NATIONAL MUSEUM *of* NATURAL HISTORY

Smithsonian

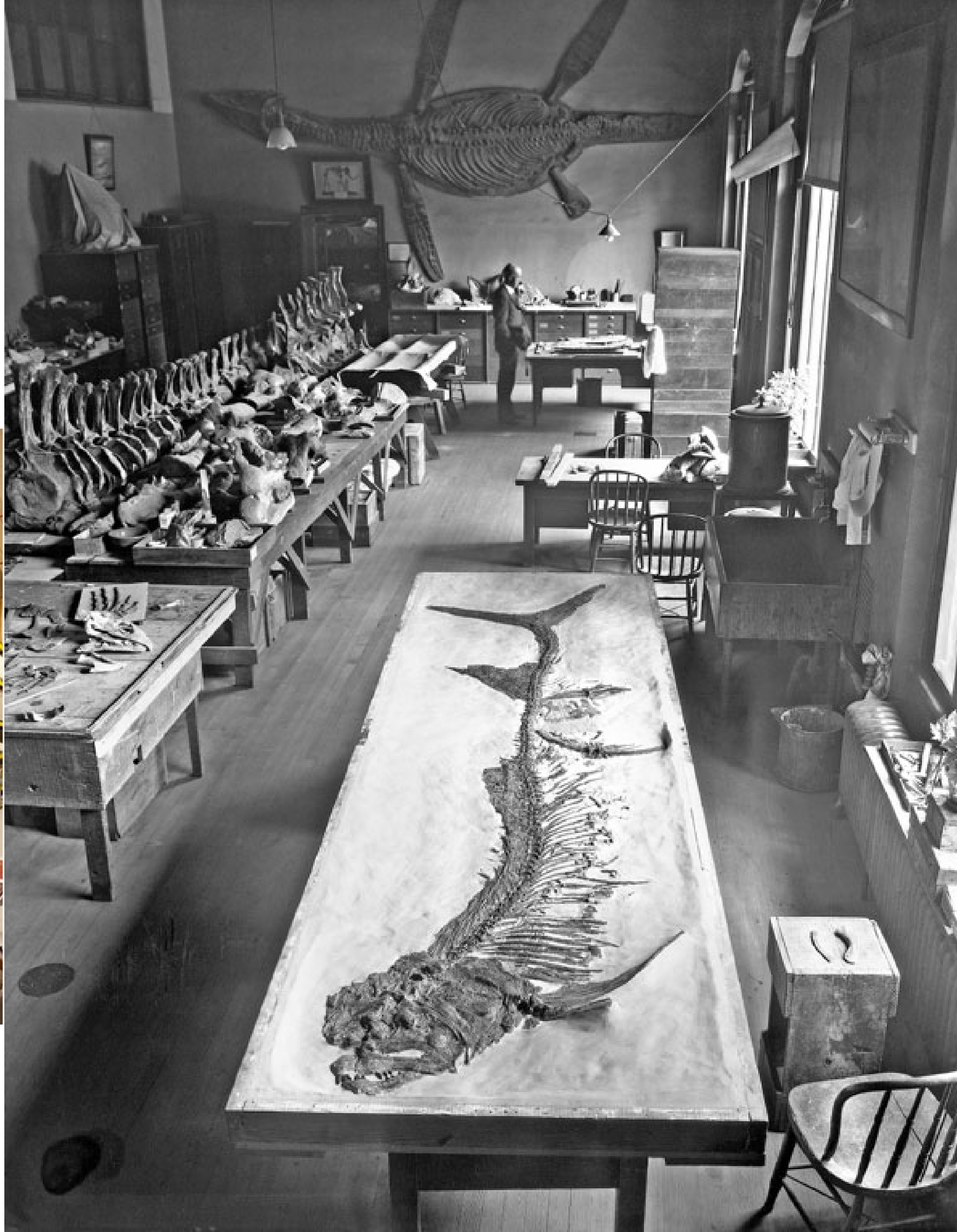
Advancing Digitization of Biodiversity Collections (ADBC)

Smithsonian NMNH efforts to integrate collections based research

Rebecca N. Johnson AM PhD

Chief Scientist and Associate Director for Science, Smithsonian NMNH

Our collections – *the* most valuable research infrastructure



Why digitize?

Benefit to Collections Management and Collections Care:

- Access to collections and collection data
- Allows data to be shared virtually saving a trip to the collection
- Provides a ‘snapshot’ of specimen condition
- Potential for less specimen handling after initial digitization

Benefit to Research and Researchers:

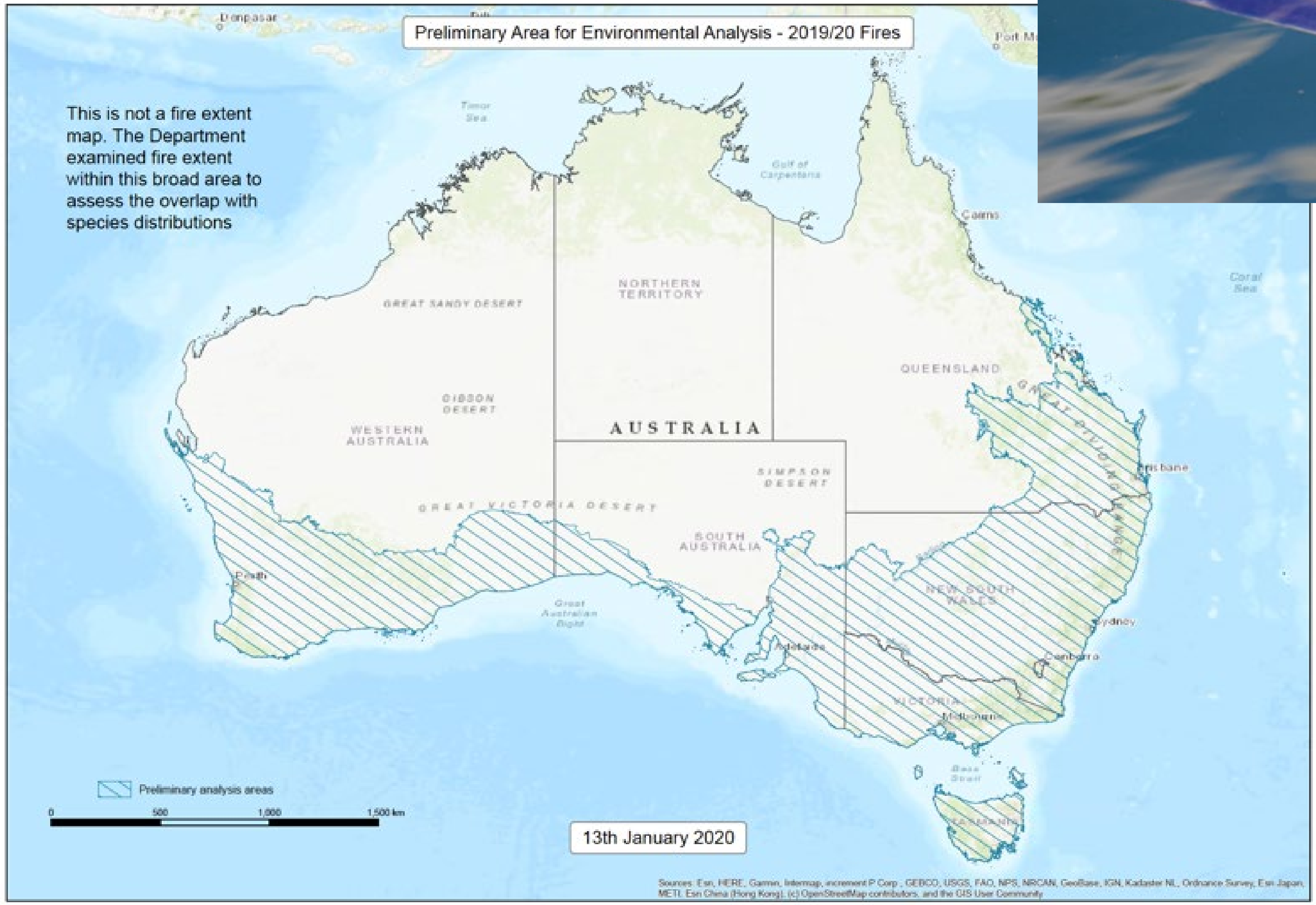
- Expansion our understanding and conservation of plant diversity
- Digital basis for floral and faunal conservation assessments
- Contribution to “Big Data” efforts - visualization, analysis, and modeling

Opportunities for Innovation:

- Improved access for the public, educators, scientific community
- Rapid Capture
- Deep Learning
- Open access Smithsonian initiative

Why digitize?

PHOTOGRAPH BY NATHAN EDWARDS



https://www.environment.gov.au/biodiversity/bushfire-recovery/research-and-resources

146.3 million reasons to digitize





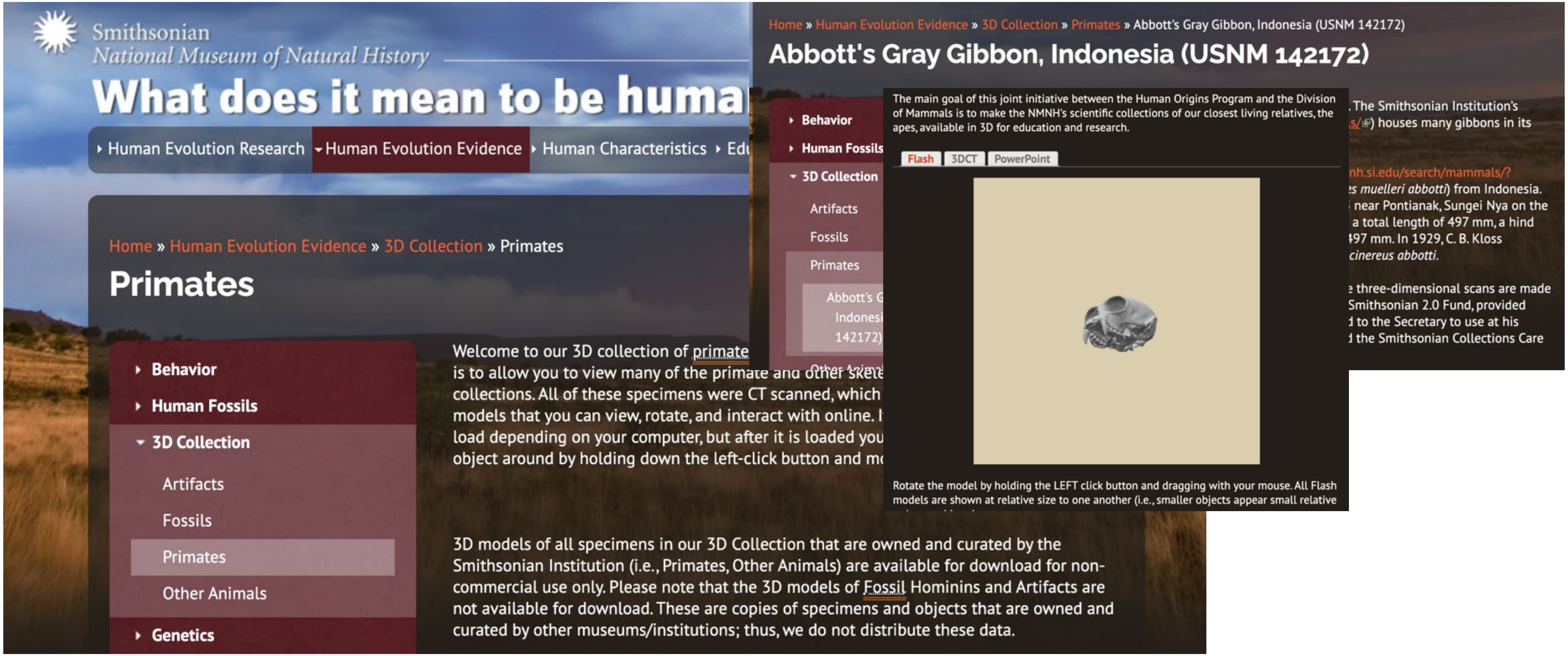



Our collections – *the* most valuable research infrastructure

- The case for digitization
- Rapid digitization supporting science
  - “Pilot” to “Production”
- Science through Deep Learning
- Lessons learned
- What’s next for NMNH?

