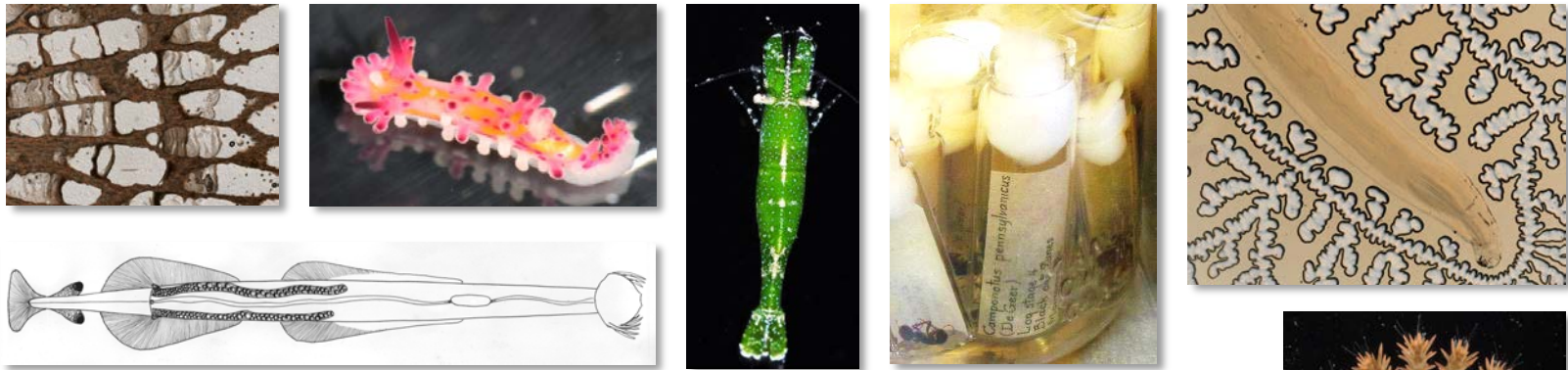
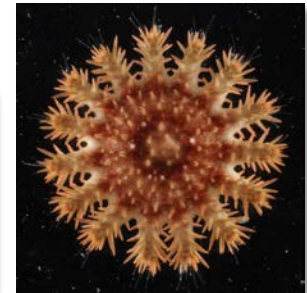




# Digitization Overview: Capturing the Past and Present for the Future



Deborah Paul, Florida State University / iDigBio  
DigIn Marine Invertebrate Digitization Workshop  
4-5 February 2019





## Digitization Overview

- Why digitize?
- Characterizing digitization
- Choosing collection management software
- Sharing data beyond the who/what/where/when
  - tissues, DNA, sequence data, images, measurements, ...
  - using extensions to Darwin Core
- Pertinent issues (data quality, research use, ...)
- Resources at iDigBio
  - share your expertise and experience please
- Discovering and addressing biodiversity data literacy needs for digitization and research



A huge untapped source of information!

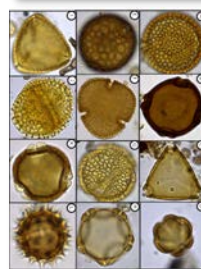
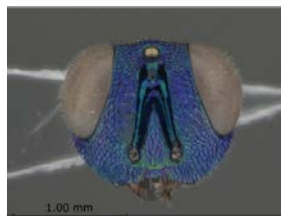
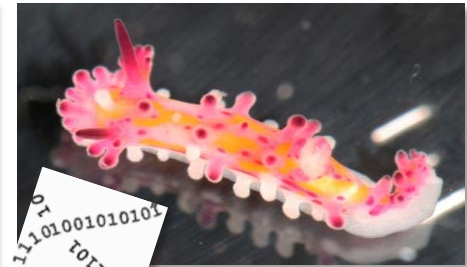
# Why digitize?

Estimates suggest **500 million and 1 billion** biological and paleobiological specimens in the United States and potentially **3-4 billion worldwide.**

