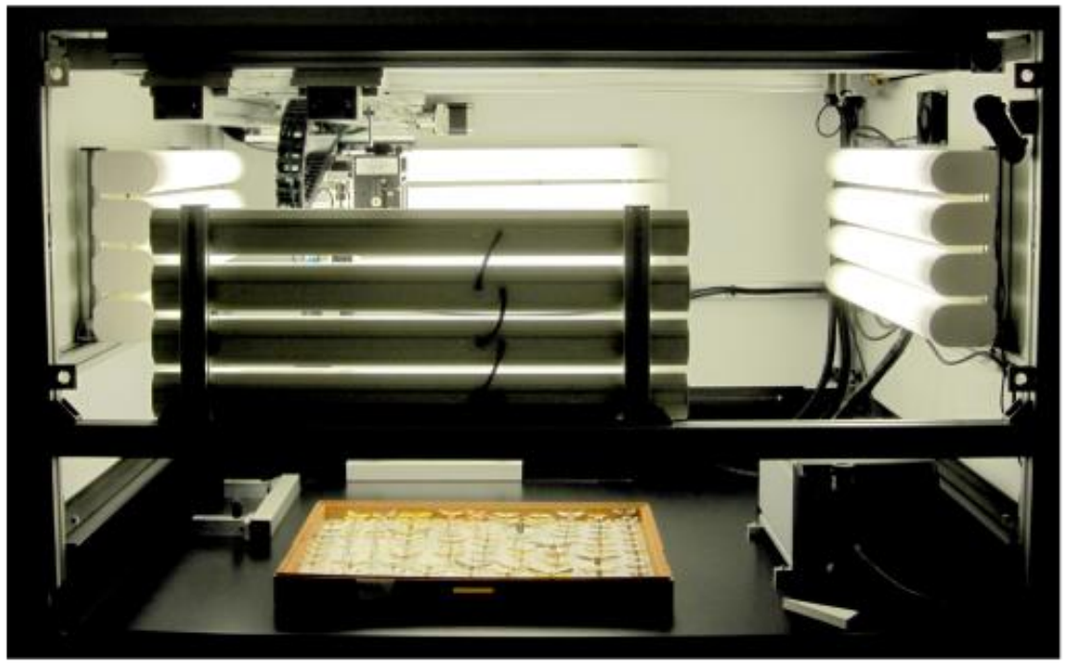


Newbold, Hudson et al, Nature doi:10.1038/nature14324

4. Enabling technologies

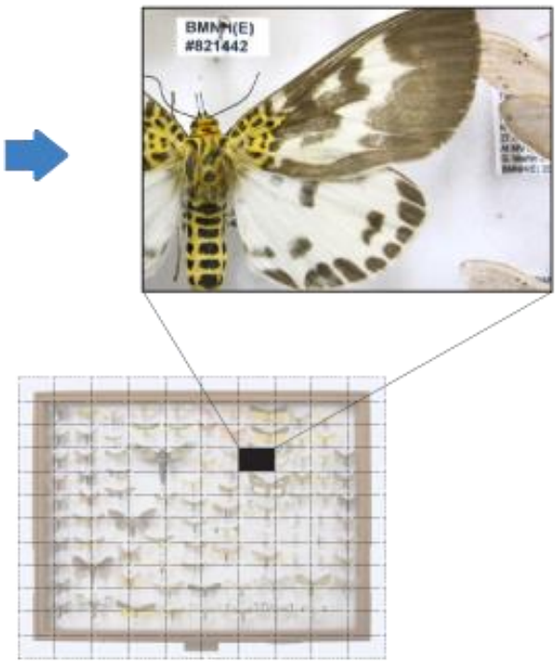
- Computer vision
- Remote sensing

Drawer level imaging is (mostly) a solved problem



1. Place drawer

2. Scan



3. Stitch

- Fast (5 mins per drawer)
- High resolution (circa 500MB per image)
- Enables bulk databasing of specimens
- Potentially reduced specimen handling

