Integration, Attribution, and Value in the Web of Natural History Museum Data

Andrew Bentley

Biodiversity Collections Network (BCoN)

and

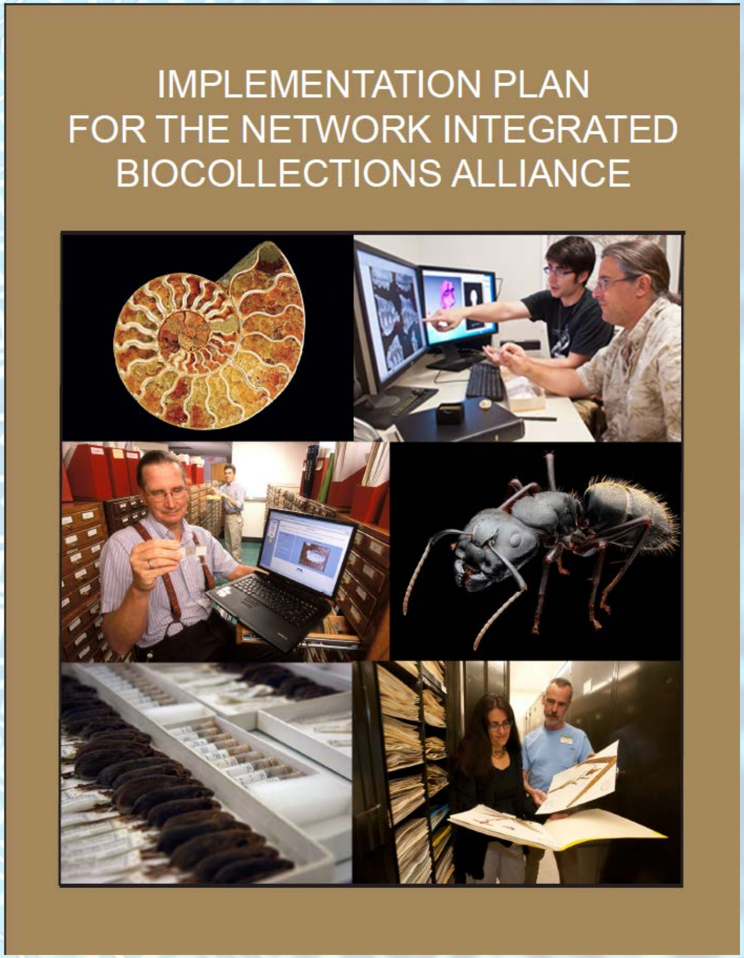
University of Kansas, Lawrence, KS

KU BIODIVERSITY INSTITUTE The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

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.



IMPLEMENTATION PLAN
FOR THE NETWORK INTEGRATED
BIOCOLLECTIONS ALLIANCE

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY
COLLECTIONS NETWORK

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.

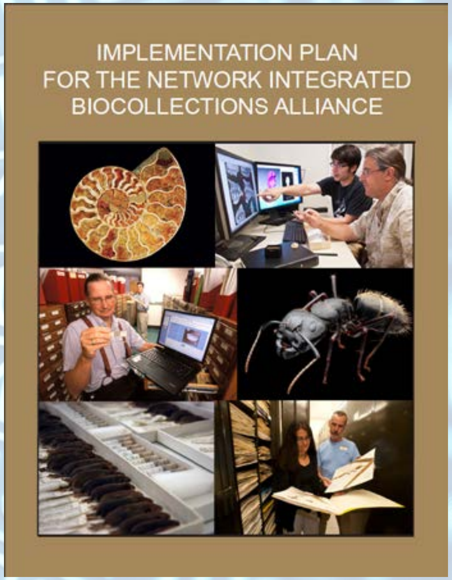
2.3 Complete development of required standards and protocols.

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.



