from the GBIF Task Force on

Accelerating Discovery of Biocollections Data

– a Malacology Point-of-View

or

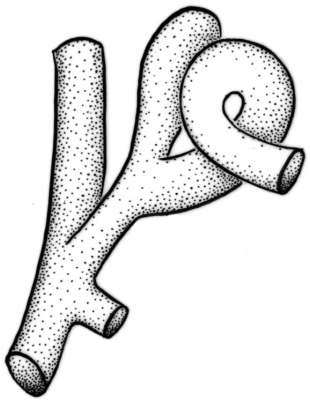
The Importance of Metadata for Discovery and Digitization

Deborah Paul, Shari Ellis

iDigBio Mollusk Workshop July 2017, Newark, DE

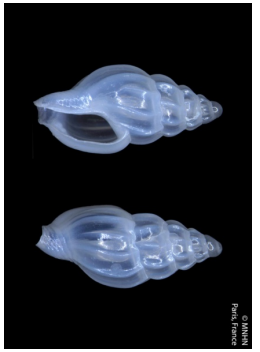
tw @idbdeb @iDigBio

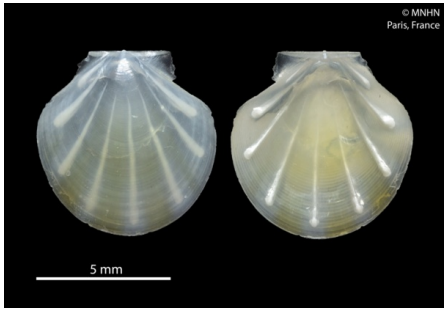
iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

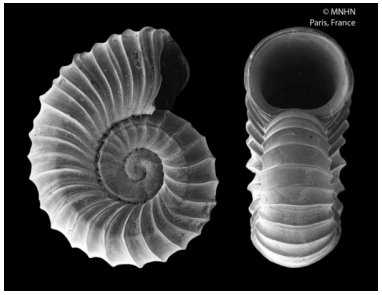


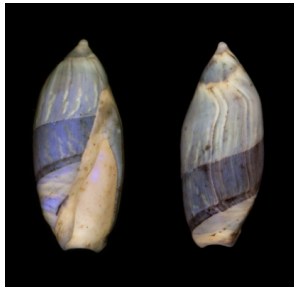

© MNHN
Paris, France

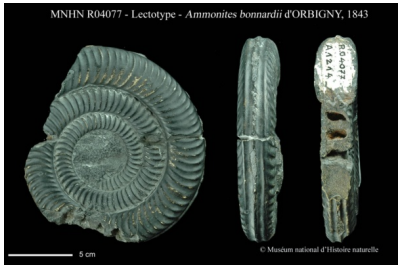

© MNHN
Paris, France

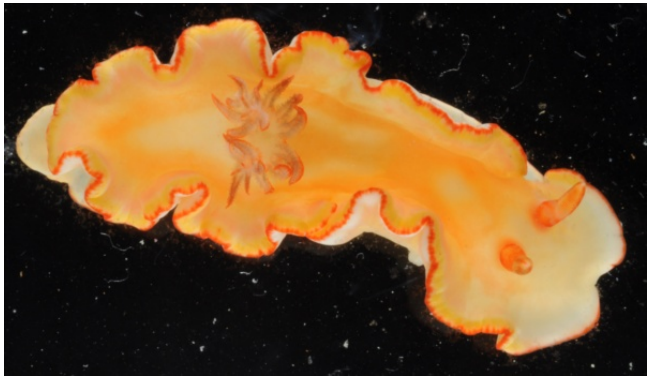

© MNHN
Paris, France


© MNHN
Paris, France
5 mm

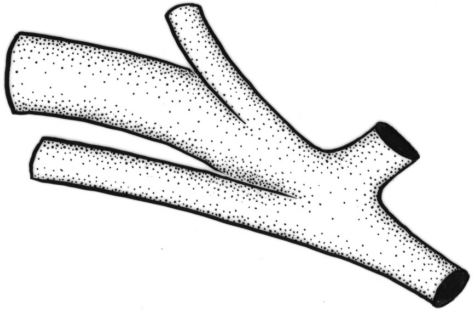

© MNHN
Paris, France



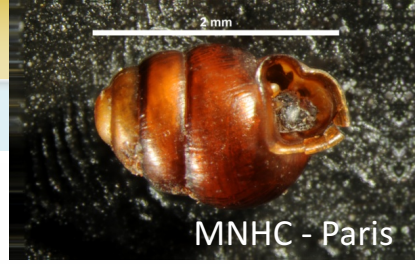

MNHN R04077 - Lectotype - *Ammonites bonnardii* d'ORBIGNY, 1843
R.04077
A.1.8.4.4
5 cm
© Muséum national d'Histoire naturelle




IGNM 121108
20mm



iDigBio
Integrated Digitized Biocollections



Digitisation Maturity?

Figure 4 Digitisation maturity model

http://tinyurl.com/digitizationala

GBIF TF: *Accelerating* Discovery of Biocollections Data

#mustGoFaster!
#needMetadata!

Why do this global survey?

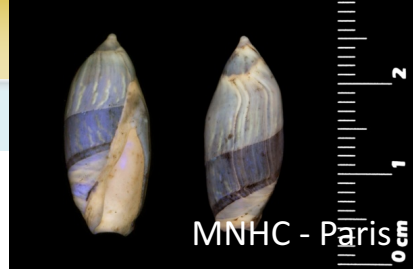
- Look at *digital readiness*
- Update *current baseline*
- Identify *benefits of digitization*
- Document *key impediments*

Photo by S. Masinde: GBIF TF meeting, 3 Nov 2015, Washington

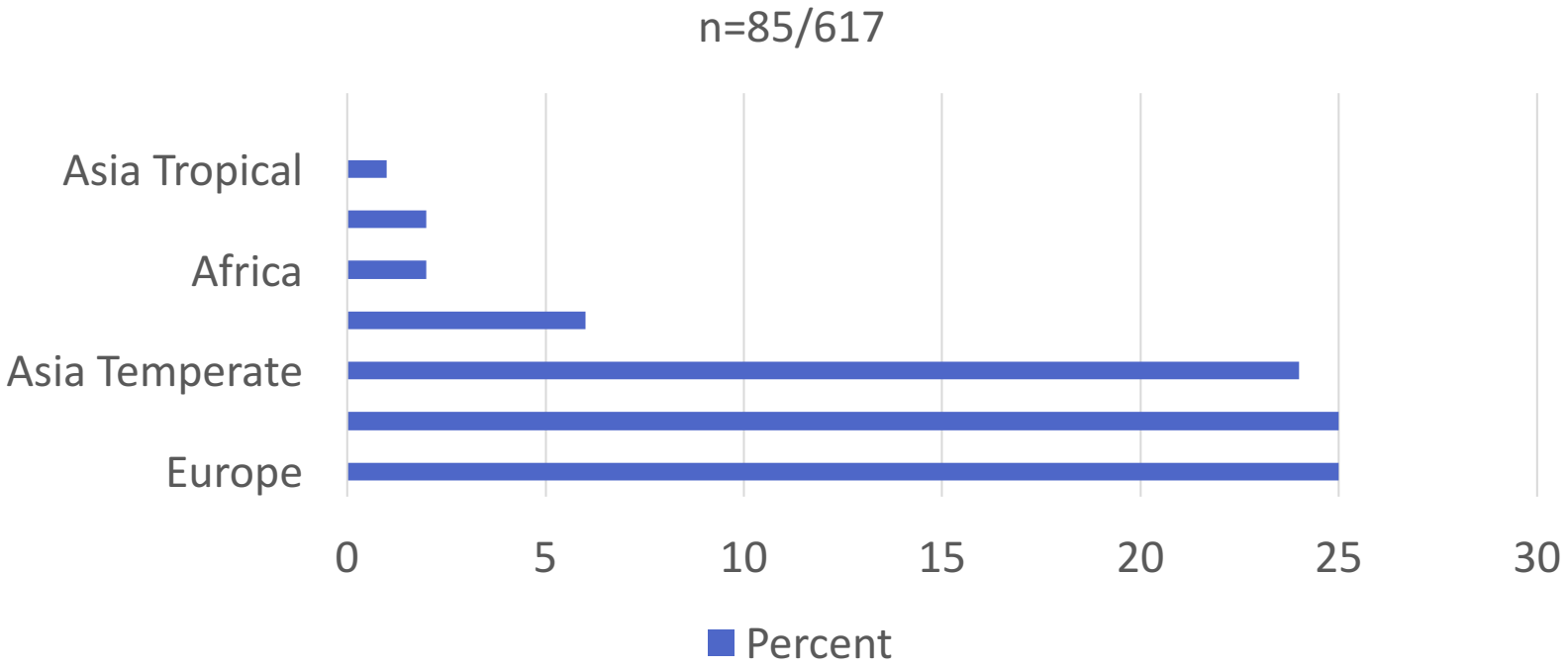
GBIF Task Force Members

1. Leonard Krishtalka, USA, Chair
2. Barbara Thiers, USA
3. Deborah Paul, USA
4. Eduardo Dalcin, Brazil
5. Masanori Nakae, Japan
6. Ian Owens, UK
7. Jean Ganglo, Benin
8. Marc Pignal, France

- Siro Masinde, GBIF's TF Coordinator
- Shari Ellis, Consultant to TF & iDigBio External

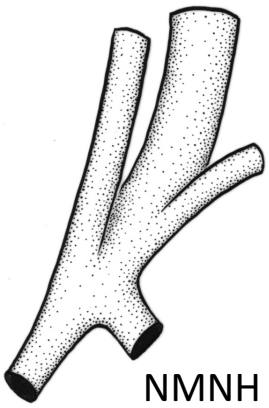


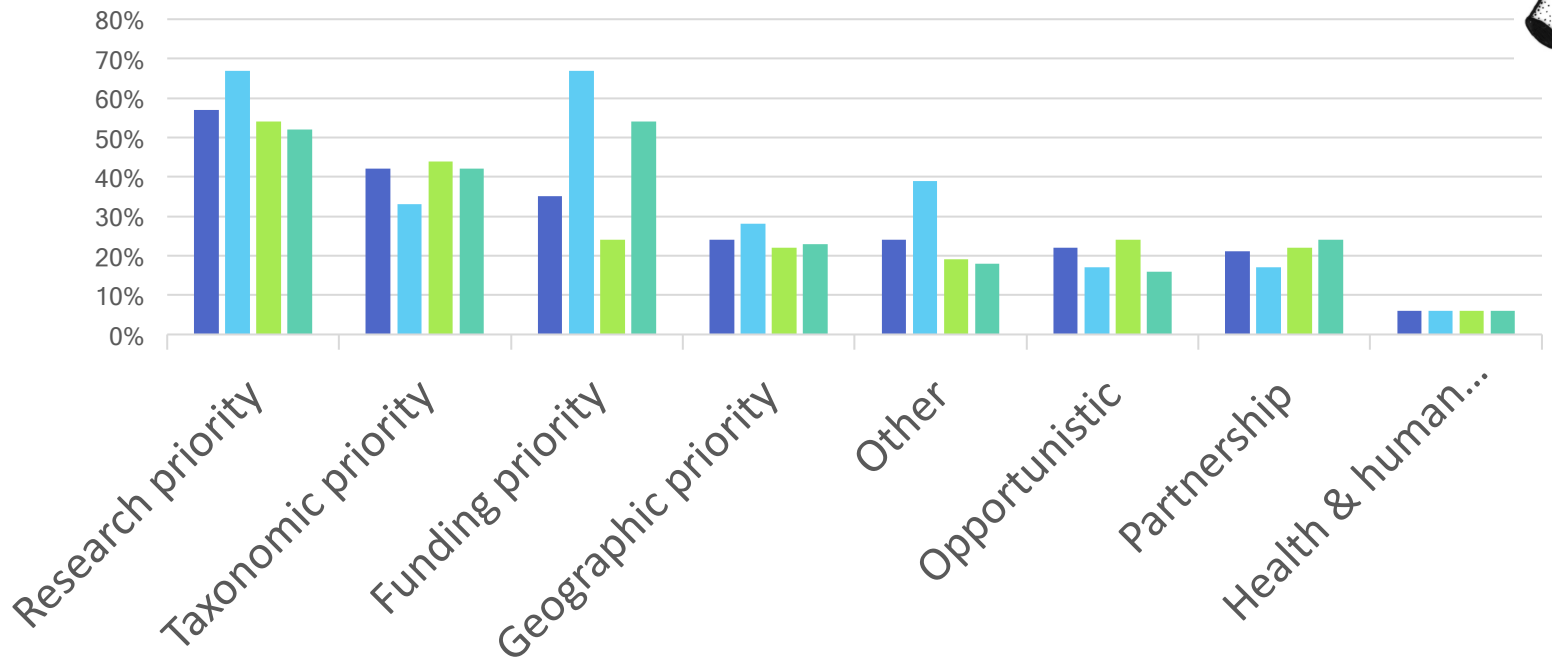
Respondents Saying They Have a Malacology Collection - by Region

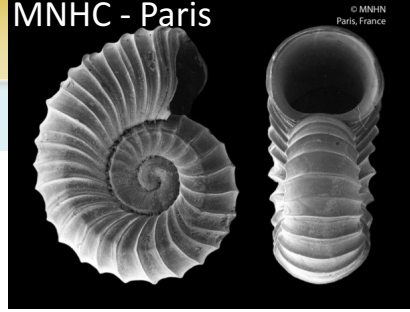


- 85/617 or 14% indicate they have a malacology collection.
- 20 of the 25 North American Collections responding are based in the US.