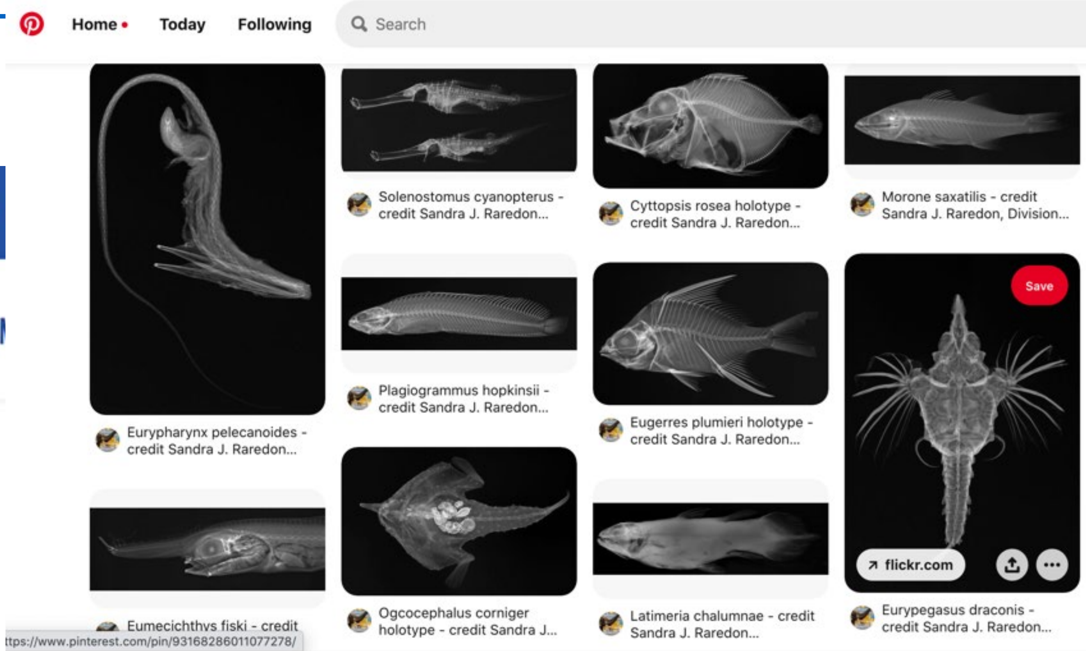
3D Primate Collection

https://humanorigins.si.edu/evidence/3d-collection/primates



X-Ray vision – Fish inside out (images prepared for research)

https://ocean.si.edu/ocean-life/sharks-rays/x-rays-fish-reveal-diversity



Smithsonian Language

OCEAN FIND YOUR BLUE

OCEAN LIFE ECOSYSTEMS PLANET OCEAN THROUGH TIME CONSERVATION HU

SLIDESHOW

X-Rays of Fish Reveal Diversity

Scientists in the **Division of Fishes** at the Smithsonian's National Museum of Natural History use X-ray imaging to study the complex bone structure and diversity of fish. This image gallery showcases X-ray images of sharks, their relatives, and bony fish, revealing how some fish have skeletons built from cartilage while others are built from bone.

In 2012, the National Museum of Natural History displayed "**X-Ray Vision: Fish Inside Out**," a temporary exhibit that showcased fish evolution and diversity through 40 black and white X-ray images prepared for research purposes. Each X-ray is paired with a photograph of the preserved fish specimen, demostrating the value of radiography as a means of study that does not damage or destroy the specimen. **See the touring schedule** to find out where this exhibit will be shown next, through 2015.

To see even more photos from the exhibit, visit **Encyclopedia of Life's X-Ray Vision Collection** .

Case Studies for Digitizing collections with Rapid Capture
https://shiny.si.edu/massdigi/

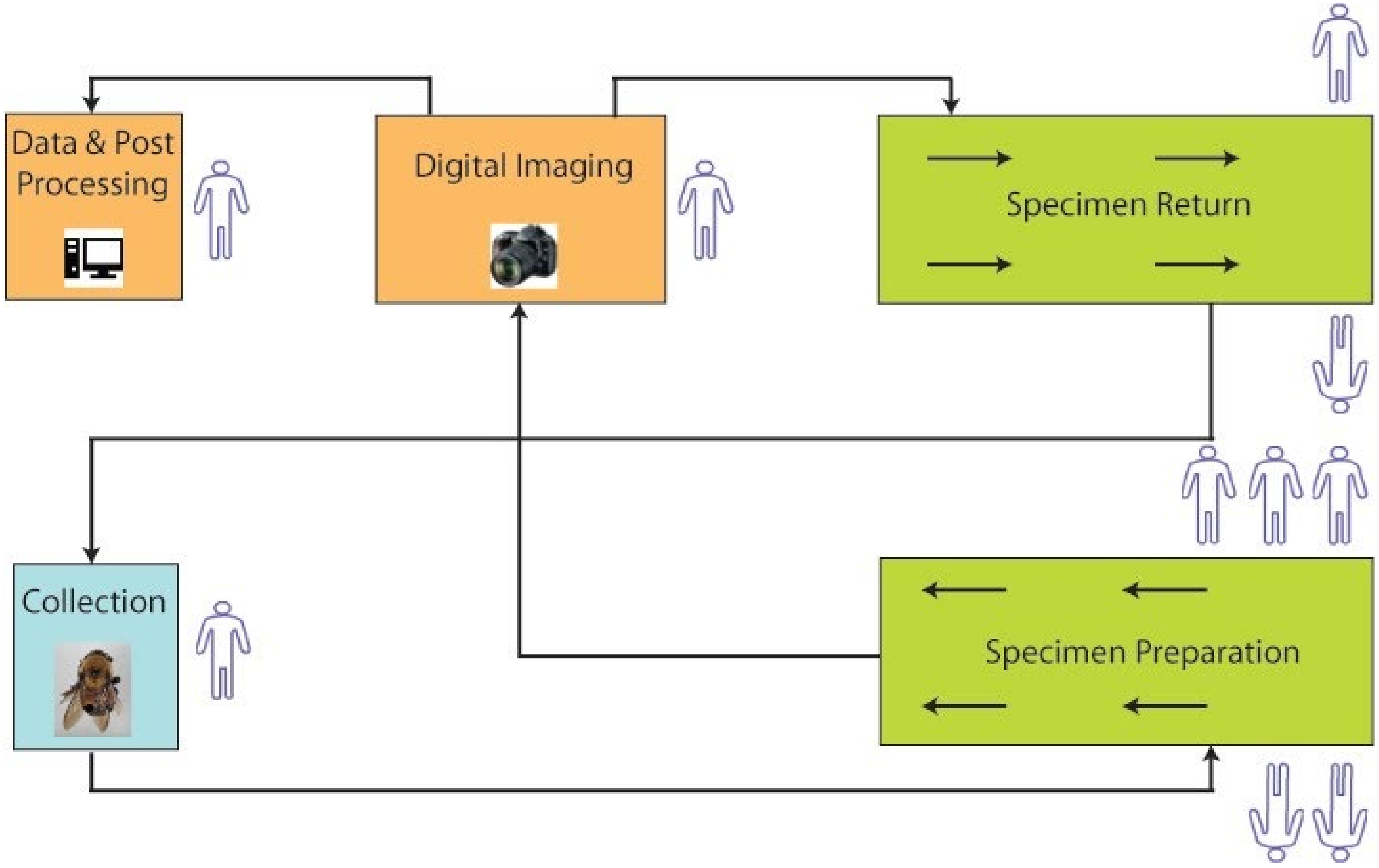
1. Bumblebees and Carpenter Bees (Pilot 2014; Production 2019)
2. NMNH Herbarium rapid capture initiative (2015-)
3. Paleobiology (2017-)
4. Invertebrate Zoology initiatives (2017-)
5. Lessons learned and future projects

Case Study 1:
Bumble bee and Carpenter Bee Rapid Capture

Pilot Goal:
Develop a **Workflow** to speed the process of digitization of Entomology pinned collections; image specimens (including labels) to **create complete records** using high resolution photographic equipment **for 14,400 bumble bees.**

Rapid Capture

- Staff intensive, 10 FTE staff, interns, and volunteer
- Expensive, $4.50 - $8.50 per specimen with imaging done on contract
- Unsustainable at that pace/larger collection



Bumble bee Rapid Capture: Pilot + Production

- Pilot (2014) 44,047 specimens captured in 8 weeks
- Production (2020) 30,020 specimens
- 34 million specimens in Entomology collection……

Case Study 2:
Digitizing the US Herbarium:
5 Million Botanical Specimens in 5 years

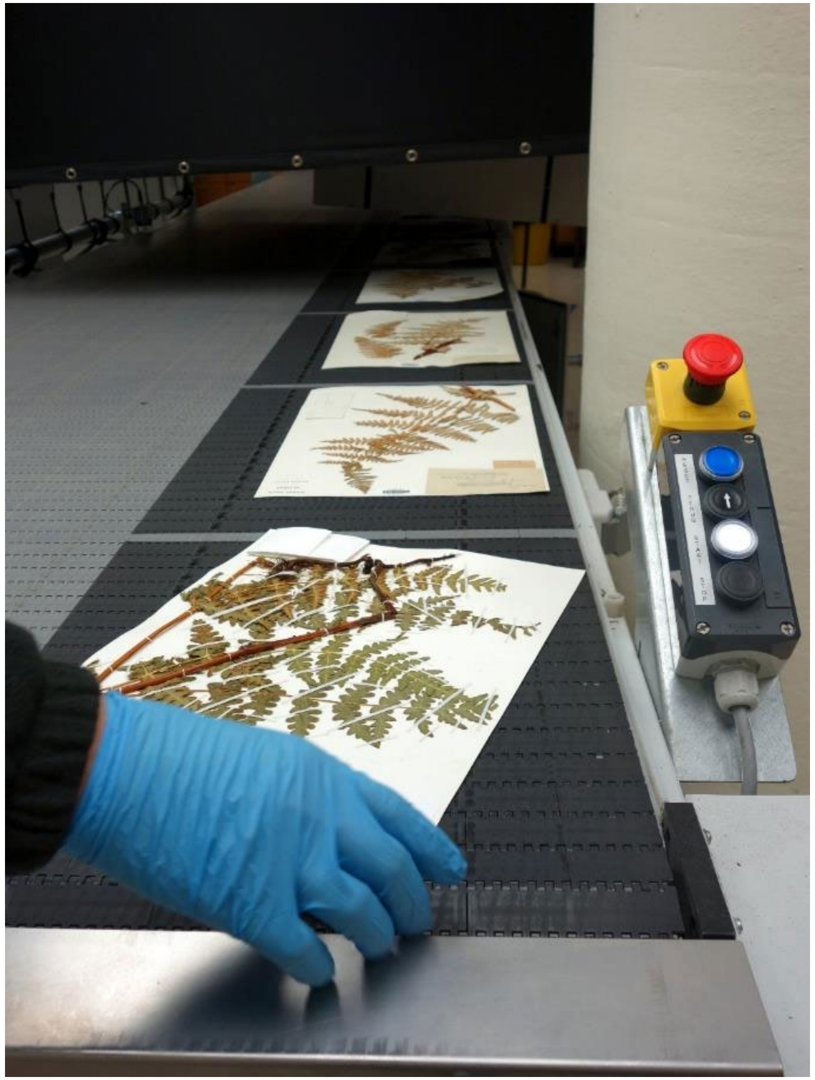
Sylvia Orli
US National Herbarium, NMNH
IT Manager
Botany

The Botany 'Conveyor Belt'

- Start date – Sept. 15, 2015
- One Herbarium sheet every 4-6 seconds
- Goal is 500,000 sheets in 6 months
- Including all secondary costs it is approximately $1.00 US per sheet (not including staff time)
- 4 technicians are currently preparing the specimens
- **2,938,390 Specimens Digitized (76.12 % of the Project Goal of 3,860,000)**

Deep Learning with Botanical Specimen Images

Deep learning models (convolutional neural networks) can be trained to rapidly classify morphological characters from digitized herbarium sheets.