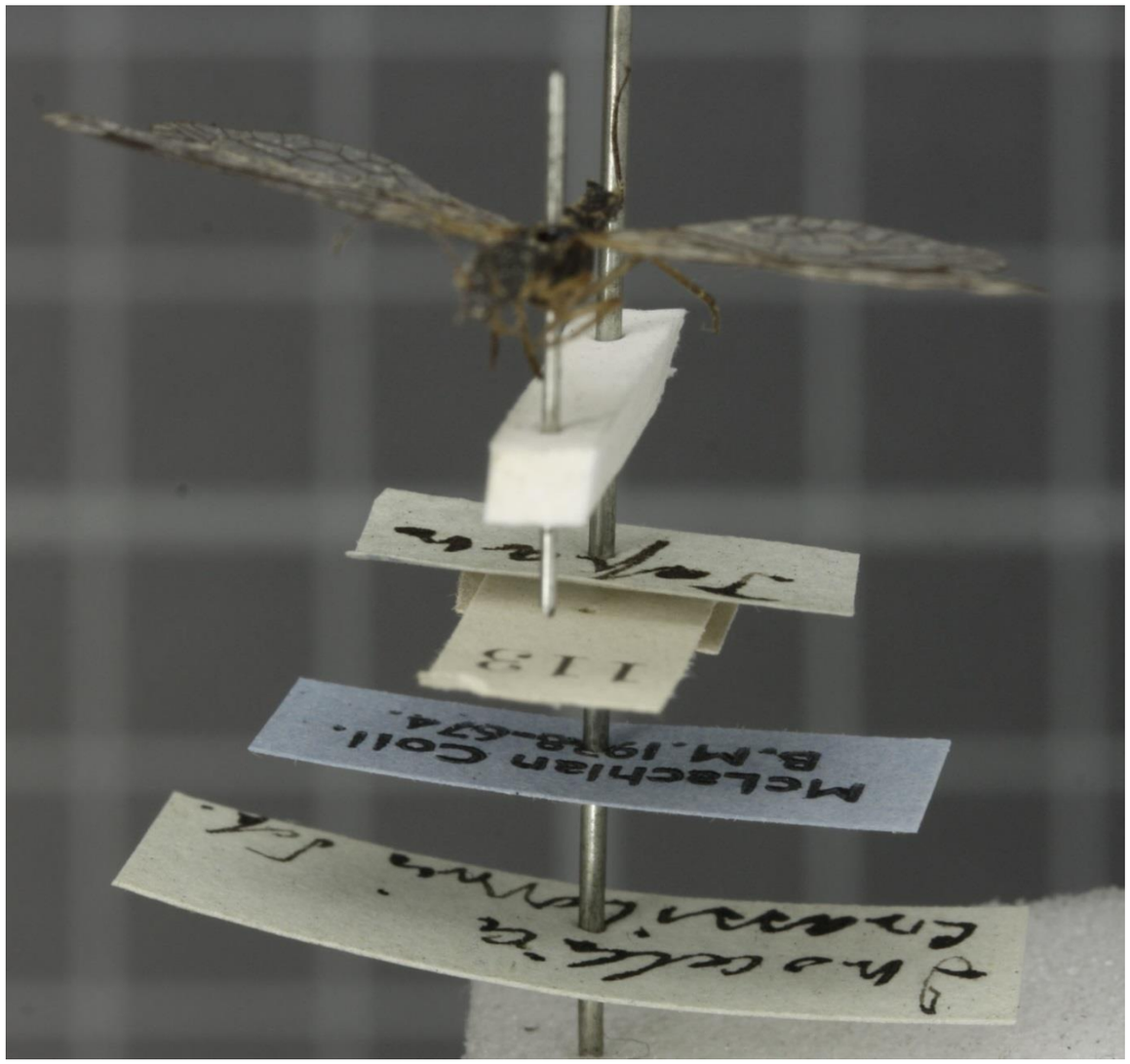
NATURAL HISTORY MUSEUM

But, two key problems remain…

1. Synchronisation

Keeping the physical & digital copies in sync

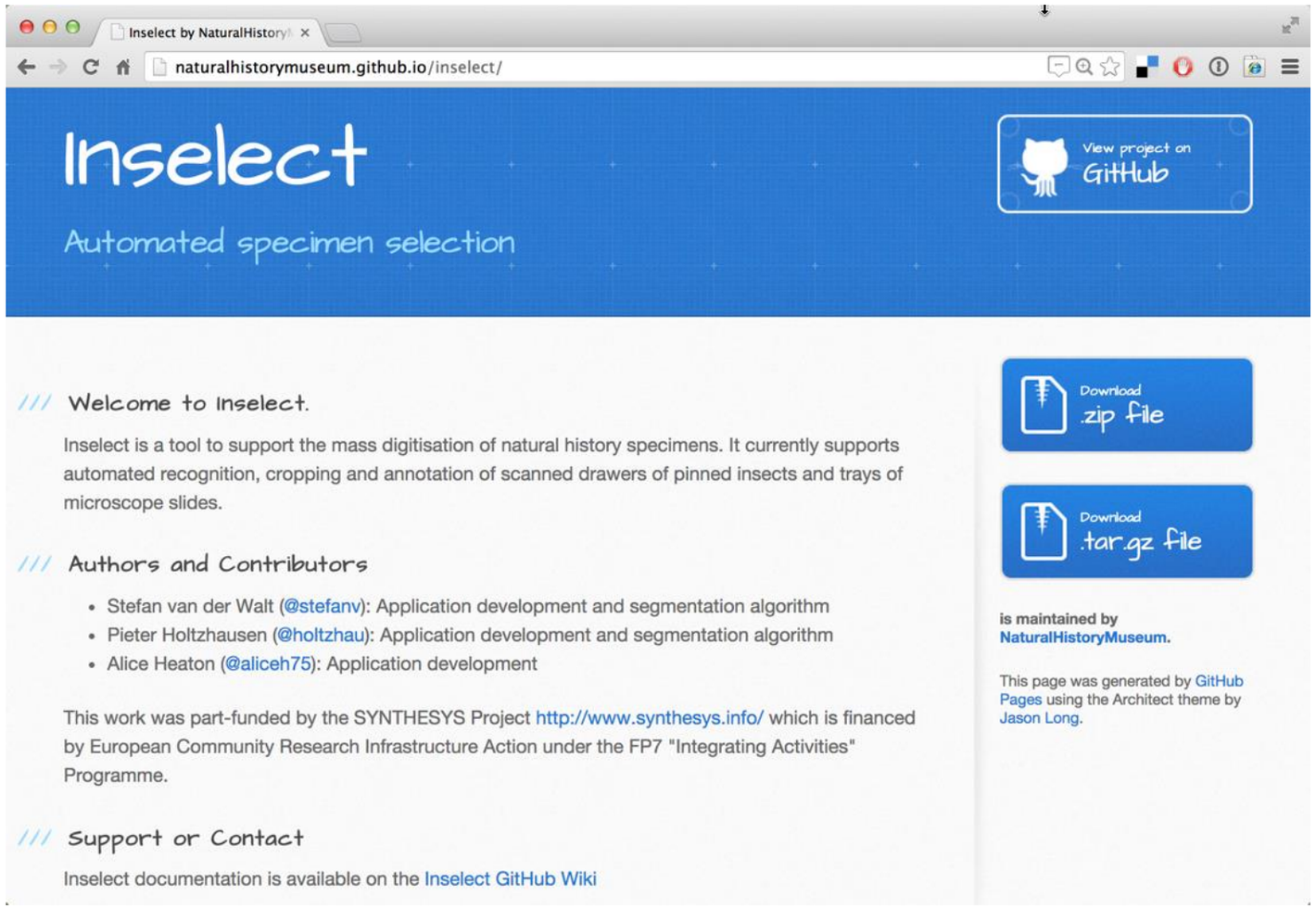
2. Label data



Capturing data from multiple pinned labels

Inselect

Automated recognition, cropping and annotation of specimens



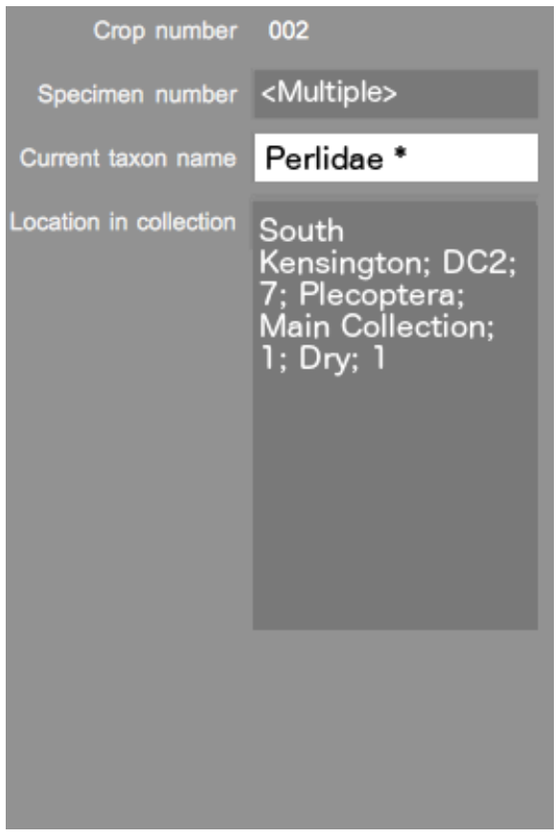
- Currently beta-release
- Automatically detects specimens
- Creates bounding boxes for cropping and exporting images
- Read barcodes
- Rapid annotation interface
- Persistent settings & keyboard shortcuts
- Data export in JSON format
- Open source & modular
- Python based (OpenCV, scikit-image libraries)
- Windows, OSX & Linux