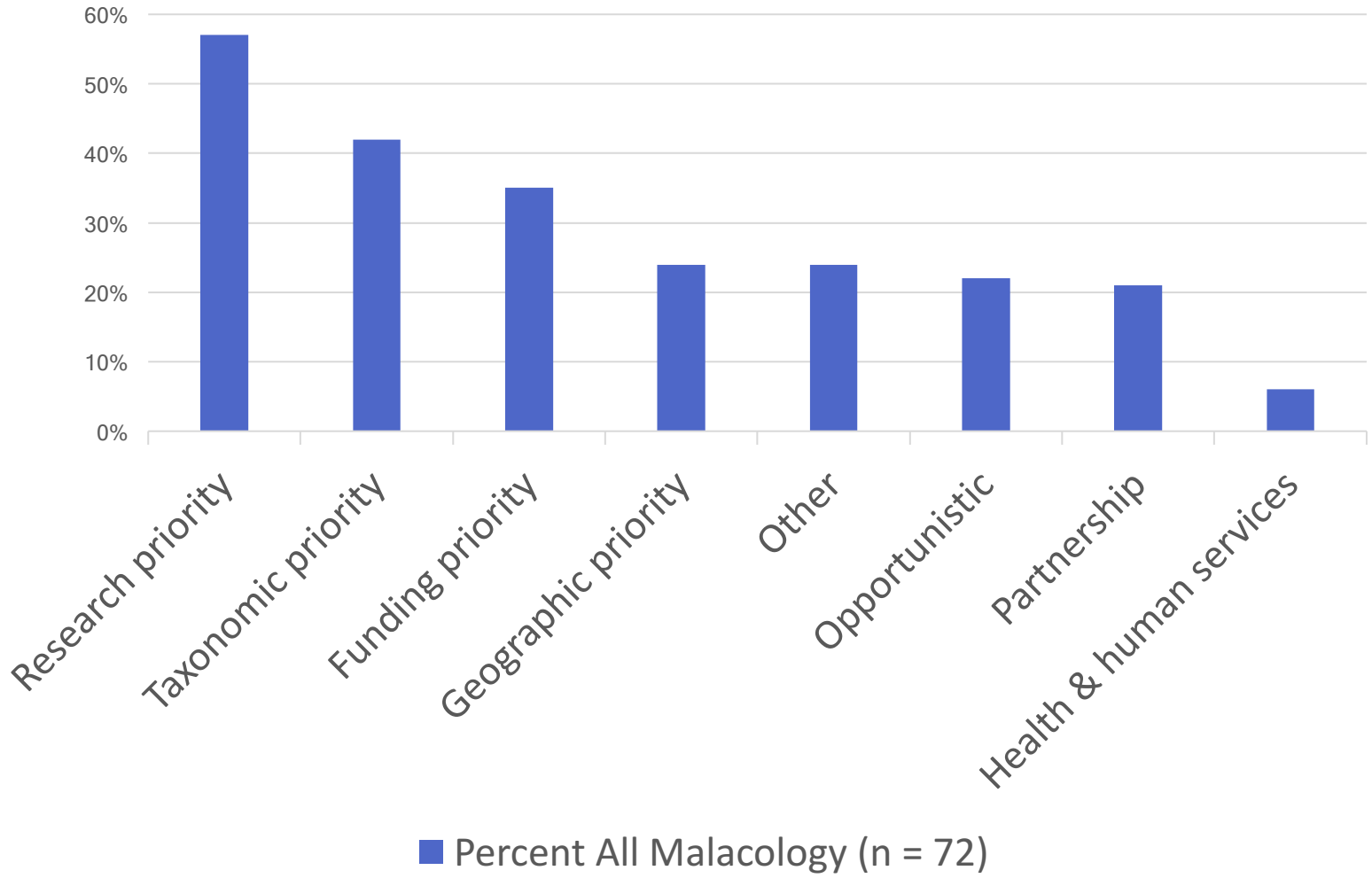
Setting Priorities for Digitization



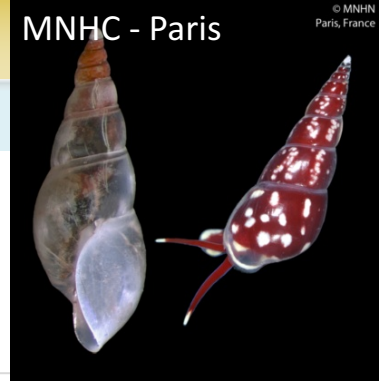




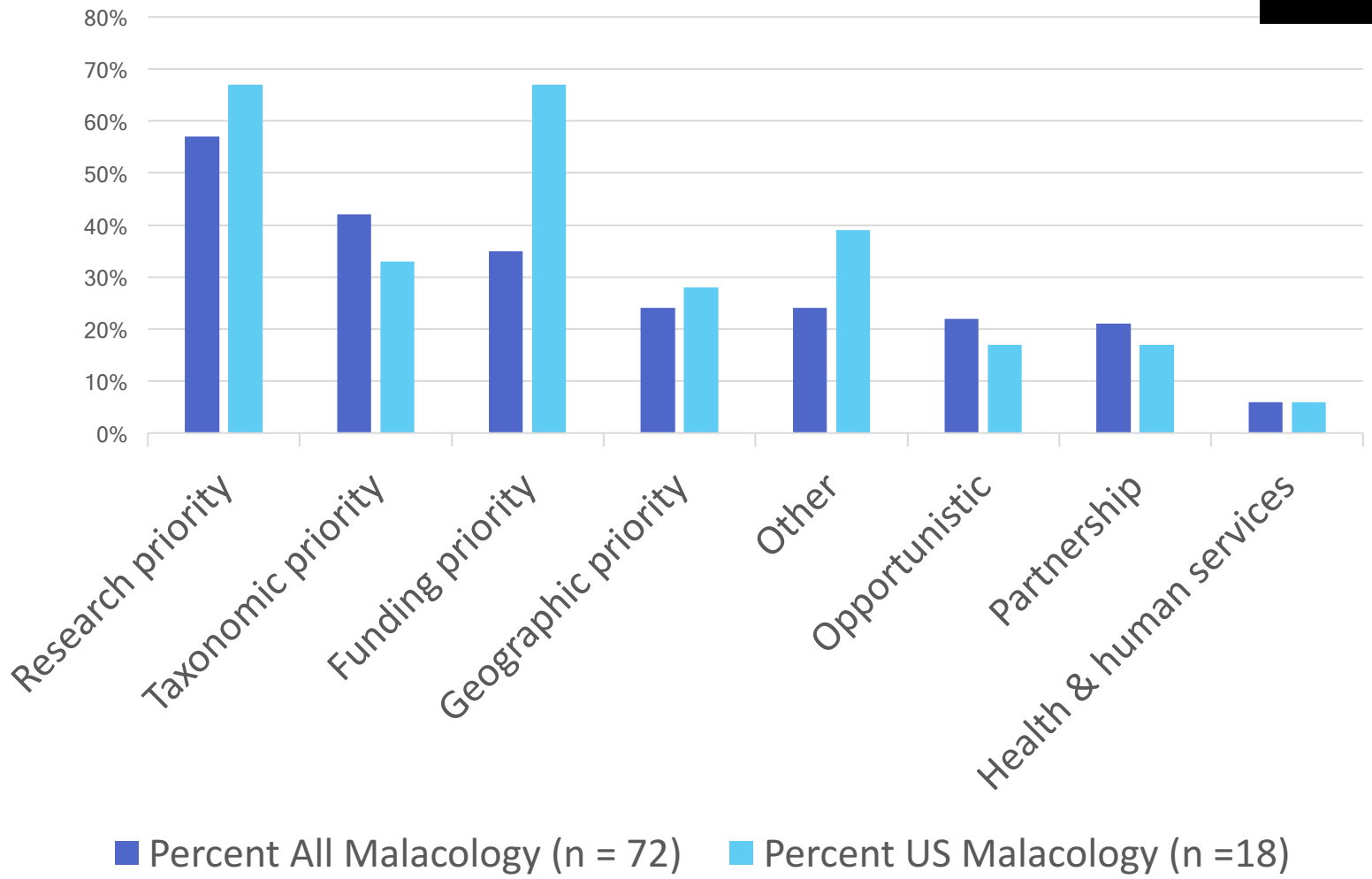
Setting Priorities for Digitization



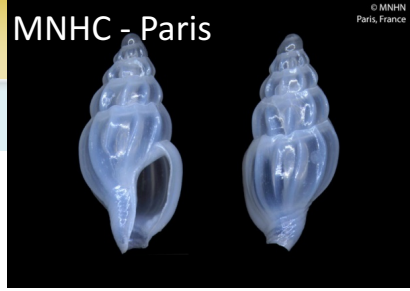
iDigBio
Integrated Digitized Biocollections