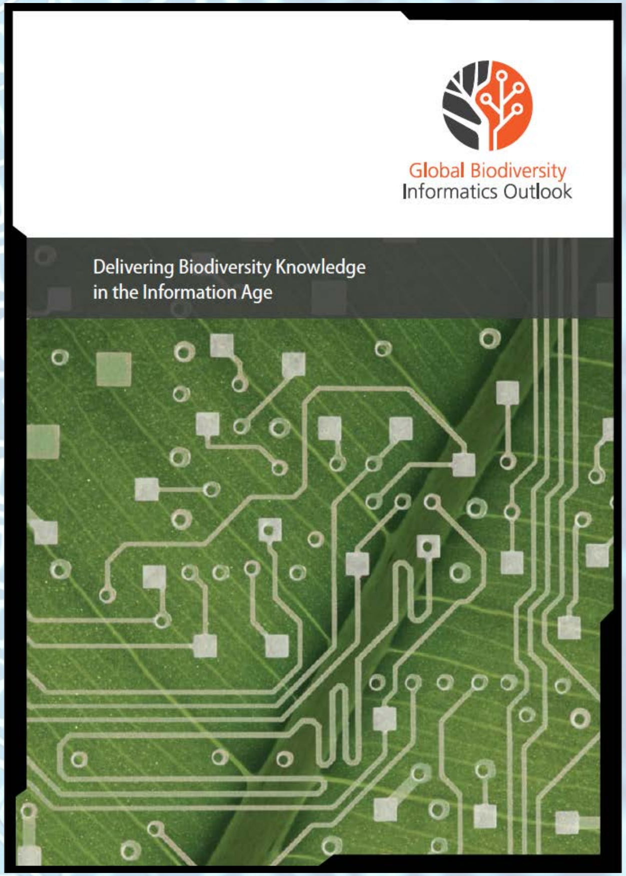
IMPLEMENTATION PLAN
FOR THE NETWORK INTEGRATED
BIOCOLLECTIONS ALLIANCE

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

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.



Global Biodiversity Informatics Outlook

Delivering Biodiversity Knowledge in the Information Age

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

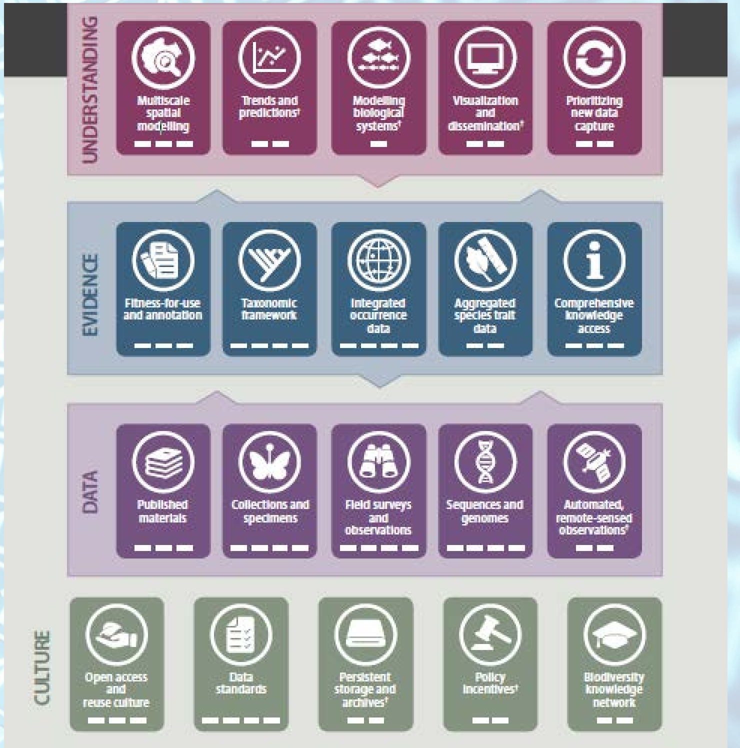
Global Biodiversity Informatics Outlook

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.


UNDERSTANDING
Multiscale spatial modelling
Trends and predictions†
Modelling biological systems†
Visualization and dissemination†
Prioritizing new data capture
EVIDENCE
Fitness-for-use and annotation
Taxonomic framework
Integrated occurrence data
Aggregated species trait data
Comprehensive knowledge access
DATA
Published materials
Collections and specimens
Field surveys and observations
Sequences and genomes
Automated, remote-sensed observations†
CULTURE
Open access and reuse culture
Data standards
Persistent storage and archives†
Policy incentives†
Biodiversity knowledge network

