

# Welcome!

## And a few logistical details

**Wiki:** ([https://www.idigbio.org/wiki/index.php/Paleo\\_Digitization\\_Workshop](https://www.idigbio.org/wiki/index.php/Paleo_Digitization_Workshop))

**Adobe Connect (Kevin Love):** <http://idigbio.adobeconnect.com/paleo>

Being broadcast and recorded

Be observant of remote audience; use microphone to make comments, ask questions

Chat box for remote participants

**Efficiency:** Starting on time; staying on track

**Lunch:** 1.25 hours/day

**Meals:** On your own; no receipts required

This material is based upon work supported by the National Science Foundation under Cooperative Agreement EF-1115210. 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.



# Integrated Digitized Biocollections (iDigBio) An Introduction

Gil Nelson

Institute for Digital Information and Scientific Communication  
Integrated Digitized Biocollections  
Florida State University

Paleo Digitization Workshop  
23-25 September 2013  
Yale Peabody Museum

This material is based upon work supported by the National Science Foundation under Cooperative Agreement EF-1115210. 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.





The U.S. National Science Foundation estimates there may be as many as 1.8 billion biological and paleontological specimens stored in U. S. museums and academic institutions (perhaps as many as 3 billion worldwide). But, no one really knows!

In an effort to make these collections universally accessible to taxonomists, ecologists, researchers, and the general public, in 2011 NSF launched a \$100 million, 10-year Advancing Digitization of Biodiversity Collections program and named Florida State University and University of Florida jointly as the national resource for digitization.

# Advancing Digitization of Biodiversity Collections



## Integrated Digitized Biocollections (iDigBio) University of Florida Florida State University Florida Museum of Natural History

The goal is to digitize and make available via the Web at least 1 billion biological and paleontological records over the 10-year life of the project.

# Mandate and Responsibility

- Provide/facilitate portal access to collections data
  - Make information available and discoverable
  - Label Data and images
- Enable digitization and research
  - Facilitate digitization workflows
  - Oversee implementation of standards and best practices for digitization
  - Allow for data discovery across organismal groups
- Be a client of digitization projects/networks
  - Actively seek partners and data sources
  - Respond to cyberinfrastructure needs
- Engage communities
  - Collections
  - Research
  - Citizen science and education
- Support ADBC goals
  - Access to information
  - Support for collections
  - Sustainability



# Mandate and Responsibility

- Provide/facilitate portal access to collections data

- Make information available to all

- 

- Enable

- Develop a cloud computing infrastructure that links biological data from collections across the U.S. through one or more

- unified web interfaces to overcome the

- Be a

- limitations of “data silos.”

- Engage

- 

- Research

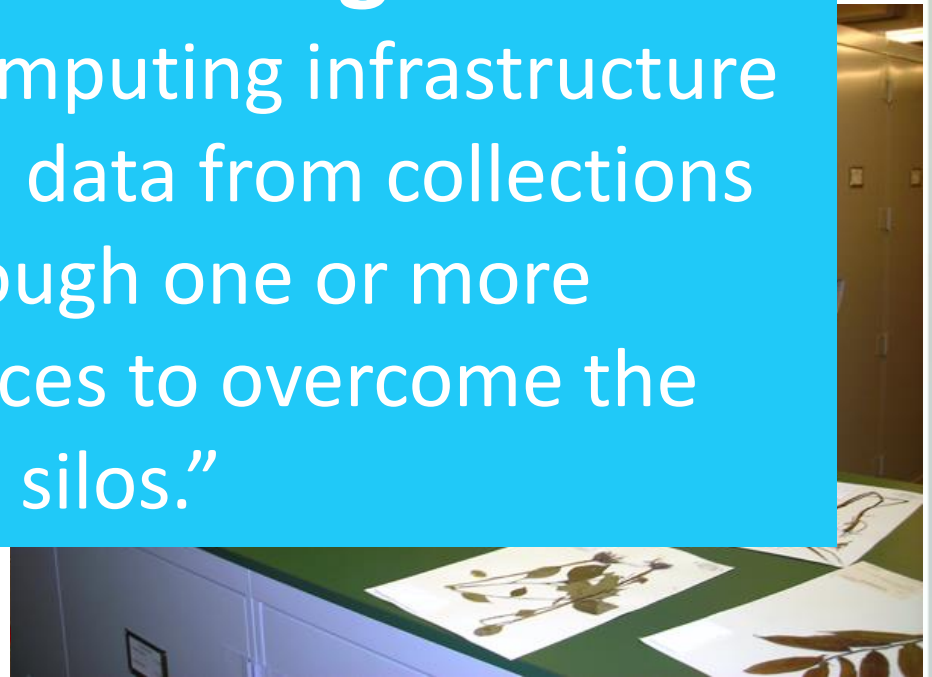
- Citizen science and education

- Support ADBC goals

- Access to information

- Support for collections

- Sustainability



# Mandate and Responsibility

- Provide/facilitate portal access to collections data

- Metadata
- 

- Enable

- Develop
- that

across

- Be a

- unified

limitations of “data silos.”

- Engage

- Research
- Citizen science and education

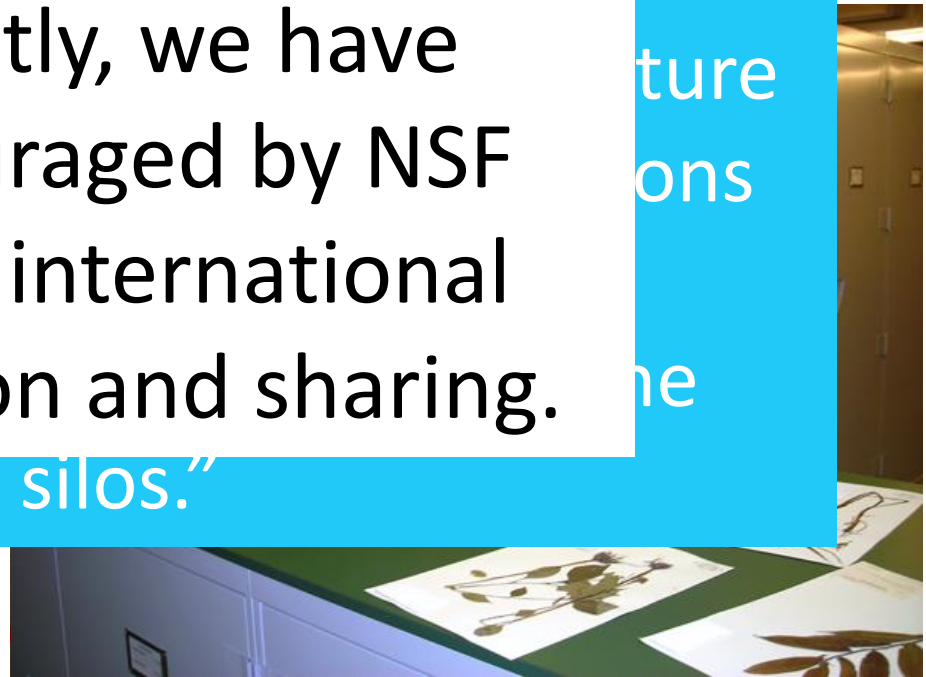
- Support ADBC goals

- Access to information
- Support for collections
- Sustainability



## Grand Challenge

More recently, we have been encouraged by NSF to enhance international collaboration and sharing. The limitations of “data silos.”





