NATURAL HISTORY DATA PIPELINES:
THE GOOD, THE BAD AND THE UGLY

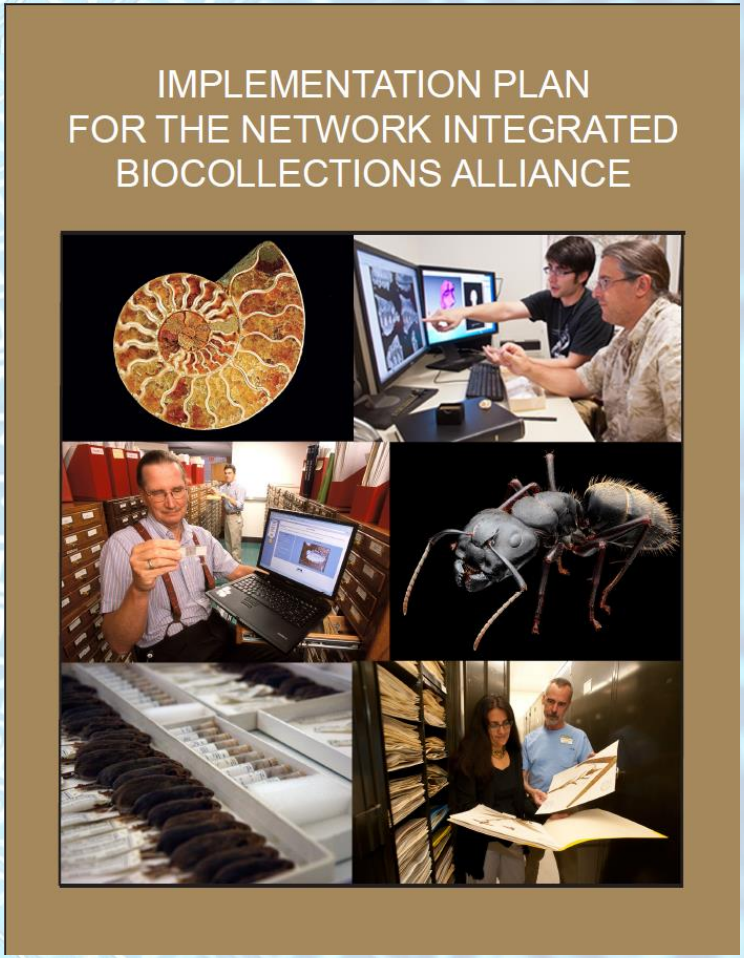
Andrew Bentley

University of Kansas, Lawrence, KS

KU BIODIVERSITY INSTITUTE
The University of Kansas

NIBA Implementation Plan 2012

Goal 2: Advance engineering of the US biodiversity collections cyberinfrastructure. Implement adaptive technology strategies around core discipline standards to enable efficient digitization workflows, effective data management, permanent data archives, innovative and synthetic research, effective biodiversity policy, and ubiquitous educational engagement.



KU BIODIVERSITY INSTITUTE
The University of Kansas

NIBA Implementation Plan

2.1. Create a national database of all digitized specimen records from US institutions and agencies.

2.2. Establish a research and development environment to deliver new specimen digitization workflow methods, tools, and techniques.

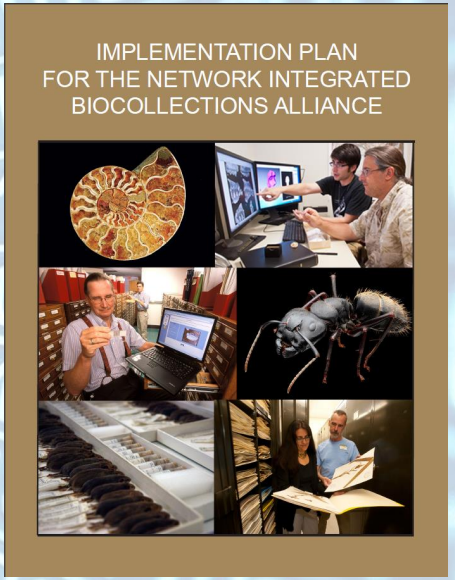
2.3 Complete development of required standards and protocols.