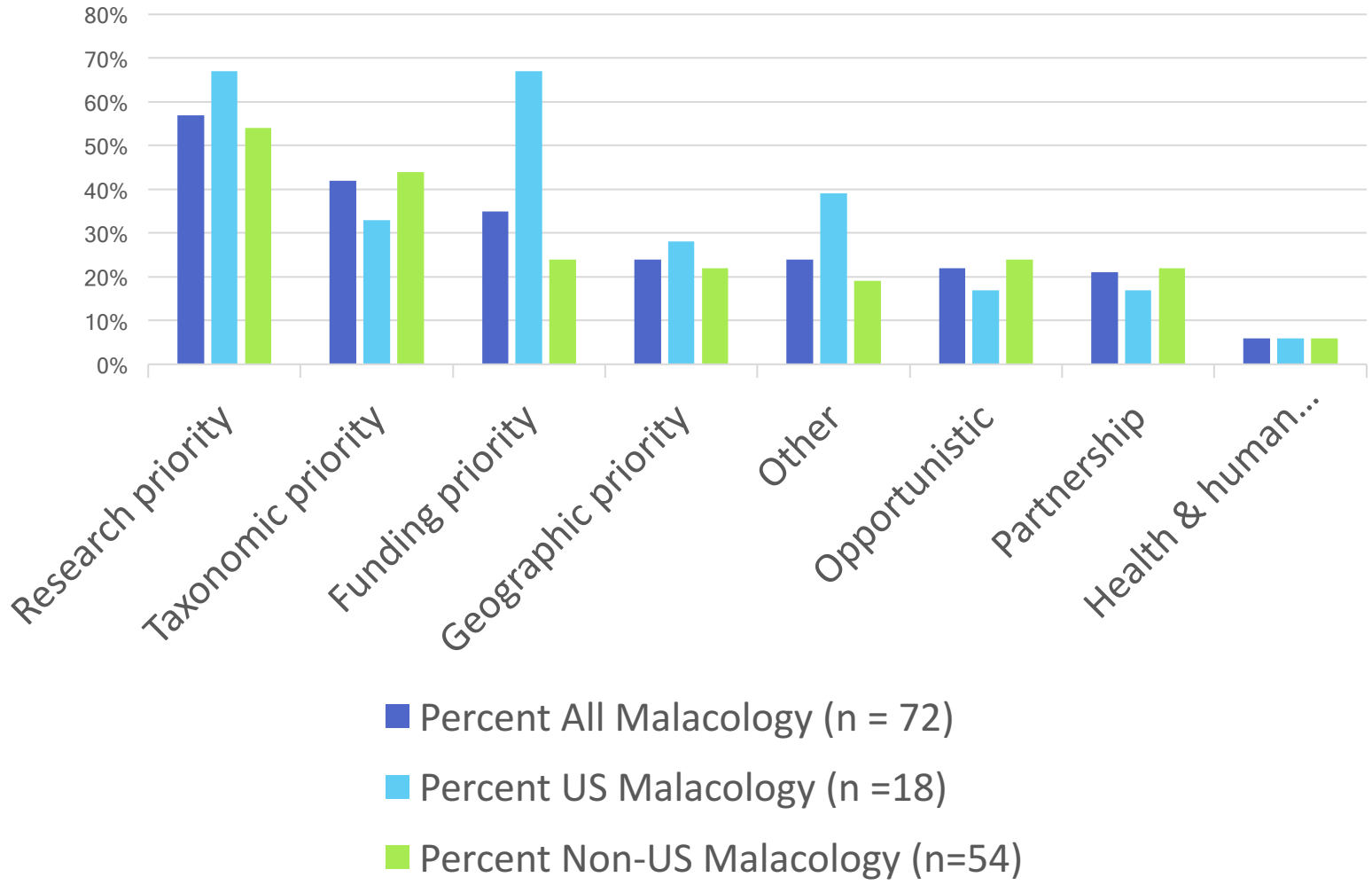
Setting Priorities for Digitization



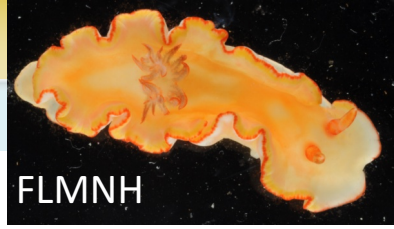
iDigBio
Integrated Digitized Biocollections



