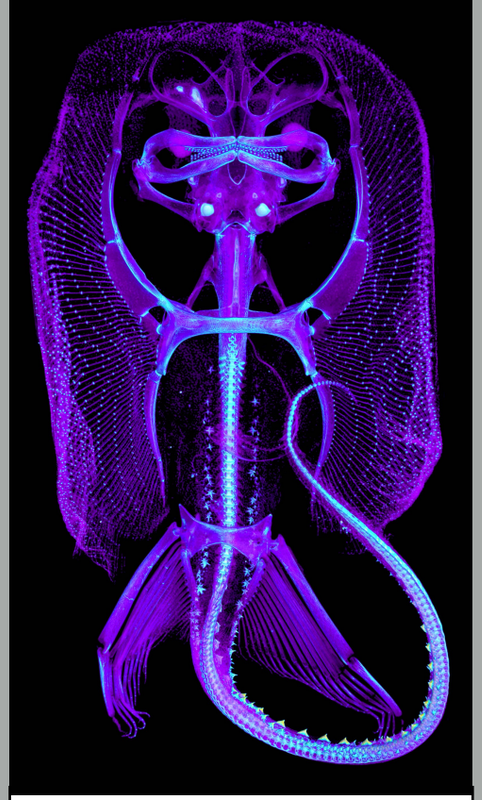
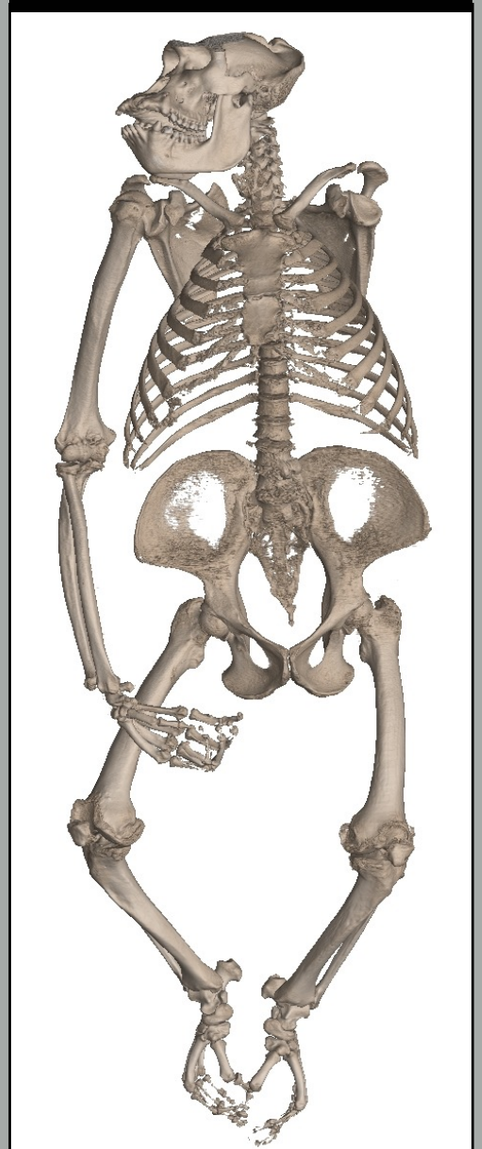




Initial Insights and Area for Further Exploration

David C. Blackburn
Associate Curator of Herpetology
Florida Museum of Natural History
University of Florida

dblackburn@flmnh.ufl.edu

Integrating Institutional Archives with Disciplinary Web Repositories
Duke University

