iDigBio Data Use:
Enabling Research Through Training, Integration, and Tools

Pamela S. Soltis
University of Florida
Overview

- Training workshops
- Sample publications
- Developing research programs: data integration
- Promoting research use through symposia, workshops, and special publications
- Enabling research through tool development
- Enabling research through tool availability
- Publication metrics
Overview

• Training workshops
• Sample publications
• Developing research programs: data integration
• Promoting research use through symposia, workshops, and special publications
• Enabling research through tool development
• Enabling research through tool availability
• Publication metrics
Training Workshops

• Specific skills in data use - iDigBio 2 (July 2016)
  – Soltis team
  – Georeferencing, ecological niche modeling, etc.
  – Venues
    • Botany 2016 (~50)
    • Denver (CU-Denver) (~25)
    • Botany 2017 (~45)
    • Applied to Ecological Society of America 2017
  – Webinar series planned for spring, 2018
  – Very positive reviews
  – Using the materials to train undergraduates, e.g. ‘mini-REU site program’, summer 2017
‘Mini-REU Site’ Program - Summer 2017

• 5 students
  – 3 from UF, 1 from UCF, 1 from USC
  – 3 women, 2 men
  – Summer research
  – Final symposium
  – 2 presentations @ Botany 2017

https://www.idigbio.org/content/research-experiences-undergraduates-digitized-collections-data
Training Workshops

• Data/Software Carpentry - iDigBio 2 (July 2016)
  – M. Collins, D. Paul et al.
  – TDWG 2016
  – TDWG 2017
  – Florida State University 2017
  – ADBC Summit 2017
  – 8 workshops @ UF; UFBI/UFII

• Developing Data Carpentry lessons using iDigBio

• Plans:
  – Establishing group to share Carpentries membership
  – Instructor training at TDWG/SPNHC
  – Talk by Carpentries development manager at TDWG/SPNHC
Sample Publications

- Marchant et al. 2016 - ecology of polyploid plants
- Willis et al. 2017 - phenology review
- James 2017 - new species
Developing Research Programs: Data Integration - Genomics

• GenBank
  – Accession numbers added where available
  – Discussions with GenBank about changing their voucher requirements to include GUIDs rather than triplets

• Linking genomics with environmental data through specimen records: Amborella

• Spatial distribution of genome sizes in plants
Selective sweeps and habitat variation in *Amborella*
Genome-scale SNPs & Environment
Selective sweeps and habitat variation in *Amborella*
Genome-scale SNPs & Environment

Selective sweeps and habitat variation in *Amborella*

Environmental Data

Allele frequencies positively correlated with habitat suitability

- Metabolic processes
- Embryo development during seed dormancy
- rRNA processing
Developing Research Programs: Data Integration - Genomics

• GenBank
  – Accession numbers added where available
  – Discussions with GenBank about changing their voucher requirements to include GUIDs rather than triplets

• Linking genomics with environmental data through specimen records: *Amborella*

• Spatial distribution of genome sizes in plants
Spatial Distribution of Genome Sizes

- Nucleic acids: large amounts of N and P; large genomes costly to build, replicate, transcribe, etc.
- Plants with large genomes should be selected against on N- and P-poor soil, favored on high-N/P soil
- Park Grass Experiment used to test this hypothesis:
  - GS of plants in high N+P plots higher than in control, N, P plots
- Continental scale, GS related to soil geochemistry?

Guignard et al. 2016
Spatial Distribution of Genome Sizes

GeoEcoEvo: USGS Powell Center Working Group
E. Bui, M. Goldhaber, PIs
I. Pearse, J. Cartwright, I. Leitch, A. Leitch, D. Soltis

Plant C-values Database
Developing Research Programs: Data Integration - Phenology

- Willis et al. (2017) *TREE*
- Plant Phenology Ontology
- Use of images - machine learning
  - Stucky, Guralnick, Soltis et al.
  - Pl@ntNet - Pierre Bonnet et al.; new collaboration with Nelson, Ellwood, Soltis et al.
Review

Old Plants, New Tricks: Phenological Research Using Herbarium Specimens

Charles G. Willis,¹,* Elizabeth R. Ellwood,²,* Richard B. Primack,³ Charles C. Davis,¹ Katelin D. Pearson,² Amanda S. Gallinat,³ Jenn M. Yost,⁴ Gil Nelson,² Susan J. Mazer,⁵ Natalie L. Rossington,⁵ Tim H. Sparks,⁶,⁷ and Pamela S. Soltis⁸
Willis et al. (2017) TREE
Plant Phenology Ontology: standards, terms, links

Willis et al. 2017

Figure 1. Simplified Representation of Ontological Classes and Logical Structure. In a complete ontology, each term or ‘class’ has a specific definition and is linked to any and all related classes via ‘relation terms’ such as ‘is_a’ or ‘part_of’. These structured linkages between classes allow integration among different methods of measuring a class (represented in blue), different subclasses within a class (white), and other types of data (yellow), which are subclasses of the general term ‘quality’ currently defined by the Phenotypic Quality Ontology.
Phenology: Engaging Citizen Scientists

Rob Guralnick

Julie Allen  Austin Mast
Phenology: Data from Multiple Sources

B. Stucky et al.
Free, web & mobile app dedicated to plant identification and gathering of botanical observations

Pl@ntNet is based on an innovative crowdsourcing workflow

Promoting Research Use
Workshops, Symposia, Special Pubs

• Phenology Workshop @ UC Berkeley, March 2016
  – Willis et al. (2017) review; TREE
  – Yost et al. (in review) methods; APPS


• Green Digitization Symposia @ Botany & IBC 2017

• Phenome 2018: promoting use of images
Save the Date

PHENOME 2018
TUCSON, AZ FEBRUARY 14-17
CONNECTING BIOLOGY, SYSTEMS, AND TOOLS
Promoting Research Use
Workshops, Symposia, Special Pubs

• Special issue of *APPS* (2018) based on symposia
  – Nelson, James, Soltis (eds)

• Special issue of *Conservation Biology* (2018)
  – Ellwood, Klein, Soltis (eds)

• Invited essay *Amer. J. Bot.* - Soltis (2017)
  – *On the Nature of Things* Series
  – Digitization of herbaria and research

• iDigBio/ADBC grad students
  – paper on research use underway
Enabling Research Through Tool Development

• GUODA: Global Unified Open Data Access (ACIS)
• Jupyter: Python & R notebook infrastructure (ACIS)
• FreshData: Notifications of new data records (ACIS)
• Effechecka: Taxonomic Checklist Generator (ACIS)
Enabling Research Through Tool Development

- Interfaces with Open Tree of Life and Lifemapper (Soltis)
- USVH as a model - linking resources (Soltis/Nelson)
Enabling Research Through Tool Availability

• Coming soon!
  – Descriptions of and links to recommended and vetted software and tools for use with specimen data
  – Accessible via the Portal and the Research page
Other Collaborations, Partnerships, Efforts

• Proposals to ABI for tool development
• Proposals to enhance research use through cyberinfrastructure development
• Proposals to enhance training in biodiversity and data science
• RCNs (e.g., Enhancing Participation; Biodiversity Literacy in Undergraduate Education)
Publication Metrics

• Compiled by Shari Ellis
• Number of publications
• Mentions of iDigBio vs. portal use in publications
• iDigBio vs. TCN portal use
Publications: Mentions vs Portal Use by Year

Number of Publications

- ADBC mentions
- Portal use
Publications Using idigbio.org or TCN portals or resources
Thank you!

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.