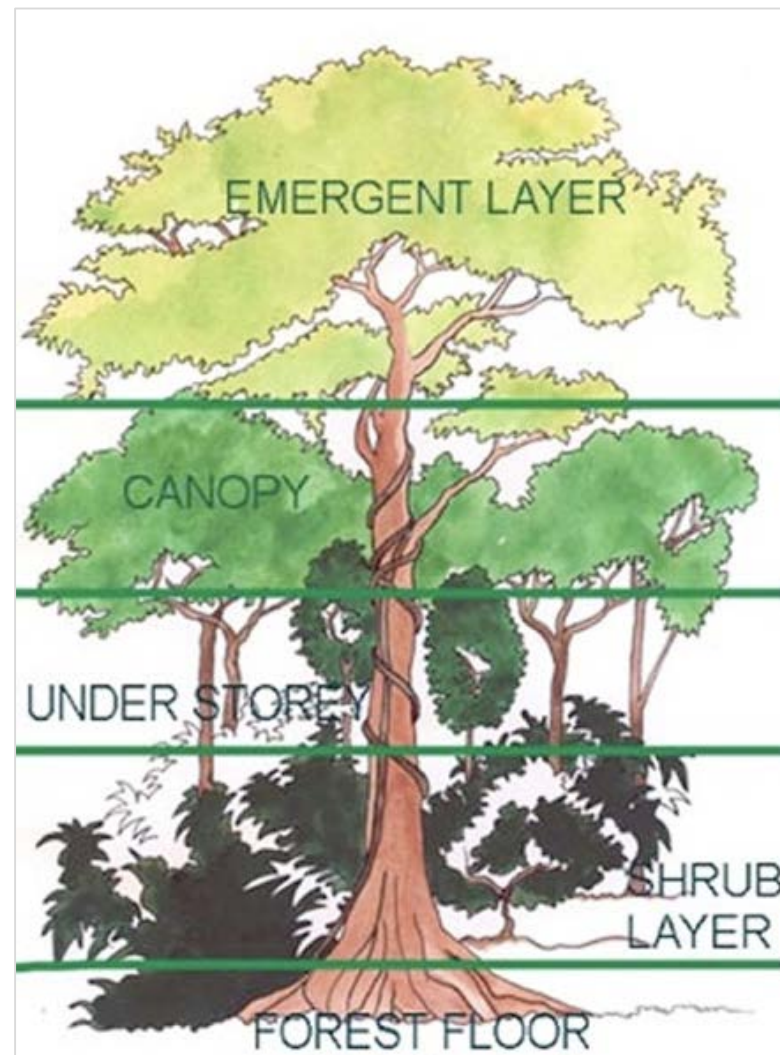
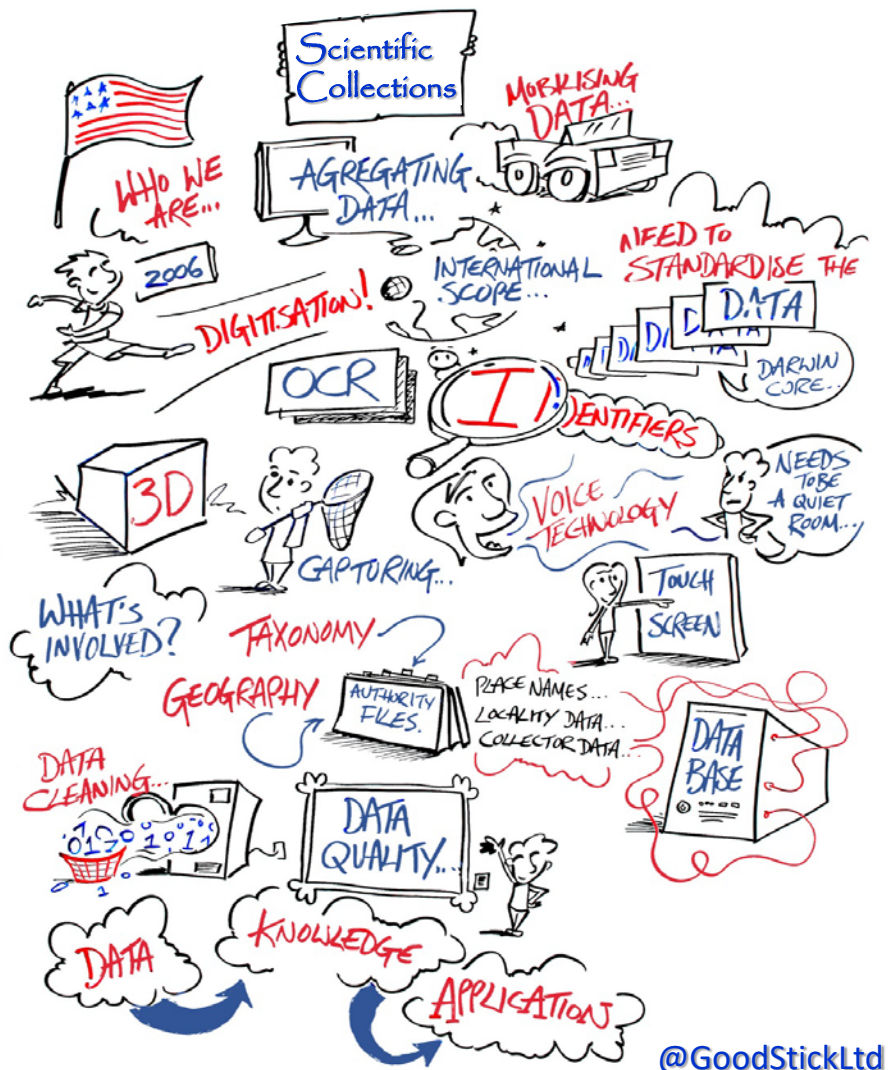
*Many are digitized, but most are not.*

ORCID ID

<https://orcid.org/0000-0003-2639-7520>









# Observing digitization across collection types

## Five task clusters that enable efficient and effective digitization of biological collections

Gil Nelson, Deborah Paul, Gregory Riccardi, Austin R. Mast

- 28 Collections
- 10 Museums
- Spanning biological and paleontological collections
- Insects and other invertebrates, plants, birds, mammals
- Wet, dry