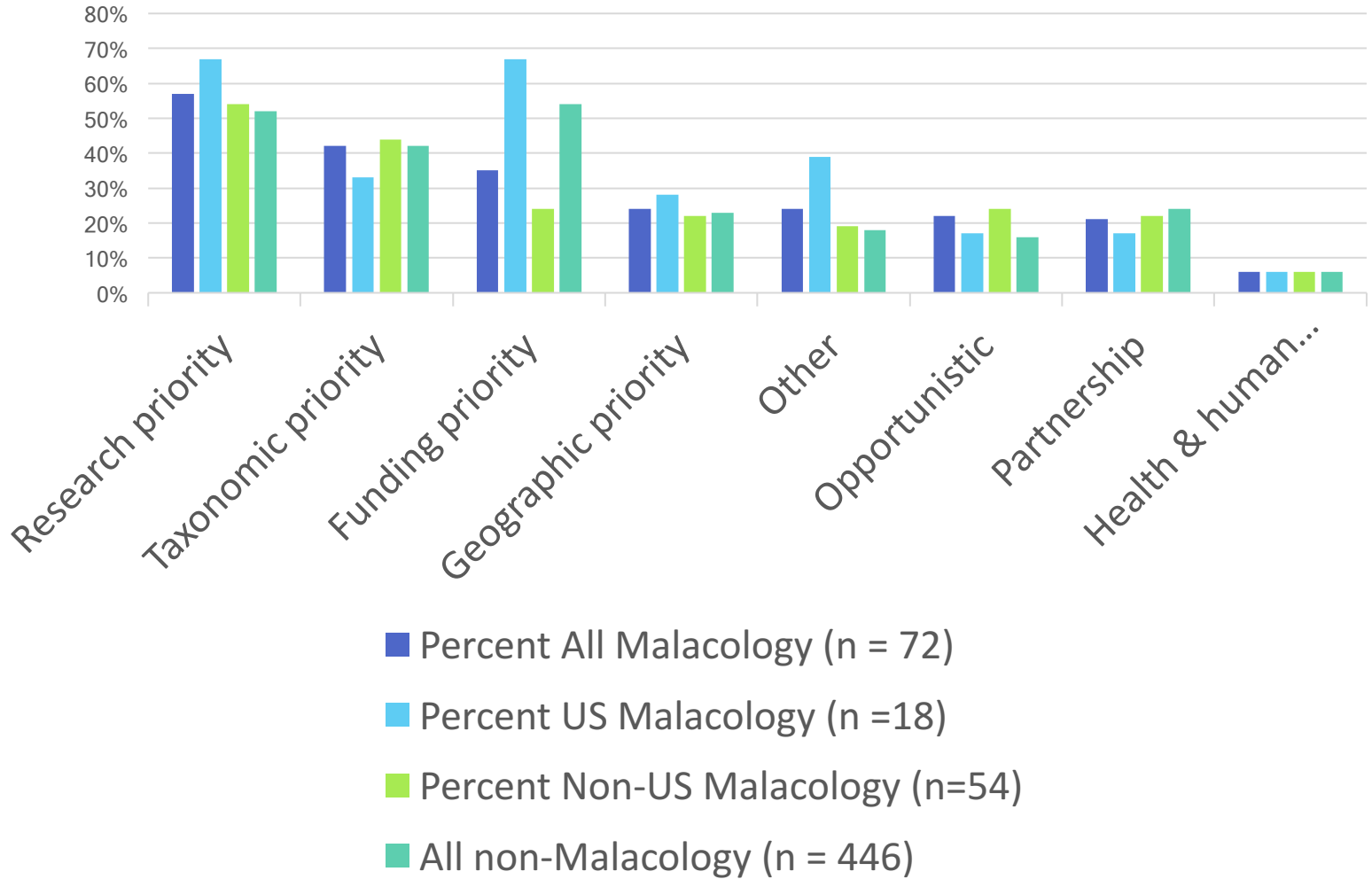
Setting Priorities for Digitization

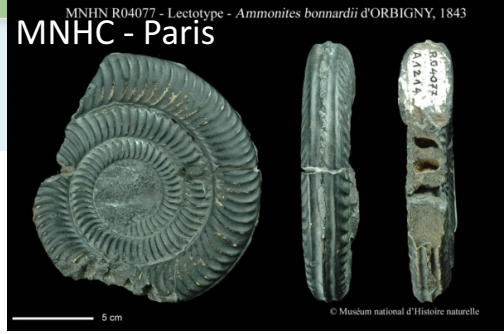


iDigBio
Integrated Digitized Biocollections



Setting Priorities for Digitization

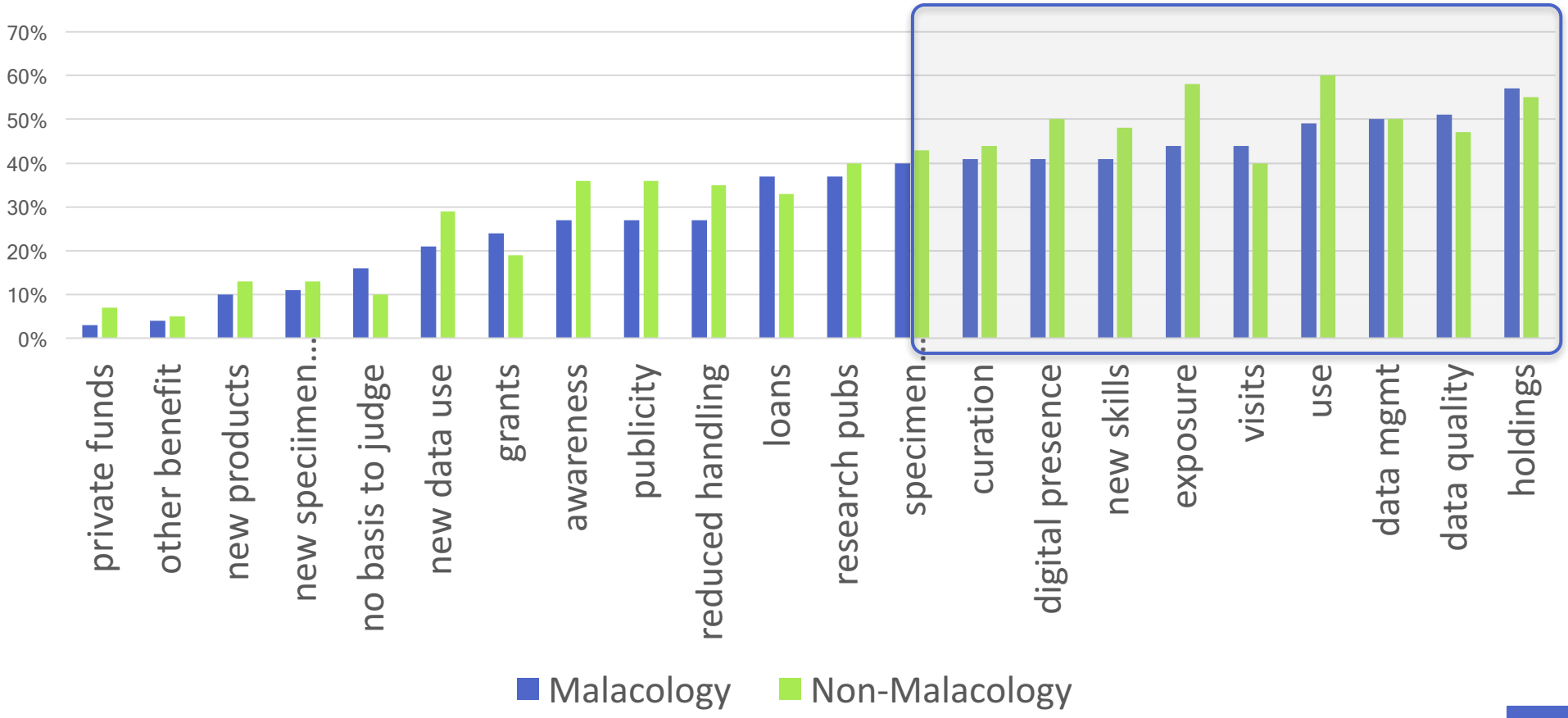




Reported benefits of digitization

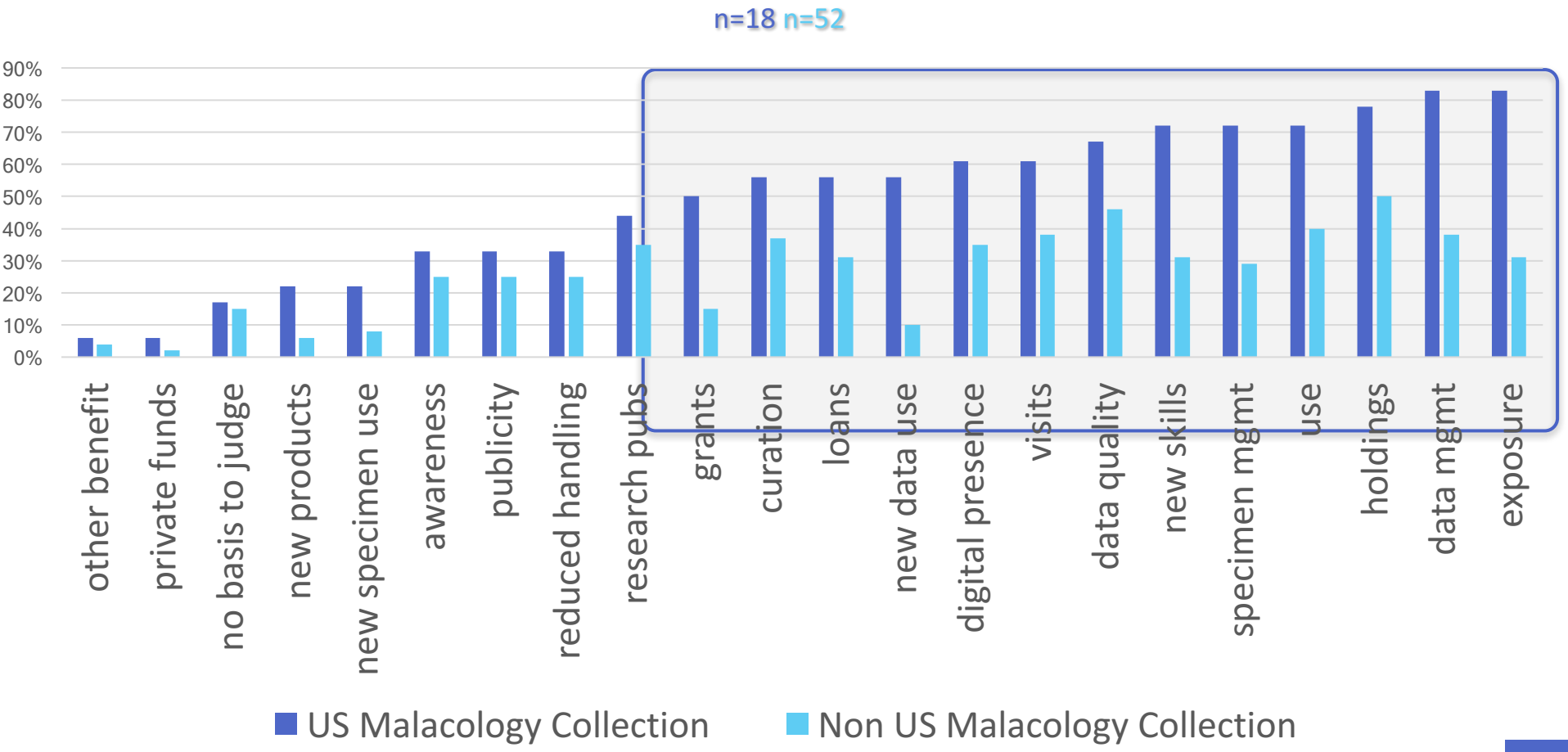
Malacology / Non-malacology

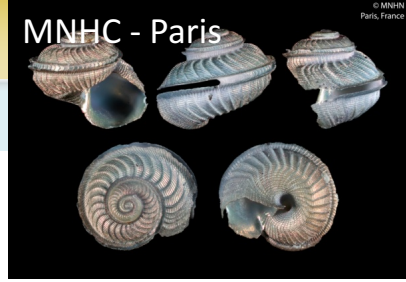
n=70 n≅445



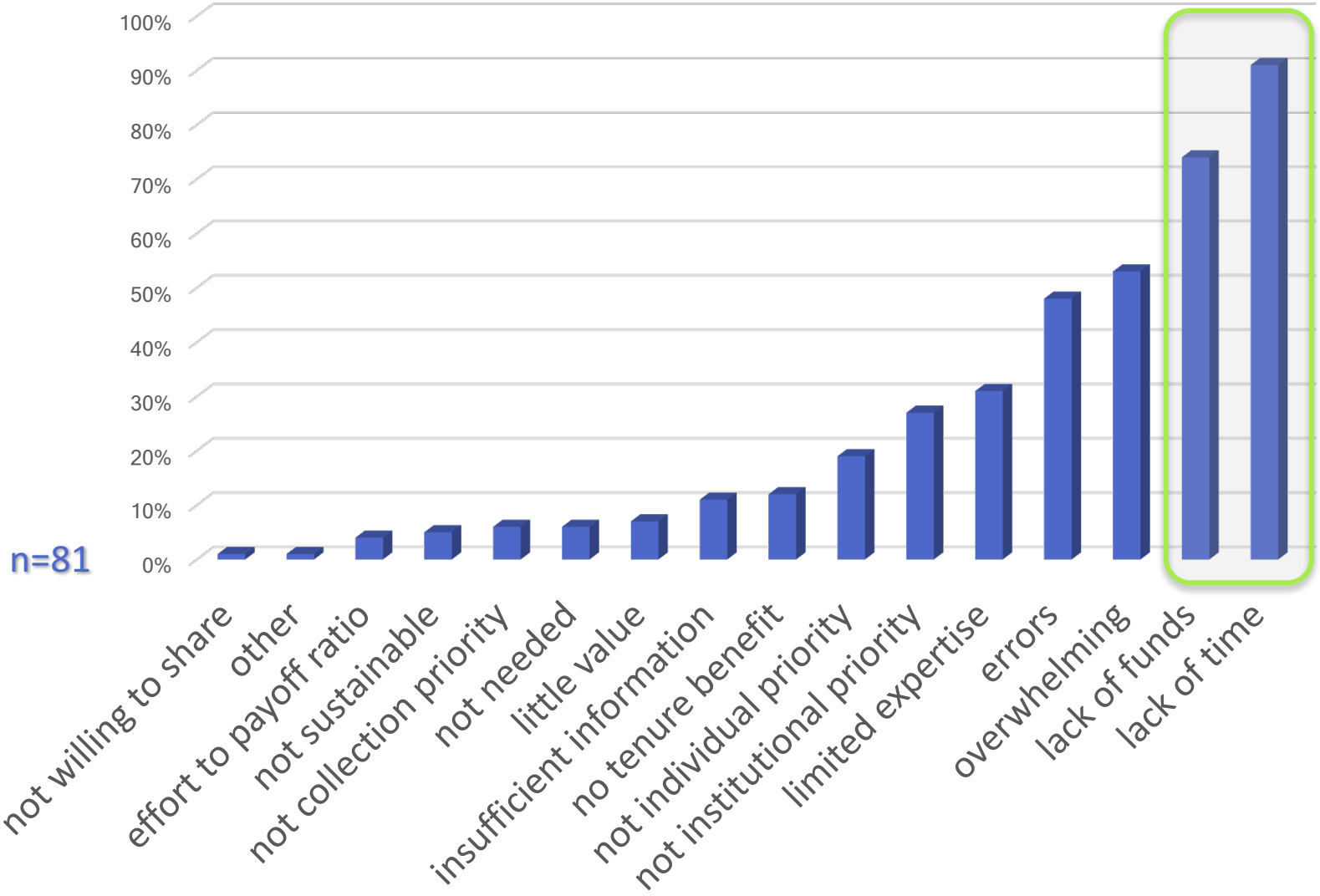
Reported benefits of digitization

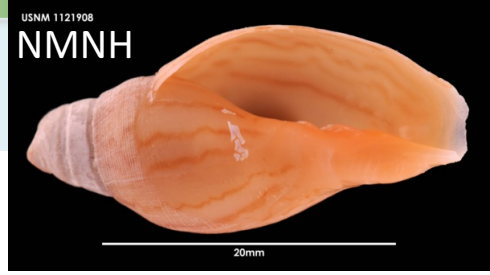
US Malacology / Non-US Malacology





Obstacles to digitization



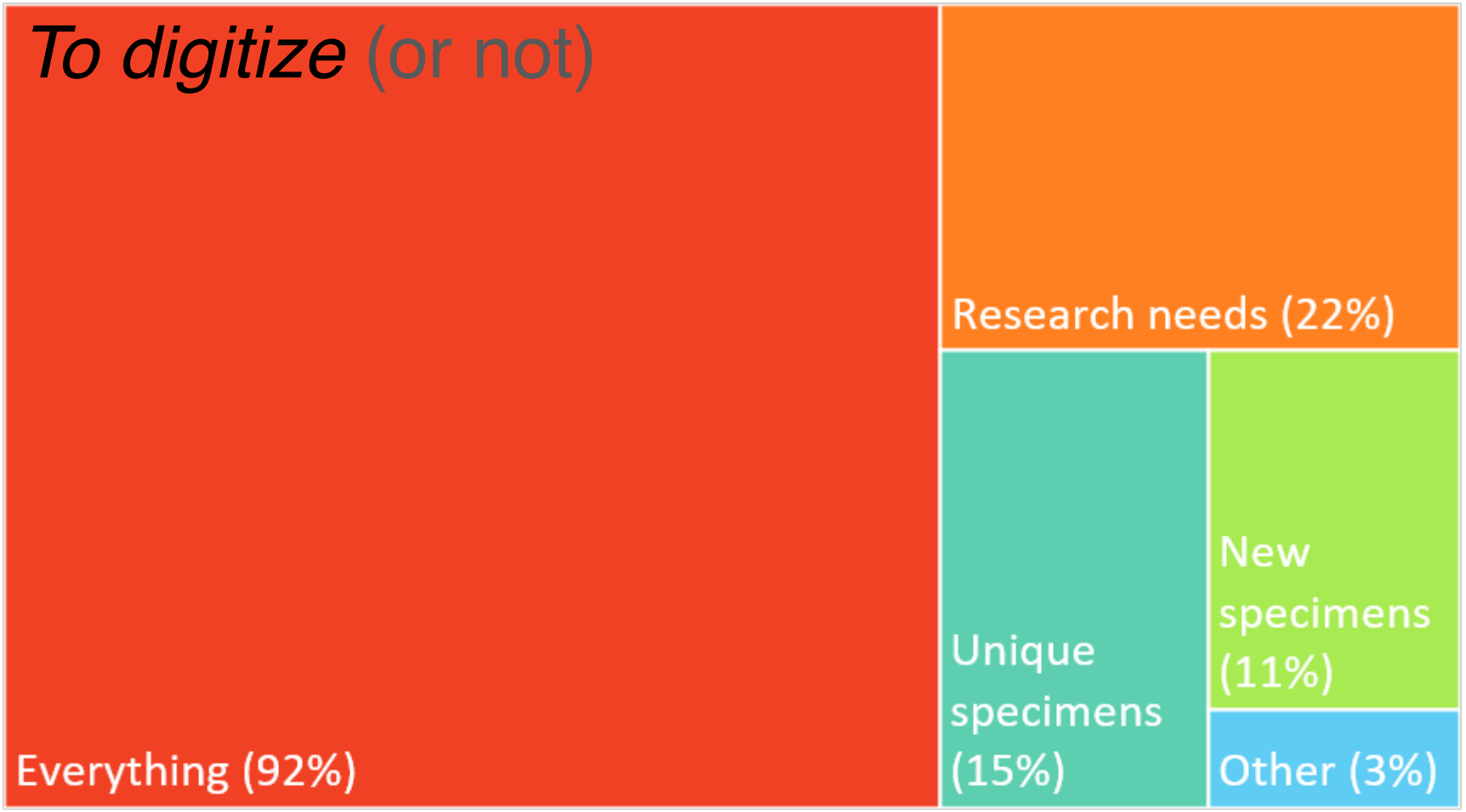


Grouping digitization obstacles

n=81

On overcoming obstacles - beyond lack of time and funds

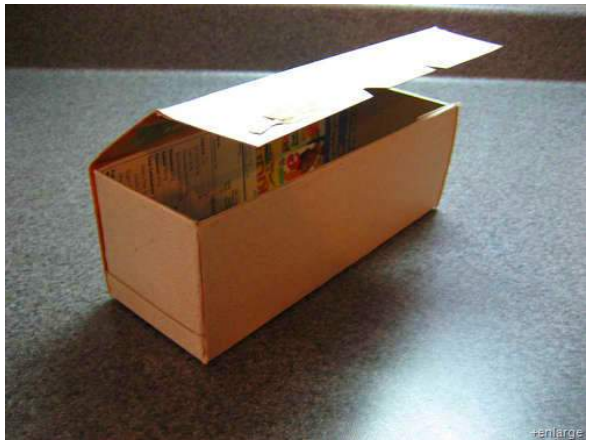
Grouping obstacles to digitization	Possible responses to consider
• size of task is overwhelming	• A given – start somewhere
• not an institutional / departmental priority, not good effort / payoff ratio, lack of perceived need, deemed not valuable or beneficial	• Focus on administrators and directors – provide statistics and use cases. SPNHC 2017 example.
• data has errors, limited expertise, lacking information on the digitization process	• In year 6 of the ADBC program, at least in the USA, human and online resources are now abundant. Spread your knowledge.



n=72

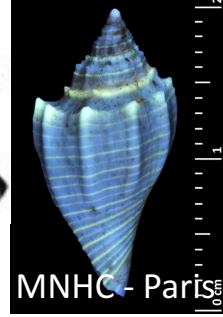
- Do not plan to digitize (n=0)
- Other (3%)
- New specimens (11%)
- Unique specimens (15%)
- Research needs (22%)
- Everything (92%)

What is Metadata? Why is it key?