Developing models to identify characters in the genus *Prunus* associated with key evolutionary events

Prunus avium
(sweet cherry)

Photo credit: Native Plant Trust

Prunus serotina
(black cherry)

Photo credit: Western New Mexico University

Prunus persica
(peach)

Photo credit: Native Plant Trust

Richard Hodel, PhD
Peter Buck Postdoctoral Fellow in Botany
National Museum of Natural History

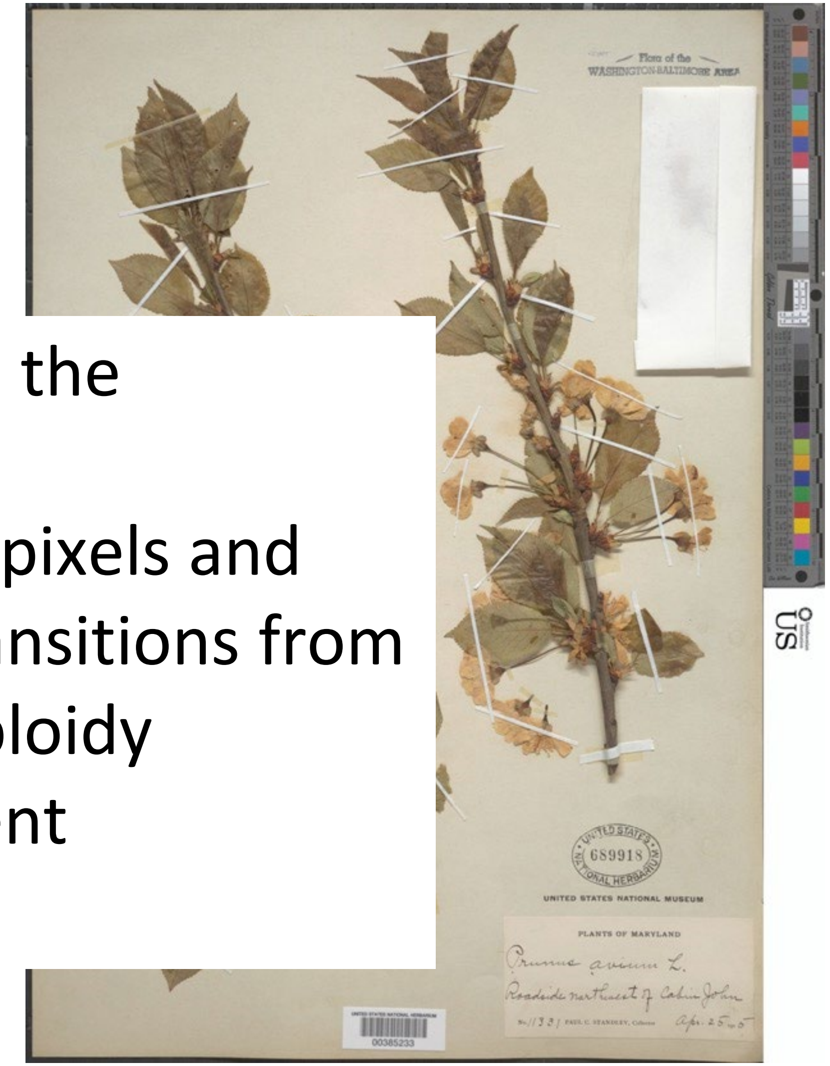
Prunus avium
(sweet cherry)

Photo credit: Native Plant Trust

Prunus serotina
(black cherry)

~1 trillion nucleotides sequenced

61 million pixels



Prunus persica
(peach)

Photo credit: Native Plant Trust

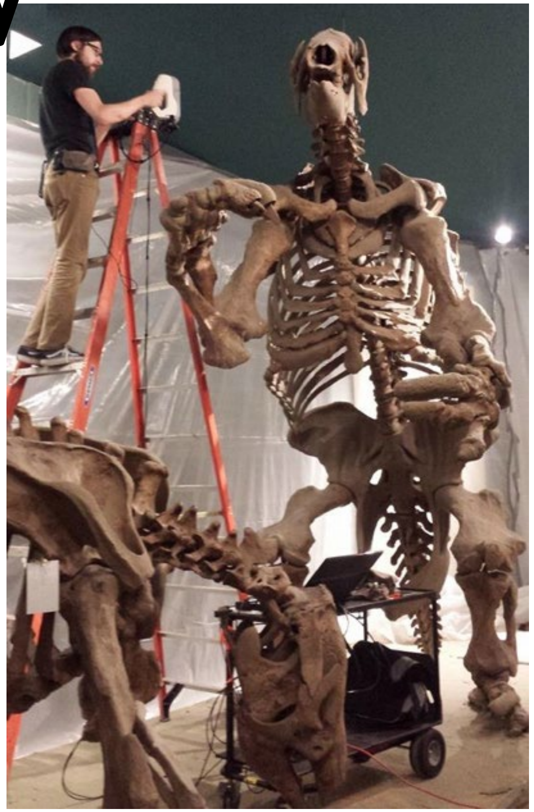
Extra 'eyes' looking at morphology in the collections…….?

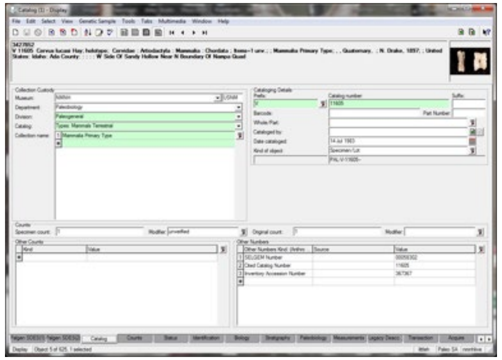
Through deep learning -> combining pixels and nucleotides – can analyze at scale transitions from tropical to temperate climates, polyploidy (genome doubling) events, and ancient hybridization.

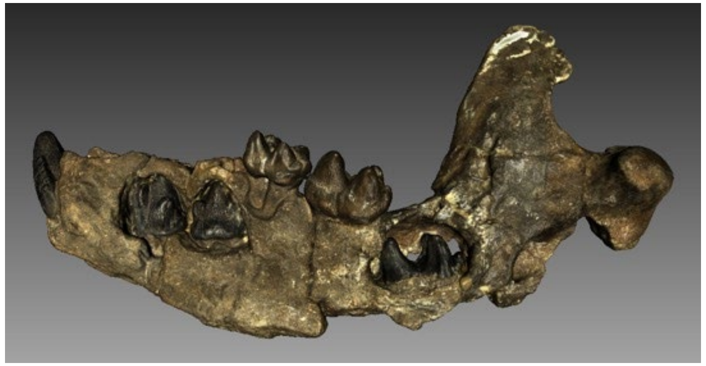
400 species x 61 million = ~24.4 billion pixels

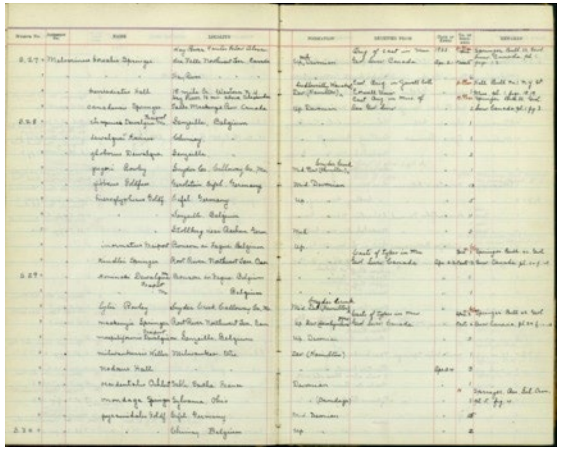
Case Study 3: Paleobiology

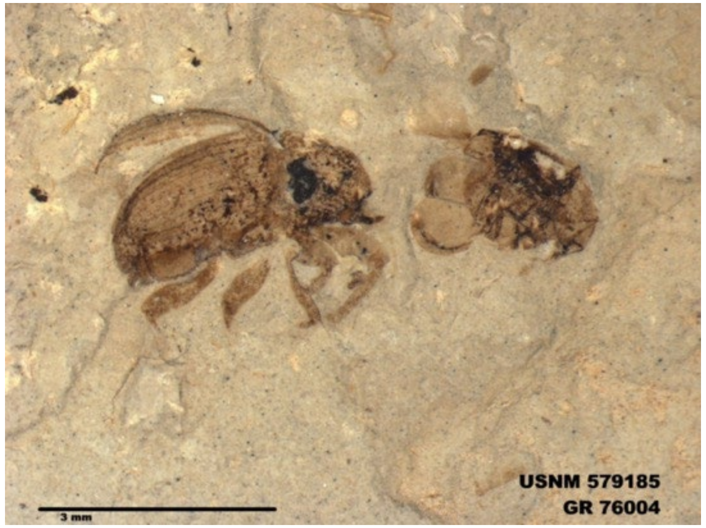
~44 million specimens = ~14 million digital records











Case Study 3:
Paleobiology
- Mass Digi & Data interoperability
- How does our collection contribute globally?

EMu

fossil insect COLLABORATIVE

EPICC

DwC hour

TDWG
ESP IG

Smithsonian
DIGITIZATION PROGRAM OFFICE

Holly Little
Informatics Manager
Paleobiology
Smithsonian NMNH

EPICC TCN Mass Digitization with DPO

Pilot Project (January 2017) - Production Project (Fall/Winter 2017 + Fall/Winter 2019)

- Eastern Pacific Invertebrate Communities of the Cenozoic, NSF ADBC Thematic Collections Network
- 9 Partner Institutions
- 75,077 specimen lots documented
- 145,348 images captured
- >6000 locality records created

http://epicc.berkeley.edu/

EPICC

Case Study 4: Invertebrate Zoology (~1/3 of NMNH collection!!)

Bill Moser

Collections Manager, Department of Invertebrate Zoology

Carol R Butler

Assistant Director for Collections

Anna Phillips

Research Zoologist and Curator of Parasitic Worms, Department of Invertebrate Zoology

Some IZ Digitization projects

- Mollusca Inventory (3+ years, 6 contractors per year): 560,942 records
- Brachyura Inventory (4 years, 2-3 contractors per year): 48,715 records
- Dry Porifera Inventory (1 contractor): 1,000 records
- R/V Albatross Inventory (1 contractor): 5,000 records
- USDA Parasites (validate & import data, curation, rearrangement (7 contractors)): 88,105 records
- Station Data Digitization (1 contractor): 32,000 records
- Transcription from 3x5 taxa cards (staff and volunteers): 20,000 records
- ULL Crustacean collection - Felder donation (1 staff): 17,000 records
- Digitization of NCI collection (1 staff): 20,662 records
- BOEM collection (vouchers, barcoding, and tissue samples (3 staff)): 8,000 records per year
- General staff cataloging (vouchers, collections backlog, etc. (CM staff)): ~12,000 per year

MSC Slide collection

NHB Slide collection

MSC POD 5 (fluid)

Biorepository

MSC POD 2 (Dry)

NHB Dry Collection

Using "the most comprehensive spatially explicit data set available for parasites, projected range shifts in a changing climate, and estimated extinction rates for eight major parasite clades."