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

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

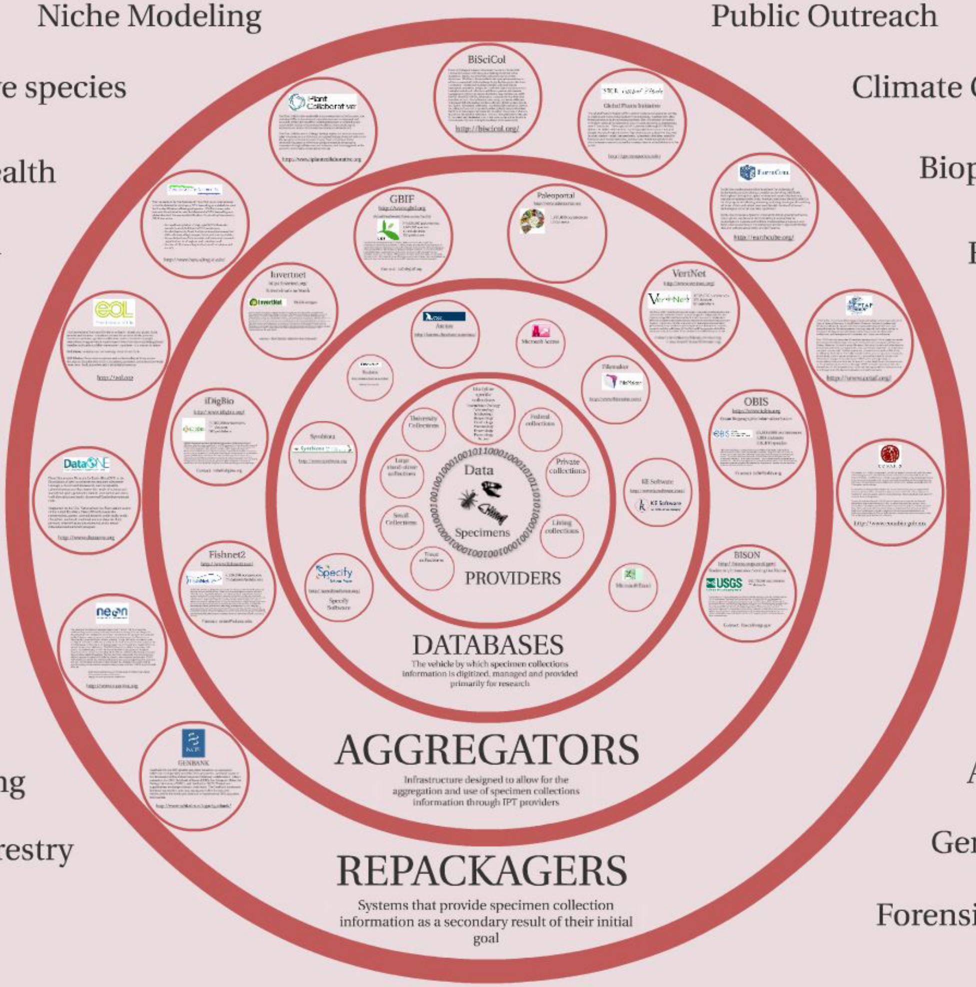
Mining

Agriculture

Forestry

Genomics

Forensics



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

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

The Good - Making great strides

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

These endeavors need to continue with higher output through innovation.