2.4. Promote a consensus for the adoption of standards.

2.5. Anticipate the future of biodiversity specimen data integration.

2.6. Develop a strategy for long-term data archiving of specimen information, including 2D and 3D images, text information, and metadata about digitization processes.

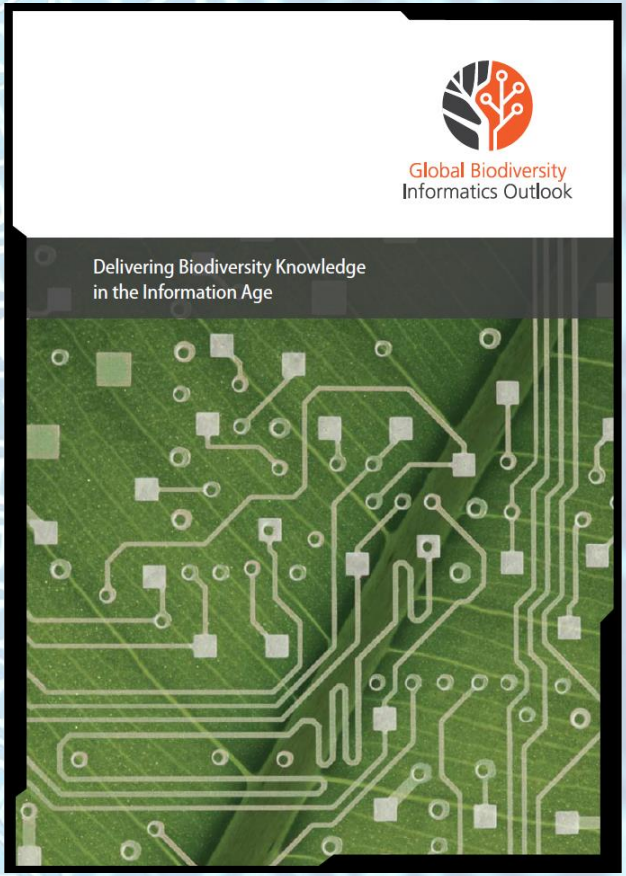
2.7. Support the development of a robust, Web-services-based architecture for handling taxonomic names applied to specimens as determinations and annotations.



KU BIODIVERSITY INSTITUTE
The University of Kansas

Global Biodiversity Informatics Outlook

The Global Biodiversity Informatics Outlook helps to focus effort and investment towards better understanding of life on Earth and our impacts upon it. It proposes a framework that will help harness the immense power of information technology and an open data culture, to gather unprecedented evidence about biodiversity and to inform better decisions.