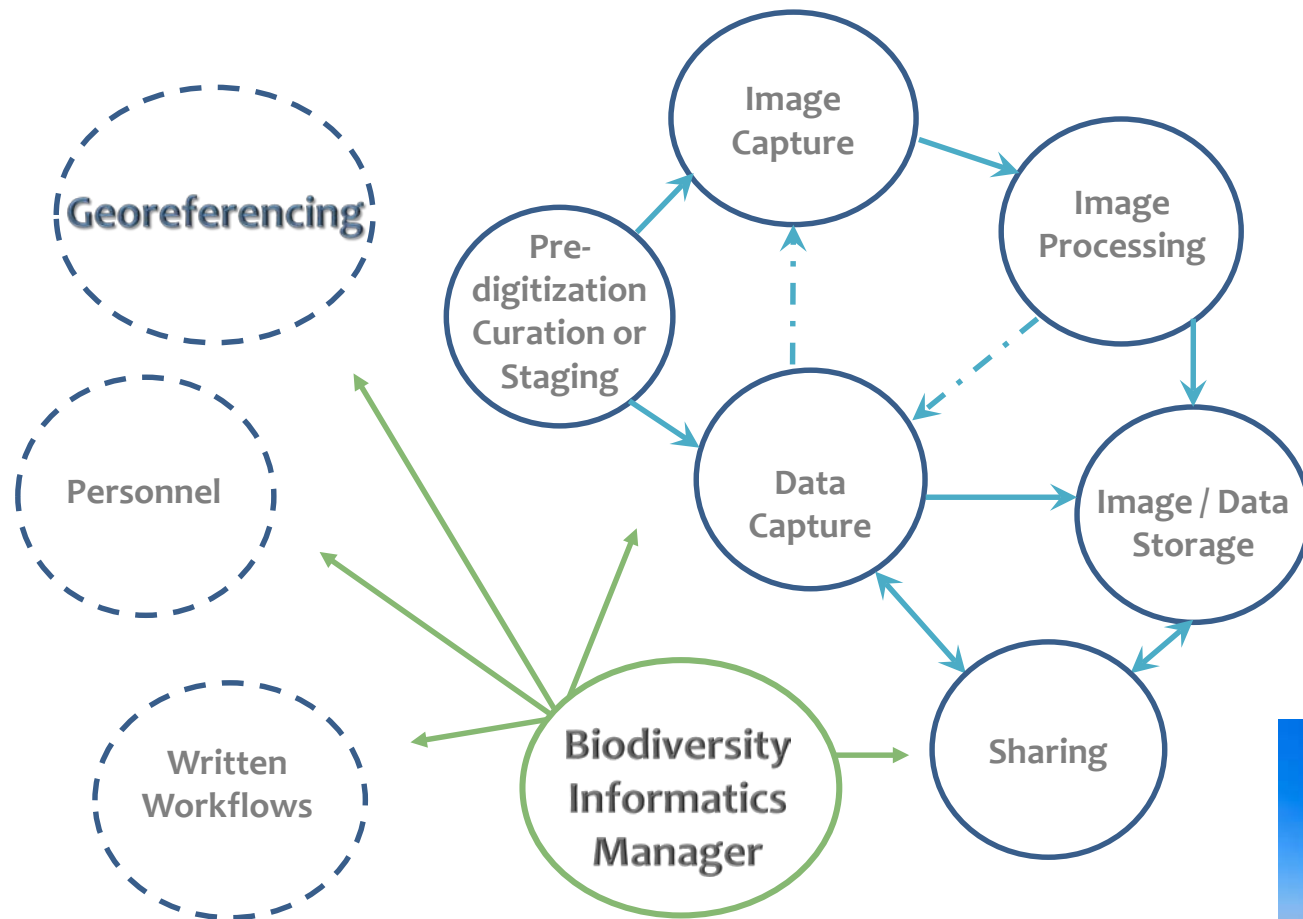


A peer-reviewed open-access journal  
**ZooKeys**  
Launched to accelerate biodiversity research





# Digitizing Collections: a task cluster framework





## Getting started

- Workflows and protocols
- Selecting and installing a database
- Imaging
- Image processing
- Pre-digitization preparation and curation
- Plan for data enhancement activities, e. g. georeferencing



# Tracks to Digitization

- **Taking the inside track** is often based on stretching the institution's resources.
  - user-initiated digitization
  - primary focus to get the job done quickly and to fill the user's request.
- **Taking the middle track** has the widest range of options, standards, and results. This is the most flexible of the tracks, where decisions often fall in gray areas.
- **Taking the outside track** focuses on the collections themselves. While users may initiate digitization, it is undertaken to deliver materials to a greater public.
  - may lead to comprehensive digitization, such as an entire book, series, or collection.
  - goal is to create maximum access to special collections
  - usually involves thought and planning that is more in-depth than the fulfillment of day-to-day digitization requests.