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

The Bad - Integration…

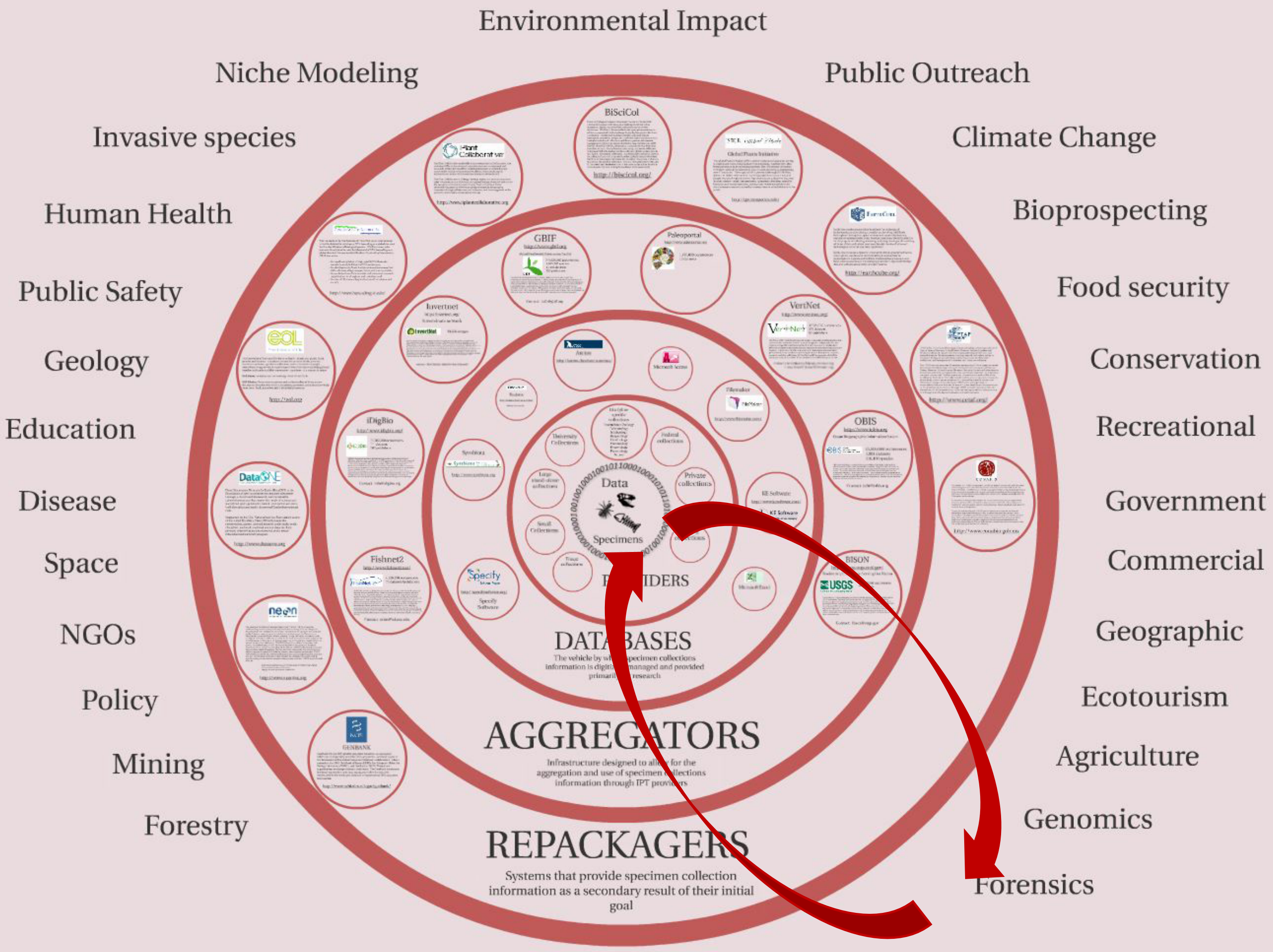
The missing piece is the integration of data components into a seamless data pipeline that can make more robust data available to the ever increasing and varied external user community as rapidly as possible and in the correct format.

Integration - We need a flow of data from the field, to collection, to aggregator, to re-packagers, to research publication and other external users with as little human intervention or impediments as possible. Some progress through APIs and GUIDs. All of these individual components are essential cogs in the functioning of the collective data pipeline.

Interoperability - 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

BIODIVERSITY COLLECTIONS NETWORK



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

P IDERS

DATA BASES

The vehicle by wh specimen collections information is digiti managed and provided primaril research

AGGREGATORS

Infrastructure designed to al for the aggregation and use of specimen llections information through IPT provi rs

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

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

Collections

- Continue to digitize and publish more robust, complete data sets along with providing specimens and data for research needs and other uses.
- 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.
- Promote varied and consistent use to remain relevant and advocate for funding.

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

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 pace of digitization.**
- **Ease publishing of data**

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

Aggregators

Primarily outward focused on use of data – need to expend some of their energy looking inward toward collections providing data. Good to see some progress in some of these areas.

- **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.**
- **Geographic/taxonomic subscription services. RSS feed.**

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

WHY SO MANY AGGREGATORS (data caches)?

KU BIODIVERSITY INSTITUTE The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

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 correction, augmentation, georeferencing, etc.**
- **Creation of dynamic GUID based linkages will facilitate this.**

KU BIODIVERSITY INSTITUTE The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

Publishers

Research deliverable conundrum can be solved by creating explicit 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 replicate 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

BIODIVERSITY COLLECTIONS NETWORK

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

BCoN activities

Data integration and attribution at the center of BCoN's remaining two year RCN.