KU BIODIVERSITY INSTITUTE
The University of Kansas

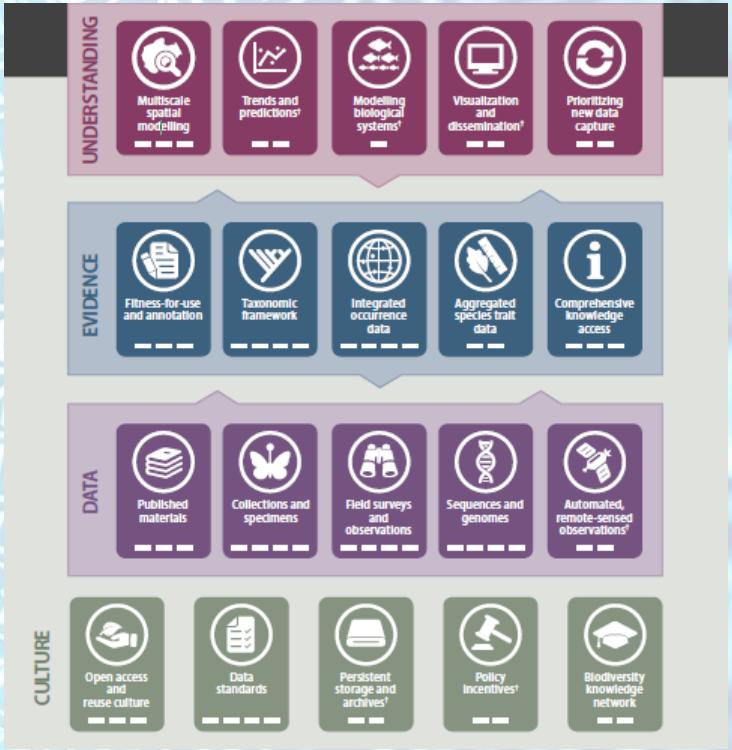
Global Biodiversity Informatics Outlook

Focus area A: Culture
Putting the foundations in place to make biodiversity data an openly shared, freely available, connected resource.

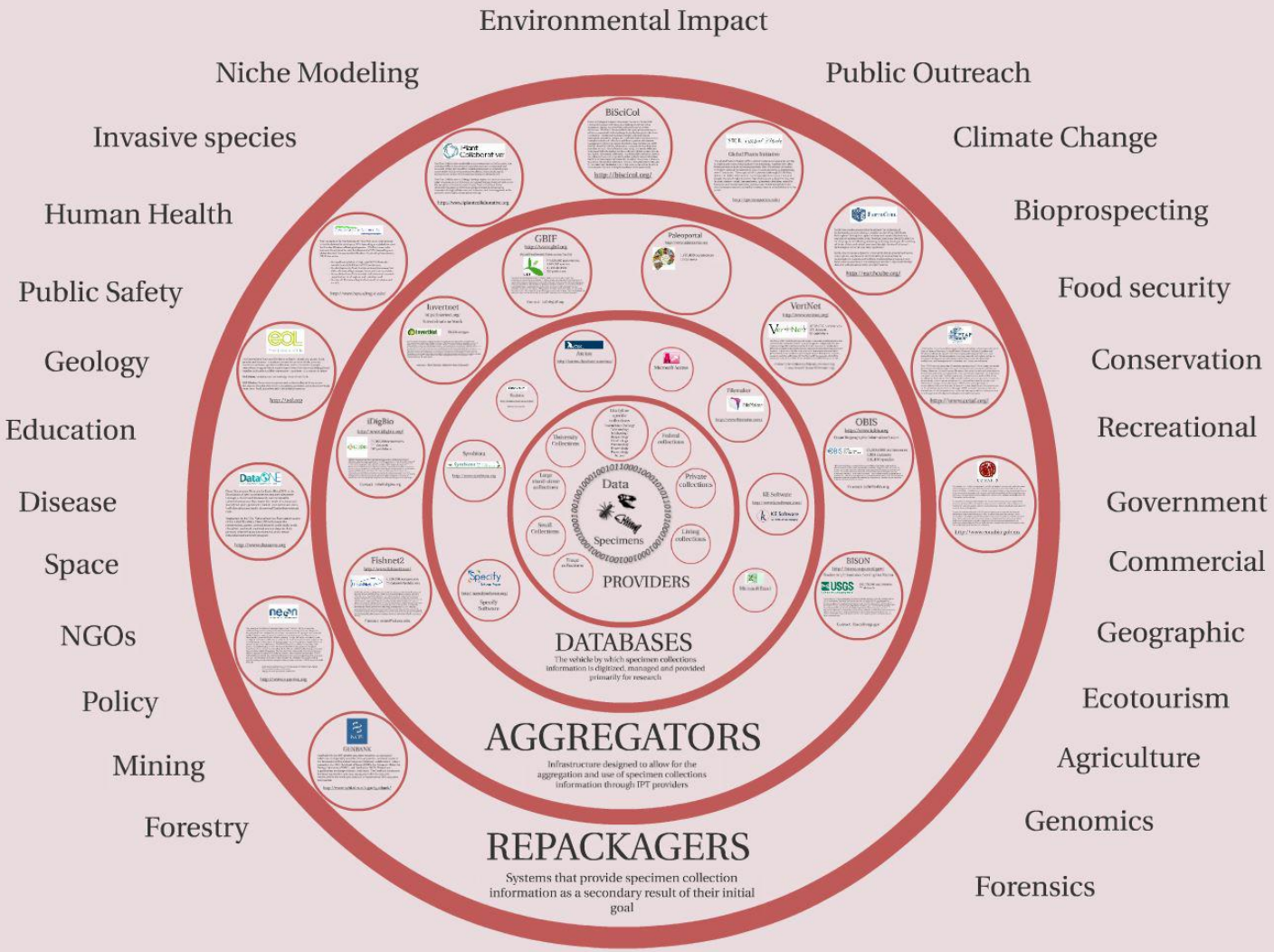
Focus area B: Data
Mobilizing biodiversity data from all sources and organizing it in forms that can support large-scale analysis and modelling.