## Long view

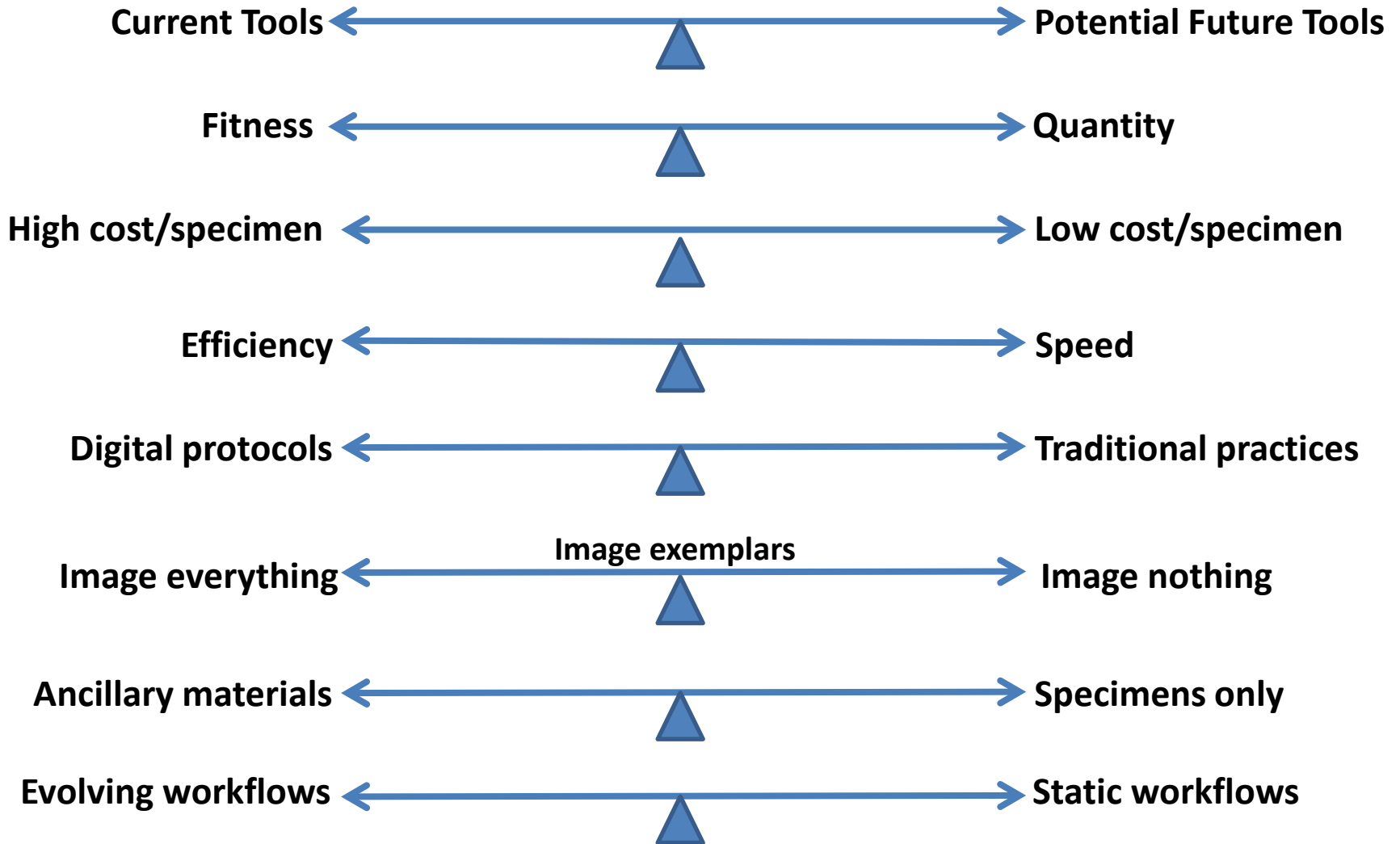
## Short view



- **Taking the long view**
  - developing doable, effective, and sustainable strategies for balancing long term goals with short term constraints, including a commitment to implementing future enhancements.
- **Pressures mitigating the long view**
  - So much data, so little time.
  - Our collections are not getting smaller.
  - The funding agencies have high output expectations.
  - We only have 3 years to get this done.
  - All of our data and all of our specimens are important.
  - Let's just use the images!
  - We'll do the minimum now and enhance it later.



# Global Digitization Continua





# Choosing a database / collection mgmt system

- Establish *institutional motivation* to digitize specimens
- Document and agree on a *priority feature set*: necessary versus desired
- Review *input/output scenarios* (*dwc, license cost, mac vs pc, security, ...*)
- Proprietary, open source, hybrid, cloud-based, (feature development, maintenance)
- *Community* advantages
- Shop vendors, *score* them (necessary and desired)
- Get a *full demo copy* and test with real data (score on ease of use - novice and expert users)
- *Evaluate costs* (up front, support, hosting, institutional support)

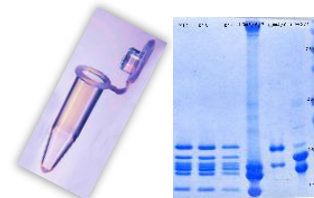


# What are some examples of standards used for sharing rich(er) biodiversity data? Where do they come from?

Data	Standards
specimens & observations	Darwin Core (DwC)
specimen & observation datasets	Ecological Metadata Language (EML)
media	Audubon Media Core
derivatives	Material Sample Core and GGBN Extensions, BOLD, ...