- **Presentations at TDWG meeting and here to convey ideas and encourage participation in upcoming workshops.**
- **Needs assessment workshop –Lawrence, KS - February 2018.**
- **iDigBio 2nd Digital Data in Research meeting workshop – Berkeley, CA – June 2018.**
- **SPNHC Annual Meeting workshop in Dunedin, New Zealand – August 25 – September 2, 2018.**
- **Other workshops also planned to tackle issues such as Nagoya/CBD, small collections, ecological data etc.**

Thank you

KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

Key points

- **Improve integration - Strengthen linkages between components of data pipeline using APIs, GUIDs and other informatics methods. Engage all generators and users of data in forming this data pipeline.**
- **Improve attribution - Ensure data pipeline is bi-directional to ensure advocacy and continued viability and use of collections.**
- **Improve data quality - Collections and field researchers need to continue to provide high quality data for research and other uses of data.**
- **Improve interoperability – Engage disciplines outside of our realm to ensure data is scalable with other systems.**

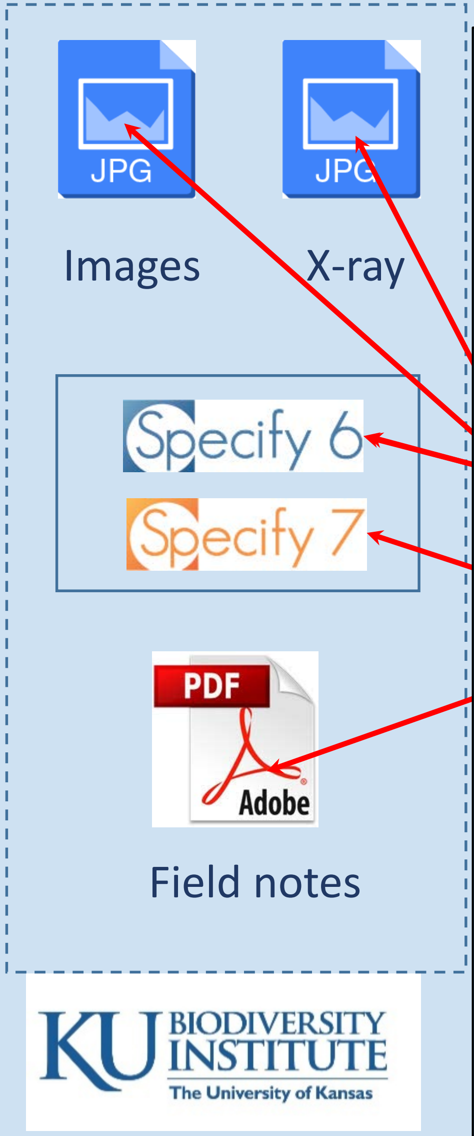
KU BIODIVERSITY INSTITUTE
The University of Kansas