Focus area C: Evidence
Providing the tools to support consistent and comprehensive global discovery and use of data from all sources about the biodiversity of any defined area over time, covering all taxonomic groups.

Focus area D: Understanding
Using the combined biodiversity data from multiple sources to generate new information, inform policy and decision makers, and help educate wider society to improve the way we manage the Earth's resources.



KU BIODIVERSITY INSTITUTE
The University of Kansas


Environmental Impact
Niche Modeling
Public Outreach
Invasive species
Climate Change
Human Health
Bioprospecting
Public Safety
Food security
Geology
Conservation
Education
Recreational
Disease
Government
Space
Commercial
NGOs
Geographic
Policy
Ecotourism
Mining
Agriculture
Forestry
Genomics
Forensics
BiSciCol
GBIF
Paleoportal
Invertnet
VertNet
iDigBio
OBIS
Fishnet2
BISON
Data
Specimens
PROVIDERS
DATABASES
The vehicle by which specimen collections information is digitized, managed and provided primarily for research
AGGREGATORS
Infrastructure designed to allow for the aggregation and use of specimen collections information through IPT providers
REPACKAGERS
Systems that provide specimen collection information as a secondary result of their initial goal
EXTERNAL USER COMMUNITY
Users outside of the collections community who utilize collections data directly from aggregators or through repackagers to facilitate research, assessment or commercial uses

The Good - Making great strides

- **iDigBio, ADBC, TCN's and collections – digitizing more specimens and more data being published through multiple data portals.**
- **Database software facilitating publication of richer data through expanded data models and integration with IPT and Darwin Core.**
- **SPNHC, iDigBio, TDWG and others – best practices, protocols and workflows being disseminated through workshops, webinars, wikis and publications to educate the workforce.**
- **BCoN, iDigBio, SPNHC and others are involved in galvanizing the community around a common cause.**
- **Others involved in training the next generation of collections personnel**

Obvious that these endeavors need to continue with innovation and higher throughput as we have a lot more to do.

The Bad - Integration...