**The challenges being pursued by iDigBio are reflective of worldwide trends in digitization**

- **Global Biodiversity Informatics Facility (GBIF)**
- **OpenUp! (European Union)**
- **Atlas of Living Australia (ALA)**
- **SYNTHESYS (20 European natural museums)**



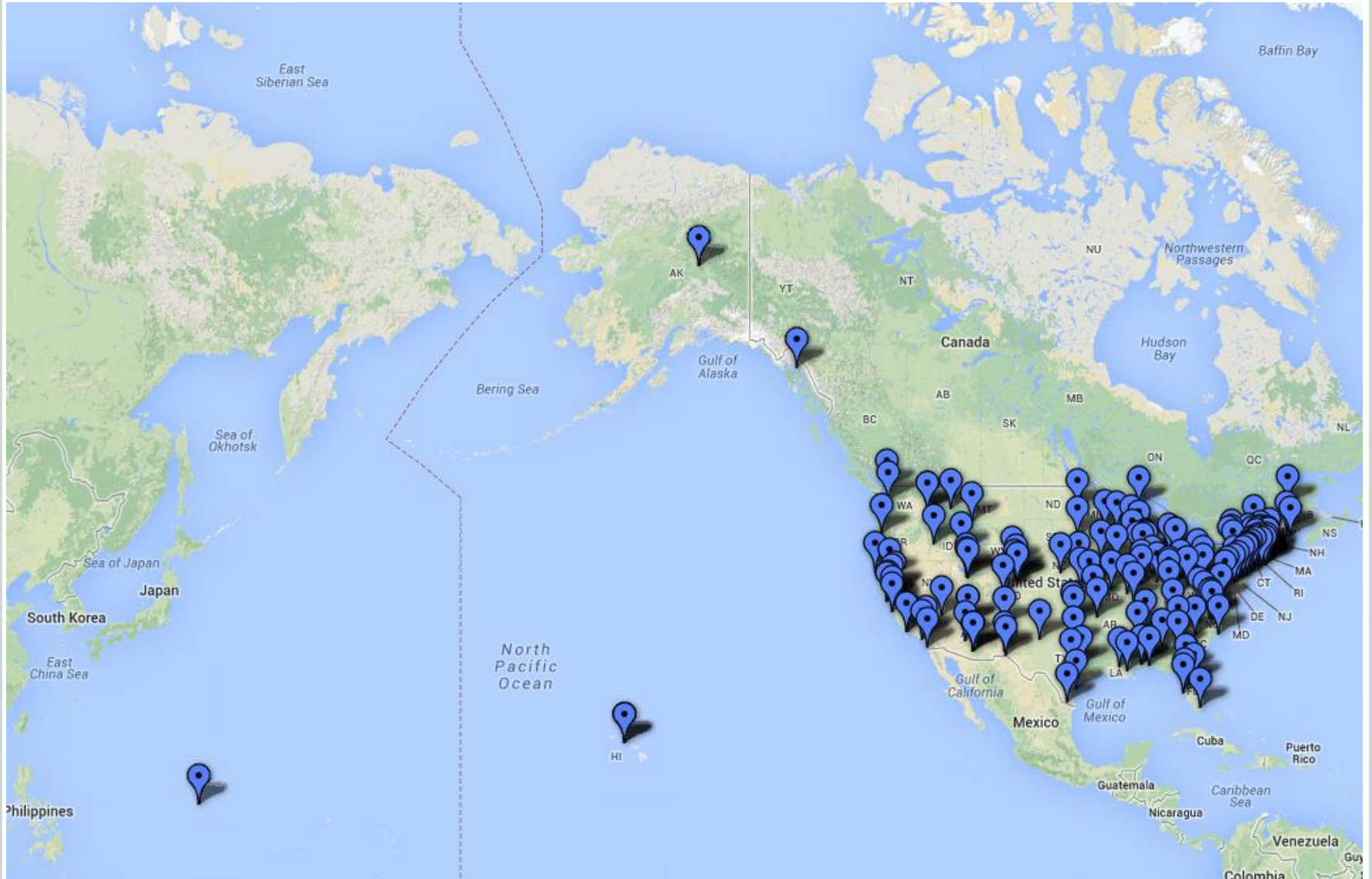
# Ten Thematic Collections Networks (TCNs) plus 2 Partner to Existing Networks (PENs)

- InvertNet: An Integrative Platform for Research on Environmental Change, Species Discovery and Identification (*Illinois Natural History Survey, University of Illinois*) <http://invertnet.org>
- Plants, Herbivores, and Parasitoids: A Model System for the Study of Tri-Trophic Associations (*American Museum of Natural History*) <http://tcn.amnh.org>
- North American Lichens and Bryophytes: Sensitive Indicators of Environmental Quality and Change (*University of Wisconsin – Madison*) <http://symbiota.org/nalichens/index.php> <http://symbiota.org/bryophytes/index.php> (plus 2 PENs)
- Digitizing Fossils to Enable New Syntheses in Biogeography - Creating a PALEONICHES-TCN (*University of Kansas*)
- The Macrofungi Collection Consortium: Unlocking a Biodiversity Resource for Understanding Biotic Interactions, Nutrient Cycling and Human Affairs (*New York Botanical Garden*)
- Mobilizing New England Vascular Plant Specimen Data to Track Environmental Change (*Yale University*)
- Southwest Collections of Anthropods Network (SCAN): A Model for Collections Digitization to Promote Taxonomic and Ecological Research (*Northern Arizona University*) <http://hasbrouck.asu.edu/symbiota/portal/index.php>

## New as of 1 July 2013

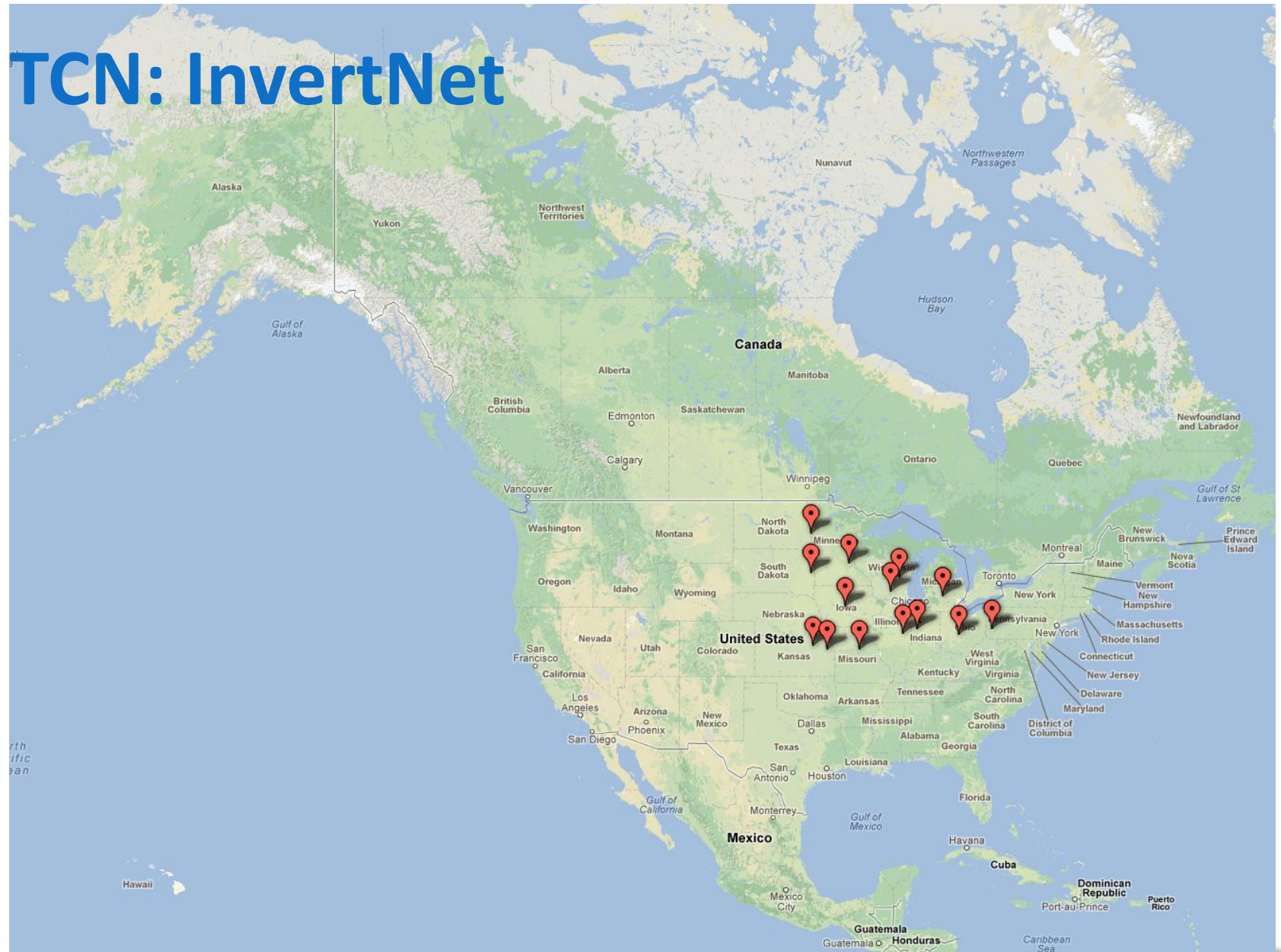
- iDigPaleo: Fossil Insect Collaborative: A Deep-Time Approach to Studying Diversification and Response to Environmental Change
- Developing a Centralized Digital Archive of Vouchered Animal Communication Signals
- The Macroalgal Herbarium Consortium: Accessing 150 Years of Specimen Data to Understand Changes in the Marine/Aquatic Environment