BIODIVERSITY COLLECTIONS NETWORK

Specimen

Publishing

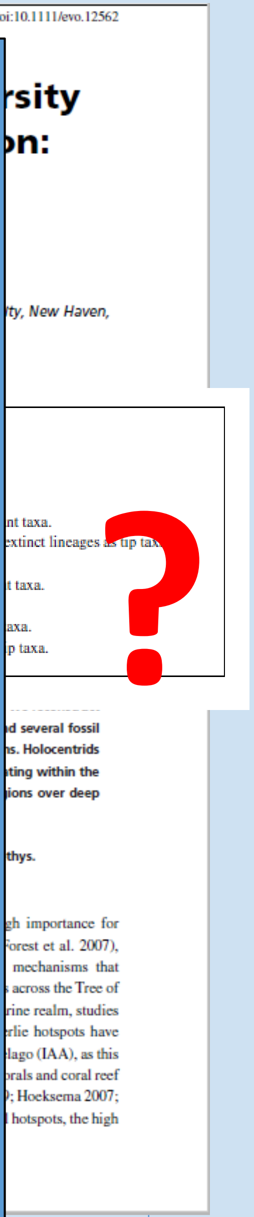
Publications



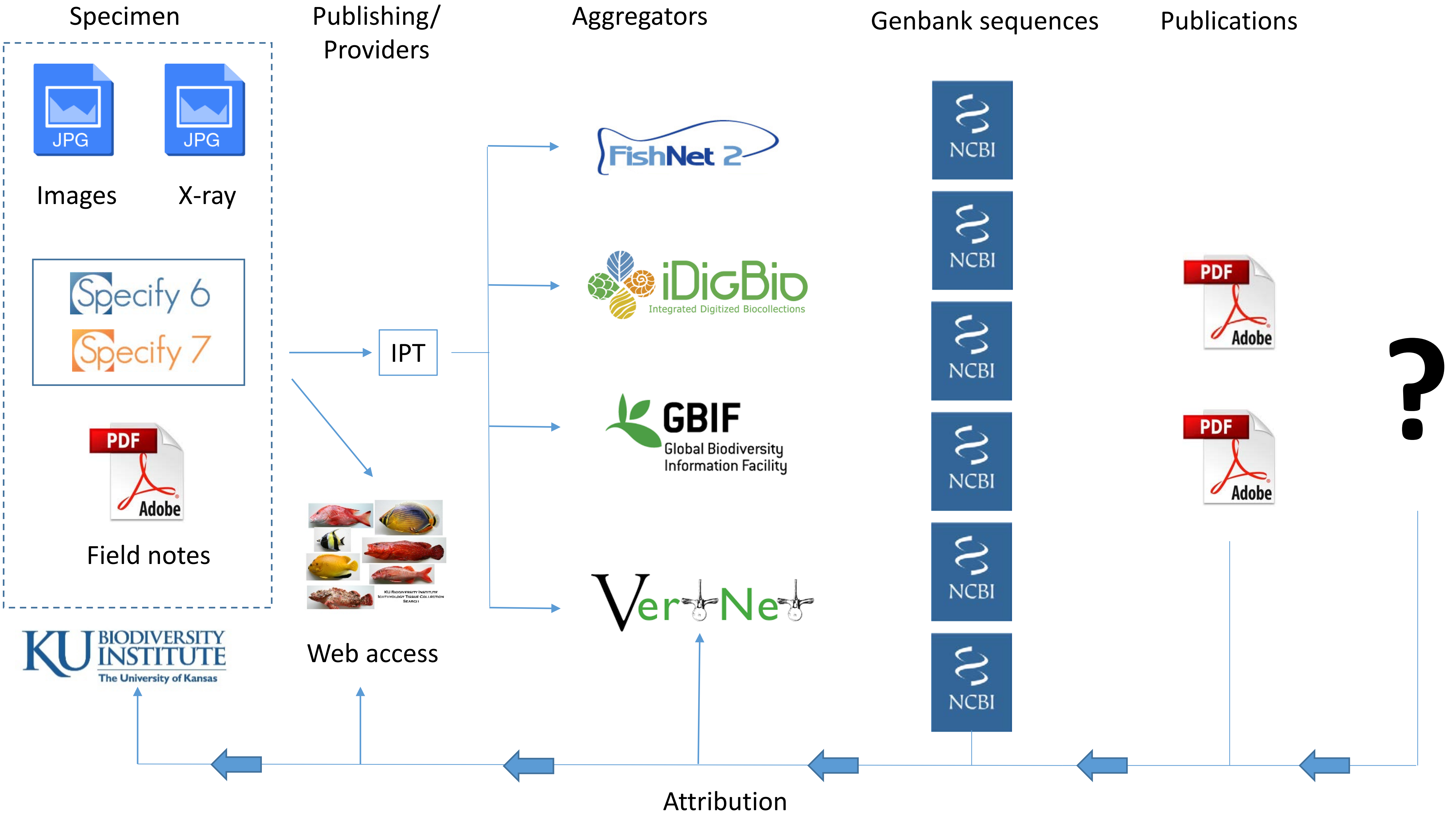
10,621 tissues
43,310 voucher lots
2,907 images
3,870 field note pages

14,265 Genbank sequences from 2,634 tissue specimens
3,483 tissue citations in 197 publications
6,145 voucher citations in 282 publications

Millions of data downloads from aggregators and web access. For what?



Attribution



Specimen

Publishing/ Providers

Aggregators

Genbank sequences

Publications

Images

X-ray

Field notes

IPT

Web access

Attribution

JPG
JPG
Specify 6
Specify 7
PDF
Adobe
KU BIODIVERSITY INSTITUTE
The University of Kansas
KU Biodiversity Institute Ichthyology Tissue Collection Search
FishNet 2
iDigBio
Integrated Digitized Biocollections
GBIF
Global Biodiversity Information Facility
VertNet
NCBI
NCBI
NCBI
NCBI
NCBI
NCBI
PDF
Adobe
PDF
Adobe
?