http://naturalhistorymuseum.github.io/inselect/

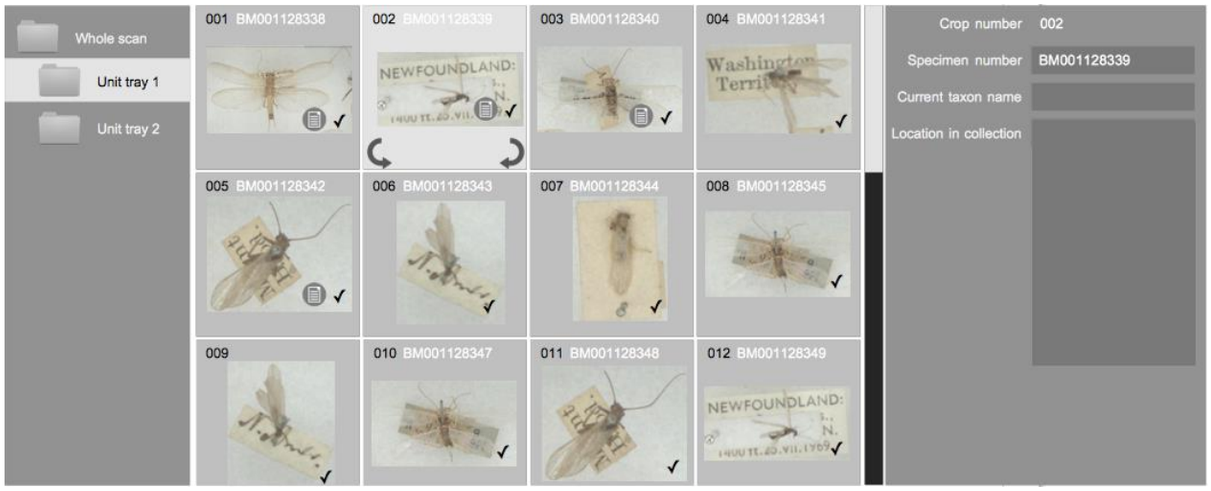
Inselect features

Slides

Vocab. services



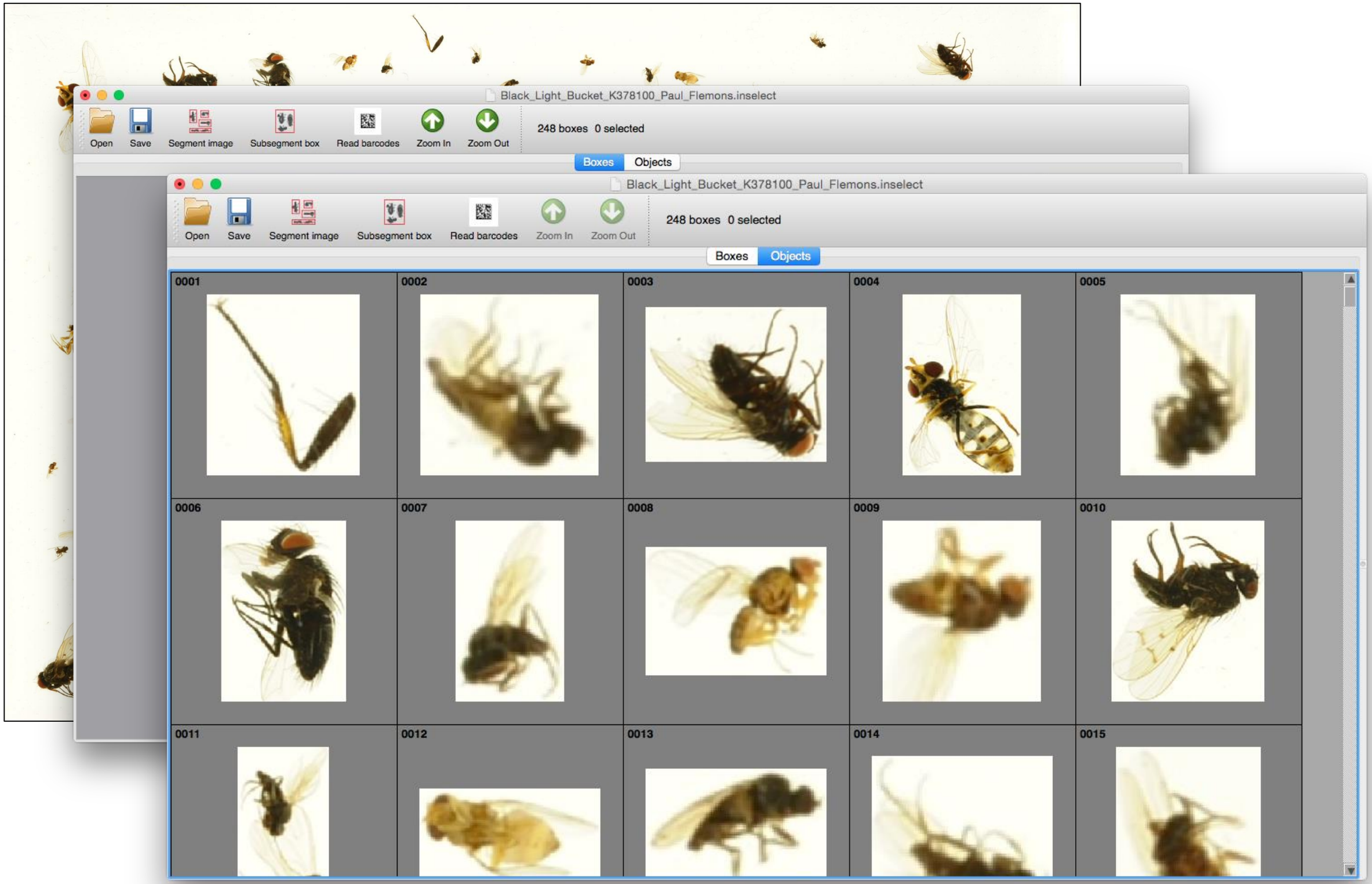
Barcode reading



Multi- specimen annotation

Unit tray recognition

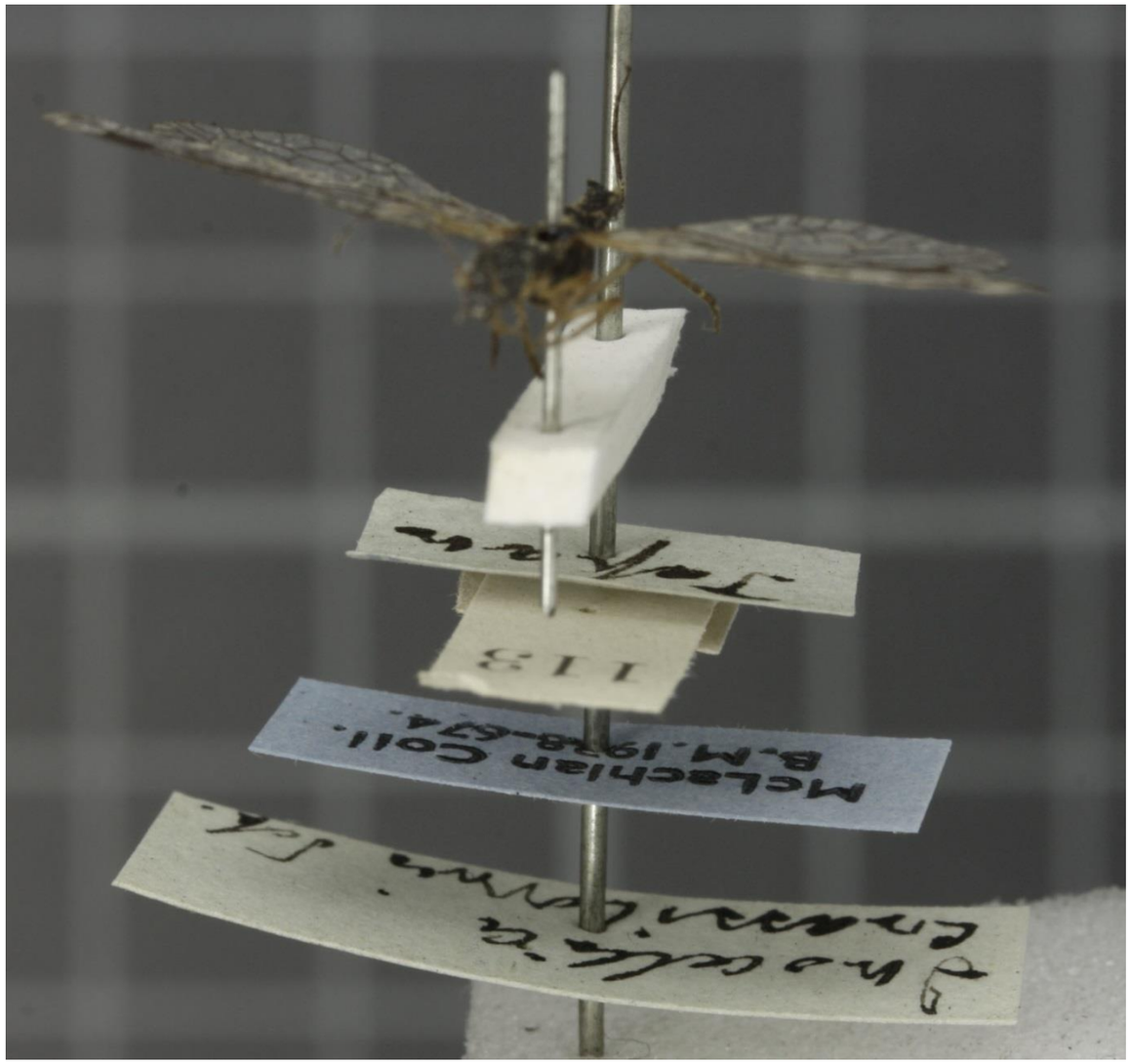
Potential applications of Inselect – “Insect Soups”