# National Resource (iDigBio), Thematic Collection Networks (TCNs)

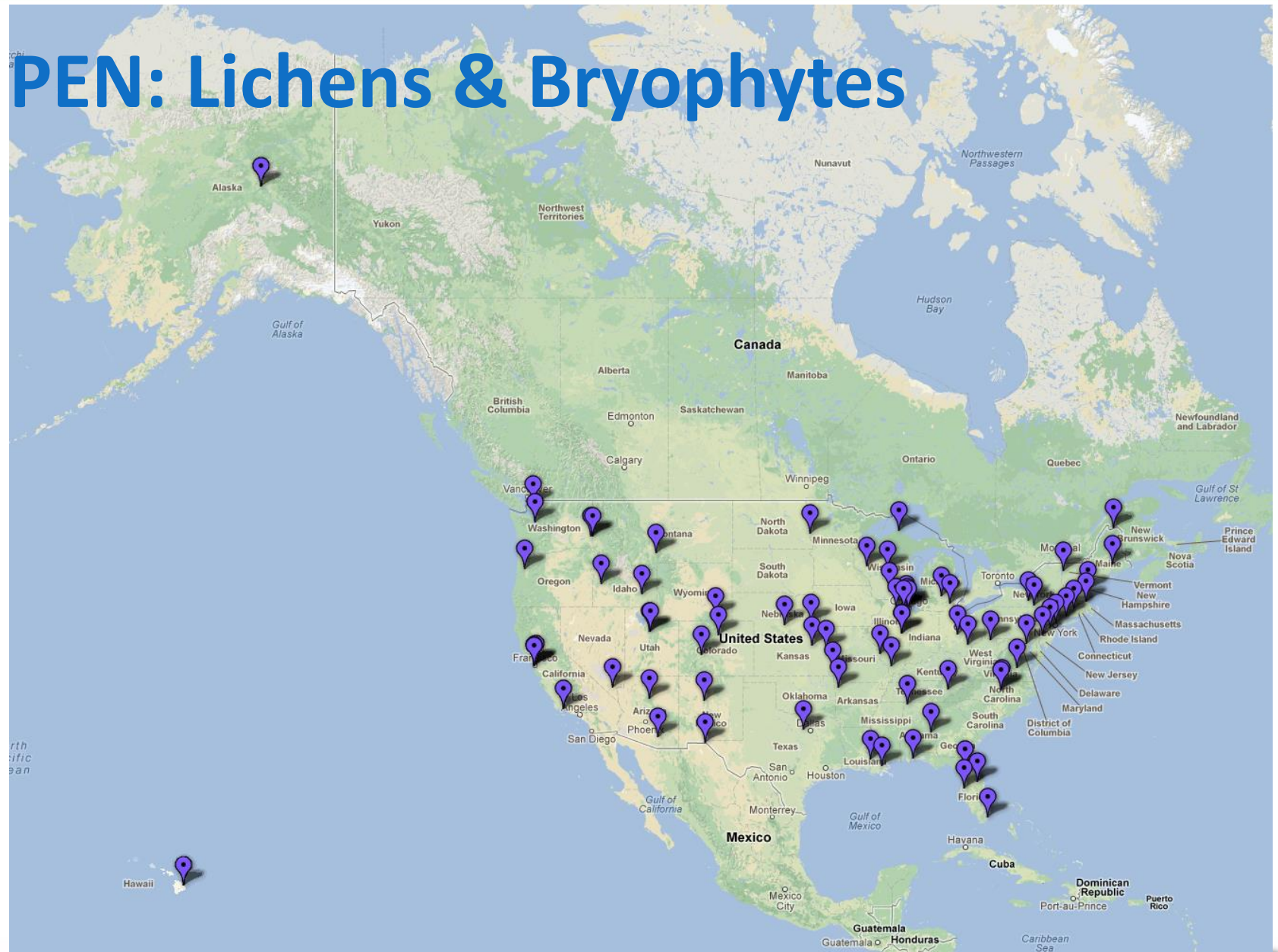


To date: 10 TCNs, 2 PENs, 160+ participating institutions, 49 states

# TCN: InvertNet



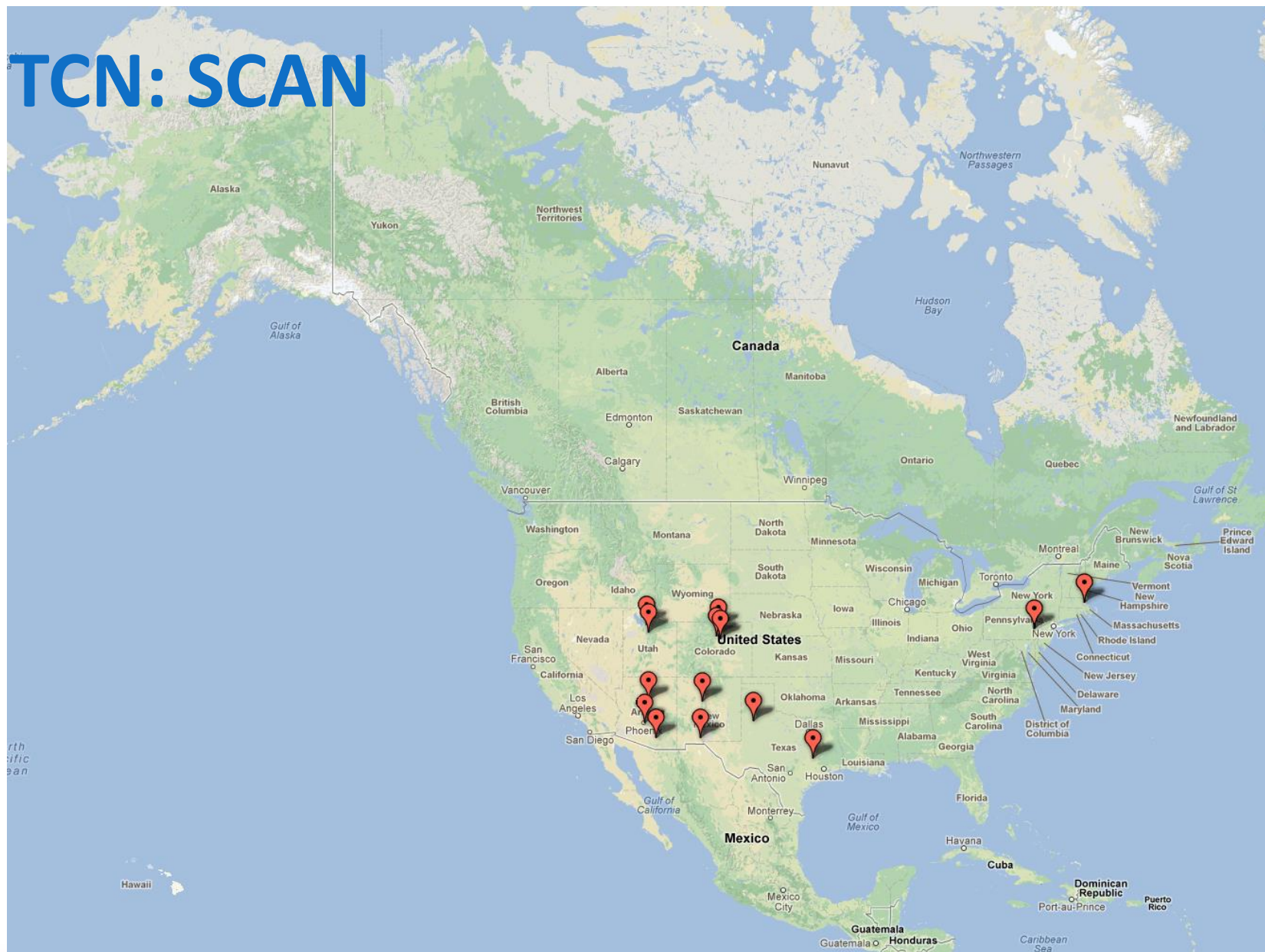
# PEN: Lichens & Bryophytes







# TCN: SCAN



# TCN: PALEONICHES

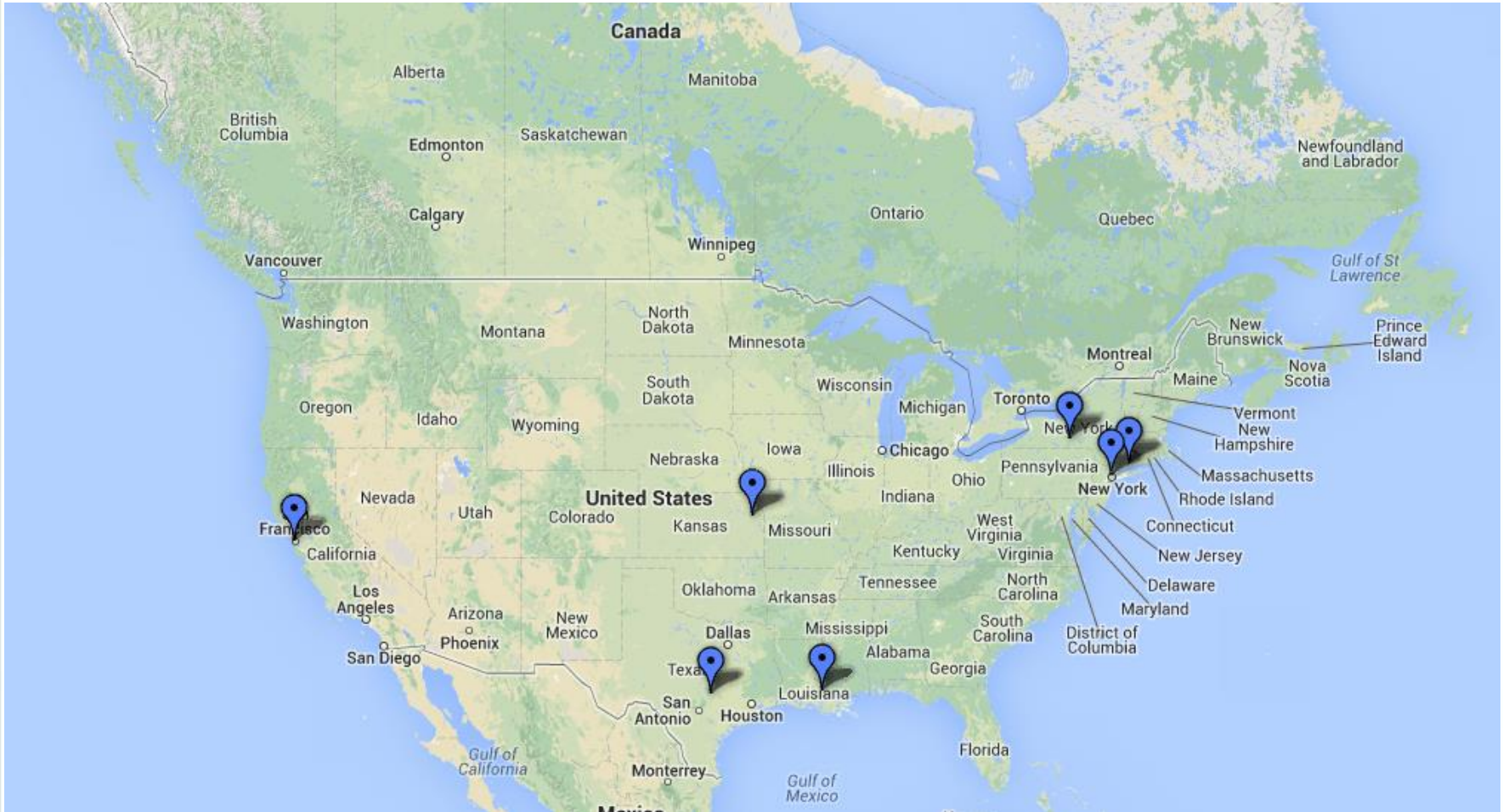