SCIENCE ADVANCES | RESEARCH ARTICLE

APPLIED ECOLOGY

Parasite biodiversity faces extinction and redistribution in a changing climate

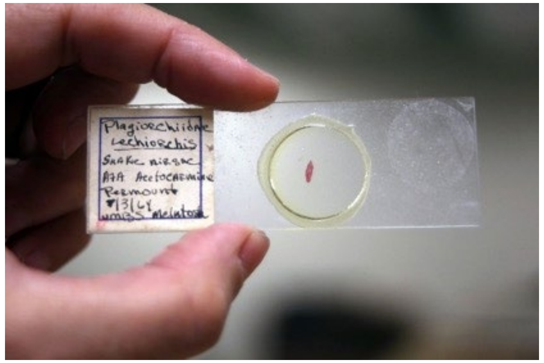
Colin J. Carlson,[1]* Kevin R. Burgio,[2†] Eric R. Dougherty,[1†] Anna J. Phillips,[3†] Veronica M. Bueno,[2] Christopher F. Clements,[4] Giovanni Castaldo,[1] Tad A. Dallas,[5] Carrie A. Cizauskas,[1] Graeme S. Cumming,[6] Jorge Doña,[7] Nyeema C. Harris,[8] Roger Jovani,[7] Sergey Mironov,[9] Oliver C. Muellerklein,[1] Heather C. Proctor,[10] Wayne M. Getz[1,11]

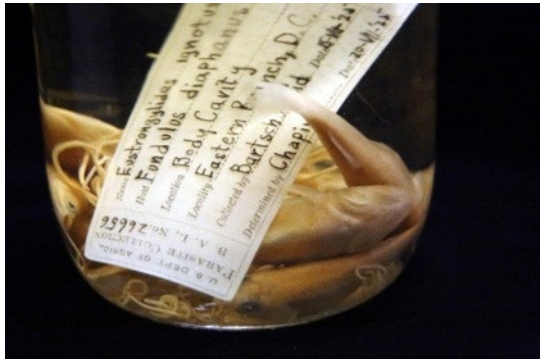
Climate change is a well-documented driver of both wildlife extinction and disease emergence, but the negative impacts of climate change on parasite diversity are undocumented. We compiled the most comprehensive spatially explicit data set available for parasites, projected range shifts in a changing climate, and estimated extinction rates for eight major parasite clades. On the basis of 53,133 occurrences capturing the geographic ranges of 457 parasite species, conservative model projections suggest that 5 to 10% of these species are committed to extinction by 2070 from climate-driven habitat loss alone. We find no evidence that parasites with zoonotic potential have a

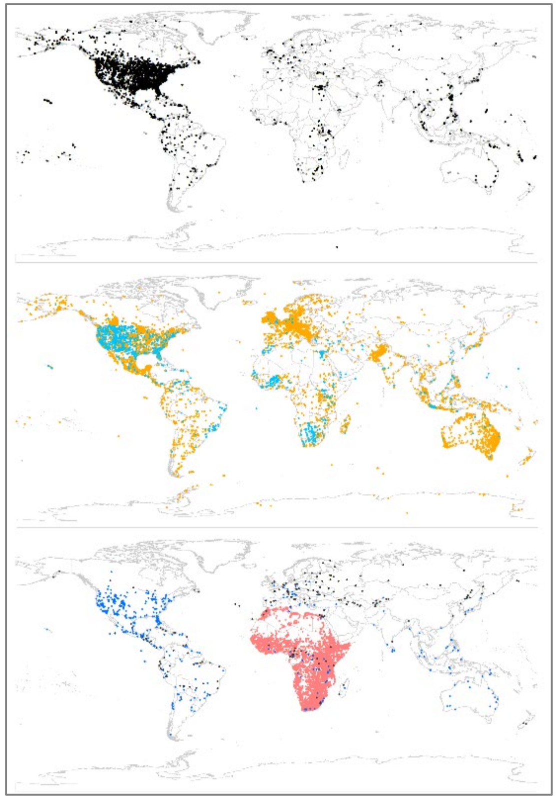
Anna Phillips

Research Zoologist and Curator of Parasitic Worms, Department of Invertebrate Zoology

Parasites Facing Extinction in a Changing Climate







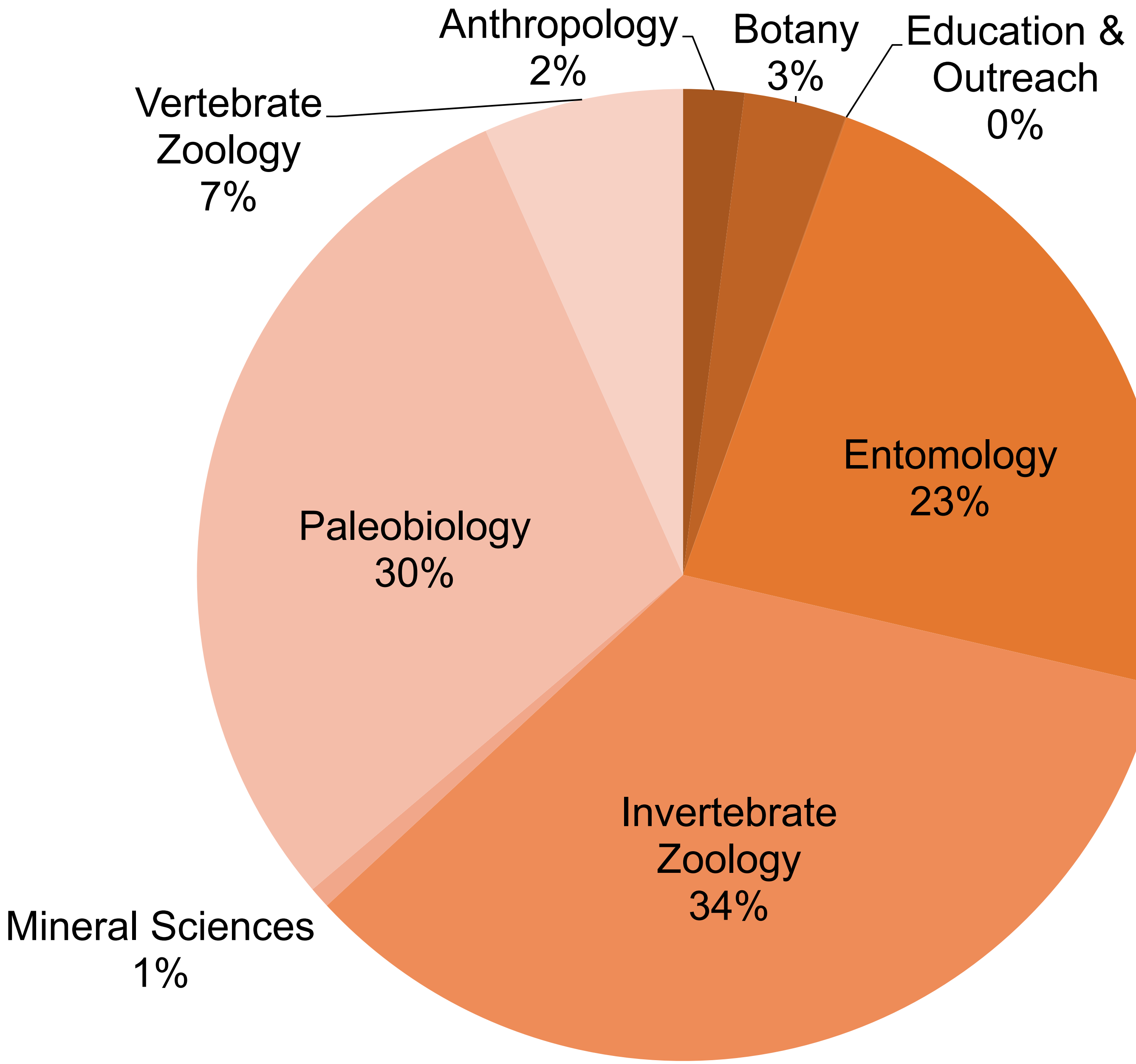
Carlson, C.J., K.R. Burgio, E.R. Dougherty, **A.J. Phillips**, et al. (2017) Parasite biodiversity faces extinction and redistribution in a changing climate. *Science Advances* 3(9): e1602422. doi: 10.1126/sciadv.1602422

- Compiled a spatially-explicit data set; 30,000+ NMNH specimen records (1/3 of the Parasite Collection)
- 5–10% parasite species could be extinct by 2070
- Full specimen records with georeferenced locality data, linkage to DNA sequence data, high-resolution imaging, etc. are the gold standard
- Specimen record completeness was a challenge

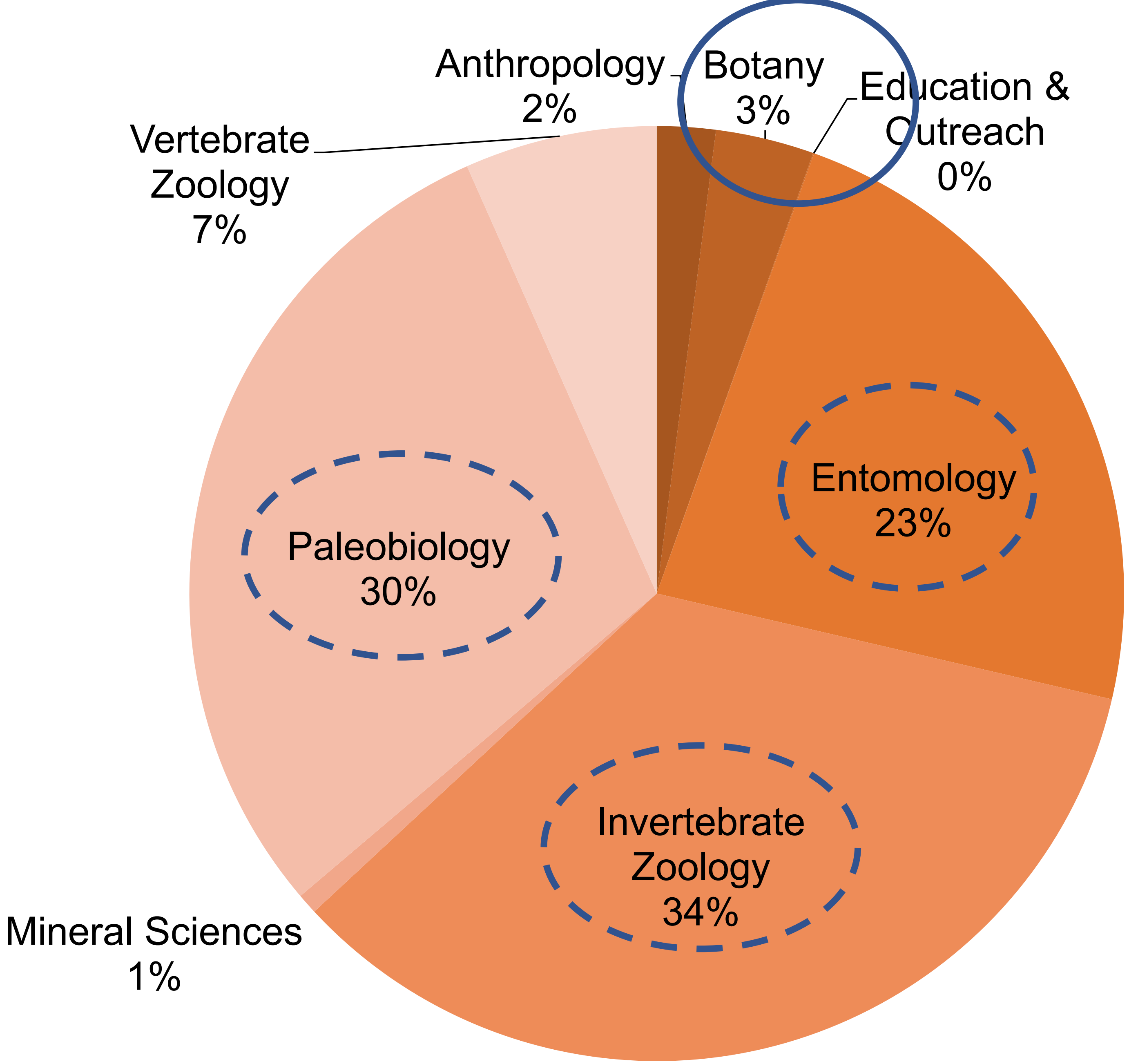
Lessons learned

- This is a resource intensive proposition (accountability is complex and important).
- Many stakeholders (good communication essential!!!).
- Maintain flexibility, and plan carefully and thoroughly.
- Frequent reevaluation and documentation to minimize cost and maximize efficiency.
- Crowd Sourcing Transcription.
- Thinking by collection type and not department……
- **The value of the output is unquantifiable………….**

NMNH Collections



NMNH Collections



What next?

- Estimated **42.8 million** descriptive and surrogate digital records are needed to adequately represent the 146.3 million objects in the collection.
- Interoperability of datasets.
- Importance of Collaboration.
- Funding……..
- Restrictions to funding severely limit what national collection can offer.

Thankyou

- The ‘Bumble bee team’
- The Botany Conveyor Belt team
- The Paleobiology team
- The Invertebrate zoology team

Smithsonian Institution

NATIONAL MUSEUM of NATURAL HISTORY

Smithsonian DIGITIZATION PROGRAM OFFICE

OCIO DATA SCIENCE LAB

Smithsonian

UF UNIVERSITY of FLORIDA

EPICC

The state and relevance of global natural history collections in the 21st Century

Ian Owens, National Museum of Natural History, Smithsonian Institution

Global natural history collections

What is the state of the global collection?

- How many specimens, and of what type?
- Where are the specimens from, and where are they held?
- Are there gaps?

How ready is it to tackle global challenges?

- Do we have collections on the key groups and regions?
- How much of it is digitized?
- How extensive are the genomic repositories?