**CC creative commons** What's in the dataset?







## Sharing extended rich(er) data

- BoLD and GenBank example
  - dwc:catalogNumber
  - dwc:otherCatalogNumbers
  - dwc:preparations
  - dwc:associatedSequences

Catalog Number	KWP:Ento:14412
Other Catalog Numbers	BoLD barcode ID=UAMIC3084-15; GenBank=KU873942
Preparations	whole organism (pinned); DNA extraction; DNA extraction



# Sharing extended rich(er) data

Identification Remarks	BOLD ID Engine
Catalog Number	BIOUG06412-F06
Other Catalog Numbers	SSBAF1317-13
Preparations	Whole Voucher
Life Stage	Adult
Institution Code	University of Guelph, Centre for Biodiversity Genomics
Collection Code	BIOUG
Occurrence ID	BIOUG06412-F06
Dataset Name	University of Guelph, Centre for Biodiversity Genomics (BIOUG)
Basis of Record	PreservedSpecimen
Associated Sequences	<a href="http://www.boldsystems.org/index.php/Public_RecordView?processid=SSBAF1317-13">http://www.boldsystems.org/index.php/Public_RecordView?processid=SSBAF1317-13</a> <a href="https://www.ncbi.nlm.nih.gov/nuccore/KM936331">https://www.ncbi.nlm.nih.gov/nuccore/KM936331</a> <a href="http://v4.boldsystems.org/index.php/Public_BarcodeCluster?clusteruri=BOLD:ACX4522">http://v4.boldsystems.org/index.php/Public_BarcodeCluster?clusteruri=BOLD:ACX4522</a>



# Data issues: time, location, authority files...

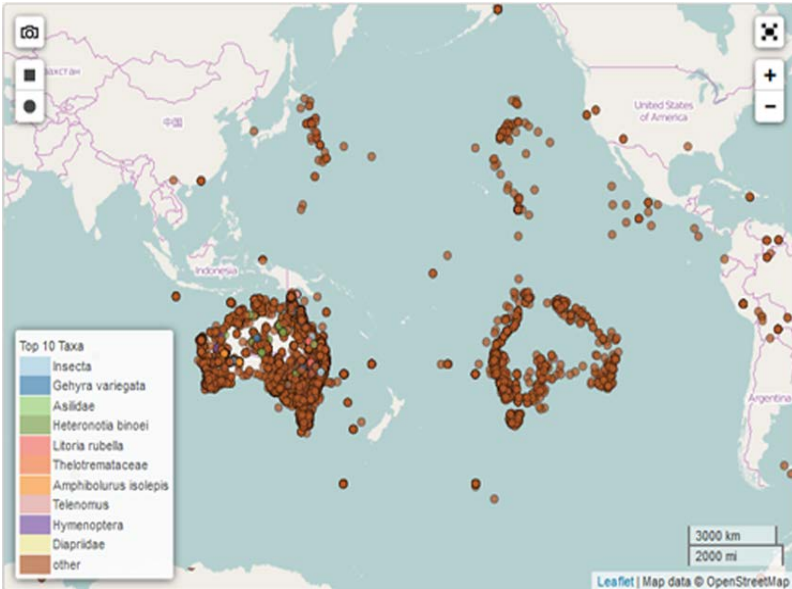


Hannah Frost  
@feefifofannah



Following

From a [@HydralNABox](#) interview: "People will put anything and their dog in the date field. It's absolutely astonishing."



Country

- united kindgom
- united king
- united kingdom**
- united kingdom (england)
- united kingdom (scotland)
- united kingdom (wales)
- united kingdom [?]
- united kingdom of great b
- united kingdom?