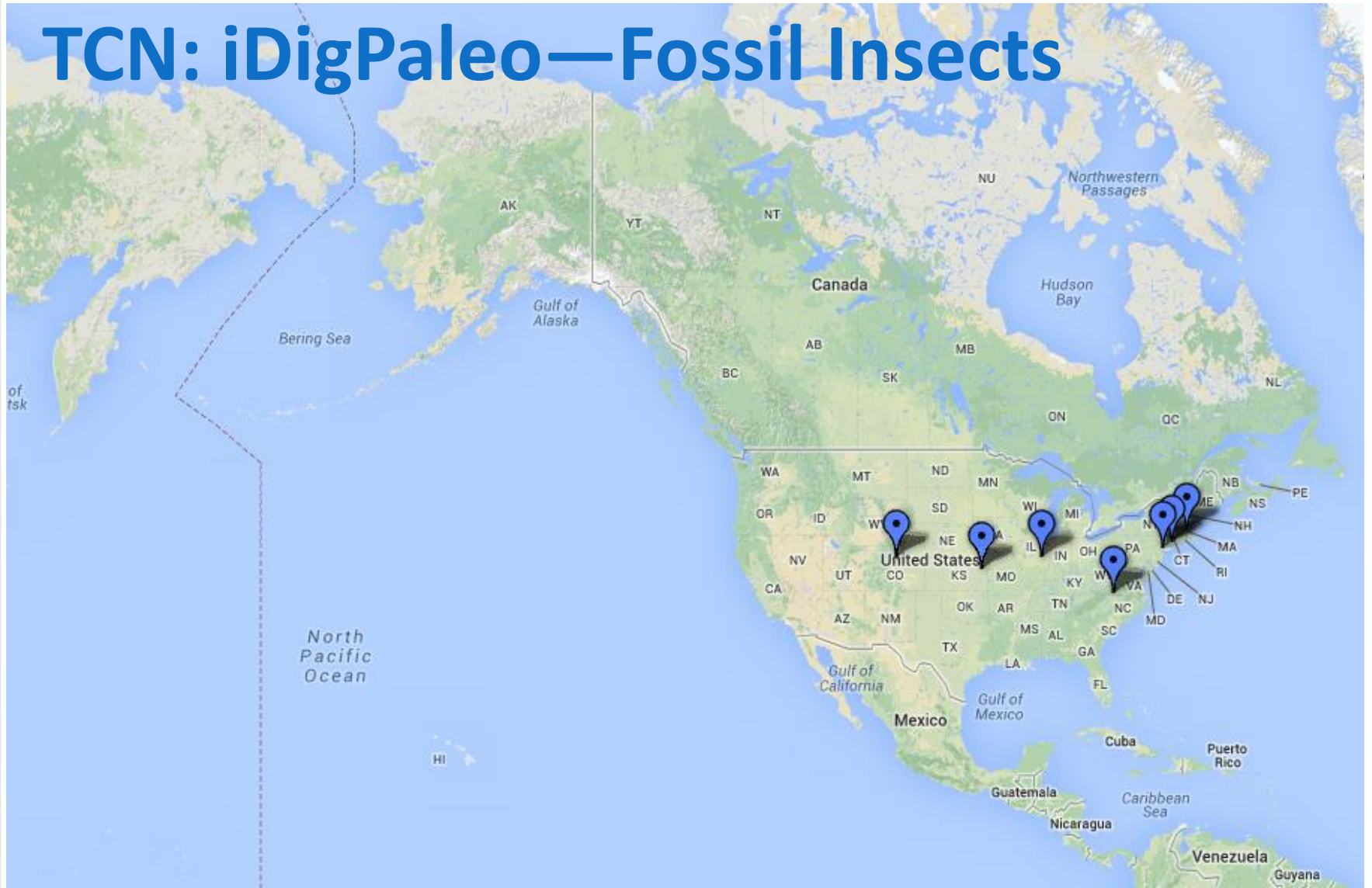
# TCN: Tri-Trophic



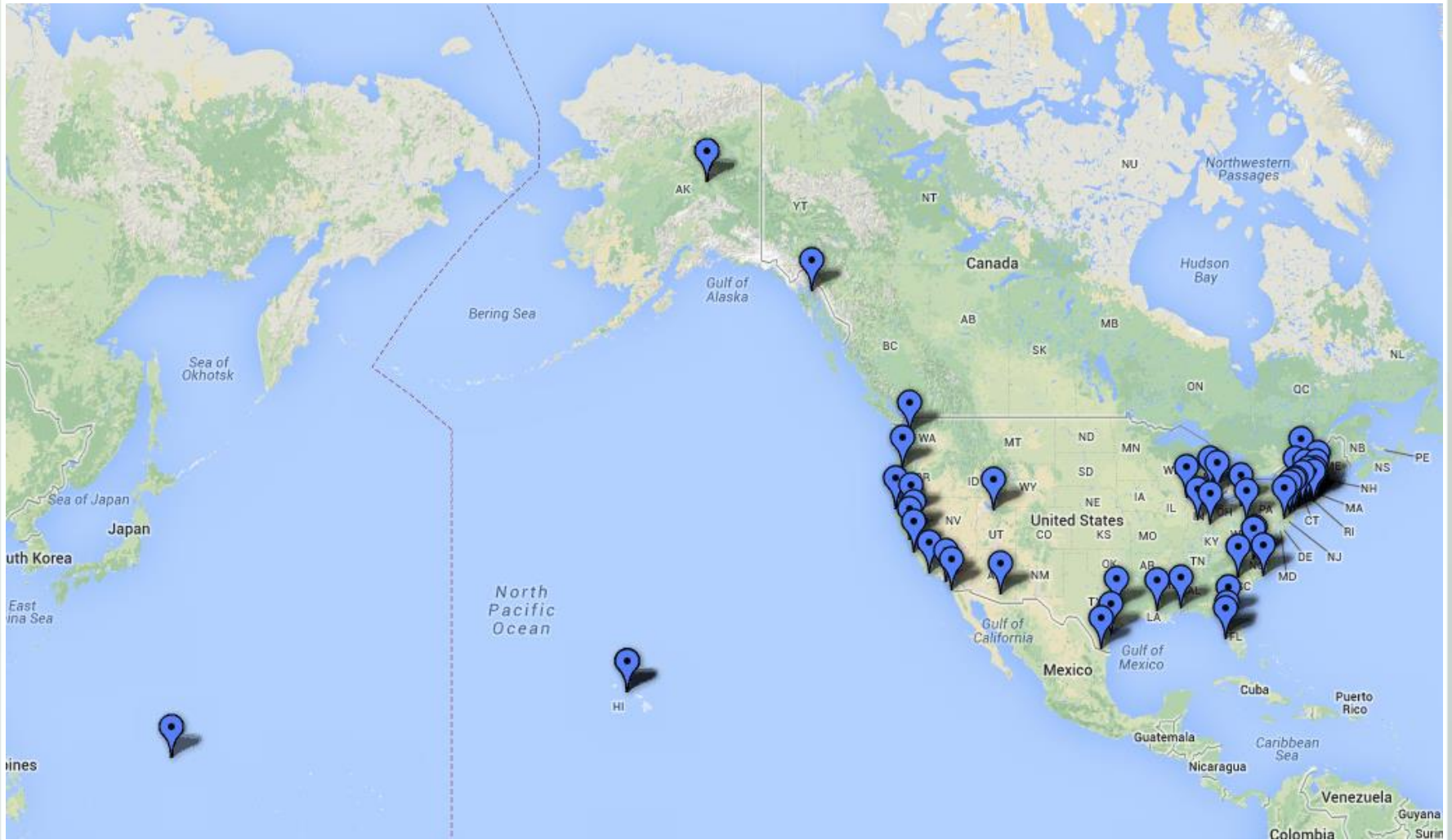
# TCN: Animal Sounds



# TCN: iDigPaleo—Fossil Insects

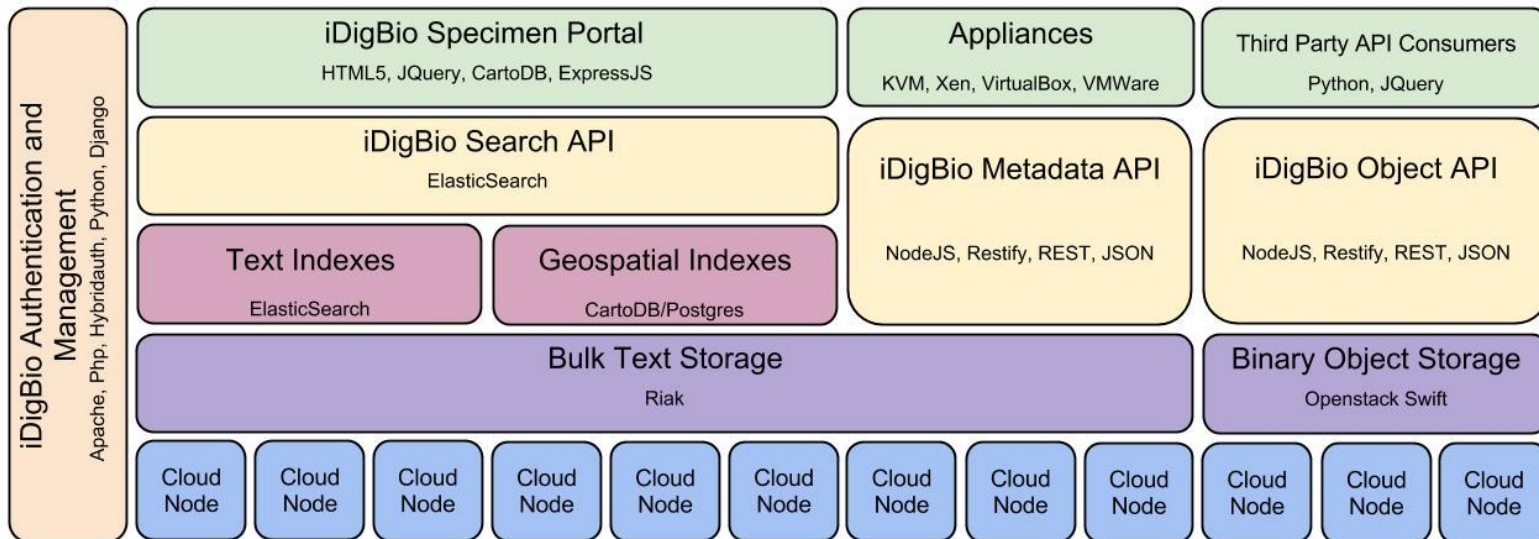


# TCN: Macroalgal



# Building the iDigBio Cloud

- Cloud-based strategy
  - Providing useful services/APIs (programmatic and web-based Application Programming Interface)
  - Federated scalable object storage and information processing
  - Digitization-oriented virtual appliances
  - Reliance on standards, proven solutions, and sustainable software
- Continuous consultation with stakeholders
  - Surveys, working groups, interest groups, workshops, person-to-person