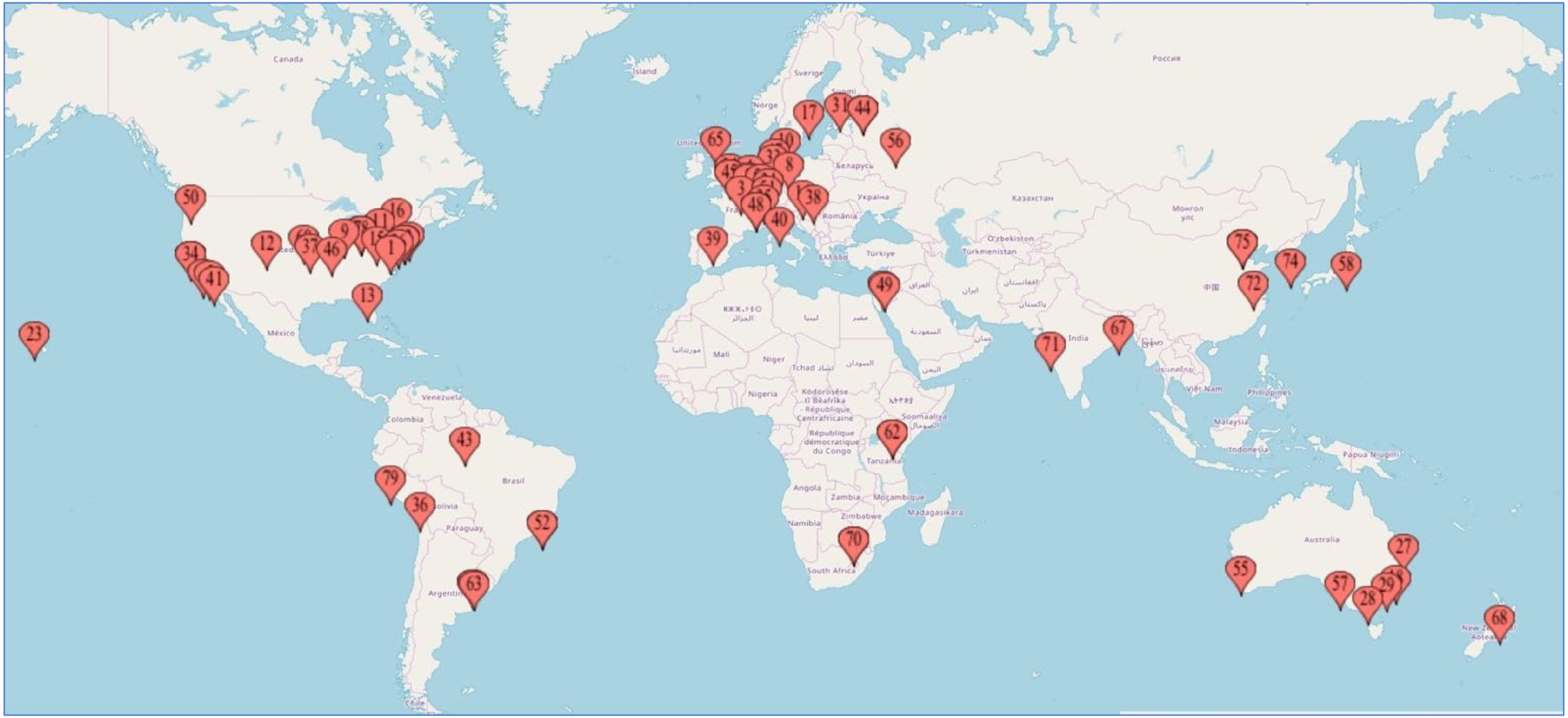
Does the expertise exist?

- Who curates the collection?
- Is there a global shortage of experts?
- Is there a demographic time bomb?

Collaborative network

72 Collection-based Institutions

- Large collections in terms of No of specimens/objects *
- Museums, Botanic Gardens, Universities, Research Inst.
- ca. 33% ‘Global South’, 33% Europe, 33% N. America



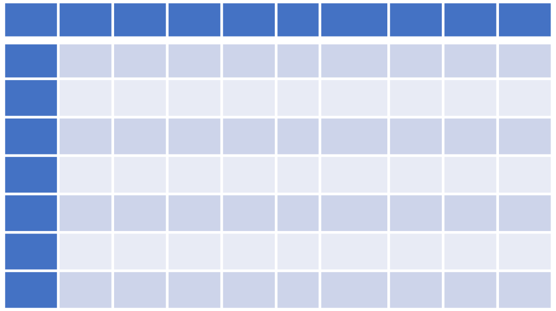
* >10 million specimens for European and N. American museums; >1 million specimens for Botanic Gardens and ‘Global South’ institutions

Mapping the global collection

No of SPECIMENS (log 10 scale)

16 Geographic regions

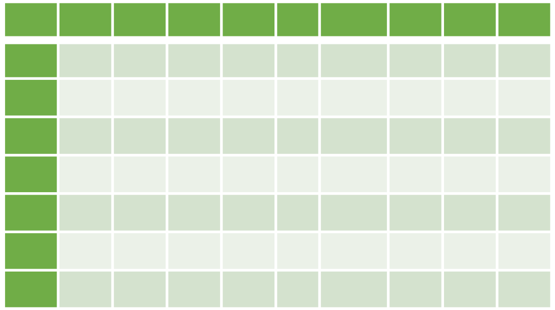
19 Collection categories



STAFF EXPERTISE

16 Geographic regions

19 Collection categories



Role - Research, Collection, Volunteer

Demography – Year PhD awarded

19 Collection categories

Botany

Entomology

Invertebrates
- Molluscs
- Arthropods
- Others

Vertebrates
- Fish
- Amphibians
- Reptiles
- Birds
- Mammals

Anthropology
- Cultural
- Archaeology
- Human biology

Paleontology
- Vertebrates
- Plants
- Invertebrates

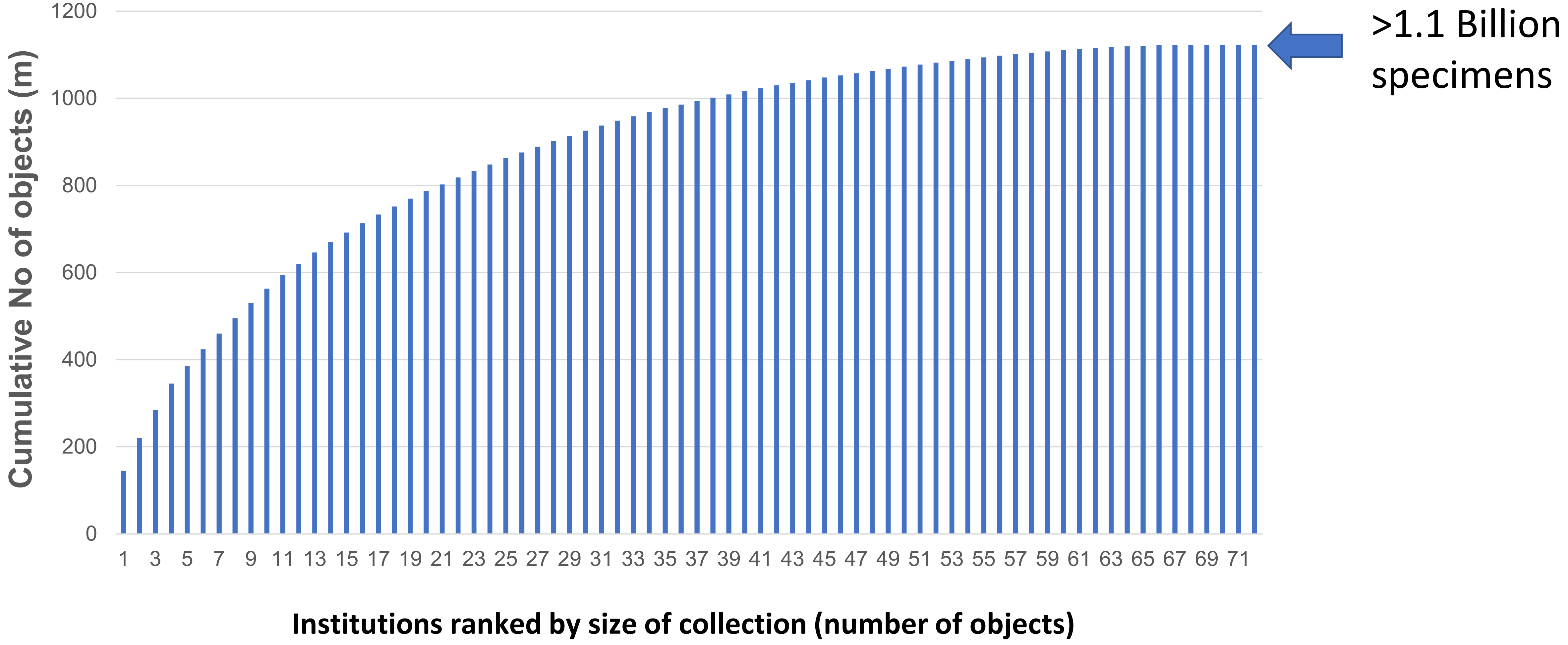
Earth science
- Minerals
- Geology
- Meteorites

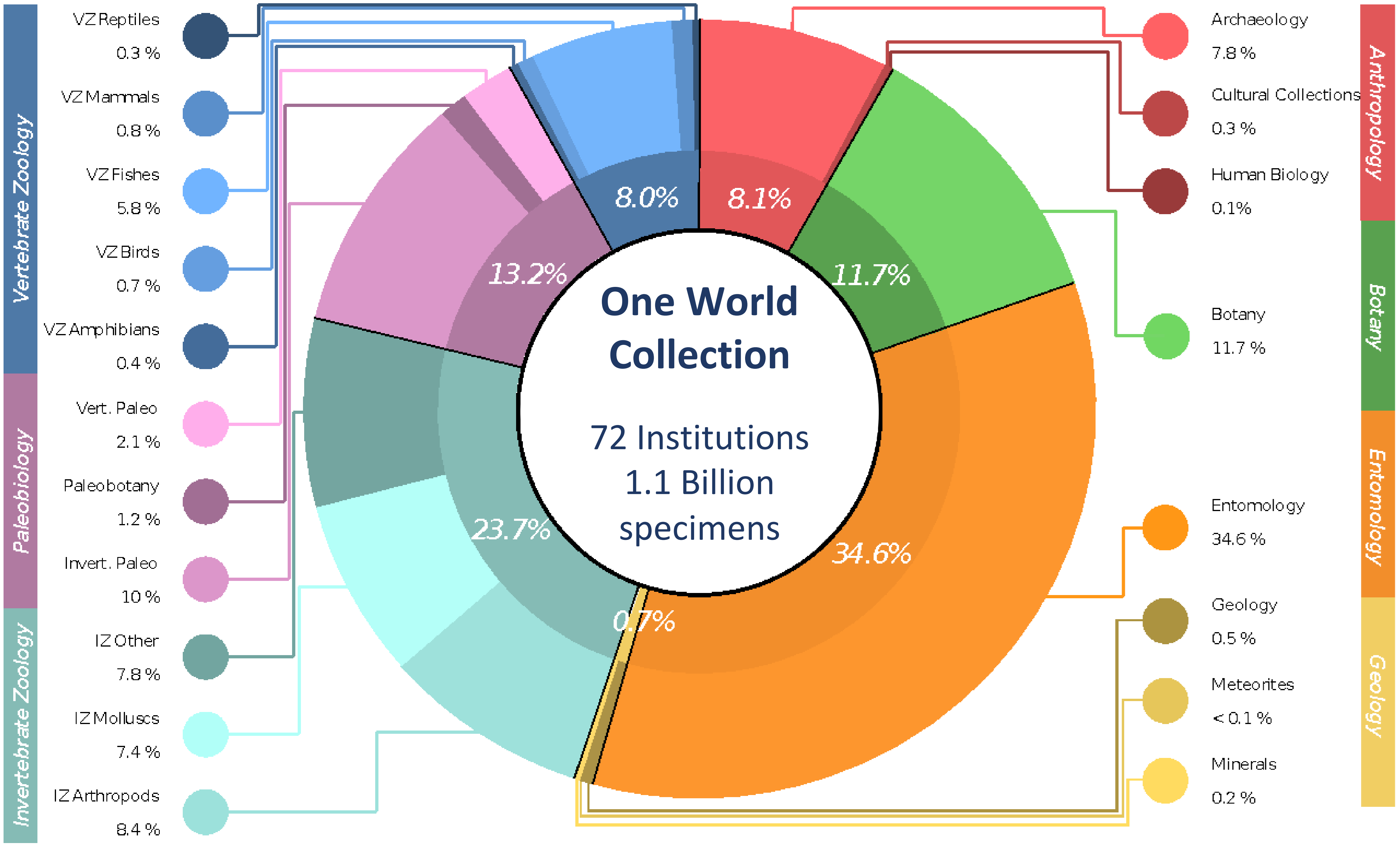
16 Geographic regions

Terrestrial
- Europe
- Africa
- Asia Temperate
- Asia Tropical
- Australasia
- Pacific
- North America
- South America
- Antarctic