January 23, 2020

audio recordings

images

preserved specimens

chemical data

genetic data

disease data

description of collecting event

audio recordings

The Cornell Lab of Ornithology
Macaulay Library

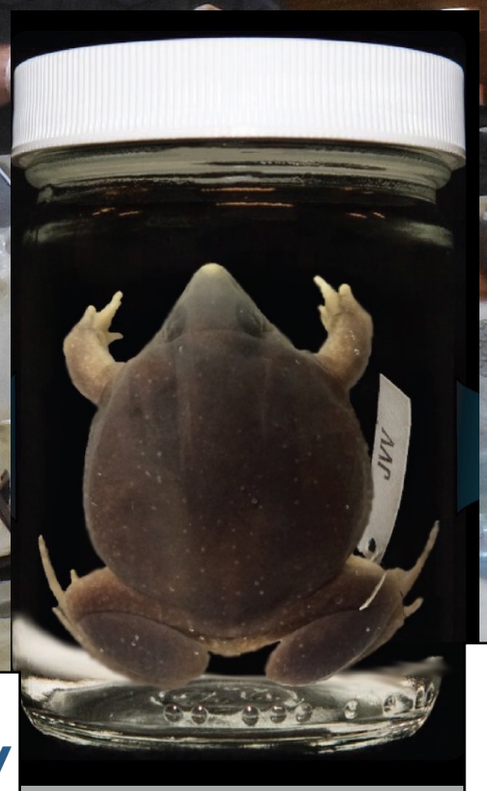
Cal Photos

images

MORPHO SOURCE

preserved specimens

chemical data



genetic data

NCBI

disease data

Specify

description of collecting event

Arctos

audio recordings

The Cornell Lab of Ornithology
Macaulay Library

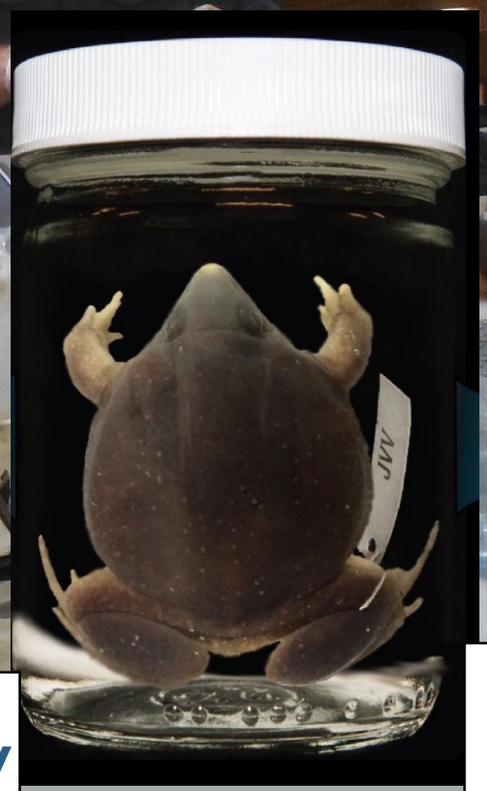
Cal Photos

images

MORPHO SOURCE

preserved specimens

chemical data



genetic data

NCBI

disease data

Specify

description of collecting event

Arctos

Why now?

oVert

openVertebrate Thematic Collection Network

18 funded institutions, including 16 museums and 6 imaging centers

CT-scan >20,000 fluid-preserved vertebrate specimens
Make both raw and processed data freely available on-line

~2+ years into project: >8,000 specimens from >42 US institutions
many specimens have two scans each; ~250 MB – 1 GB

Tracking usage of digital data

oVert-generated media on MorphoSource
- viewed >204,000 times
- downloaded >7,000 times

Downloads

nearly 50% for “non-research”

Tracking usage of digital data

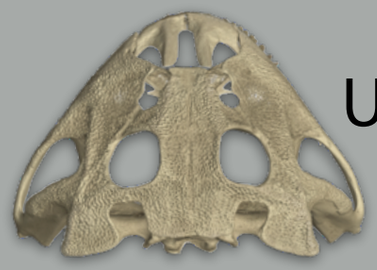
oVert-generated media on MorphoSource
- viewed >204,000 times
- downloaded >7,000 times

Downloads

nearly 50% for “non-research”

oVert

Getting info on media files and usage back to collections



UF-Herp-12345

MORPHO SOURCE

Darwin Core
structured metadata

referenceID
occurrenceID
locality
collectionDate
etc.

for each collection
(i.e., UF Herpetology)