*"If data are LEGO bricks, then **metadata are the shiny box and instructions**. They enable discovery of your collections, your datasets, and make it possible to assess relevance for particular needs, so it pays off investing some time providing them."*

From Canadensys (http://www.canadensys.net/publication/data-publication-guide)

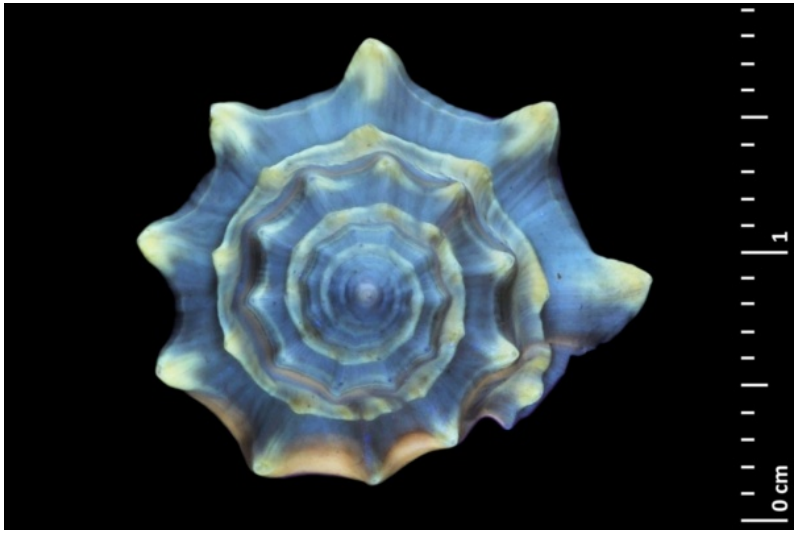


Instant Up-To-Date Collections Data? No more surveys!? (well, fewer anyway)! How? Metadata

- From GRBio, to GBIF, TDWG, Fantastic Fishes and Where to Find Them* @iDigBio's US Collections List - we're working on interoperable APIs – but need Metadata to realize the dream.
- The data you provided for this workshop!
- fast, up-to-date information
- community will need to adopt / engage
- perhaps in your database, in the future
- See GBIF TF recommendations

* Randy Singer (iDigBio Grad Student), Kevin Love (iDigBio IT)

Let's get started



idigbio.org/wiki

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/iDigBio

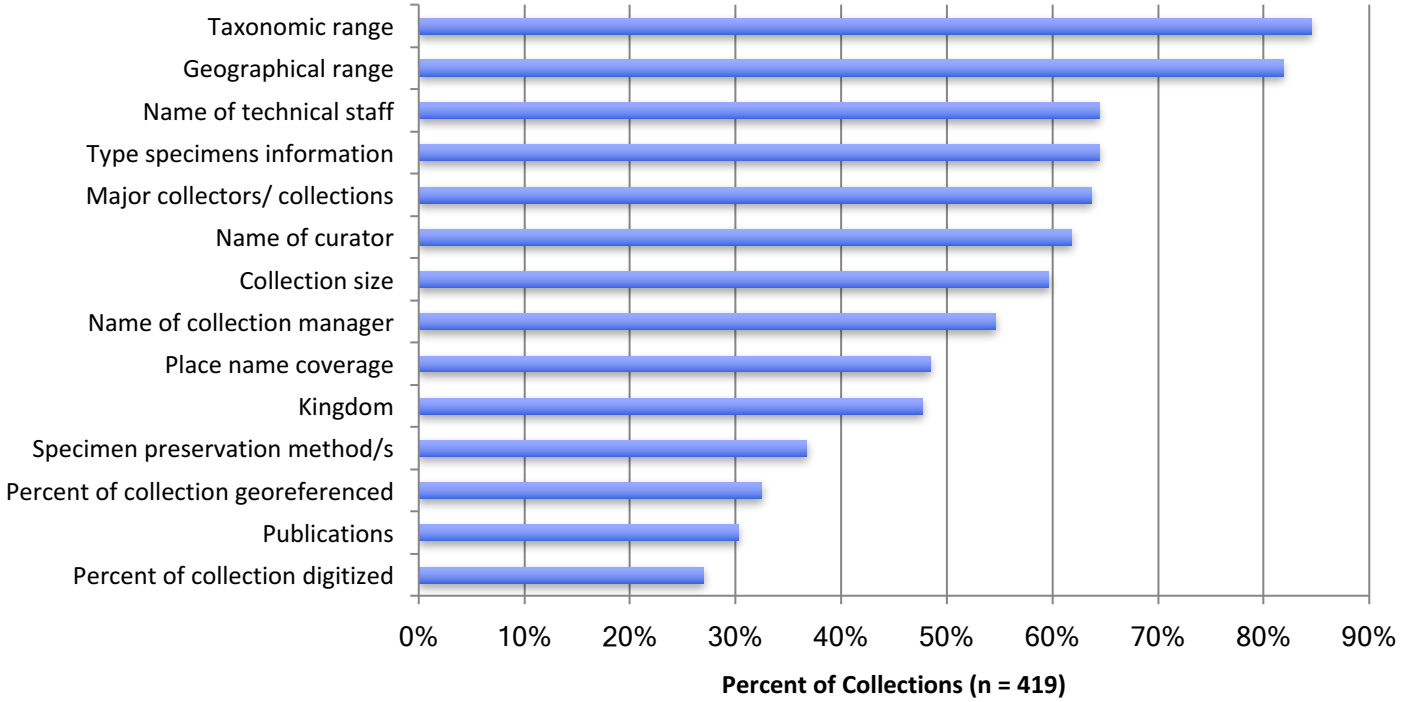
idigbio.org/rss-feed.xml

idigbio.org/events-calendar/export.ics

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Data and Metadata.
It's about discovery and data re/use.
It's about feedback and accountability.
It's about credit and attribution.
Make sure your data are not under a rock.

METADATA BEING PROVIDED



Over 80% of respondents ranked taxonomic and geographical ranges as critical metadata that should be provided. Other highly ranked elements are: name of technical staff, type specimen information, major collectors/collections, name of curator, collection size, and name of collection manager

Key Issues ~ from iDigBio's point of view

Scale

- Collections data now big data
- New approaches to enabling use -- desired

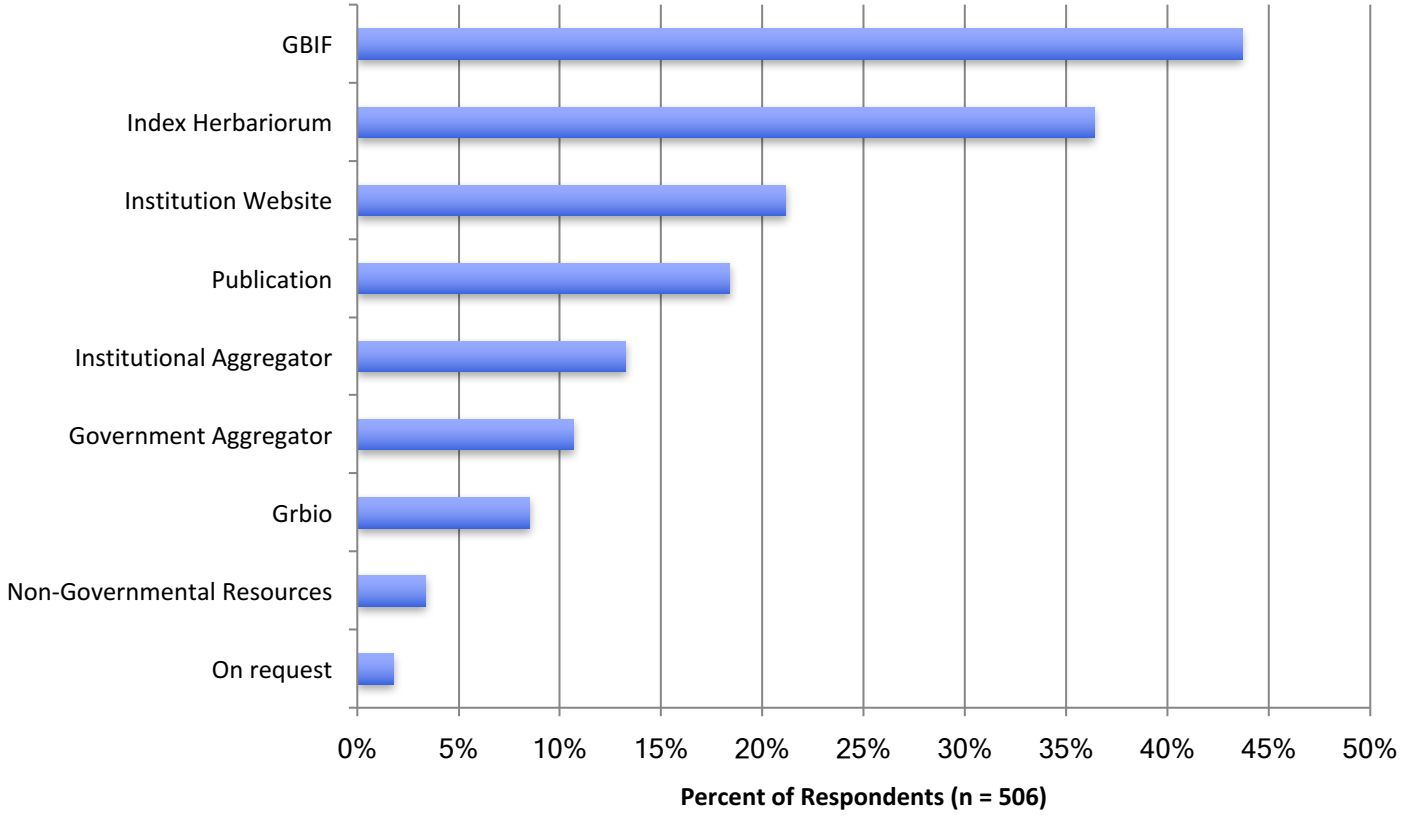
Scope

- Need more collections data used in
  - research, education, policy, *industry*
- *Need new stakeholders and new uses*
- Data have gaps and need to be *linked*
  - *But maybe this is a selling point?*

Speed

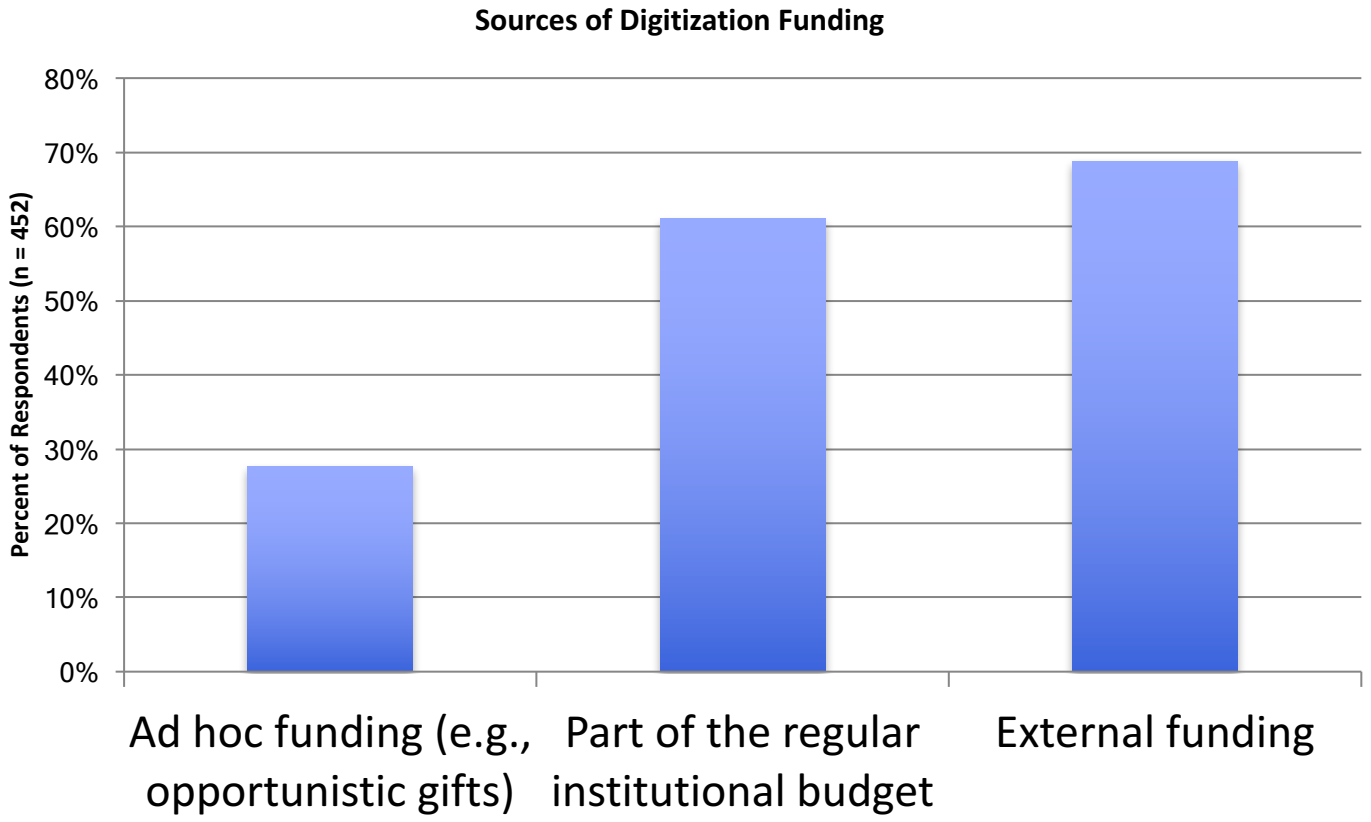
- Must find a way to speed up publishing
- Need worldwide participation
- **Required to remain relevant & funded**