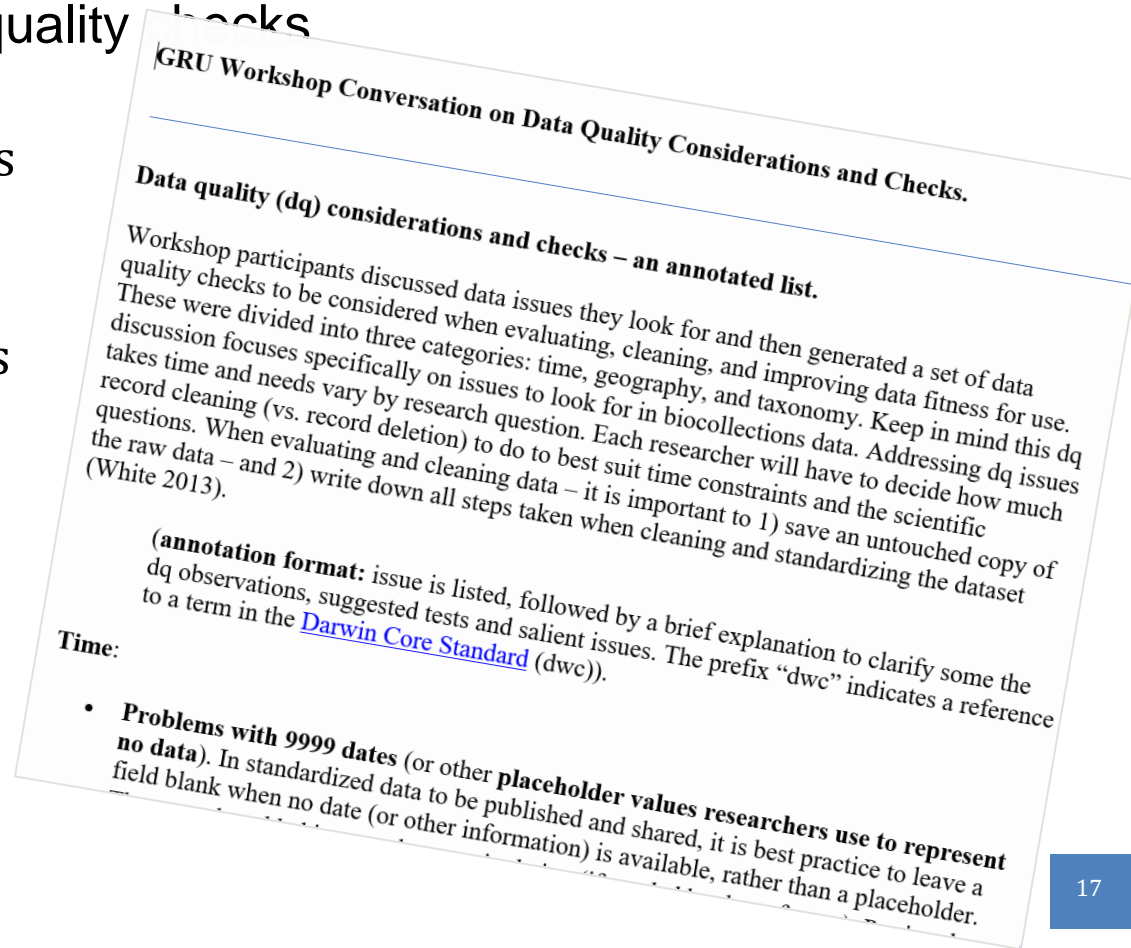
List

Family



# Researchers report back on evaluating suitability of biodiversity data for their research:

- evaluating the research fitness-for-use of these data
- creating a list of data quality checks
- Timey-wimey stuff
  - date issues like formats
- Geography
  - place name issues
  - out of expected bounds
  - missing metadata
- Taxonomy
  - taxon name issues\*
    - transparency please
  - concepts
  - authority files
  - parsing

















# *Georeferencing for Research Use (GRU): An integrated geospatial training paradigm for biocollections researchers and data providers*




Katja Seltmann,  Sara Lafia, Deborah Paul,  Shelley James,  David Bloom, Nelson Rios, Shari Ellis, Una Farrell, Jessica Utrup, Michael Yost,  Edward Davis, Rob Emery,  Gary Motz, Julien Kimmig,  Vaughn Shirey,  Emily Sandall,  Daniel Park, Christopher Tyrrell,  R. Sean Thackurdeen, Matthew Collins,  Vincent O'Leary, Heather Prestridge, Christopher Evelyn, Ben Nyberg

Workshop Report

doi: [10.3897/rio.4.e32449](https://doi.org/10.3897/rio.4.e32449)



 17-12-2018

 Unique: 517 | Total: 805

 Reprint: € 7,20

HTML

XML

PDF





# @iDigBio: many resources for digitization, data mobilization, data use

## Recommendations for the Acquisition, Processing, and Archiving of Digital Media

iDigBio has created recommendations for capturing, processing, and storing digital media.

[Recommendations for the Acquisition, Processing, and Archiving of Digital Media](#)

## Interest/Working Groups

The following links take you to Interest/Working Groups focused on Digitization. For other working groups please use the following links.

- [International Whole-Drawer Digitization Interest Group](#)
- [NANSH Working Group](#) (North American Network of Small Collections)
- [Fluid-preserved Arthropod and Microscopic Slide Imaging Working Group](#)
- [Paleontology Digitization Working Group](#)
- [Small Collections Network Working Group](#)
- [Vertebrate Digitization Interest Group](#)
- [Field Station Interest Group](#)

## Digitization Avenue

The following links provide information on the task clusters that are currently active. For more information on these clusters please read the following [Five task clusters that enable digitization](#).

- [Pre-digitization Curation and Staging](#)
- [Specimen Image Capture](#)
- [Specimen Image Processing](#)
- [Electronic Data Capture](#)
- [Georeferencing Locality Descriptions](#)
- [Digitization Workflows and Protocols](#)
- [More on digitization](#)

## Digitization Resources

This page provides resources and information for the series of digitization training materials. It includes information on how to use the resources as well as a plethora of digitization information and resources. Included is a gallery of digitization resources, including videos, presentations, and other important information related to biological collections.

[Contents \[hide\]](#)

- [1 iDigBio Introduction](#)
- [2](#)
- [3 Recommendations for the Acquisition, Processing, and Archiving of Digital Media](#)
- [4 Interest/Working Groups](#)
- [5 Digitization Avenue](#)
- [6 iDigBio Workshops, Reports, and Wikis](#)
- [7 Videos- Digitization Resources and Workflows](#)

### Researchers

[Browse our specimen portal](#)



### Collections Staff

[Learn how your collection can benefit from our work](#)



### Teachers & Students

[Learning resources & opportunities to engage](#)





# ADBC Community building

## Digitization

Workflows & Protocols  
Task Clusters  
Dissemination

## Research Use

Tool collaboration  
Portal development  
ENM workshop  
Research Spotlight  
Data quality

## Training

Biodiversity data skills  
Data literacy  
Collections software  
Imaging  
Project Management

Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	Columns
Acanthogammaridae	Acanthogammarus (Acanthogammarus)...	1915-06-26	Russian Federation	MZLU	preservedspecimen	<a href="#">view</a>
acanthogammaridae	Acanthogammarus goddardii				Specimen	<a href="#">view</a>
Acanthogammaridae	Acanthogammarus...				PreservedSpecimen	<a href="#">view</a>
Acanthogammaridae	Acanthogammarus...				PreservedSpecimen	<a href="#">view</a>
acanthogammaridae	Gammarus...				Specimen	<a href="#">view</a>