NATURAL HISTORY MUSEUM

Capturing label data

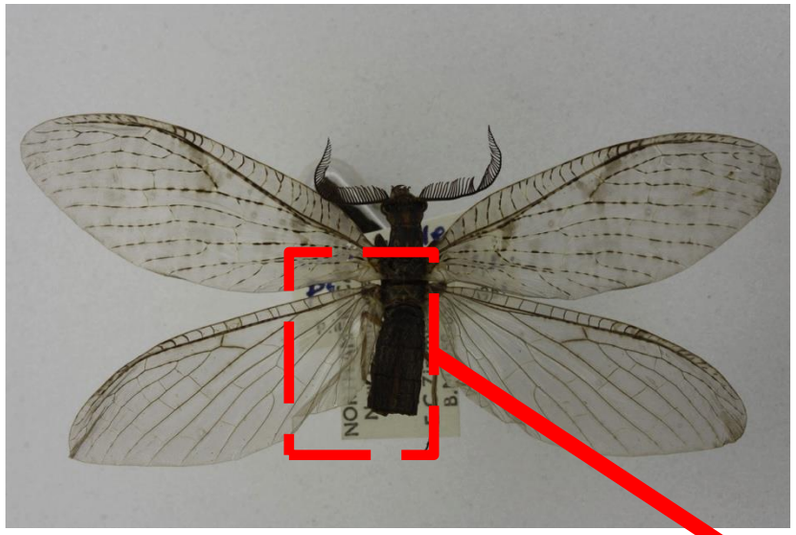
- Some data can be captured but not all
- Specimen and other labels obscure labels below them
- Holy grail method: capture label data without removing them



