Welcome!
And a few logistical details

Wiki: (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 and discoverable
- Enable digitization and research
  - Facilitate digitization workflows
  - Oversee implementation of standards and best practices for digitization
- 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

Grand Challenge

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 limitations of “data silos.”
Mandate and Responsibility

- Provide/facilitate portal access to collections data
  - Make information available and discoverable

- Enable digitization and research
  - Facilitate digitization workflows
  - Oversee implementation of standards and best practices for digitization

- 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

Grand Challenge

More recently, we have been encouraged by NSF to enhance international collaboration and sharing to overcome 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)


- Digitizing Fossils to Enable New Syntheses in Biogeography - Creating a PALEONICHES-TCN (University of Kansas)


- Mobilizing New England Vascular Plant Specimen Data to Track Environmental Change (Yale University)


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: MacroFungi
TCN: New England Vascular Plants
TCN: SCAN
TCN: PALEONICHLES
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
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)
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.
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.
Minimum Information for Scientific Collections Working Group

- Organization
- Geology
- Person
- Morphology
- Taxon
- hostFor
detHis
- CollectionObject
- Media
- Geography
- Genetics
- hostFor
Reference
CollectingEvent
Sample
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

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

http://www.talltimbers.org/museum.html#Birds:279
urn:uuid:b1495230-ac34-42ea-b6b7-7af8b9f1b212

UUID or GUID does not have to appear on the specimen itself.
MISC Phase I
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.
<table>
<thead>
<tr>
<th>ID</th>
<th>CollectionObject</th>
<th>Many/One</th>
<th>Dwc</th>
<th>Definition</th>
<th>Validation/Notes</th>
<th>Specimen or CollectionObject Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>SpecimenGUID</td>
<td>Many</td>
<td></td>
<td>Occurrence</td>
<td>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.</td>
<td>Validate uniqueness. Validate prefix.</td>
</tr>
<tr>
<td>4</td>
<td>BarcodeValue</td>
<td>Many</td>
<td></td>
<td>Occurrence</td>
<td>Machine readable alpha-numeric identifiers given to the collection object. Usually unique within a collection. If different than AccessionID</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>AccessionID</td>
<td>Many</td>
<td></td>
<td>Occurrence</td>
<td>Historical alpha-numerical identifiers given to collection objects.</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CollectionNumber</td>
<td>Many</td>
<td></td>
<td>Occurrence</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>OtherCatalogNumber</td>
<td>Many</td>
<td></td>
<td>Occurrence</td>
<td>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</td>
<td>Differs from Occurrence otherCatalogNumbers, in that this should not be a concatenated list, but a &quot;many&quot; term.</td>
</tr>
<tr>
<td>8</td>
<td>Preparation</td>
<td>Many</td>
<td></td>
<td>Occurrence</td>
<td>How the specimen has been prepared or presented.</td>
<td>Differs from Occurrence preparations, in that this should not be a concatenated list, but a &quot;many&quot; term.</td>
</tr>
<tr>
<td>9</td>
<td>CollectionDateInterval</td>
<td>Many</td>
<td></td>
<td>Occurrence</td>
<td>Verbatim date and time when the object was collected, either actually by the collector in field book, or for the interpreted date, consult CollectionEvent record</td>
<td></td>
</tr>
</tbody>
</table>