MorphoSource RSS Feed
(via referenceID)
containing
1) Audobon Core metadata
2) usage statistics

FLORIDA MUSEUM

add Audobon Core
to IPT

GBIF

iDigBio
Integrated Digitized Biocollections

https://github.com/FLMNH/MorphoSourceRSSDownloader

Getting info on media files and usage back to collections

Data reporting for MorphoSource media

MorphoSource provides summary reports of media, download usage, and download requests for media that represent specimens that have been reported to iDigBio. The media report is formatted according to the Audubon Core metadata standard, and so can be incorporated into publisher reporting software, such as an IPT. All report files are linked and described in a single RSS feed, which can be used to receive regular report updates via automated download. Additionally, all report files are individually listed below, sorted by iDigBio publisher and recordset. For each report, there is a link to: **1)** a Comma Separated Values (CSV) spreadsheet with the primary media, download usage, or download request metadata; and **2)** an XML file encoded in Ecological Markup Language (EML) providing metadata about the CSV spreadsheet. These reports are updated as necessary on a daily basis.

RSS Feed: https://www.morphosource.org/rss/ms_rss.xml

Yale Peabody Museum IPT Service (0bdd6e08-91e3-4ef0-a14f-7a987f9e9362)

Recordset	Media	Downloads	Download Requests	Pub Date
Invertebrate Paleontology Division, Yale Peabody Museum (137ed4cd-5172-45a5-acdb-8e1de9a64e32)	CSV EML	CSV EML	CSV EML	Fri, 08 Mar 2019 16:16:42 -0500
Vertebrate Paleontology Division, Yale Peabody Museum (0220907a-0463-4ae0-8a0b-77f5e80fff40)	CSV EML	CSV EML	CSV EML	Fri, 08 Mar 2019 16:17:29 -0500
Vertebrate Zoology Division - Herpetology, Yale Peabody Museum (cf60ed8a-2c79-4b85-a259-15a8e216dae4)	CSV EML	CSV EML	CSV EML	Fri, 08 Mar 2019 16:16:55 -0500
Vertebrate Zoology Division - Ichthyology, Yale Peabody Museum (30ab9c2a-0b54-4c04-84ca-bc7abdd90b52)	CSV EML	CSV EML	CSV EML	Fri, 08 Mar 2019 16:18:27 -0500

Why now?

New questions have arisen:

- Who owns these data?
- Who should store these data?
- What data should be stored?
- How do we keep track of derivatives (and associated information)?
- Who owns and stores derivative data?
- What should be the life cycle of data for generators and users?

GOALS

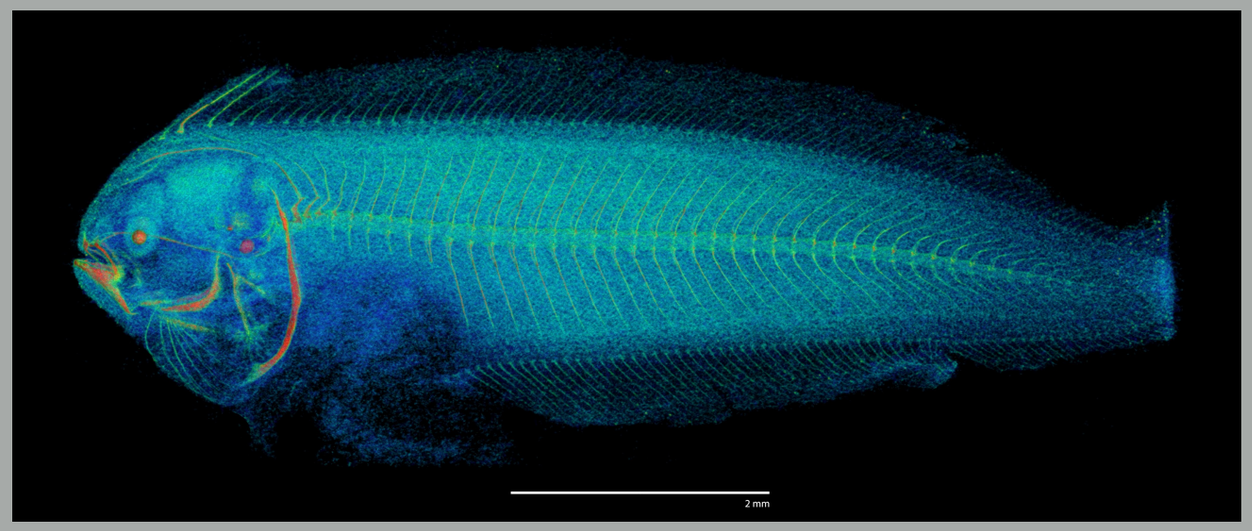
(1) Survey current needs, workflows and trajectories of the community