Label imaging & text recognition

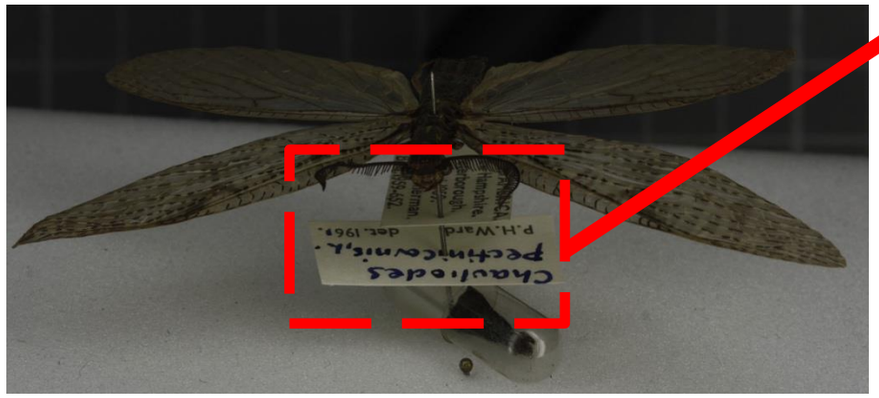
Chauliodes pectinicornis

Dorsal

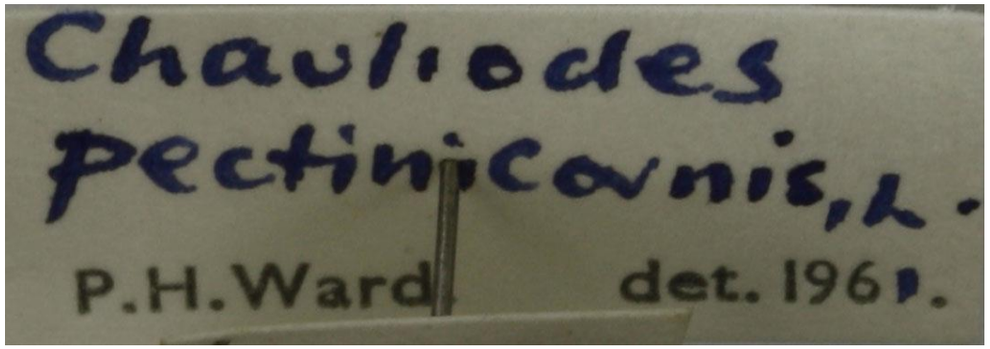


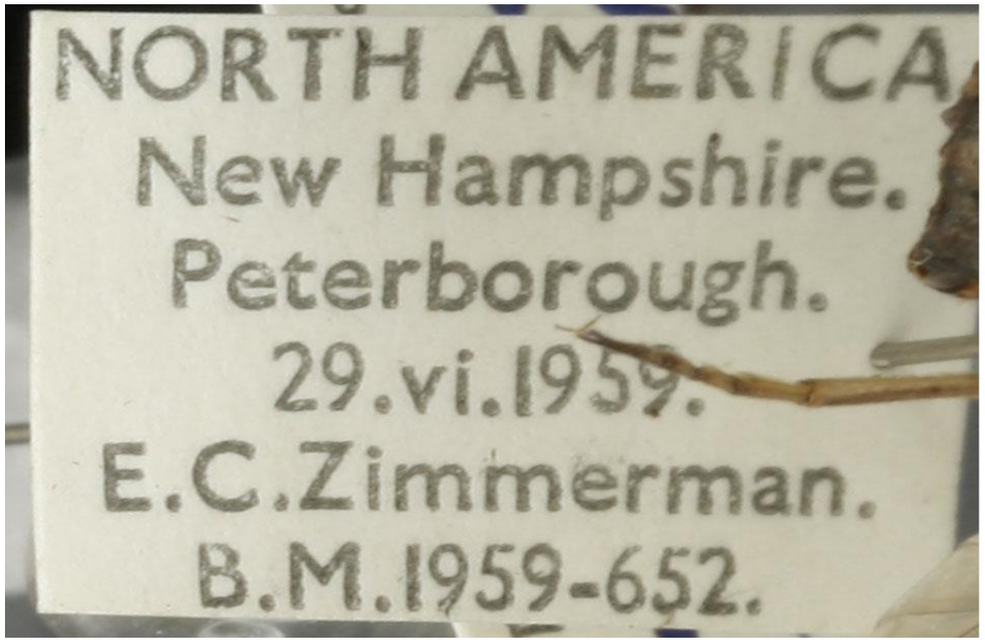
Caudal

Frontal



Reconstructed labels



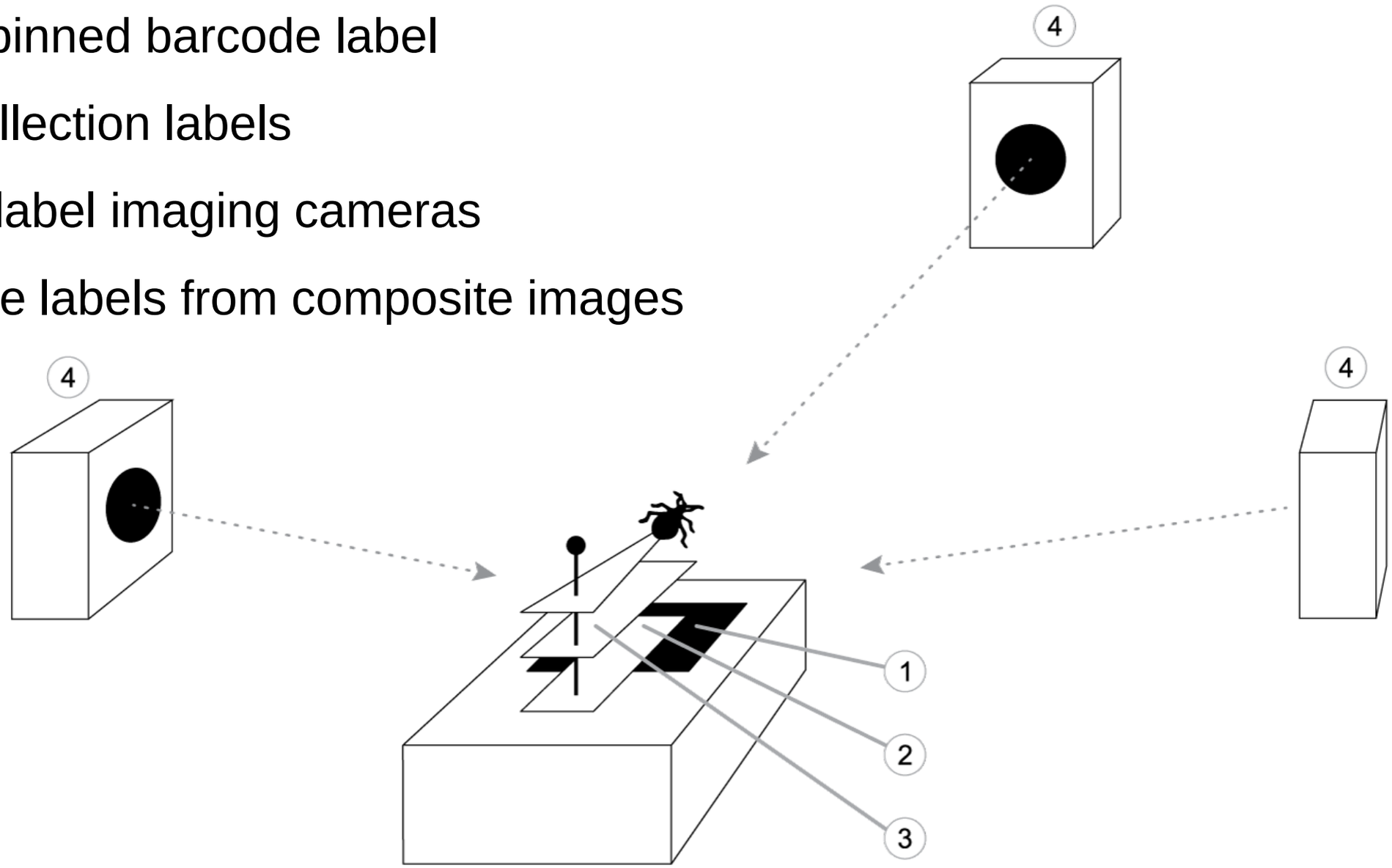


Whole Drawer image

Approaches to label imaging pinned specimens

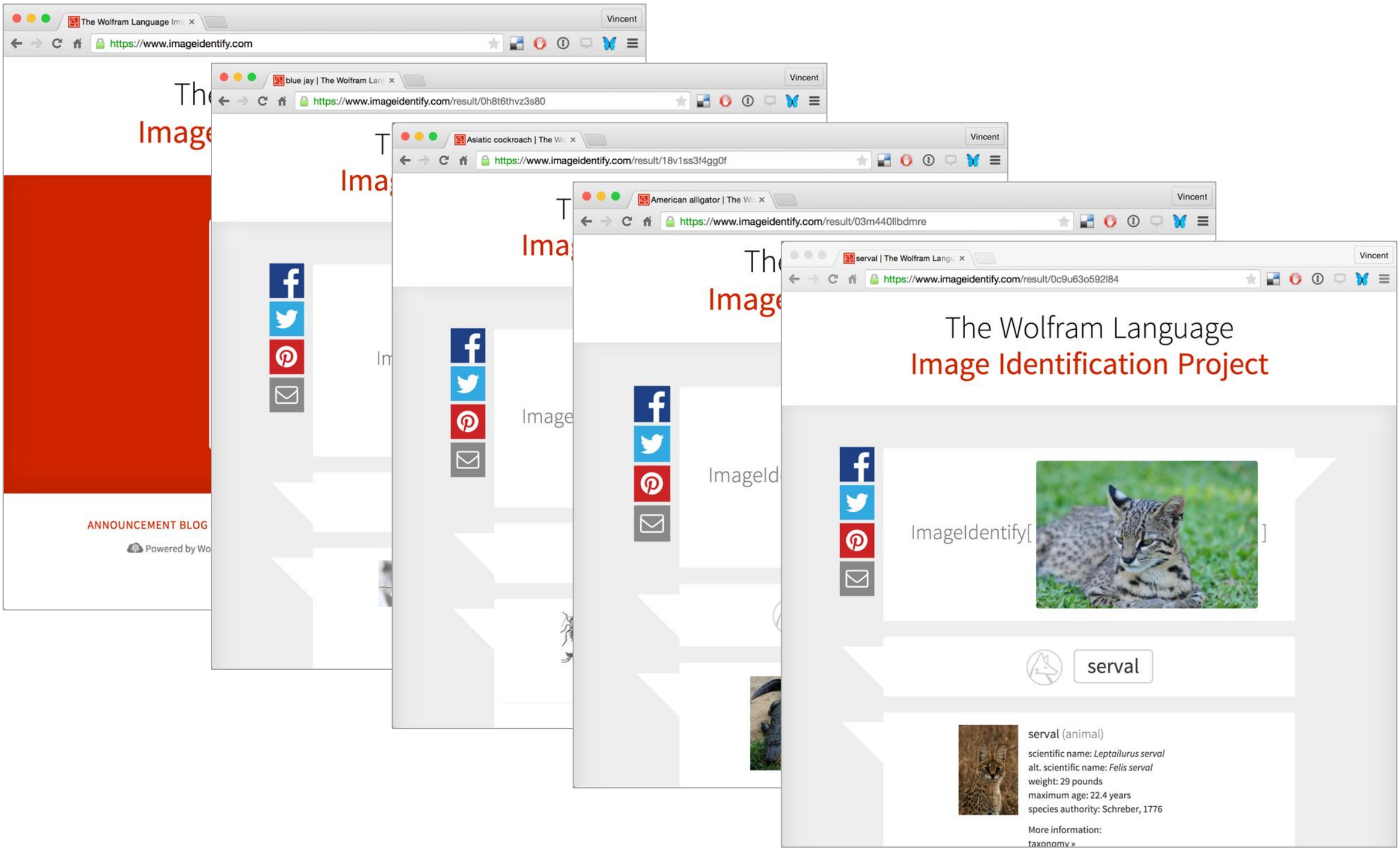
Could be incorporated as part of the barcode dispensing process

1. Barcode dispenser & scanner
   (two-sided barcode labels)
2. Freshly pinned barcode label
3. Other collection labels
4. Multiple label imaging cameras
5. Assemble labels from composite images