## Education Outreach


Citizen Science  
K-12 materials  
Undergraduate  
Fossil Clubs  
Mentor teachers

## Methods

Workshops  
Webinars  
Symposia  
Conferences  
Working Groups  
Short Courses  
Adobe Connect  
Listservs  
Publications  
Social Media



## Workshops reveal patterns - skills needs and knowledge gaps

- Digitisation workflow workshops
  - Flat Sheets and Packets, Pinned Specimens in Trays and Drawers, Things in Spirits, 3D objects in Trays, Imaging, ...
- Capacity building needs revealed
  - software
  - standards
  - data cleaning and management
  - spreadsheets, text files
  - data visualization and synthesis
  - recognizing automatable tasks
  - limited number of people in the community with the necessary skills
- Actions
  - Partner in developing and implementing Data Carpentry, now
  -  THE CARPENTRIES
  - Biodiversity Informatics Workshop Series at iDigBio
    - Data Carpentry
    - Managing NHC Data
    - Demystifying Data Standards and the IPT
    - Field to Database
  - Partner in [Biodiversity Informatics 101](#) at SPNHC
  - Partner in Darwin Core Hour
  - See Ethan White's Semester Data Carpentry Course on GitHub







## What is most important for developing *National Biological eCollections for Research?*

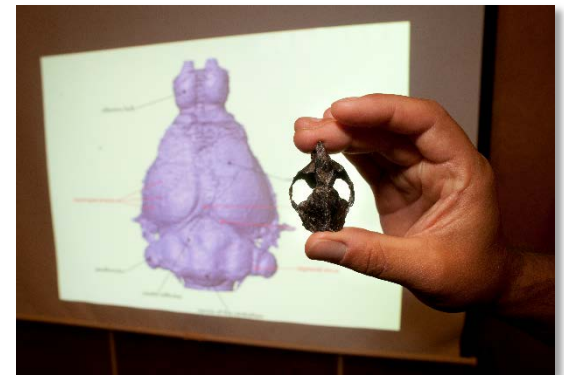
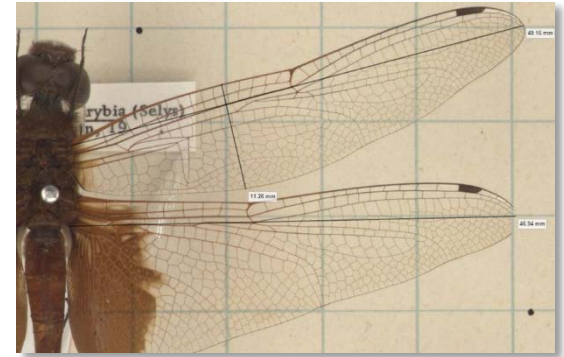
“Arguably the highest resource requirement of research infrastructure development is human capacity and capability.”

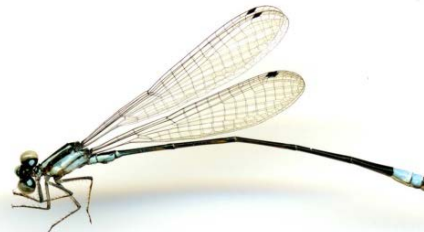
“2016 National Research Infrastructure Roadmap Capability Issues Paper.” CSIRO. Toni Moate, Director, National Collections and Marine Infrastructure. *On building National Biological eCollections*



# We want to engage with you!

- iDigBio provides access to data with the means to answer research questions
- Many opportunities for collaboration and potential funding:
  - **Public participation** in digitization
    - e.g., host a [WeDigBio](#) event, host a DwC hour
  - **Research** using the data already in the portal
    - e.g., niche modeling, conservation, etc.
  - **Data mining** the portal for new discoveries
    - e.g., extract measurements/characteristics from images or 3D models
  - Gathering all of the “**dark data**”
    - e.g., proposals for TCN, PEN, CSBR, IMLS, etc.
  - **Enhancing** and **enriching** the data
    - e.g., data linking, field notes, etc.





# Did you know about...

- ... Society for the Preservation of Natural History Collections
  - the SPNHC Emerging Professionals Group (EPG)
  - SPNHC Collections Club Network



- ... Biodiversity Information Standards (TDWG)



- ... Small Collections Network (SCNet)



- ... Biodiversity Literacy in Undergraduate Education

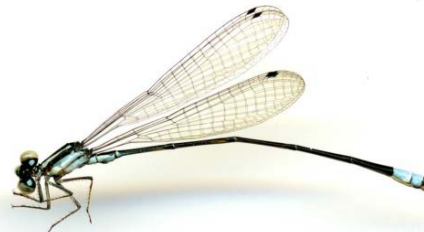


- ... The Carpentries



- ... Darwin Core Hour





# Did you know about...

... Biodiversity Collections Network



... Global Biodiversity Information Facility GBIF



... Distributed System of Scientific Collections (DiSSCo)