# Key Features of iDigBio

- Ingest all contributed data with emphasis on use of GUIDs, no restrictions
- Maintain persistent datasets and versioning, allowing new and edited records to be uploaded as needed while preserving existing records
- Ingest textual specimen records, plus associated still images, video, audio, and other media (or links to these resources as determined by the provider)
- Ingest linked documents and associated literature, including field notes, ledgers, monographs, related specimen collections, etc.
- Provide virtual annotation capabilities and track annotations back to the originating collection (collaborating with FilteredPush)
- Facilitate sharing and integration of data relevant to biodiversity research
- Provide computational services for biodiversity research

## Recent, Ongoing, Upcoming Activities

- Assessment of common and effective digitization practices (paper in *ZooKeys*)
- Working groups
  - Minimum information for scientific collections working group (MISC)
  - Digitization workflows working groups
  - Georeferencing
  - Optical character recognition (OCR)
  - Biodiversity Informatics Manager working group
- Workshops - year 2:
  - > 150 institutions, 9 workshops, 3 symposia
  - 368 sponsored participants
  - Video archives on Vimeo, live streaming for remote participation
  - New model this year: train the trainer
  - Series of digitization training workshops (herbaria, wet collections, entomology, paleontology, fluid-preserved invertebrate imaging, small herbaria, )
- Server hosting: 8 virtual machines, TCN support
- Specimen data portal and website – continuous improvements
- Call for appliances, frequent opinion surveys


# Digitization Workshops

In March 2012, the Steering Committee established a series of preparation-specific digitization training workshops focused on helping collections managers get started with and/or enhance local digitization programs, all to be held at host institutions.



- DROID (Developing Robust Object->Image->Data, May 2012)
- Herbarium digitization (Valdosta State, September 2012)
- Fluid-preserved collections digitization (U. Kansas, March 2013)
- Dried insect collections digitization (Field Museum, April 2013)
- Collections Digitization (West Virginia, ASB, April 2013)
- Imaging fluid-preserved invertebrates (U. Michigan, September 2013)
- Paleontology digitization (Yale Peabody Museum, September 2013)
- Small Herbarium Digitization (Florida State University, December 2013)
- Broadening Biodiversity in the Biodiversity Sciences (Atlanta, January, 2014)
- Original Source Materials Digitization (Yale Peabody Museum, March 2014)
- Digitization in the South Pacific (Honolulu, March 2014)
- Recruiting and Retaining Small Collections in Digitization (Mt. Pleasant, MI, April 2014)





Home Wiki Community portal Current events Recent changes Random page Help

## Digitization Resources

This page provides resources and information for the series of digitization training workshops being conducted by iDigBio as well as a plethora of digitization information and resources. Included is a growing list of links to documents, websites, videos, presentations, and other important information related to biological collection digitization.

### Contents

[hide]

- 1 iDigBio
- 2 Interest Groups
- 3 Preparation-specific Workshop Wikis
- 4 Workshop Summaries
- 5 General Digitization Resources
- 6 Leveraging the Library and Other Institutional Resources
- 7 Example Digitization Protocols
- 8 Imaging Documents and Resources
- 9 Image File Types and File Specifications
- 10 Imaging Station Equipment and Specifications
- 11 Camera Manuals & Specifications
- 12 Workflows and Protocols
- 13 Georeferencing Resources
- 14 Database Resources and Tools
- 15 Identifiers
- 16 Videos

**iDigBio** [edit]

- Introduction to iDigBio Slide Set
- Intro to iDigBio pdf file

**Interest Groups** [edit]

- International Whole-Drawer Digitization Interest Group

**Preparation-specific Workshop Wikis** [edit]

- Herbarium Workshop Wiki
- Wet Collections Workshop Wiki
- Dried Insect Digitization Workshop Wiki
- Paleo Collections Digitization Workshop Wiki

**Workshop Summaries** [edit]

- iDigBio Workshop Summary Page
- Herbarium Digitization Workshop Report
- Wet Collections Workshop Report

**General Digitization Resources** [edit]

- No specimens left behind: mass digitization of natural history collections (ZooKeys Special Issue)
- Five task clusters that enable efficient and effective digitization
- Gil Nelson: Herbarium Digitization Tasks and Components Overview
- iDigBio's Intellectual Property Rights statement

**Views**

- Page
- Discussion
- Edit
- History
- Move
- Watch

**Personal tools**

- Gnelson
- My talk
- My preferences
- My watchlist
- My contributions
- Log out

**Navigation**

- Main page
- Community portal
- Current events
- Recent changes
- Random page
- Help

**Toolbox**

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

Developed a community-oriented digitization resources wiki in support of our workshops and to serve digitization-related information across all preparation types.

Established a digitization list serv to promote workshop follow-up as well as community discussion and sharing.



# iDigBio

Integrated Digitized Biocollections

# Minimum Standards for Scientific Collections (MISC)

Gil Nelson

Institute for Digital Information and Scientific Communication  
Integrated Digitized Biocollections  
Florida State University

Paleo Digitization Workshop  
23-25 September 2013

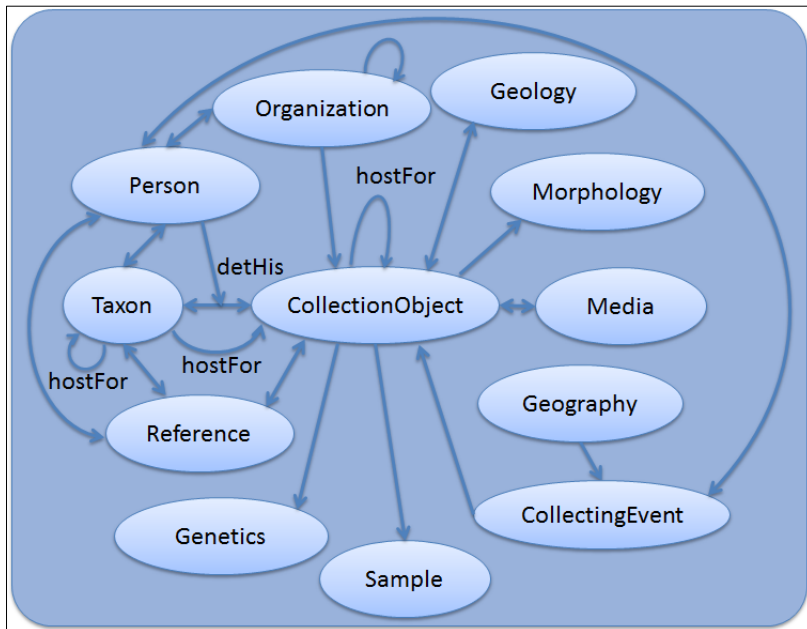
This material is based upon work supported by the National Science Foundation under Cooperative Agreement EF-1115210. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



# iDigBio Informatics and Cyberinfrastructure Workshop

28–30 March 2012

## Minimum Information for Scientific Collections Working Group



| IDigBio          | Many | One                      | Definition  | Validation/Notes  | Specimen or Collection-Object References |
|------------------|------|--------------------------|---|---|--|
| SpecimenGUID     |      | Occurrence:occurrenceID  | GUID for a specific physical specimen (collection object) given by the provider. Contributors encouraged to use an identifier created at the source to avoid duplication of records as data is shared with aggregators. The GUID should not change when a specimen is moved/donated/gifted to another collection. | Validate uniqueness. Validate prefix: <a href="http://rs.tdwg.org/func/terms/">http://rs.tdwg.org/func/terms/</a> <a href="https://www.idigbio.org/content/guid-statement">https://www.idigbio.org/content/guid-statement</a> |  |
| BarcodeValue     |      |                          | Machine readable alpha-numeric identifiers given to the collection object. Usually unique within a collection.  | if different than AccessionID   |  |
| AccessionID      |      | Occurrence:catalogNumber | Historical alpha-numerical identifiers given to collection objects.   |   |  |
| CollectionNumber |      | Occurrence:recordNumber  | Collector's number, the identifier given by the collector to a specimen or sample in the field and which is likely to have been written in associated field notes. The CollectionNumber isn't the same as the AccessionID, which is usually only applied once the specimen gets accessioned into a collection.    |   |  |



HOME ABOUT ENGAGE CONTRIBUTE

### MISC-Authority-File-Working-Group

This is the Wiki for the MISC/Authority File Working Group.