NATURA

Wolfram automated image identification



NATURAL HISTORY MUSEUM

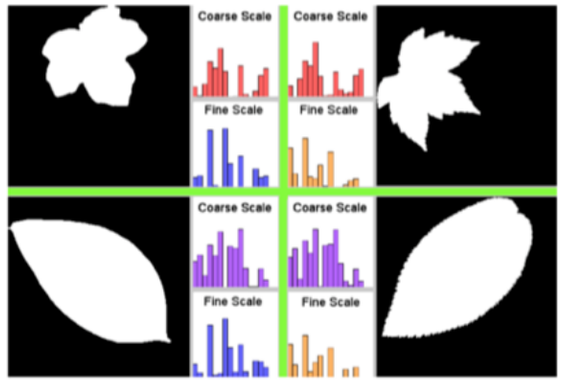
https://www.imageidentify.com/

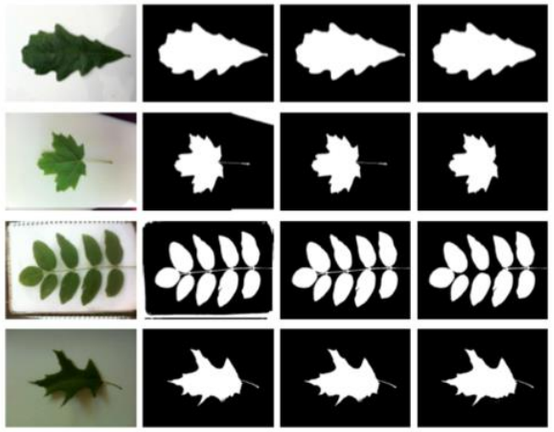
Smart phone image analysis for UK / US tree identification

- Computational analysis of pictorial information to support species ID
- Match shape and motion with machine learning to build vision systems
- Relies on large validated image databases

- Snap leaf against a white background
- Image processed & cross referenced with database to produce species ID
  - discarding non-leaf images
  - segmenting the leaf from background
  - extracting leaf curvature features over multiple scales

Apps being developed for bird song

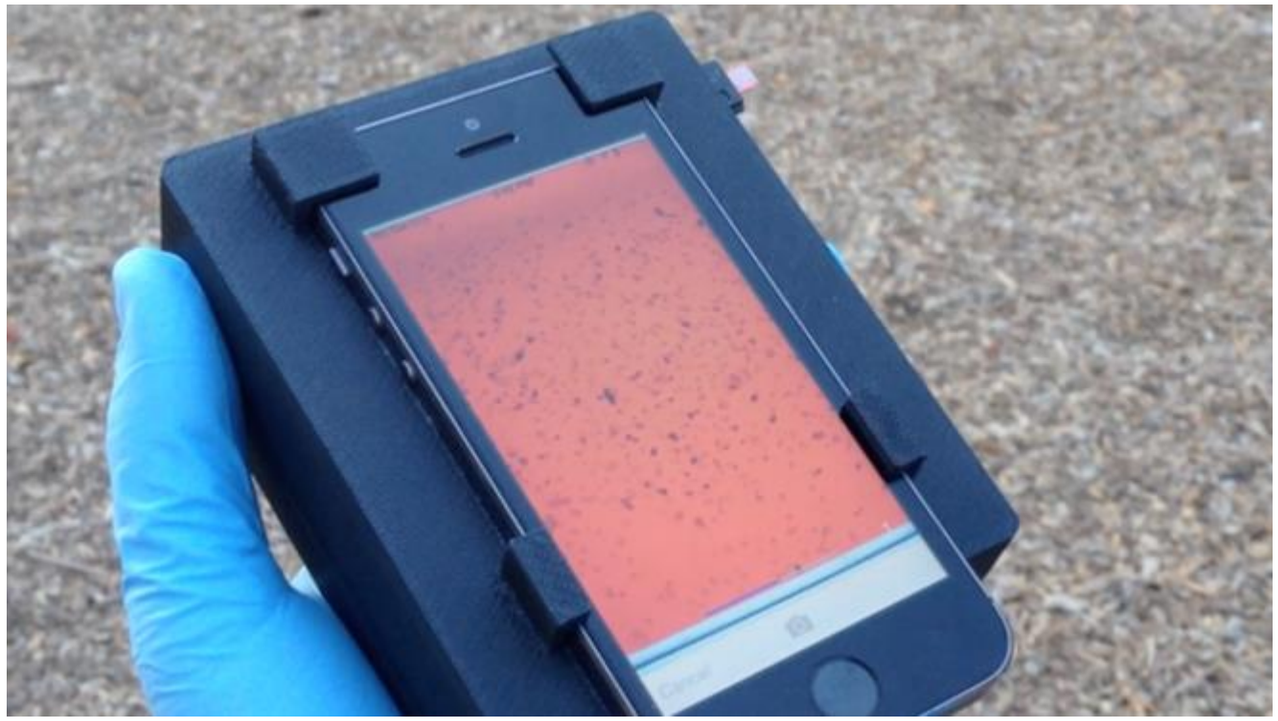




LeafSnap (UK & US edition)

NATURAL HISTORY MUSEUM

Smart phone video analysis for parasite identification



- Pin drop of blood collected
- Loaded into a handheld box
- App analyses movement
- Software predicts the number of parasites
- Informs healthcare worker whether they are suitable for drug treatment

CellScope System

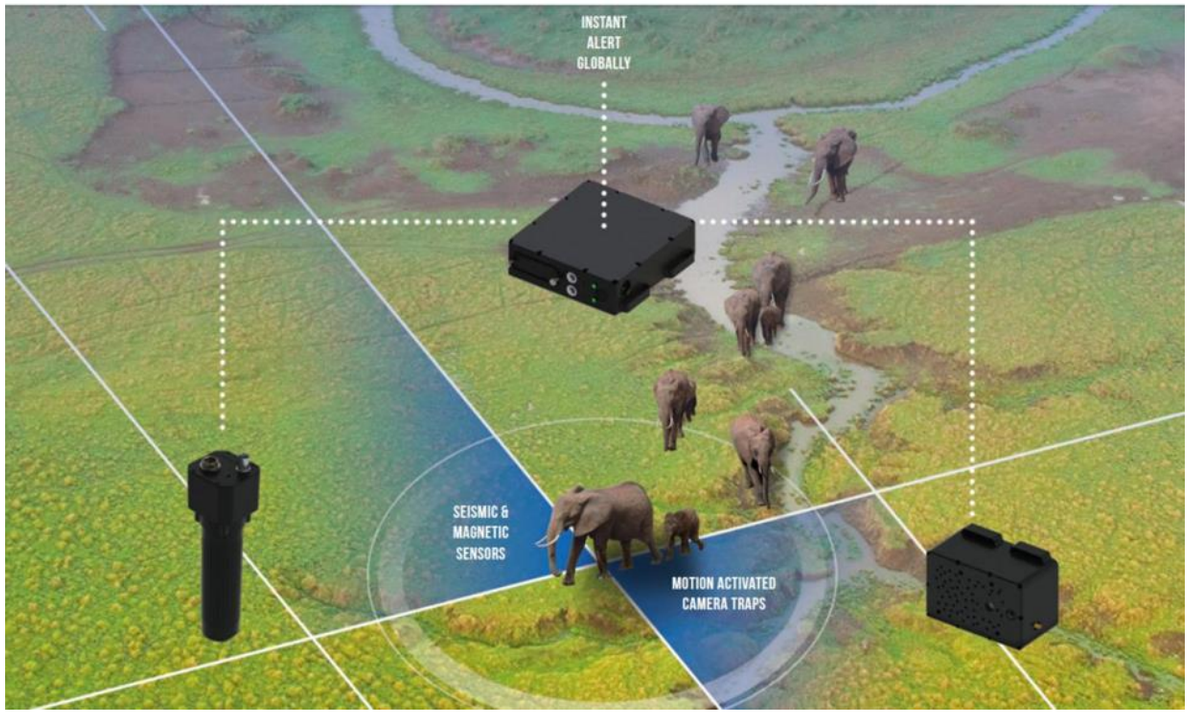
- Required very little training
- Potential applications for many parasitic helminths
- e.g. river blindness and elephantiasis, even malaria & TB
- Used in trials on *Loa loa* ("eye worm") in Cameroon