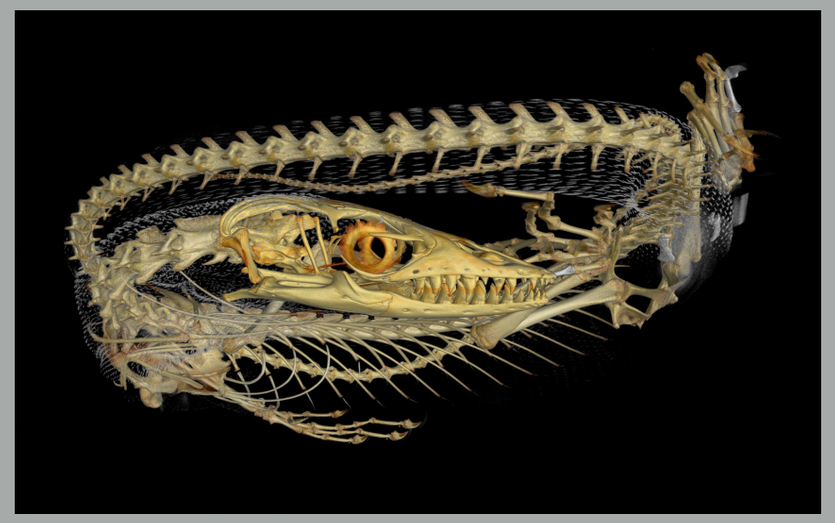
(2) Identify common ground among institutions

(3) Explore potential for unifying approaches

(4) Assess role of domain-specialized repositories

(5) Articulate an overarching plan





Participants

43 respondents from 19 institutions

Museum/Collection Staff	19.15%	9
Museum/Collection IT staff or Software Developer	14.89%	7
Domain Specialized Third Party Repository or Database Software Provider Staff/Administration	6.38%	3
Generic Third Party Repository or Database Software Provider Staff/Administration	4.26%	2
Institutional Repository IT Staff Or Software Developer	10.64%	5
Museum/Institution Higher Administrator (Sets Policies Affecting Both IT and Collections Practices)	29.79%	14
Other (specify below)	14.89%	7

All have strong interests in curating, digitizing, and using collections for research, education, and outreach

Participants

Range of roles and responsibilities related to biodiversity data

Maintain or create digital data on collections objects	18.93%	32
Research and/or develop solutions to digital data infrastructure for a museum collection (i.e., informatics)	17.16%	29
Make decisions about informatics solutions for museum collections	17.16%	29
Research and/or develop policies for hosting and management of museum collection digital data	18.34%	31
Make decisions about policies for hosting and management of digital data in my department/institution	20.71%	35
Other roles you feel are important to indicate (specify below)	7.69%	13

Representational Data

Many collections create and store 2D representational data
Growing creation and storage of 3D representational data

Field	Collect/Create		Store		Unknown	
Specimen photographs	51.79%	29	48.21%	27	0.00%	0
Field photographs	47.46%	28	49.15%	29	3.39%	2
Histological slide images	48.72%	19	41.03%	16	10.26%	4
Sound recordings	50.00%	21	45.24%	19	4.76%	2
Videos	45.65%	21	43.48%	20	10.87%	5
3D photogrammetry models	46.34%	19	46.34%	19	7.32%	3
3D synchotron scans	28.57%	8	28.57%	8	42.86%	12
3D laser scans	38.89%	14	36.11%	13	25.00%	9
3D structured light scans	40.00%	12	33.33%	10	26.67%	8
3D CT scans	49.02%	25	49.02%	25	1.96%	1
3D MRI scans	25.00%	5	25.00%	5	50.00%	10