#### Contents

- 1 Working Documents
- 2 Data Model and MISC Placement
- 3 Data Element Lists by Data Model Concept
- 4 Name Sources

#### Working Documents

- MISC/Authority Files Way of Work
- MISC/Authority Files Working Document
- First Meeting Notice
- Agenda for first merged MISC/Authority files meeting
- Agenda 2012-10-16

#### Data Model and MISC Placement

- Working Data Model
- MISC Process

#### Data Element Lists by Data Model Concept



# MISC Goals

- **Ensure that data elements ingested by iDigBio are relevant and understandable to biologists and collections managers.**
- **Ensure that iDigBio data elements reflect, as much as possible, the content expressed by terms common to widely used biodiversity databases, schemas, and standards.**
- **Categorize data elements into those that are:**
  - **required for minimum scientific value,**
  - **highly desired for maximum scientific value,**
  - **complementary/supplementary (expendable, but important).**
- **Ensure flexibility by:**
  - **being open to all contributed data, regardless of whether currently included in MISC, DwC, AC, or other standards,**
  - **preserving opportunities to expand and refine MISC and the elements we ingest in the face of changing needs, standards, and contributions.**

# Why Standards?

Data translation and interchange

Interoperability

Efficient data mapping

Common vocabulary

Foster the development of ontologies (data relationships)

## Common Standards

Darwin Core (<http://rs.tdwg.org/dwc/terms/>)

Access to Biological Collections Data (ABCD) (<http://www.bgbm.org/tdwg/codata/schema/>)

Audubon Core (<http://vocabularies.gbif.org/node/126782>)

Dublin Core (<http://dublincore.org/documents/dcmi-terms/>)

Data elements that make a specimen minimally scientifically useful:

- Sense of identity of the organism
- Sense of time collected
- Sense of time period lived
- Sense of place collected
- Sense of stratigraphy
- Globally Unique Identifier (GUID)

# Identifying Objects



| ID         | 1565              | TSN        | 176580                            |
|------------|-------------------|------------|-----------------------------------|
| ROUND      |                   | 228.0      | IDORSAL <input type="checkbox"/>  |
| CATNO      | 279               |            | IVENTRAL <input type="checkbox"/> |
| SCIENTNAME | Scolopax minor    |            |                                   |
| Accepted   | Scolopax minor    |            |                                   |
| COMMONNAME | American Woodcock |            |                                   |
| Accepted   | American Woodcock |            |                                   |
| SEX        | #                 | Subspecies |                                   |
| MONTH      | 01                |            |                                   |
| DAY        | 04                |            |                                   |
| YEAR       | 1962              |            |                                   |
| COLLNAME   | Stoddard, Sr.     |            |                                   |

Record: 1 of 361 of 2945 | No Filter | Search



UUID or GUID does not have to appear on the specimen itself.

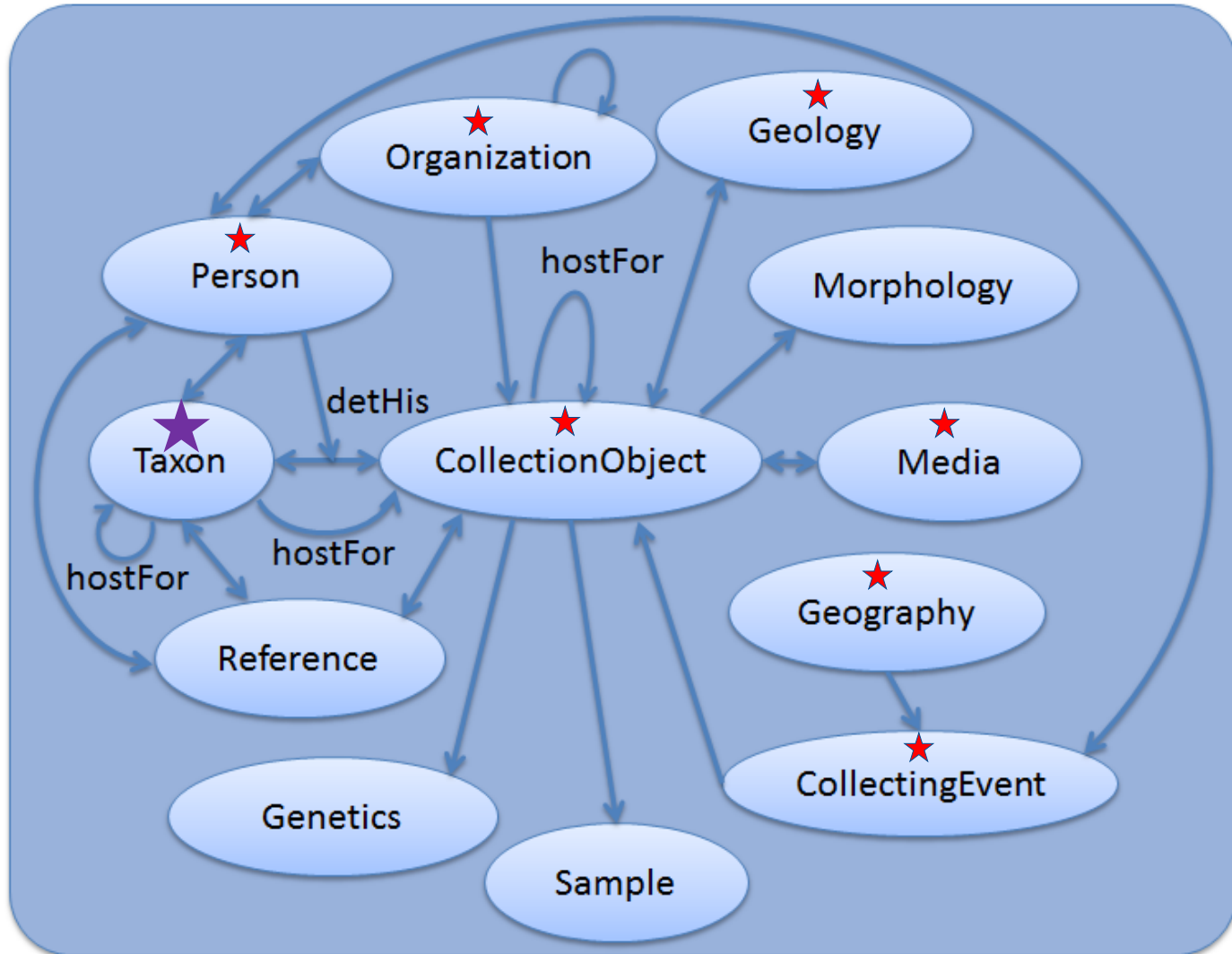
Add column to data record for a globally unique, persistent identifier.

<http://www.talltimbers.org/museum.html#Birds:279>  
[urn:uuid:3Ab1495230-ac34-42ea-b6b7-7af8b9f1b212](http://www.talltimbers.org/museum.html#Birds:279)



# MISC Phase I

<https://www.idigbio.org/wiki/index.php/MISC-Authority-File-Working-Group>



## MISC Product

At its core, the product of the MISC working group is iDigBio's attempt to:

- put flesh on the bones of the data model presented earlier,
- bring a biologist's or collection manager's perspective to the data elements iDigBio ingests,
- ensure that we account for all data currently or potentially stored in collections databases (hence, MISC may be a misnomer),
- narrowly (and perhaps selfishly?) focus on data elements iDigBio should be prepared to ingest over the long haul, to prioritize these elements with respect to whether they should be treated as required, highly desired, or supplementary, and to recognize that the list of these elements might grow over time,
- take a scientific perspective on data fitness,
- start with Darwin Core as a foundation and augment this standard from the many other schemas currently in use in our community,
- map MISC data elements to as many existing vocabularies/schemas as possible to facilitate ingestion.

## MISC-Authority-File-Working-Group

This is the Wiki for the MISC/Authority File Working Group.