WHERE METADATA IS SHARED

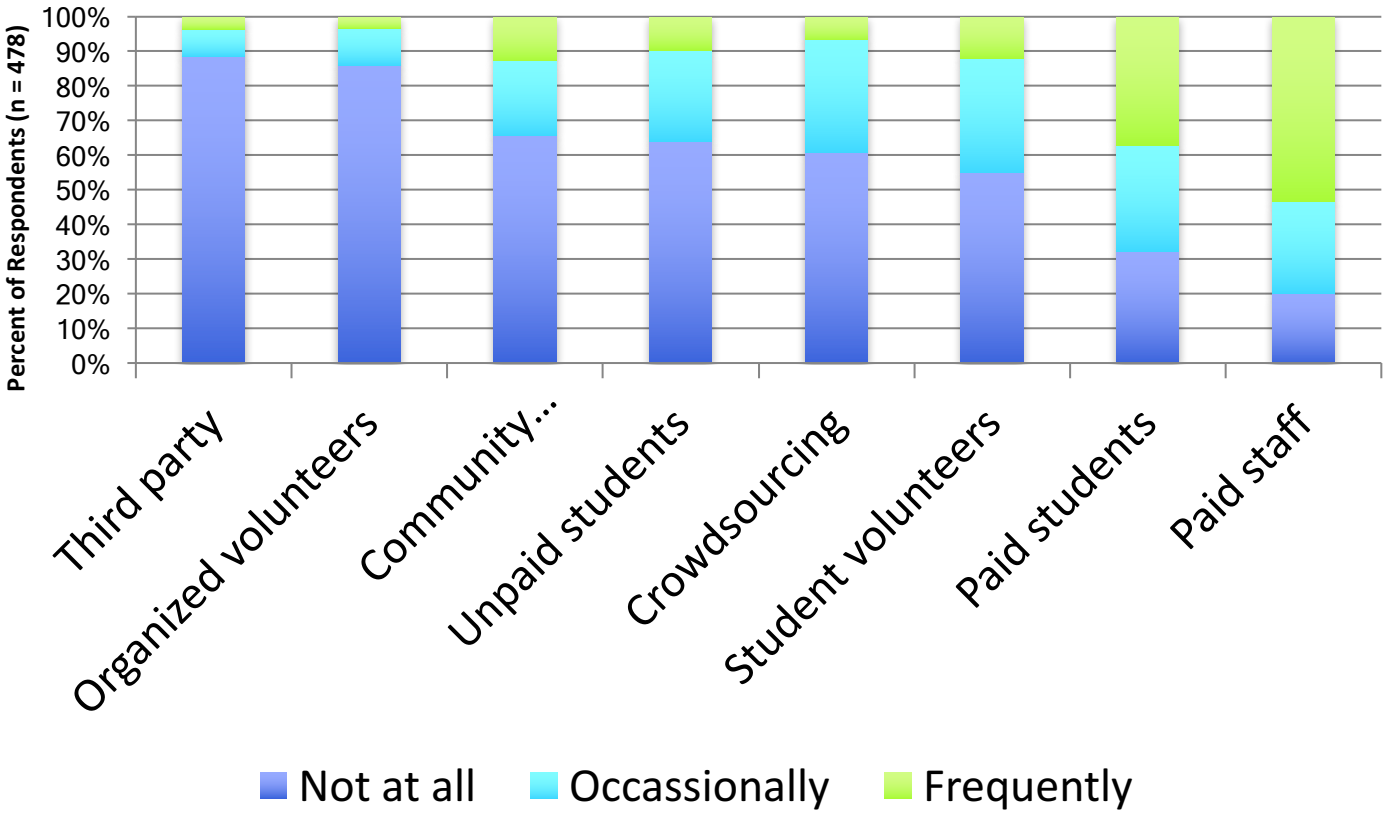


Over 60 locations where metadata is shared, some still analog

SOURCES OF FUNDING



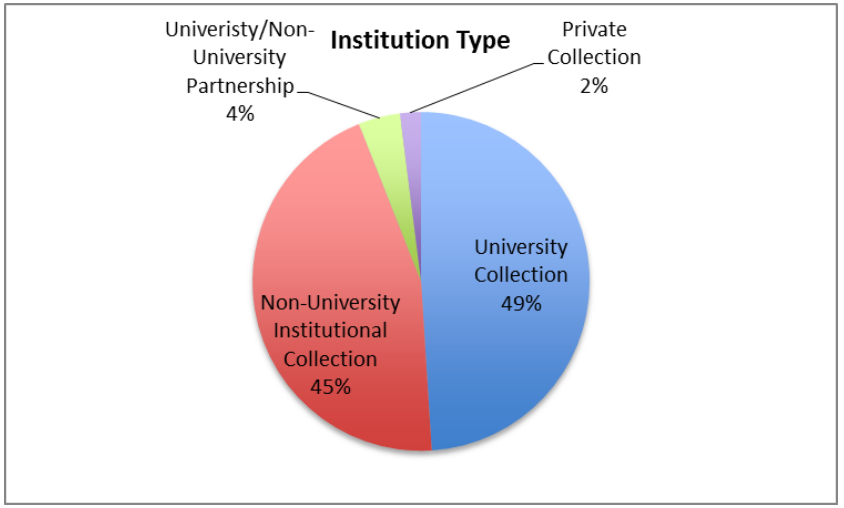
WHO IS DOING THE DIGITISATION?



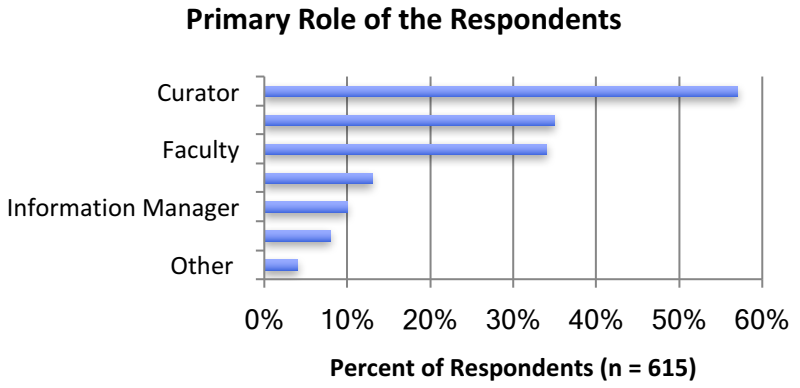
Digitisation is mostly carried out by paid staff (53%) and students (paid, unpaid, or volunteers) (59%), and rarely, third party organizations (2 to 10 %)

INSTITUTION TYPE ROLES OF RESPONDENTS

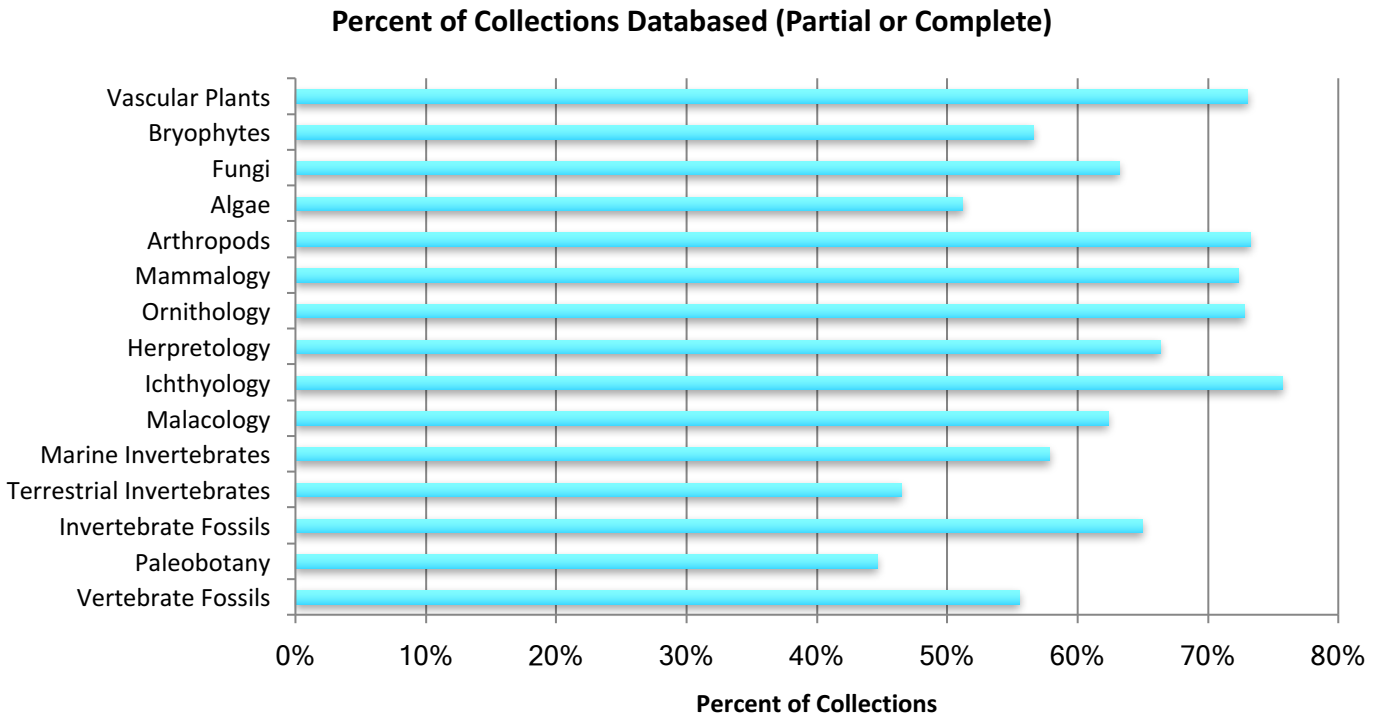
- 76% individuals based at publicly funded institutions:
  - 40% universities
  - 36% non-university



- C. 92% primarily curators or collection managers with 10% as head of research and collections