Representational Data

High demand for photographs and CT scans
Lower demand for field photographs, sound, and video recordings

General feeling of insufficient resources for 3D representational data

Common Problems

Insufficient staff for curation of and requests for representational data

Decentralized and unsynchronized data across repositories
leading to duplication of effort and waste of limited staff time

Representational Data

We get a lot of requests and it is a time consuming challenge to deal with...

We receive thousands of requests per year and it is time consuming to deal with them. At present, some file formats require substantially more time and labor than others.

Media accessed through Arctos is easy as many researchers can download from there; high res images, especially the historic ones are time consuming since each request is different and usually requires followup and other requests. Researchers may find these beyond Arctos, like the Ecoreader for field notes and field images or CalPhotos.

The issue is also creating more access which creates more demand

We get lots of requests, but not overly time consuming to deal with them.

We get a lot of requests and it's trivial to deal with them- not sure why this isn't an option?

There has been a significant uptick in requests to image specimens either by us or by visitors and more damage occurs as a result.

this is decentralized

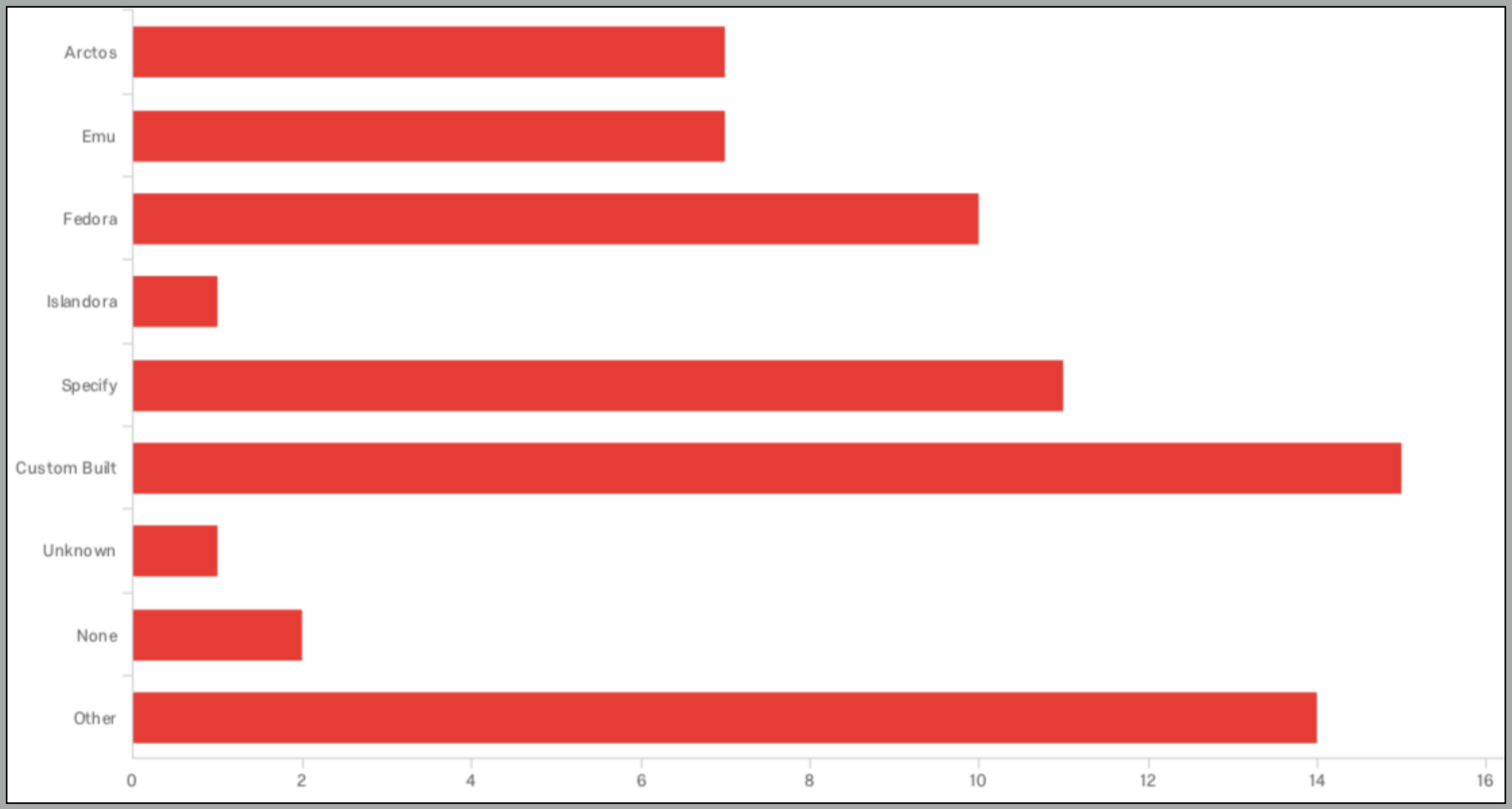
Burden is growing.