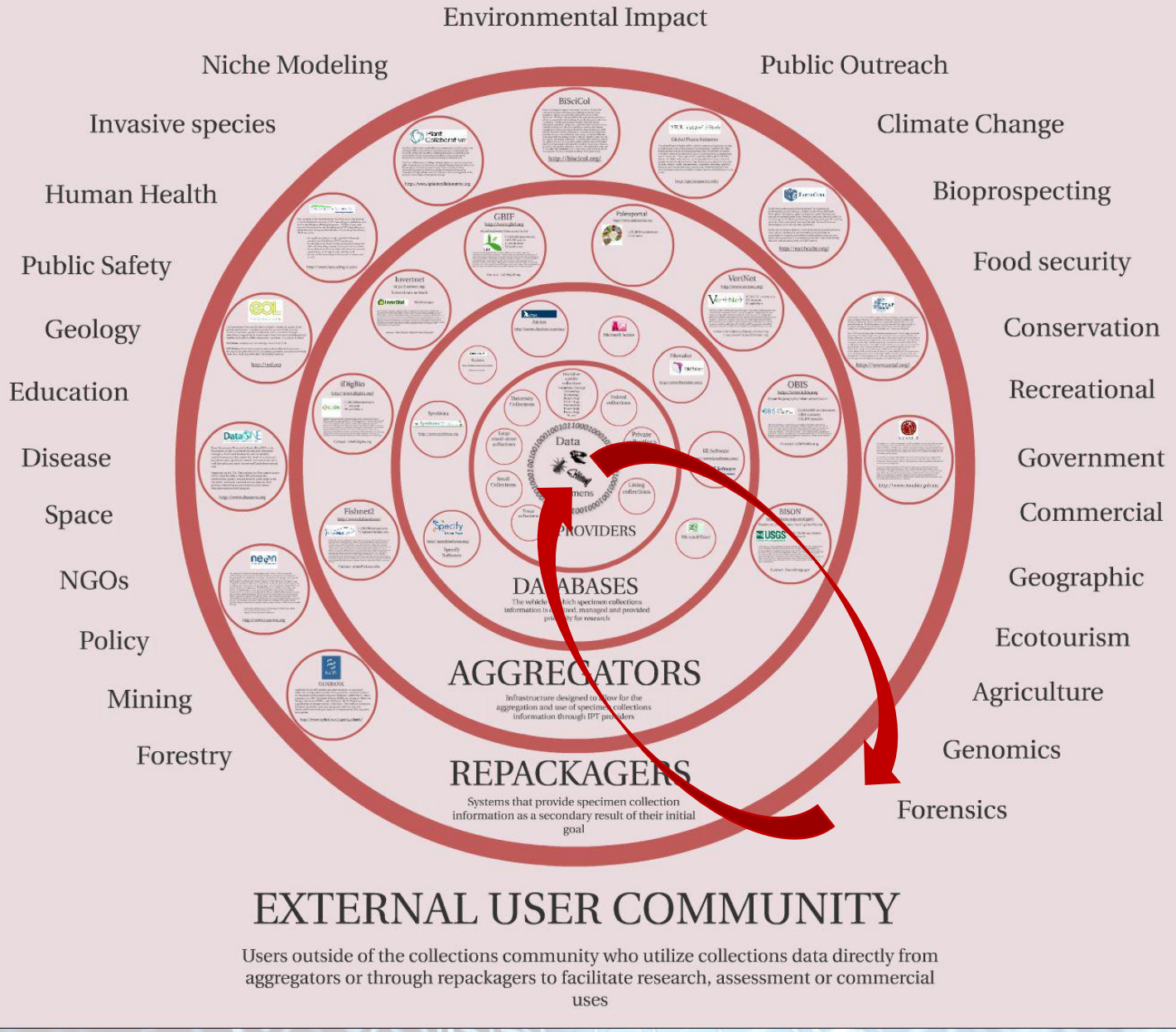
The missing piece is the integration of data components into a seamless data pipeline that can make more robust data available to the ever increasing and varied external user community as rapidly as possible and in the correct format.

We need a flow of data from collection in the field, to natural history database, to aggregator, to re-packagers, to research publication and other external users with as little human intervention or impediments as possible.

All of these individual components are essential cogs in the functioning of the collective data pipeline.

Our data needs to integrate with and be scalable to other sources of data from outside of our immediate community.