Marine
- North Pacific
- South Pacific
- North Atlantic
- South Atlantic
- Indian
- Southern
- Arctic Marine

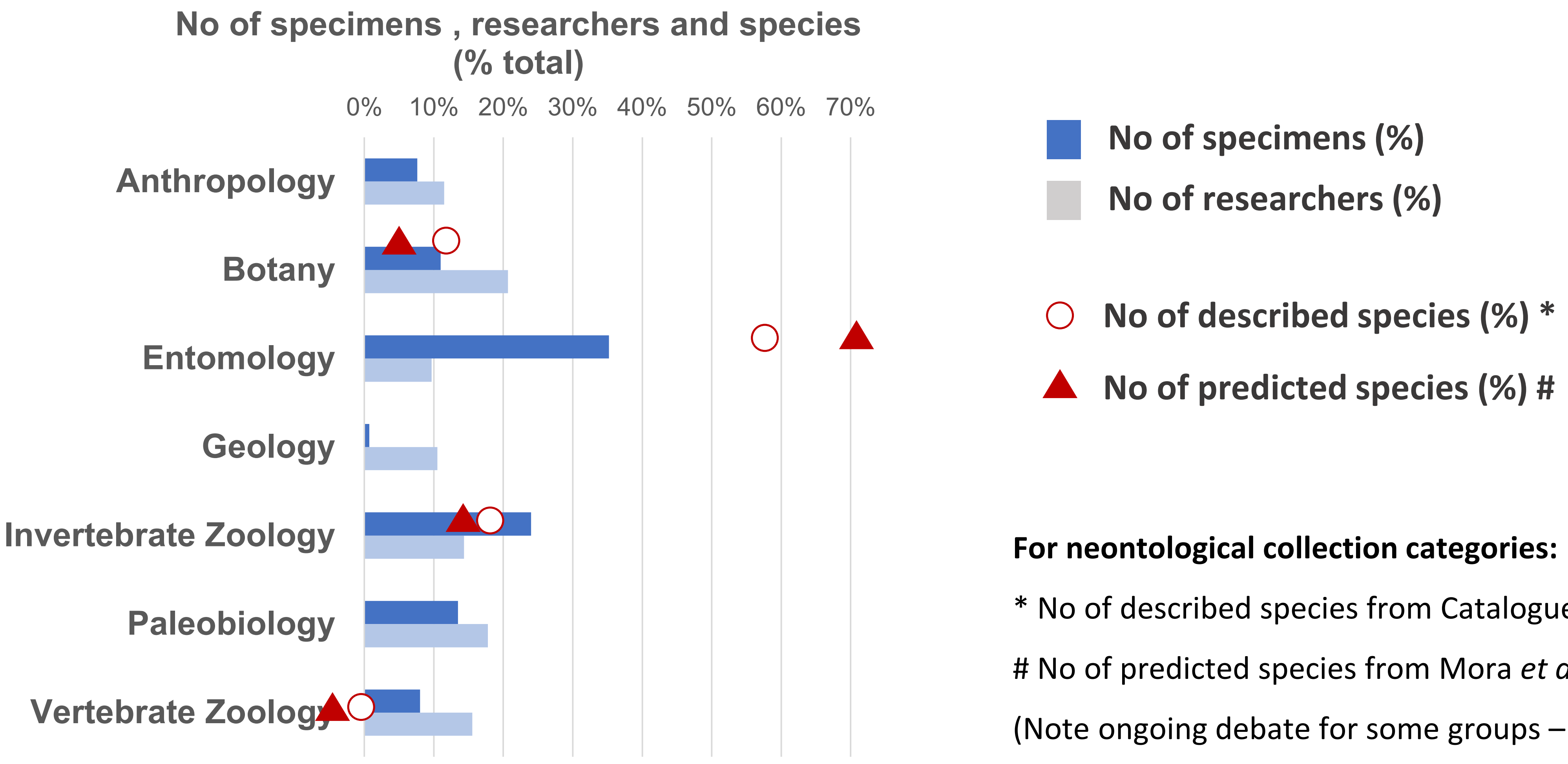
Cumulative holdings of the world's largest Natural History collections





Vertebrate Zoology
VZ Reptiles 0.3 %
VZ Mammals 0.8 %
VZ Fishes 5.8 %
VZ Birds 0.7 %
VZ Amphibians 0.4 %
Paleobiology
Vert. Paleo 2.1 %
Paleobotany 1.2 %
Invert. Paleo 10 %
Invertebrate Zoology
IZ Other 7.8 %
IZ Molluscs 7.4 %
IZ Arthropods 8.4 %
8.0%
8.1%
13.2%
11.7%
23.7%
34.6%
0.7%
One World Collection
72 Institutions
1.1 Billion specimens
Anthropology
Archaeology 7.8 %
Cultural Collections 0.3 %
Human Biology 0.1%
Botany
Botany 11.7 %
Entomology
Entomology 34.6 %
Geology
Geology 0.5 %
Meteorites < 0.1 %
Minerals 0.2 %

Specimens, researchers and species



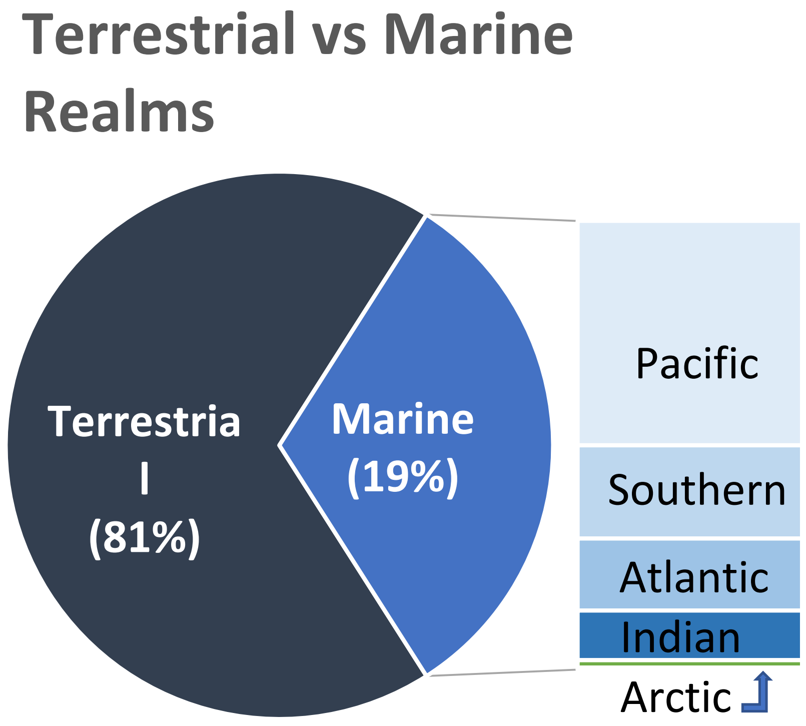
For neontological collection categories:

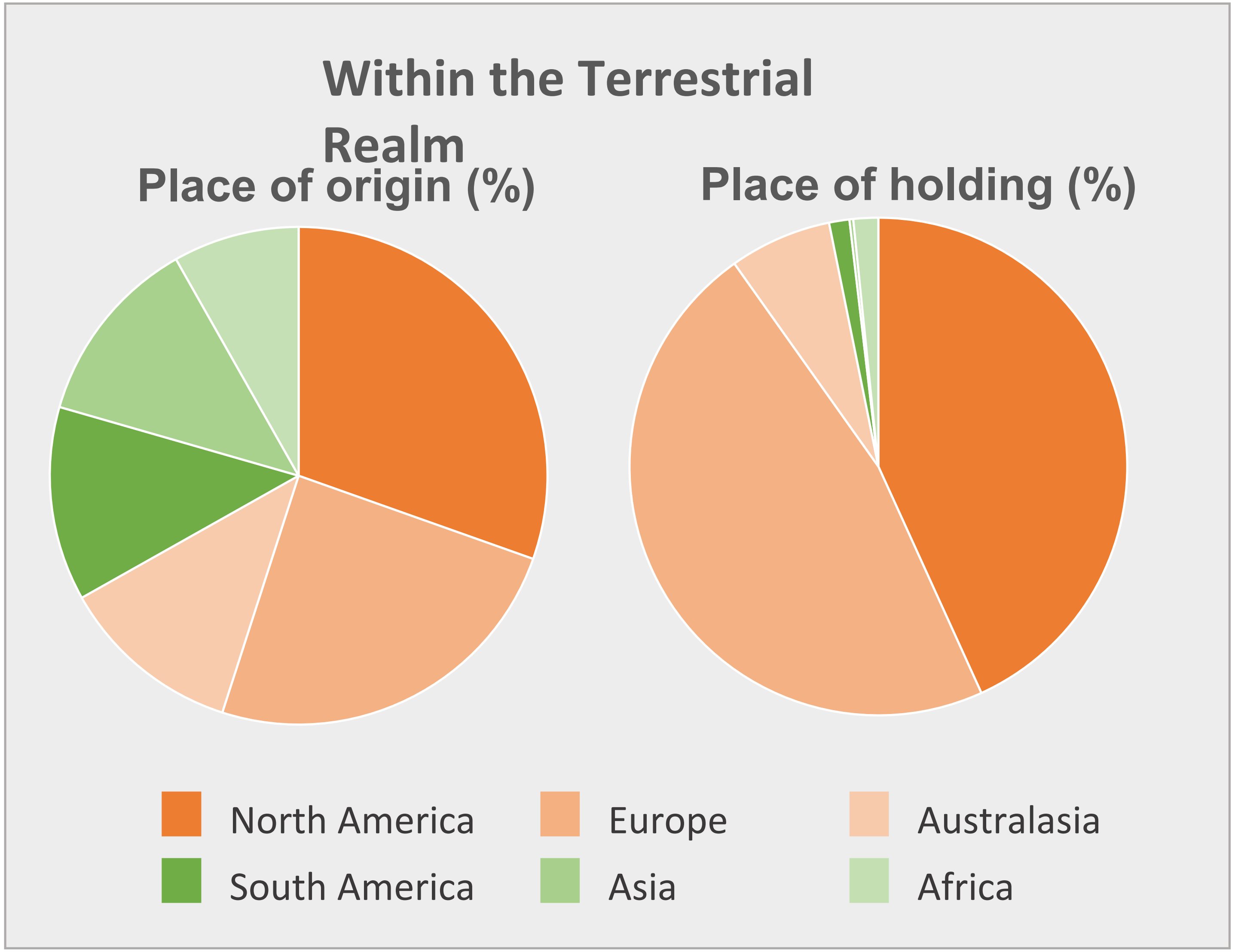
* No of described species from Catalogue of Life

No of predicted species from Mora *et al.* (2011) *PLoS Biol*

(Note ongoing debate for some groups – eg, fungi, nematodes)