DAMs and Databases

Many use multiple DAMs, including custom-built

Most common collection database software among participants:

Arctos, Emu, and Specify (few using Symbiota)

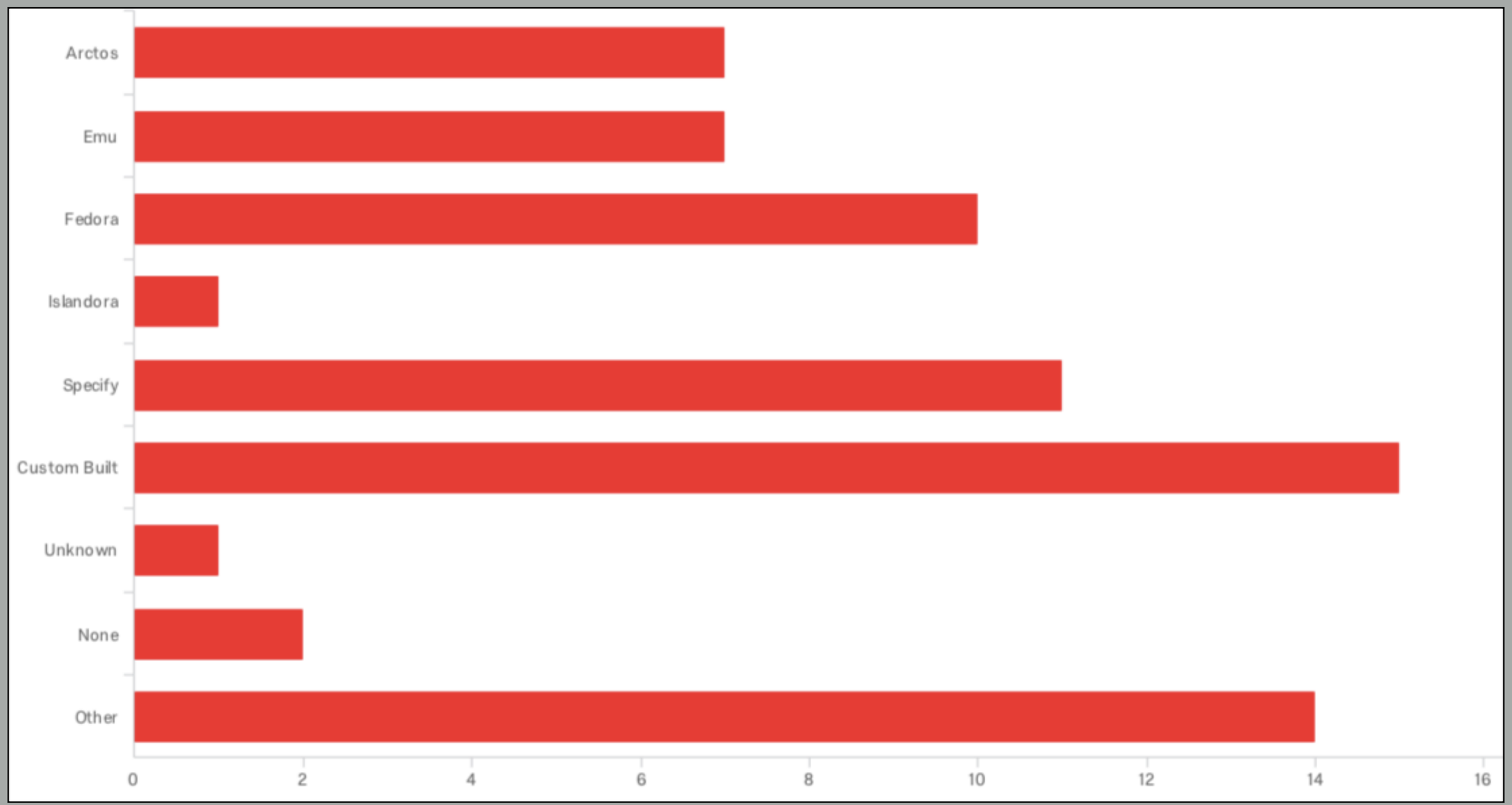


DAMs and Databases

Needs from collection databases:

better tools to describe relationships among representational datasets

APIs and support for IPT manifests



Data Storage

Most institutions have access to a suite of data storage solutions
- preference for networked storage solutions
- many unable to use cloud storage solutions
- only one-third have access to free institutional archival storage

#	Field	Never	Infrequently	Frequently	Primary Solution	N/A	Unknown
1	Storage on Local Computers	35.71% 10	25.00% 7	17.86% 5	10.71% 3	3.57% 1	7.14% 2
2	Storage on External Hard Drives	32.14% 9	17.86% 5	17.86% 5	7.14% 2	10.71% 3	14.29% 4
3	Storage on Network Attached Storage Devices Managed By Your Lab or Your Department	20.00% 5	8.00% 2	24.00% 6	28.00% 7	8.00% 2	12.00% 3
4	Replicated Network Attached Storage Managed By Your Institution	3.33% 1	6.67% 2	23.33% 7	50.00% 15	0.00% 0	16.67% 5
5	Cloud Storage Subscriptions (AWS S3, Wasabi or other HTTP storage)	16.67% 4	12.50% 3	4.17% 1	16.67% 4	20.83% 5	29.17% 7
6	Services Allowing Personal File Folder Organization (Box, Dropbox, GoogleDrive)	25.00% 7	21.43% 6	32.14% 9	14.29% 4	0.00% 0	7.14% 2

Data Storage

Divergent opinions on best options for storing representational data
- some prefer third-party solutions
- some prefer institutional solutions
- some prefer an overarching federal solution similar to NCBI

Field	strongly disagree		disagree		Somewhat disagree		No opinion		Somewhat agree		Agree		Strongly agree	
It is important to store representational data of collection objects in an accessible, discoverable, and manageable way	2.27%	1	0.00%	0	0.00%	0	2.27%	1	0.00%	0	6.82%	3	88.64%	39
Because third party repositories tend to have the most domain specialized tools, they often present the best solutions for access, discovery, and management of representational data.	2.33%	1	0.00%	0	9.30%	4	18.60%	8	37.21%	16	18.60%	8	13.95%	6

Points of Agreement

All are interested in recovering and preserving data created by third-party contributors

All want to facilitate
discovery of representational data
connections to other related data for collection objects
reporting on usage of representational data

Institutions want to determine their own data structure, maintain security, and have low costs for storage