... Synthesys+





2019 OPEN DIGITAL SCIENCE WEEK ON BIOLOGICAL & GEOLOGICAL DIVERSITY

# biodiversity\_next

better Data - better Science - better Policies

Jointly organised by

Biodiversity  
Information  
Standards  
T D W G



DISCO  
Distributed System of Scientific Collections



21-25 October 2019  
Leiden, NL

Hosted by  
  
Naturalis  
Biodiversity  
Center



## 115 National Facilities 21 Countries

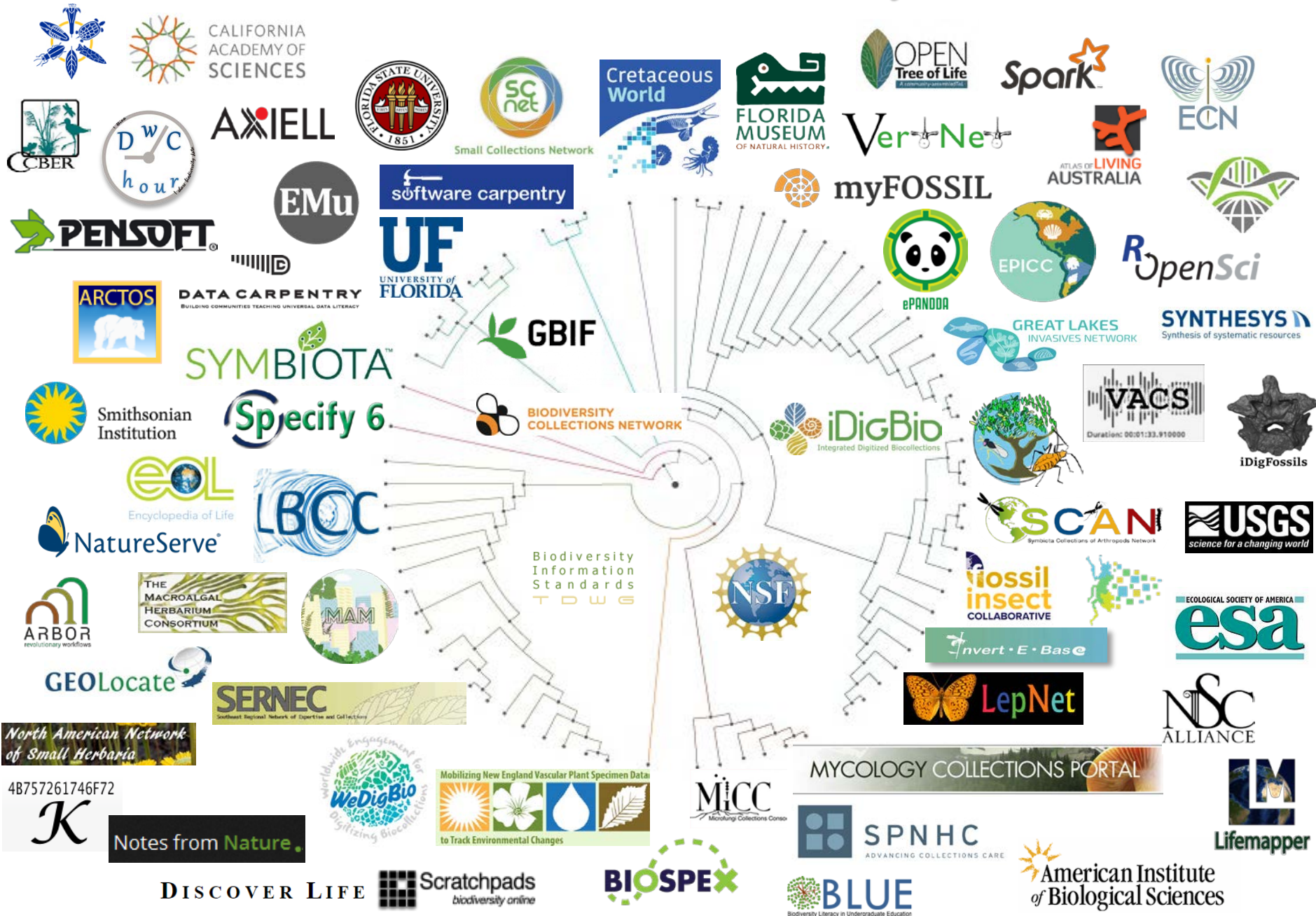
- Largest ever formal agreement between natural science collection facilities
- A system of distributed facilities
- Centralised shared governance model in place
- Supporting network of **working groups**

## a new business model: ONE EUROPEAN COLLECTION

- One European Collection of scientific assets
- Common Collections development strategy
- Economies of scope and scale
- Monitoring impact of collections (documenting ROI)
- Specialisation strategies (e.g. in alignment with national priorities, e.g. Smart Specialisation Strategies)
- Joint Research Agendas

Find out more at [www.dissco.eu](http://www.dissco.eu)

# Collaboration is the key!





# Thanks, any questions? thoughts?



[www.idigbio.org](http://www.idigbio.org)



[facebook.com/iDigBio](https://facebook.com/iDigBio)



[twitter.com/iDigBio](https://twitter.com/iDigBio)



[vimeo.com/idigbio](https://vimeo.com/idigbio)



[idigbio.org/rss-feed.xml](http://idigbio.org/rss-feed.xml)



[webcal://www.idigbio.org/events-calendar/export.ics](http://webcal://www.idigbio.org/events-calendar/export.ics)