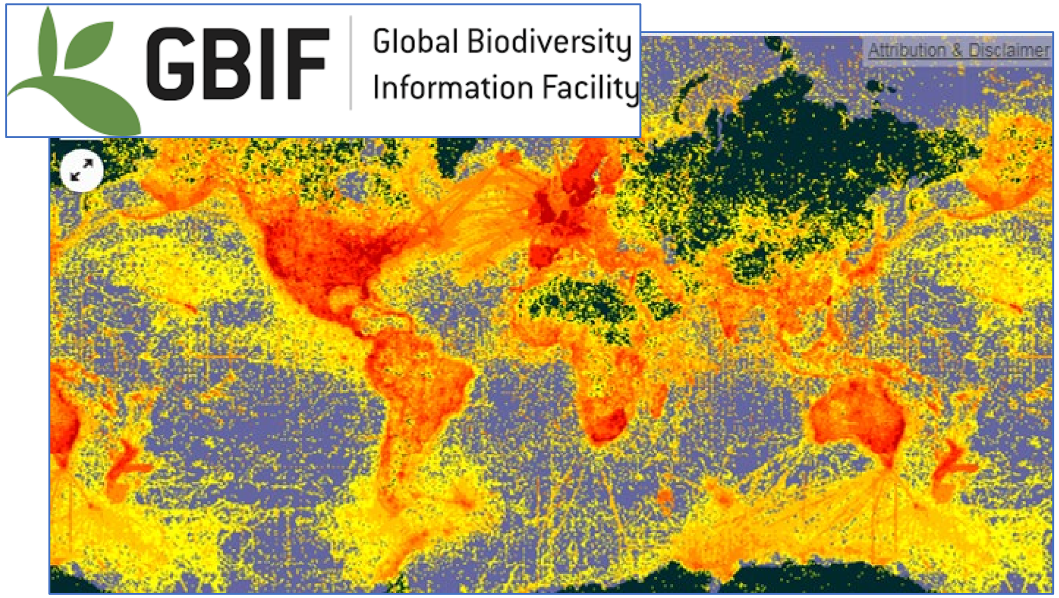
Geographical distribution of specimens





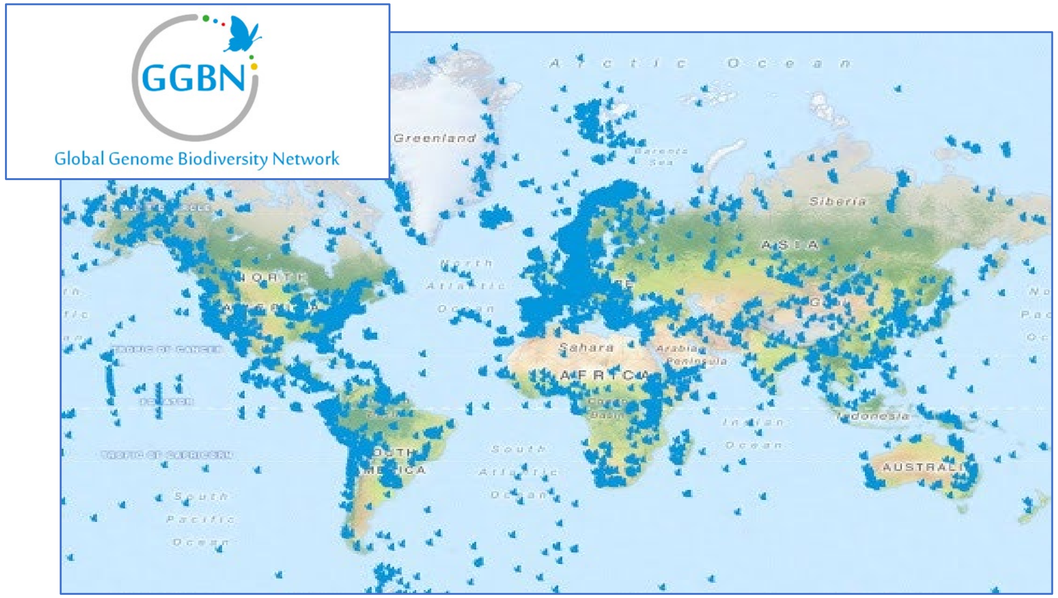
Dark data

Digital records
Global Biodiversity Information Facility (GBIF)
>1 billion records (50% bird obs from Europe & N America)



ca. 16% of specimens are digitally available (50% of these are plants)

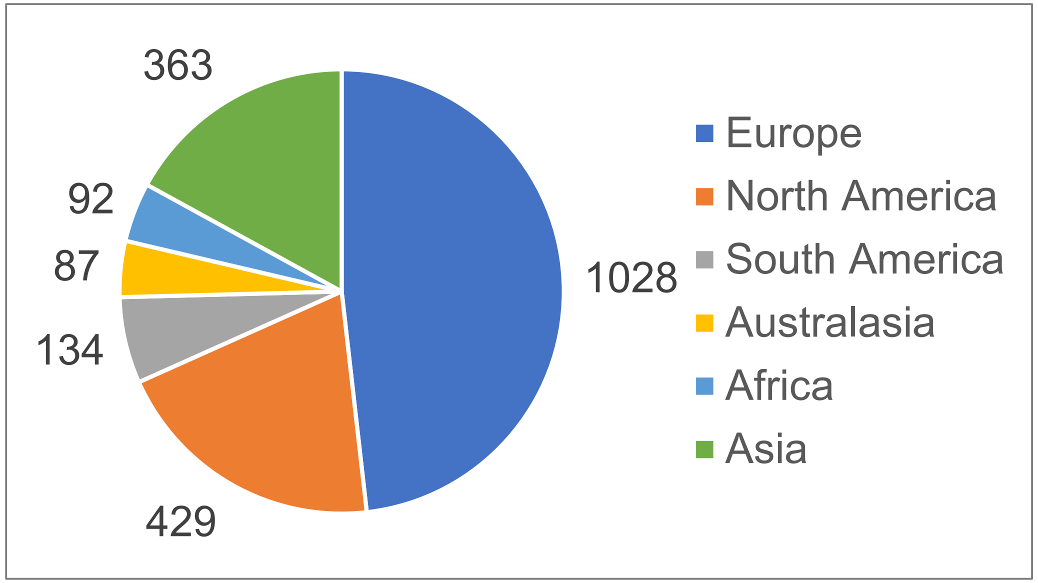
Genomic samples
Global Genome Biodiversity Network (GGBN)
3.7 million records



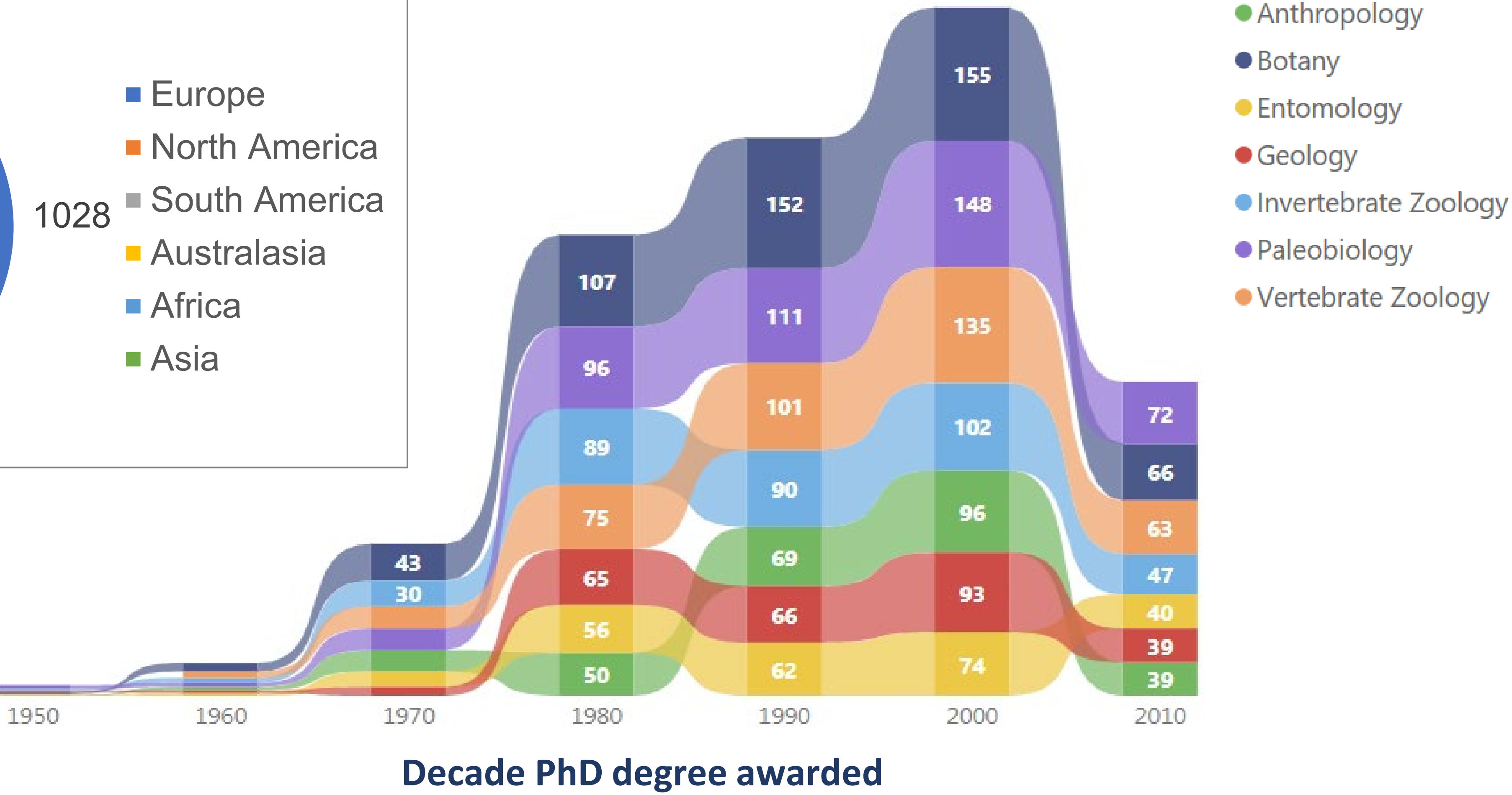
ca. 0.2% of specimens are genomically available

Researcher expertise and demography

Researchers by region of employment



Researcher fields & PhD award date



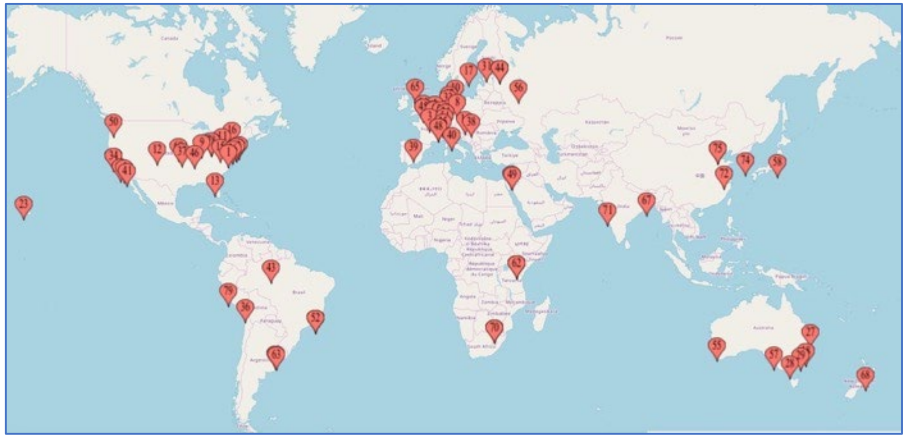
Collaborative global strategy

1. Global data platform * (see next slide)

1. Mobilizing existing data
    - Digital capture
    - Informatics

2. Collecting for the future
    - Gaps (e.g., tropical, polar & marine realms)
    - Global challenges (e.g., monitoring & key species)
    - Genomics (e.g., sampling & repositories)

3. Expertise
    - Training and capacity building
    - Community engagement



Future vision - Global data platform

GENOTYPES

(GenBank, GGBN, BOLD)

LITERATURE

(BHL, Google Scholar)

PHENOTYPES

(MorphoBank, TraitBank, ARBOR)

SPECIMENS

Taxonomy, location, date, images

(eg iDigBio, ALA, DiSSCo)

CITIZEN SCIENCE

(iNaturalist, Zooniverse, iSpot, eBird)

TAXONOMY & PHYLOGENY

(Catalogue of Life, ITIS, Open Tree of Life, EoL)