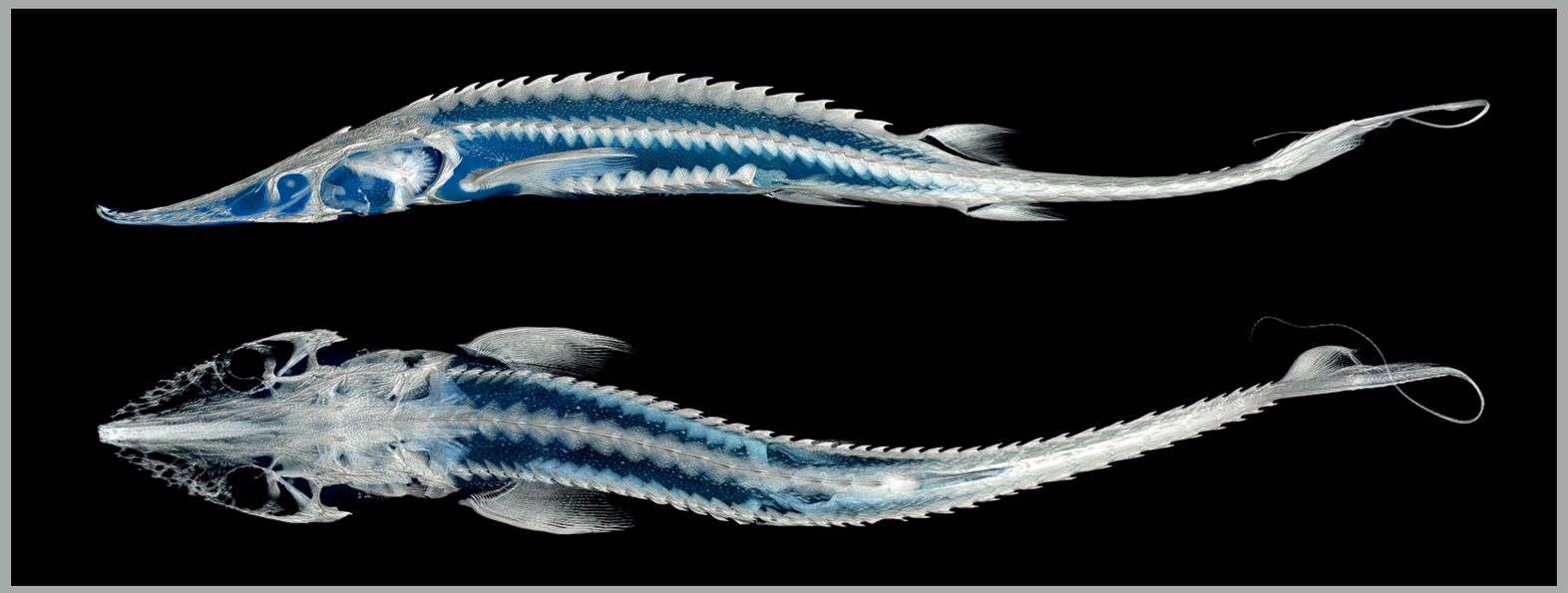
Most institutions lack strong institutional policies about representational data, and have flexibility in setting these

Major Concerns about Repositories

Security, back-up, long-term sustainability of repositories
(and related, potential loss of data when repositories disappear)

Mapping information between institutional and third-party systems

Not clear how best to control rights and access to data
Which representational data are copyrightable?
Which data are owned by the institution?
How can access be controlled to limit commercial use?
How to specify data that should not be made public?



Issues to Address

Most highly ranked among participants:

Integration and collaboration between:

- institutional repositories, domain specialized repositories, collections software, and/or third party repositories

- IT departments, libraries, and museum collections *within* an institution

Best practices for formats, data models, and metadata associated with representational data

Sustainable storage solutions for representational data

Upcoming Meetings

Digital Data in Biodiversity Research Conference
June 1–3, 2020
Indiana University
https://www.idigbio.org/content/digital-data-biodiversity-research-conference

Biodiversity Summit 2020
September 20–25, 2020
Alexandra, Virginia
https://www.idigbio.org/content/biodiversity-summit-2020