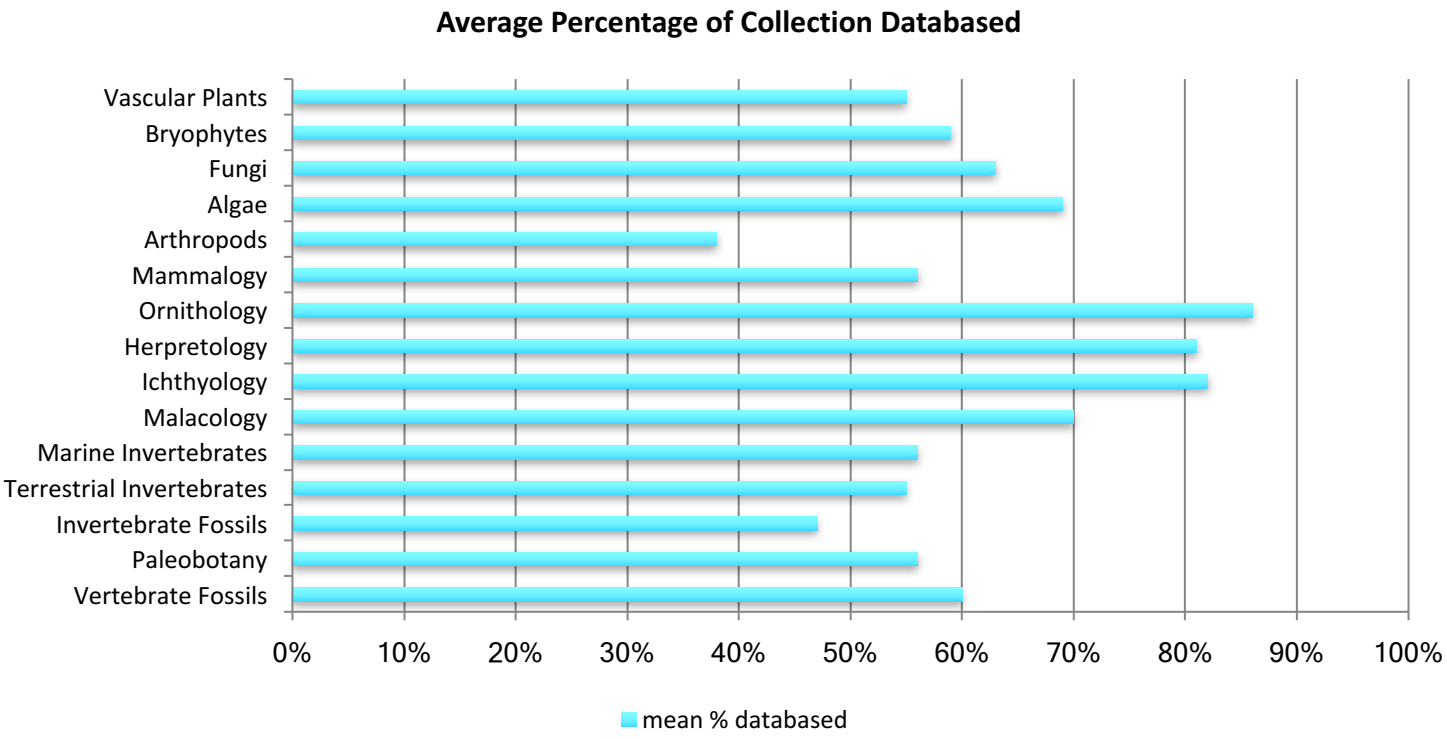
CURRENT STATUS OF DIGITISATION



86% (615 resp.) currently digitizing or completed digitizing at least some or all of their coll.

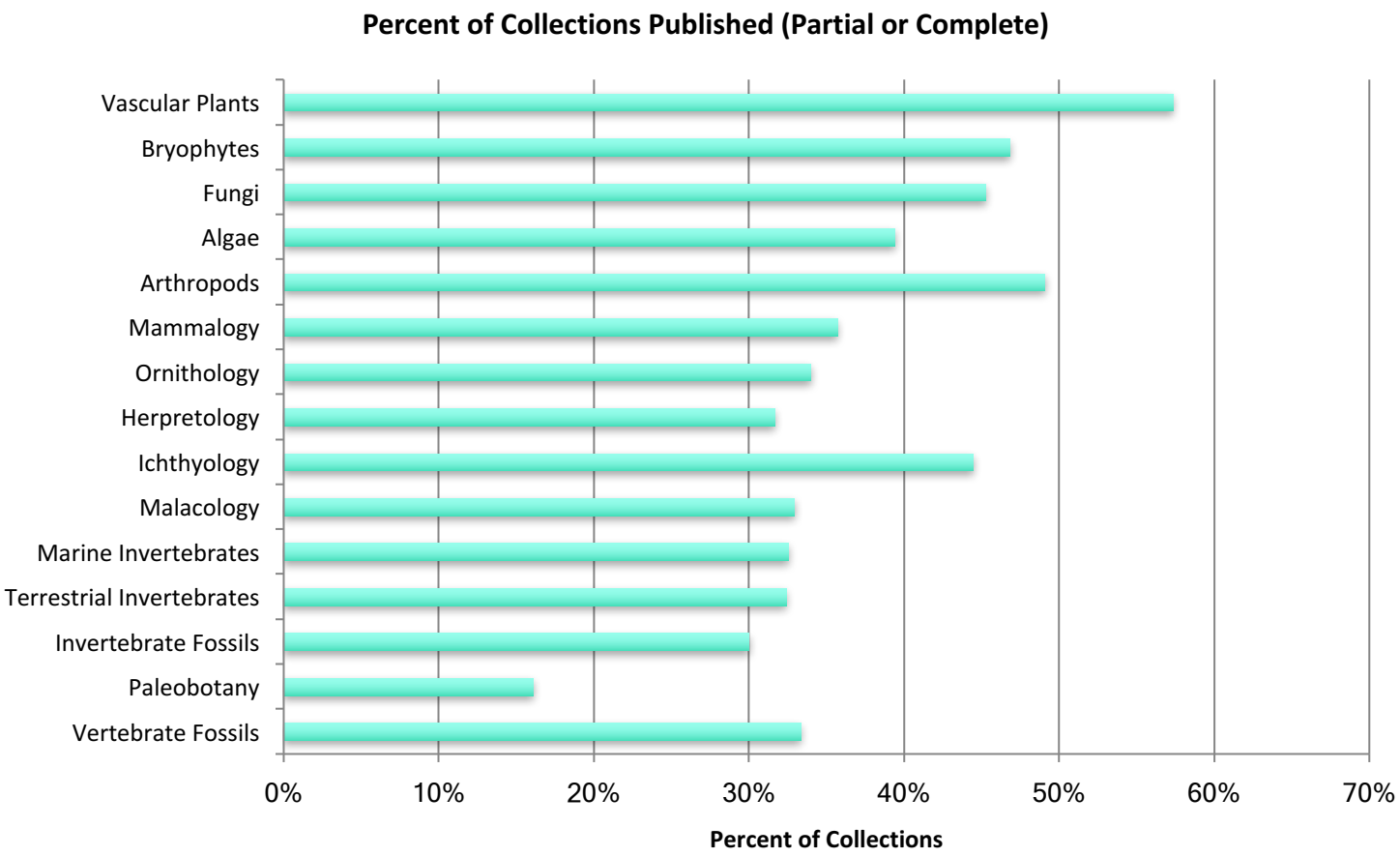
Very few respondents (1% or 5 individuals) reported not digitizing and have no plans to do so

AVERAGE % OF COLLECTION DATABASED

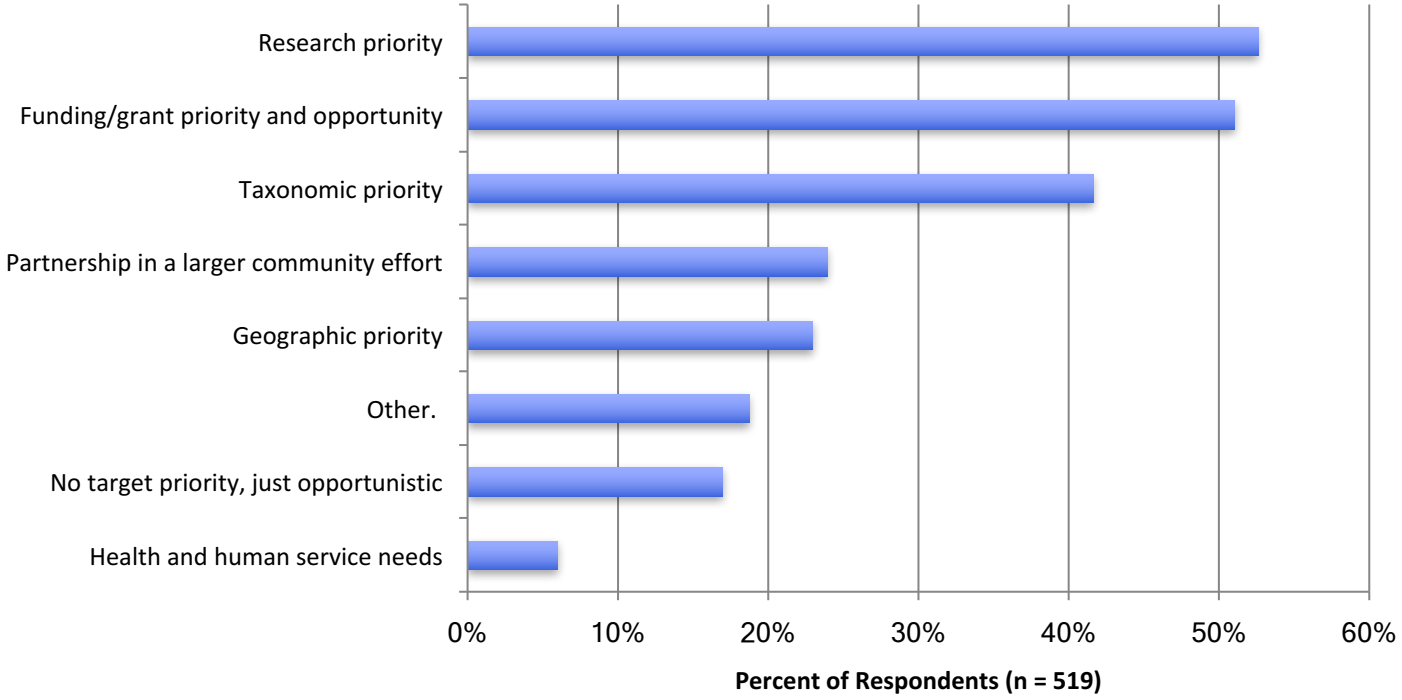


13 out of 15 collection types report their collections over 50% electronically databased