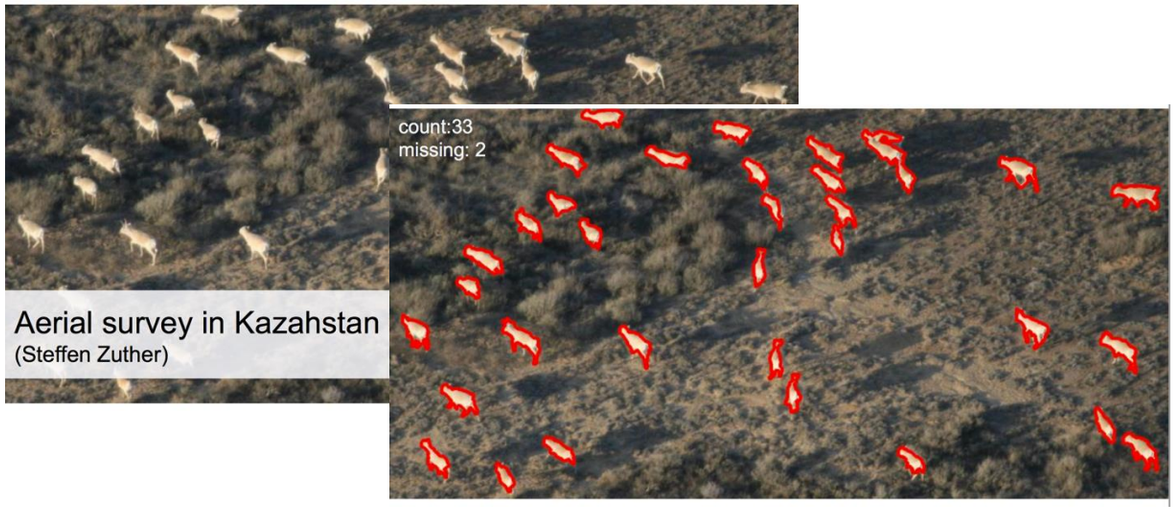
NATURAL HISTORY MUSEUM

Michael V. D'Ambrosio et al. DOI: 10.1126/scitranslmed.aaa3480

Camera traps to support real time identification

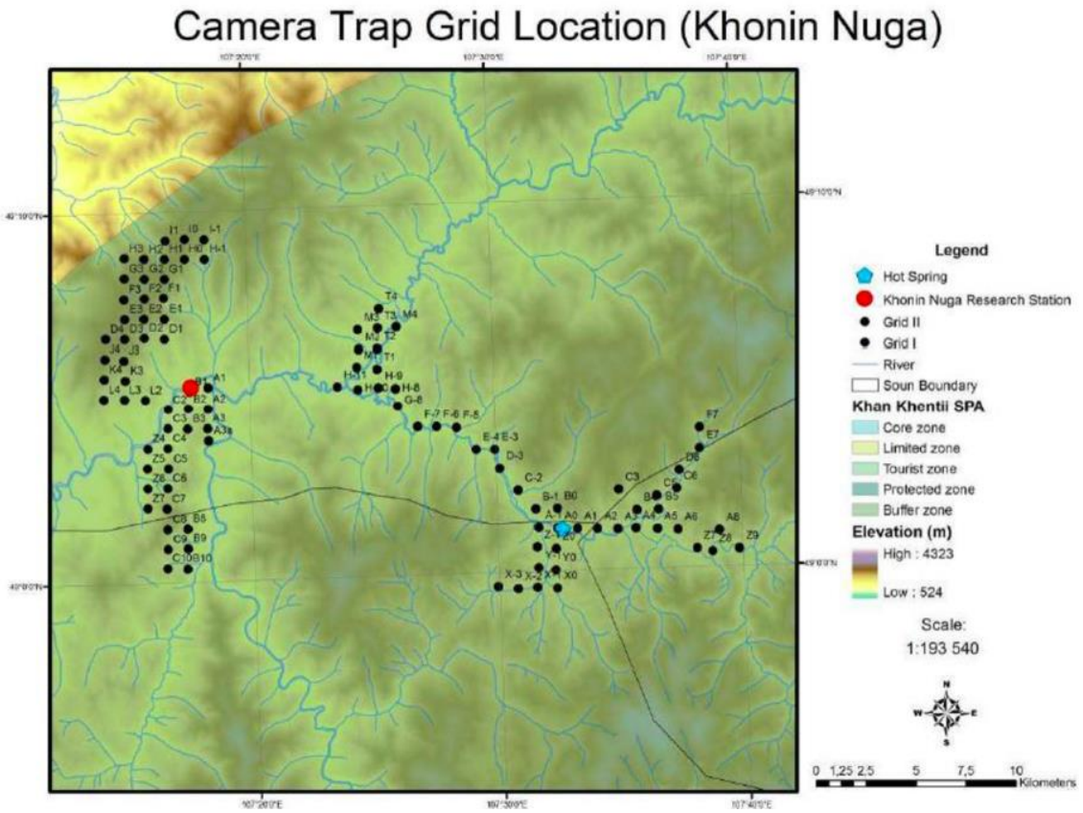
Real-time data





Aerial survey in Kazahstan
(Steffen Zuther)

- Currently applied to field based animal counts
- Usually rely on aerial surveys
- Networks of camera traps provide real time data