KU BIODIVERSITY INSTITUTE
The University of Kansas


Environmental Impact
Niche Modeling
Public Outreach
Invasive species
Climate Change
Human Health
Bioprospecting
Public Safety
Food security
Geology
Conservation
Education
Recreational
Disease
Government
Space
Commercial
NGOs
Geographic
Policy
Ecotourism
Mining
Agriculture
Forestry
Genomics
Forensics
BiSciCol
GBIF
Paleoportal
Invertnet
VertNet
iDigBio
OBIS
Fishnet2
BISON
Data
PROVIDERS
DATABASES
The vehicle by which specimen collections information is stored, managed and provided primarily for research
AGGREGATORS
Infrastructure designed to allow for the aggregation and use of specimen collections information through IPT providers
REPACKAGERS
Systems that provide specimen collection information as a secondary result of their initial goal
EXTERNAL USER COMMUNITY
Users outside of the collections community who utilize collections data directly from aggregators or through repackagers to facilitate research, assessment or commercial uses

Collections

- Continue to digitize and publish more robust, complete data sets along with providing specimens for research.
- Advocate for collections and collections research through publication, tours, outreach, social media etc.
- Make collections visible in community through joining portals and advertising collections and research on websites and social media.
- Show varied and consistent use to remain relevant and advocate for funding.

KU BIODIVERSITY INSTITUTE
The University of Kansas

Databases

- Collection management software is primary vehicle for making specimens records available.
- A number of the more commonly used platforms have been free to end users through NSF funding.
- There is an indication that this funding is coming to an end.
- We need to find a community based solution to funding these critical pieces of infrastructure.
- Software needs to continue to advance and keep pace with community driven needs of digitization and provide the necessary tools to increase the rate of digitization.

KU BIODIVERSITY INSTITUTE
The University of Kansas

Aggregators

Primarily outward focused on use of data – need to expend some of their energy looking inward toward collections providing data. Some ideas:

- **Usage statistics – collections use and advocacy. Vertnet model.**
- **Standardization.**
- **Measures of uniqueness – what do I have in my collection that no one else has – geographic and taxonomic measures.**
- **Data cleanup assistance – controlled vocabularies, georeferencing, incorrect mappings/data.**
- **Annotations and other forms of data user interaction.**
- **Indications of extent of external users of data.**
- **Geographic/taxonomic subscription services.**

Provision of these services will increase participation and publication of more data.

WHY SO MANY AGGREGATORS?

Researchers and other end users

Research products primary metric for showing collections use and advocacy. Unfortunately too often they are not "collections advocacy aware"

- **In the field - collect robust, augmented, complete data with specimens.**
- **Correct citation of voucher and tissue numbers in publications and Genbank sequences.**
- **Repatriation of these and other products created during use of specimens or data for research – images, data cleanup, georeferencing.**
- **Creation of dynamic linkages will facilitate this.**

KU BIODIVERSITY INSTITUTE
The University of Kansas

Publishers

Research deliverable conundrum can be solved by making link between research and specimens, data and research products more explicit

- **Improved, formalized citation of materials examined in publication and Genbank.**
- **Pensoft ARPHA writing tool material examined import model from aggregator is a huge step in the right direction but more publishers need to adopt it or copy it.**
- **Primary focus of this work is on streamlining research publication but has secondary (and equally important) consequences for collections advocacy, visibility and linking of data.**

KU BIODIVERSITY INSTITUTE
The University of Kansas

NSF

- NSF not funding “entire research endeavor”.
- Funding for infrastructure (CSBR) and digitization (ADBC) but little to none for pure curation and long term care.
- Proposed specimen management plan for NSF specimen-based research grants.
  - Base cost of curation and long-term care of discipline specific specimens.
  - Calculation based on number of expected specimens collected.
- Required funding in proposals – cannot be cut.
- Provide direct monetary link between research and collections

KU BIODIVERSITY INSTITUTE
The University of Kansas

Summary

- Collections and researchers need to continue to provide high quality data for research and other uses of data.
- Need to improve linkages between components of data pipeline without human intervention – APIs etc.
- Engage all generators and users of data in forming this data pipeline.
- Ensure data pipeline is bi-directional to ensure continued viability and use of collections.

KU BIODIVERSITY INSTITUTE
The University of Kansas

Thank you