DATA PUBLISHED

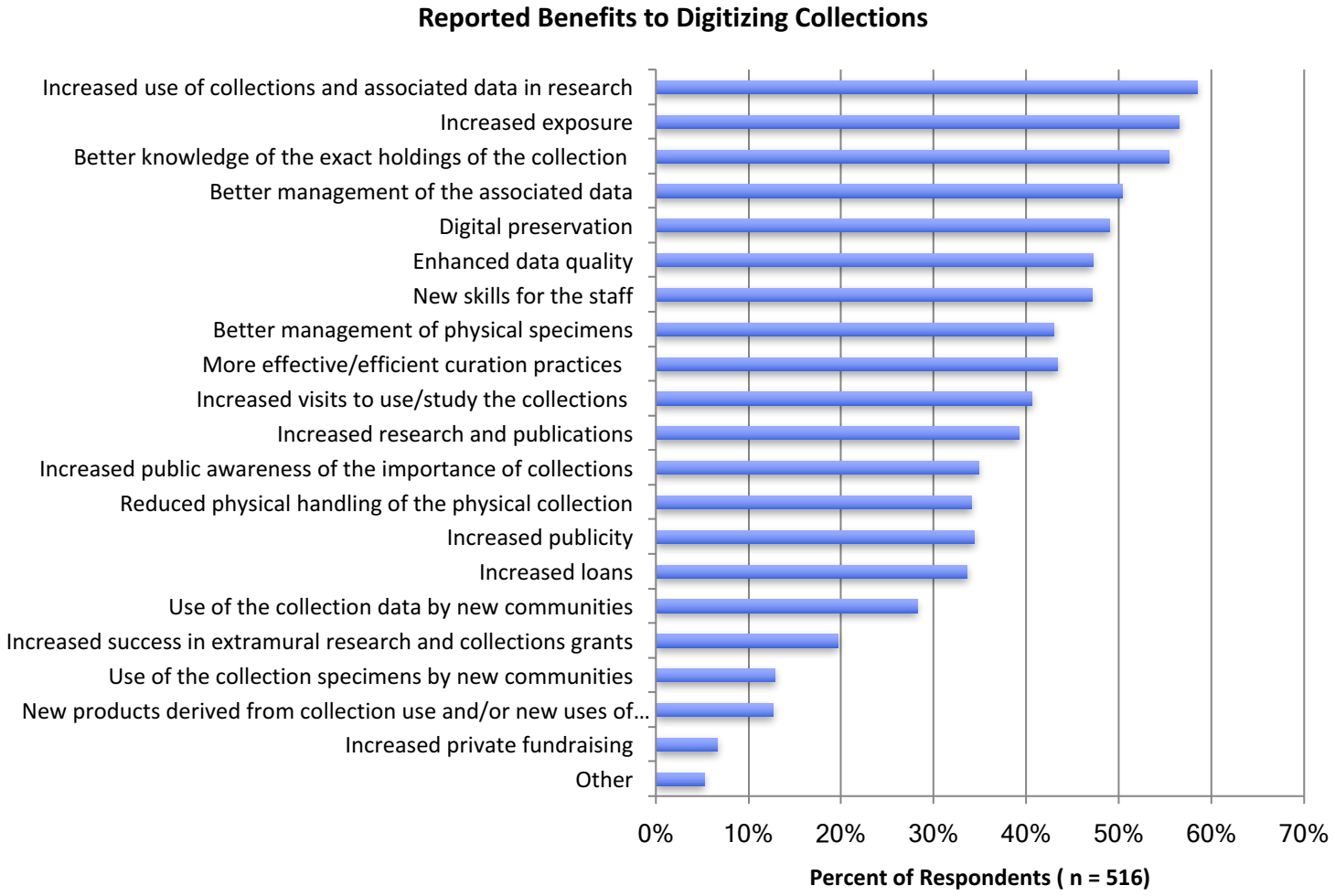


FACTORS INFLUENCING WHAT IS DIGITISED

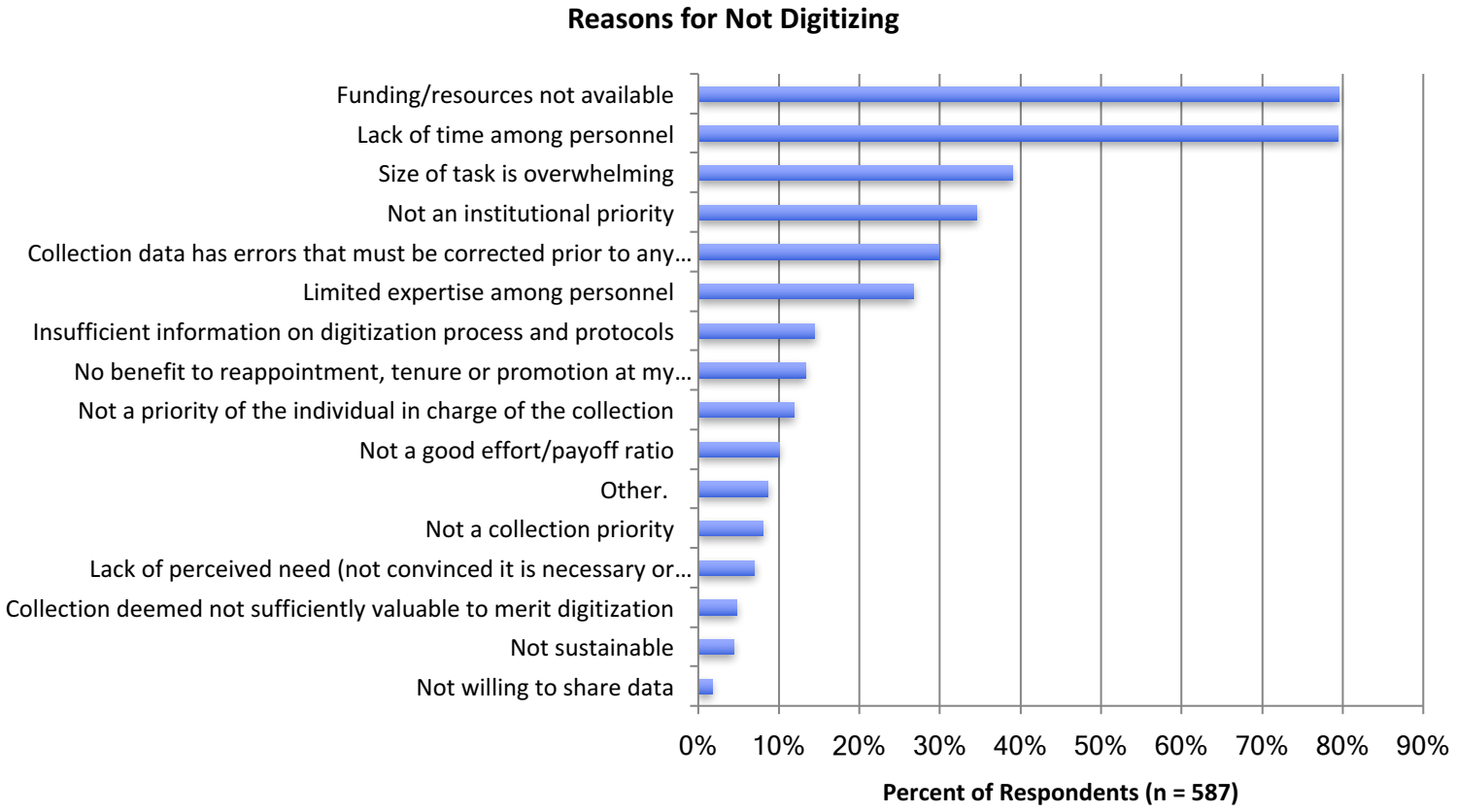


Top 3 priorities influencing what is digitised are:
Research (53%); Funding (51%), Taxonomic priorities (42%)

BENEFITS OF DIGITISATION

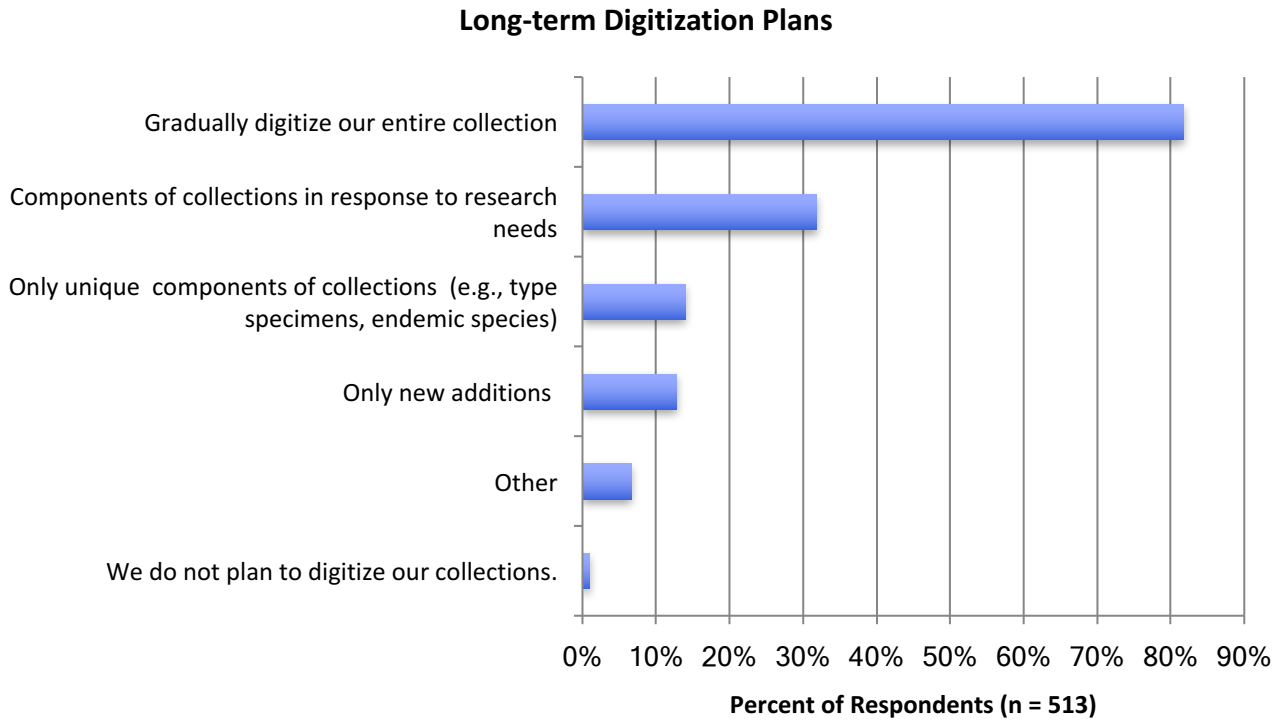


IMPEDIMENTS TO DIGITISATION



- Key impediments are lack of funding/resources & time

LONG-TERM DIGITIZATION PLANS



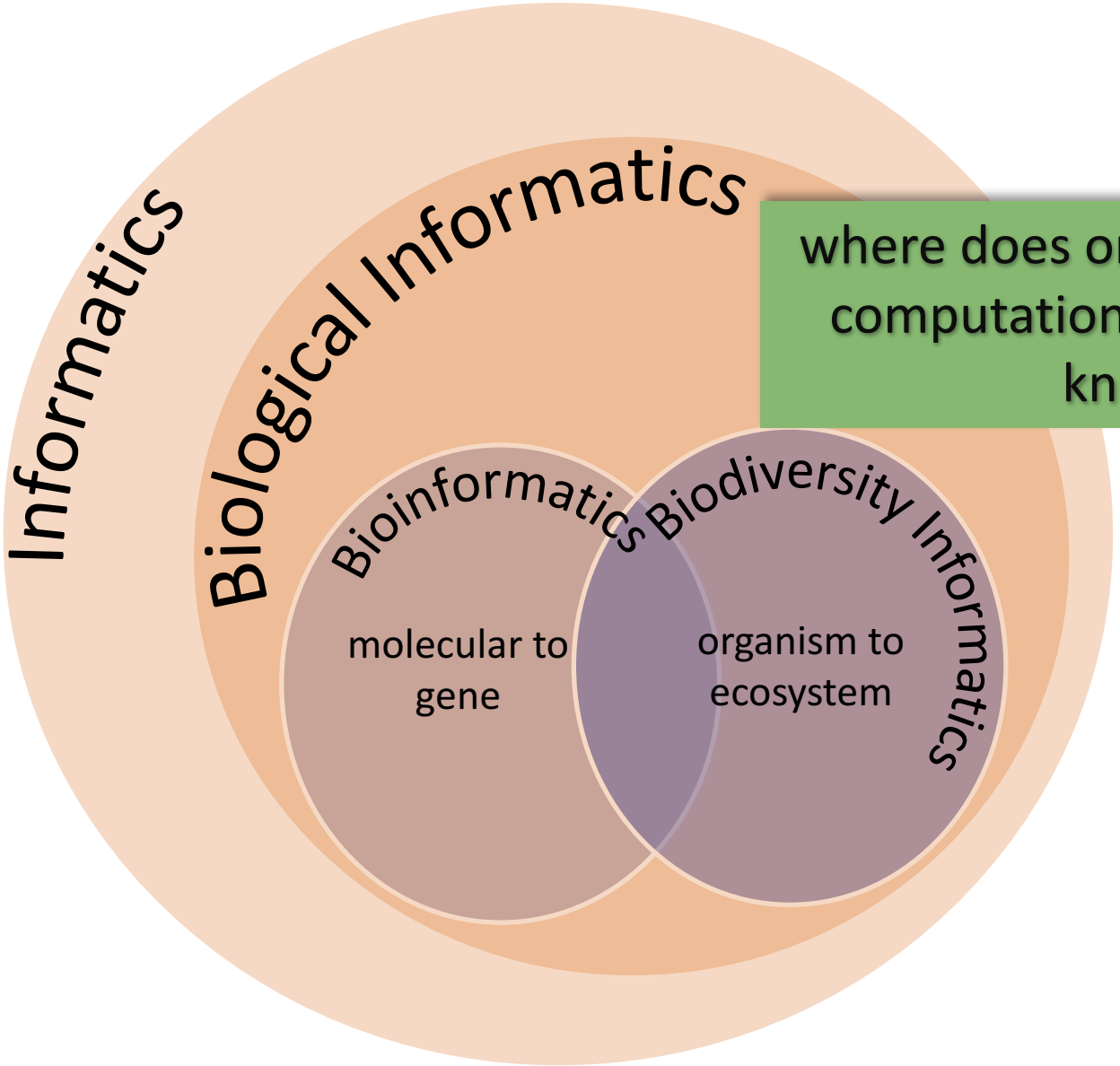
- Over 80% of 513 respondents indicate their institution / organization intends to digitize their entire collection/s

iDigBio
Integrated Digitized Biocollections

NEXT STEPS

- Further analysis of survey and a final report
- Roadmap documents for mobilising NHCs metadata
- Biocollections use-cases for specific communities (researchers, policy makers, educators, etc) to demonstrate the benefits
- Help form a closer-working cooperative network of global bio-collection entities and societies to achieve a critical mass for planning, policy impact, and generating resources
- Task Force recommendations in final report

GBIF



where does one get these data and computational literacy skills and knowledge?

For Biodiversity Informatics:
What human skills and software tools are needed to collect, manage, and do research with this specimen and related data? What infrastructure is needed (hardware and software)? What data standards are needed? What data and computational literacy skills and knowledge are required through the data pipeline from data collection to digitization to data use / re-use?