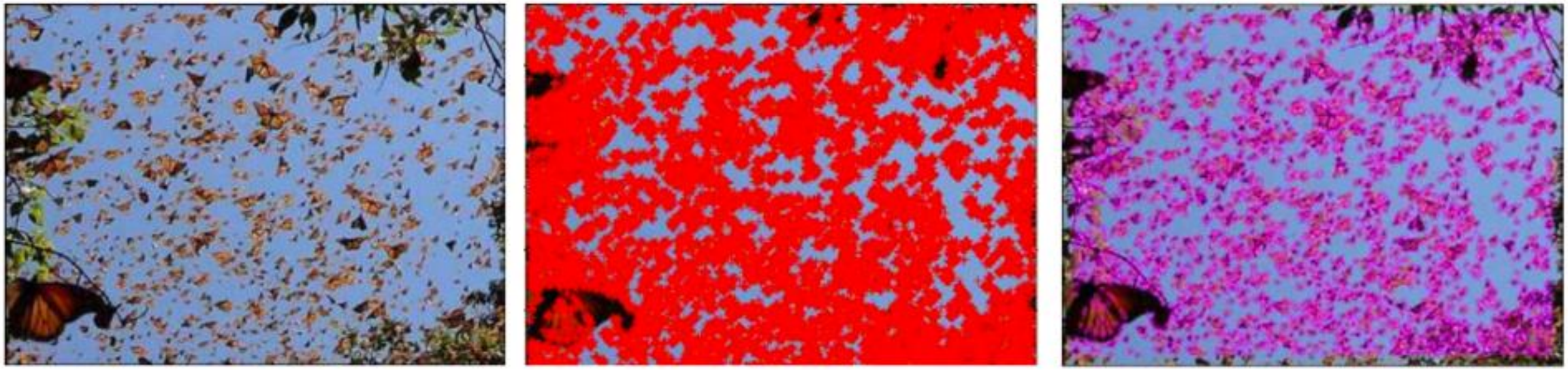


Peter Reinartz & Martin Wegmann 2014, German Aerospace Centre

Image analysis to support animal counts

Sample Images | Extracted Local Features | Automatic Detection Results

Southern right whales



Flying foxes

NATURAL HISTORY MUSEUM

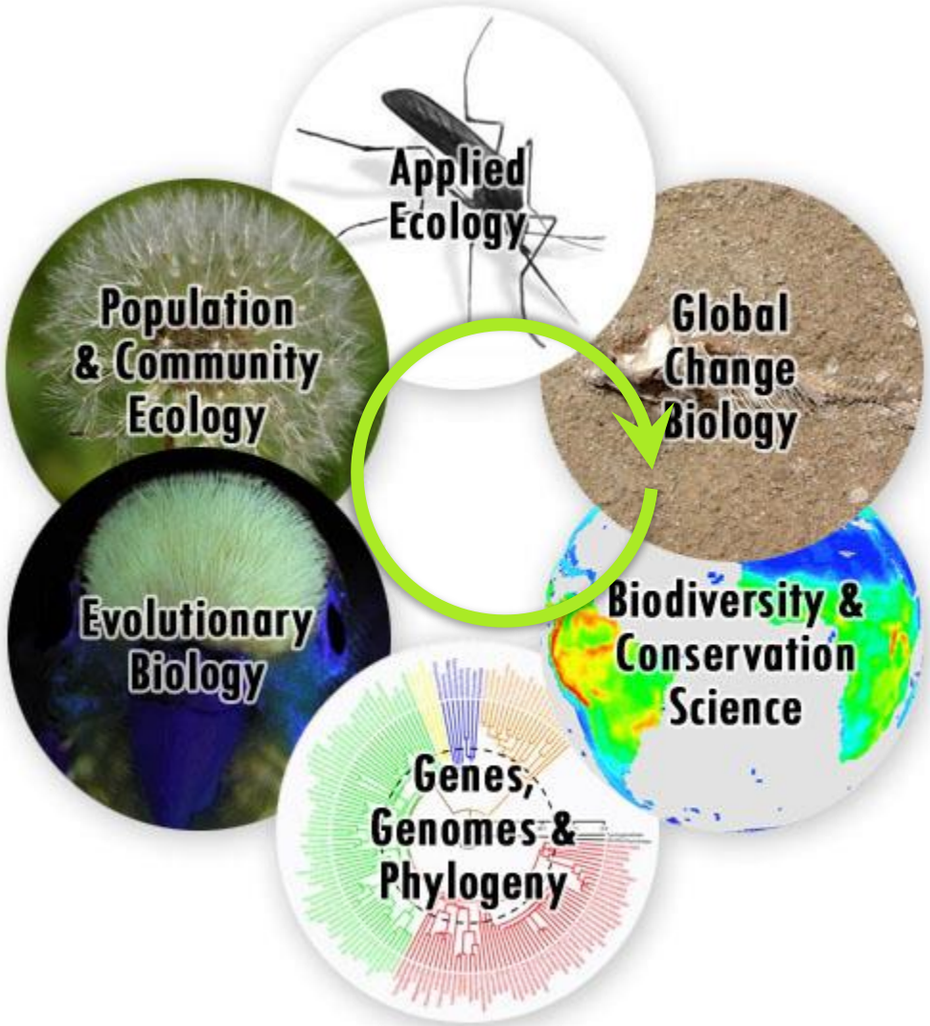
Peter Reinartz & Martin Wegmann 2014, German Aerospace Centre

5. Adapting to the future

- Lessons from past activities
- Sustainability

Lessons from past informatics activities

- **Break out of our discipline, technical & project centric activities** (it is unsustainable, inefficient & bad for science)
- **Integrate & build on exiting programmes where possible** (LifeWatch is a potential umbrella for these activities)
- **Bridge the disconnect between informaticians & users** (make the users informaticians & informaticians users)
- **Use H2020 as a mechanism to achieve integration** (where possible!)
- **Build for the long term, in partnership with our home institutions**



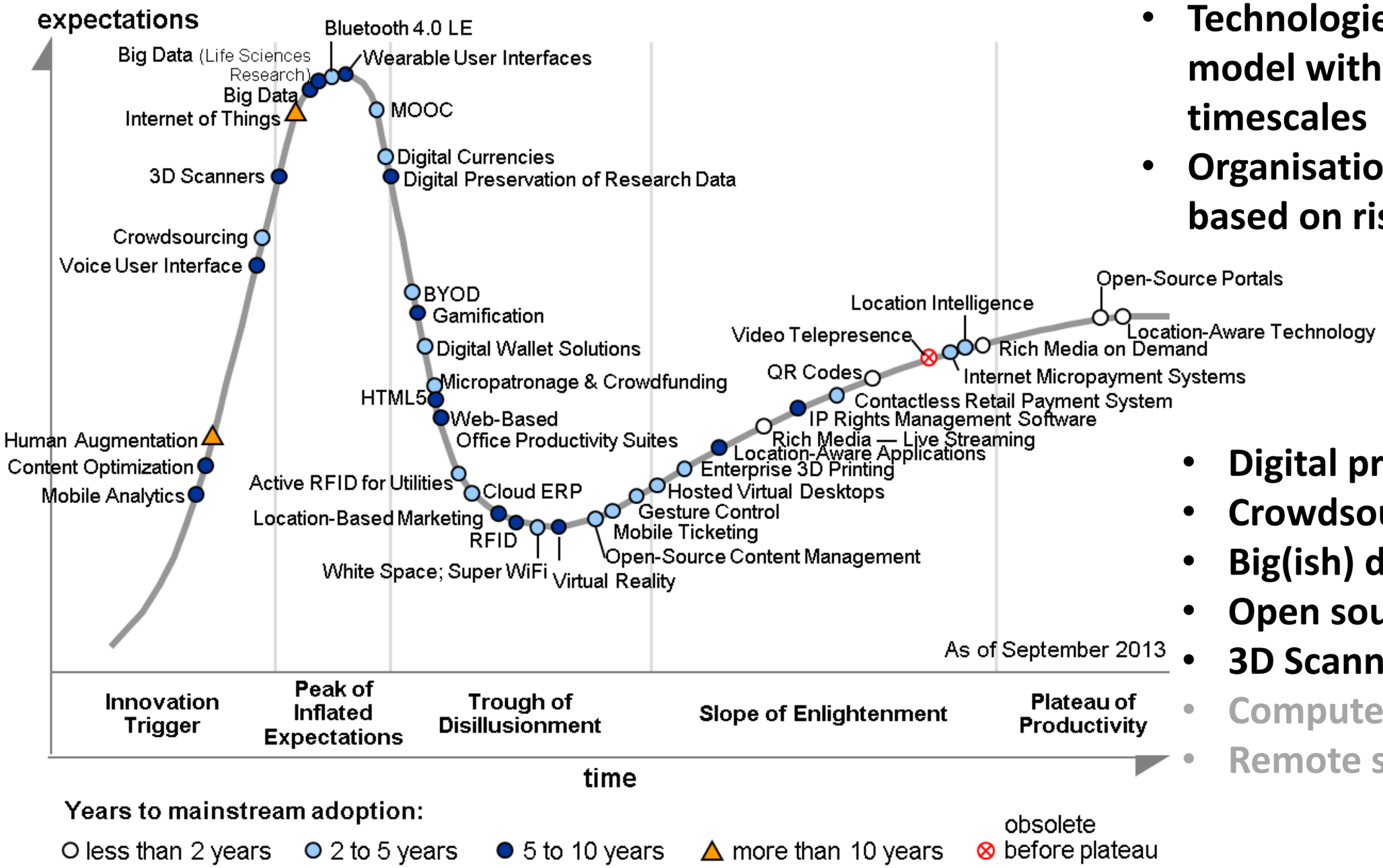
How do we join up these activities?

= recipe for sustainability

NATURAL HISTORY MUSEUM

2013/14 NHM London technology hype-cycle

Compiled every 2 years in NHM Techwatch report



- **Hype cycles used to plot emerging technologies**
- **Technologies follow this model with varying timescales**
- **Organisational adoption based on risk**

- **Digital preservation**
- **Crowdsourcing**
- **Big(ish) data**
- **Open source portals**
- **3D Scanning / printing**
- **Computer vision**
- **Remote sensing**

Using digitisation to enhance public engagement with natural history specimens:

- **Augmented reality in gallery**
- **Animated content for pre- and post-visit apps**