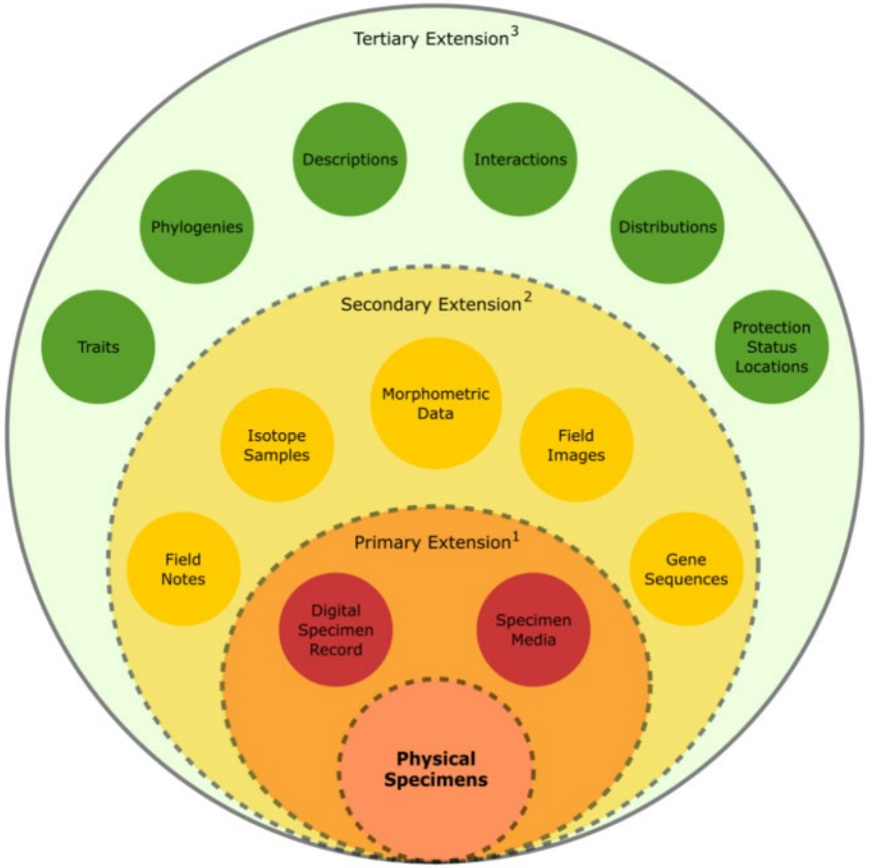
ENVIRONMENT

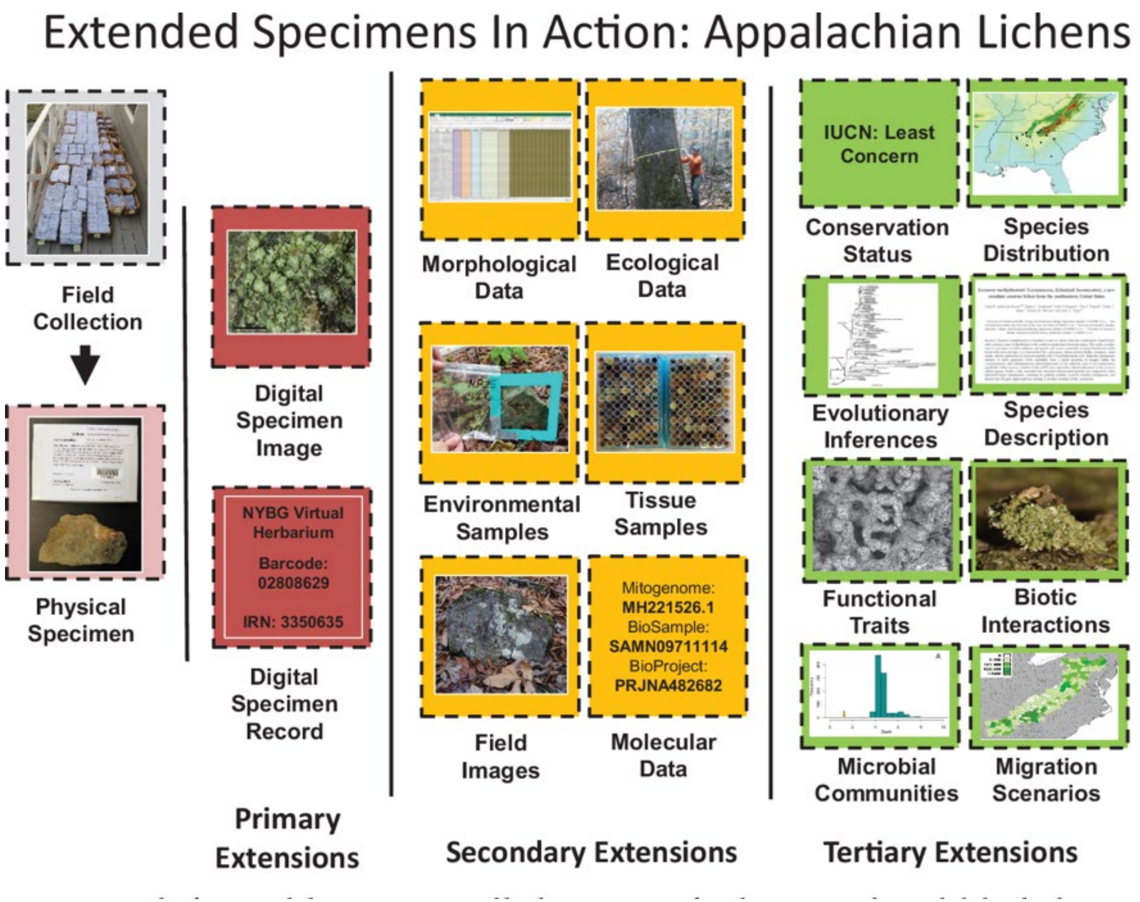
(NEON, NASA, Google Earth)

SPECIES DISTRIBITIONS

(GBIF, Lifemapper, MoL)

The ‘Extended Specimen’ Concept

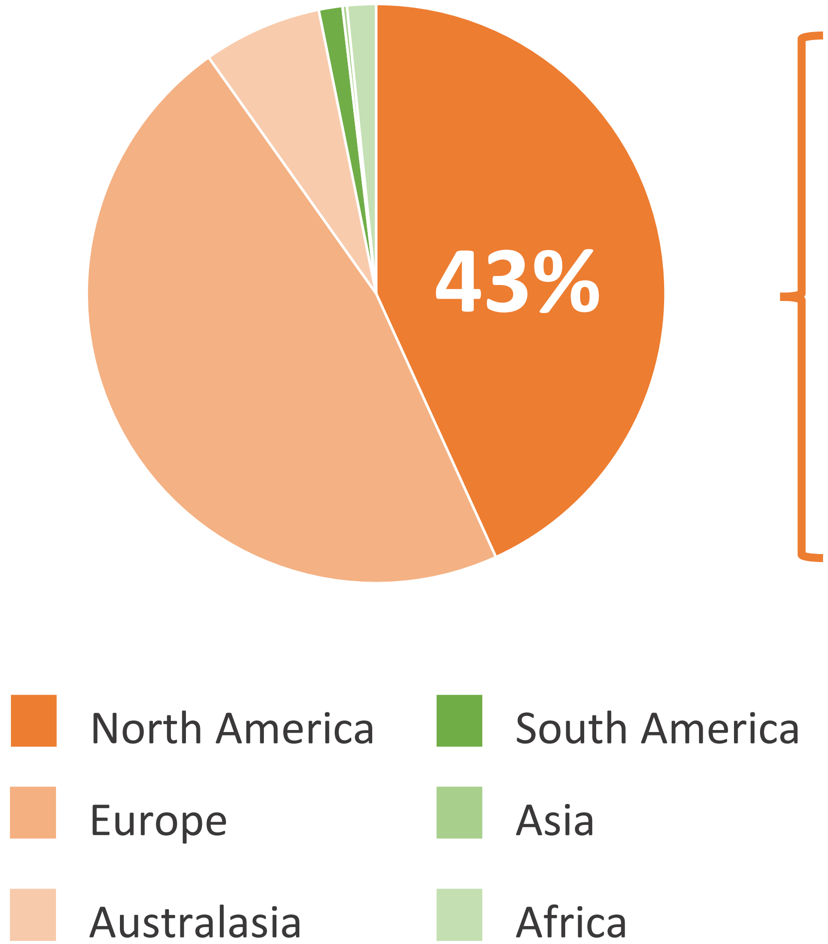




Thiers B et al. (2019) *Extending U.S. Biodiversity Collections to Promote Research and Education*. The Biodiversity Collections Network.
Lendemer J et al. (2019) The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. *Bioscience* 70, 1-8.

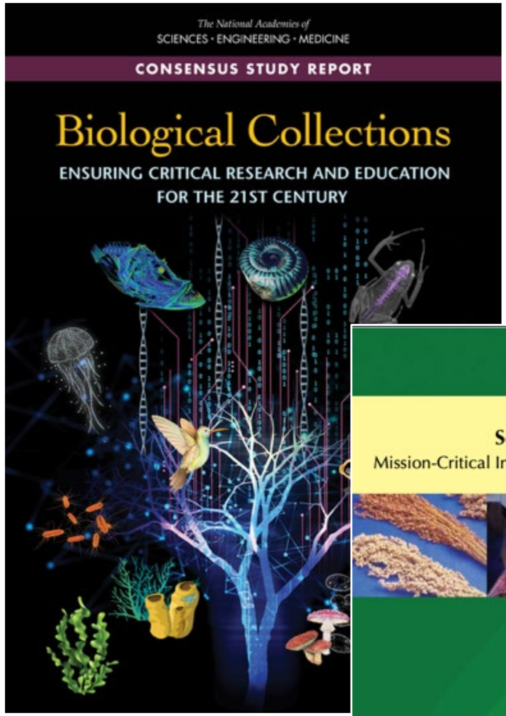
What does this mean for US collections?

Place of holding (%)



Federal museums
Federal agencies

State museums
Private museums
Universities
Research Institutes
Botanic gardens
Herbaria



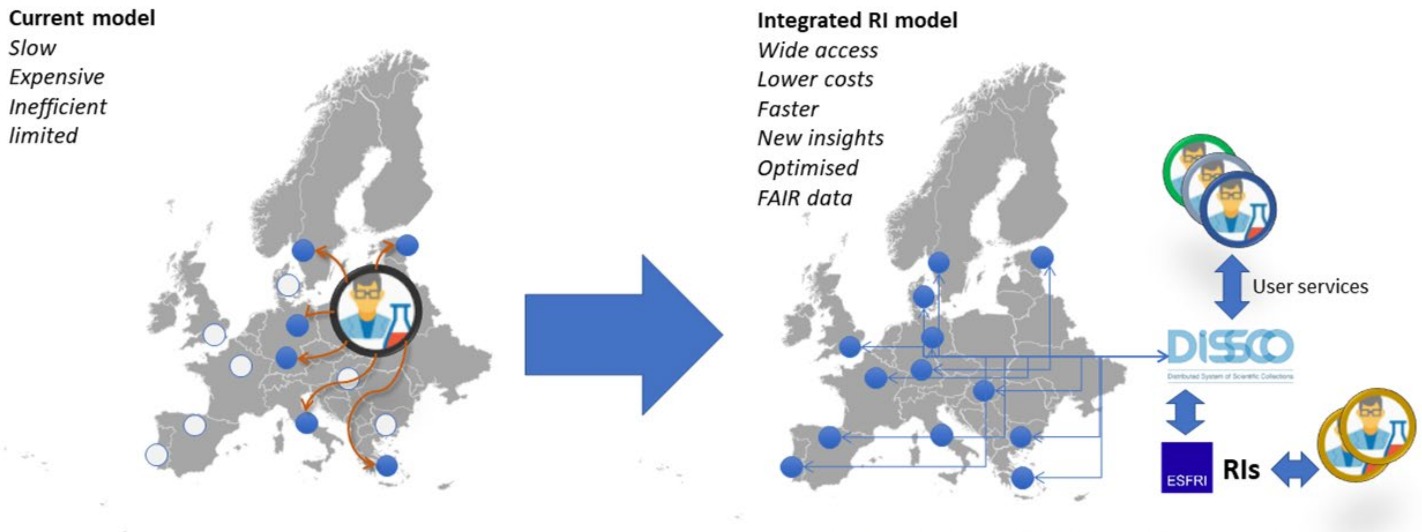
National Academies of Science (2020)

Biodiversity Collections Network (2019)

Intragroup Working Group on Scientific Collections (IWGSC) (2009)

A continental-scale collection project

Distributed System of Scientific Collections



The first mass scale initiative to re-unite and serve genomic, chemical, geographical, morphological and taxonomic information and link it to collections objects

The Idea:
The problem
The opportunity
Funding sources *
Core team

The Proposal:
Data
Users
Collaborative strategy
Writing team
Expert reviewers

Building support:
Institutions
Influencers
National networks
National funders
Other EU infrastructures

Funding:
The big idea
Project structure
Implementation plan
National support
Budget