### Contents

[hide]

- 1 Working Documents
- 2 Data Model and MISC Placement
- 3 Data Element Lists by Data Model Concept
- 4 Name Sources

#### Working Documents

[edit]

- MISC/Authority Files Way of Work
- MISC/Authority Files Working Document
- First Meeting Notice
- Agenda for first merged MISC/Authority files meeting
- Agenda 2012-10-16

#### Data Model and MISC Placement

[edit]

- Working Data Model
- MISC Process

#### Data Element Lists by Data Model Concept

[edit]

- Taxon Data Elements
- Specimen/Collection Object Data Elements
- Collecting Event Data Elements
- Geography Data Elements
- Collection Data Elements
- Geology Data Elements
- Person Data Elements
- Media Data Elements

#### Name Sources

[edit]

- Taxonomic Name Sources
- Geographic Name Sources

Retrieved from "https://www.idigbio.org/wiki/index.php/MISC-Authority-File-Working-Group"

This page was last modified on 10 October 2012, at 10:05.



#### Views

- Page
- Discussion
- Edit
- History
- Move
- Watch

#### Personal tools

- Gnelson
- My talk
- My preferences
- My watchlist
- My contributions
- Log out

#### Navigation

- Main page
- Community portal
- Current events
- Recent changes
- Random page
- Help

#### Toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

# MISC Working Group – Terms for Data Model Concepts

CollectionObject ☆

idigbioweb@flmnh.ufl.edu

Comments

Share

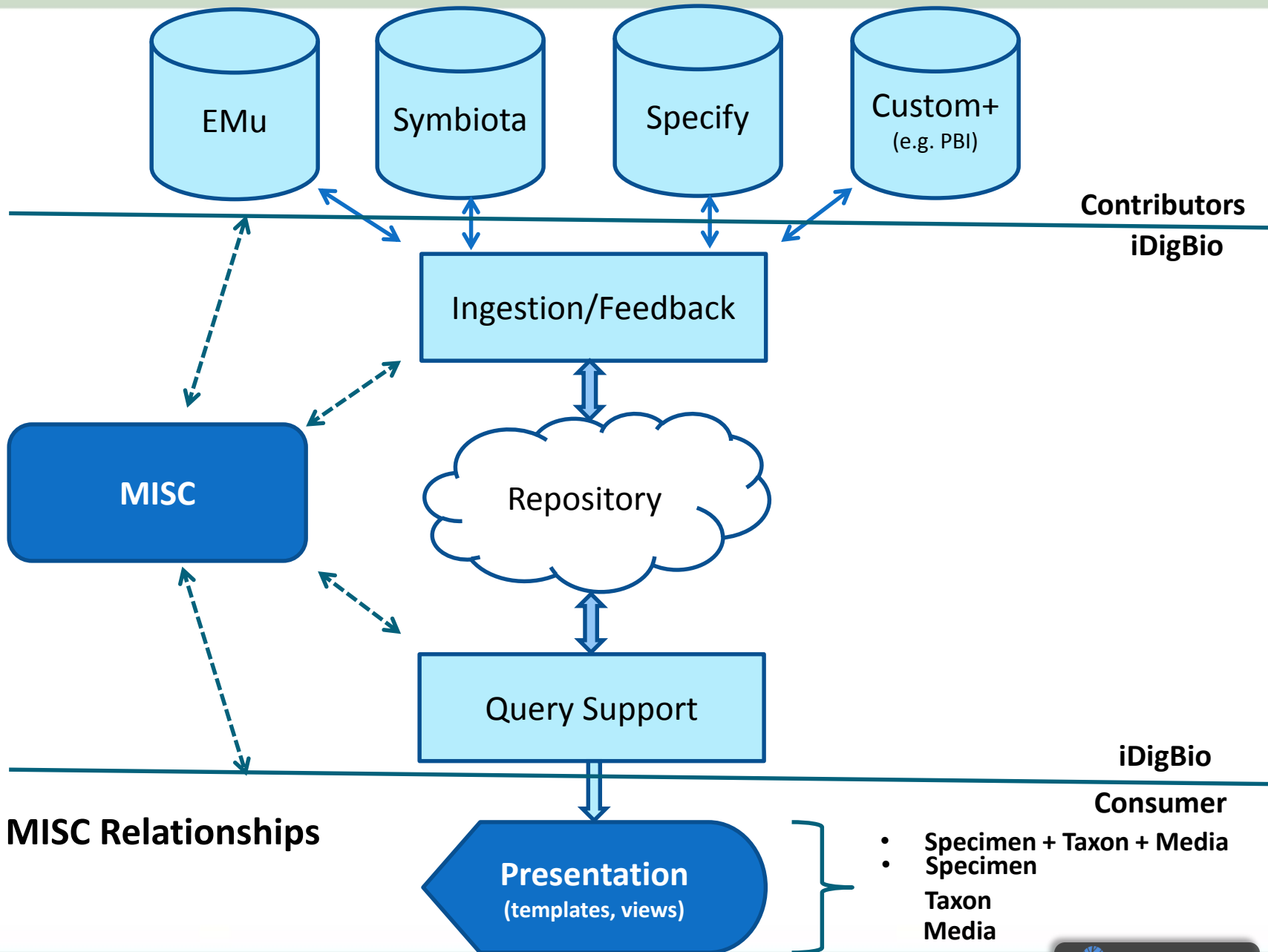
File Edit View Insert Format Data Tools Help All changes saved

fx | DWG \$ % 123 11pt B A A

|   | A                              | B    | C                        | D   | E   | F  |                  |
|---|--------------------------------|------|--------------------------|---|---|--|------------------|
|   | iDigBio                        |      | Dwc                      | Definition  | Validation/Notes  | Specimen or CollectionObject Terms   |                  |
|   |                                | Many |                          |   |   | References   |                  |
|   |                                |      |                          |   |   | User   |                  |
| 1 |                                |      |                          |   |   |  |                  |
| 2 |                                |      |                          |   |   |  |                  |
| 3 | SpecimenGUID                   |      | Occurrence:occurrenceID  | GUID for a specific physical specimen (collection object) given by the provider. Contributors encouraged to use an identifier created at the source to avoid duplication of records as data is shared with aggregators. The GUID should not change when a specimen is moved/donated/gifted to another collection. | Validate uniqueness. Validate prefix  | <a href="http://rs.tdvwg.org/dwc/terms/occurrence">http://rs.tdvwg.org/dwc/terms/occurrence</a><br><a href="https://www.idigbio.org/content/idigbio-guid-statement">https://www.idigbio.org/content/idigbio-guid-statement</a> | Specimen         |
| 4 | BarcodeValue                   |      |                          | Machine readable alpha-numeric identifiers given to the collection object. Usually unique within a collection.  | If different than AccessionID   |  | Barcode          |
| 5 | AccessionID                    |      | Occurrence:catalogNumber | Historical alpha-numerical identifiers given to collection objects.   |   |  | Accession Number |
| 6 | CollectionNumber               |      | Occurrence:recordNumber  | Collector's number, the identifier given by the collector to a specimen or sample in the field and which is likely to have been written in associated field notes. The CollectionNumber isn't the same as the AccessionID, which is usually only applied once the specimen gets accessioned into a collection.    |   |  | Collector        |
| 7 | OtherCatalogNumber             | Y    |                          | Previous or alternate fully qualified catalog numbers or other human-used identifiers for the same Occurrence, whether in the current or any other data set or collection   | Differs from Occurrence:otherCatalogNumbers, in that this should not be a concatenated list, but a "many" term. |  | Other Cat        |
| 8 | Preparation                    | Y    |                          | How the specimen has been prepared or presented.  | Differs from Occurrence:preparations, in that this should not be a concatenated list, but a "many" term.        |  | Preparati        |
|   | CollectionDateIntervalVerbatim |      | event:verbatimEventDate  | Verbatim date and time when the object was collected, exactly as reported by the collector in field book  | For the interpreted date, consult CollectionEvent concept   |  | Verbatim         |

Sheet1 Sheet2 Sheet3

Dwc





# iDigBio

Integrated Digitized Biocollections