Digitization Symposium
Association of Southeastern Biologists

- Gil Nelson (gnelson@bio.fsu.edu)
- Deb Paul (dpaul@fsu.edu)

(Florida State University)

12 April 2013

Charleston, WV

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
Integrated Digitized Biocollections

Integrated Digitized Biocollections

Introducing iDigBio: An NSF Collaborative of the University of Florida and the Florida State University

Gil Nelson
Deb Paul
Florida State University

Advancing Digitization of Biodiversity Collections

- **Facilitate use of biodiversity specimen data to address environmental, scientific, and economic challenges**
  - Biodiversity researchers and scientists
  - Educators
  - General public
  - Policy-makers

iDigBio
Integrated Digitized Biocollections

- **Enable digitization of biodiversity collections data**
  - Develop efficient and effective digitization standards and workflows
  - Respond to cyberinfrastructure needs

- **Provide access to biodiversity data in a cloud-computing environment**

- **Plan for long-term sustainability of the national digitization effort**
  - Expand participation: partners and data sources

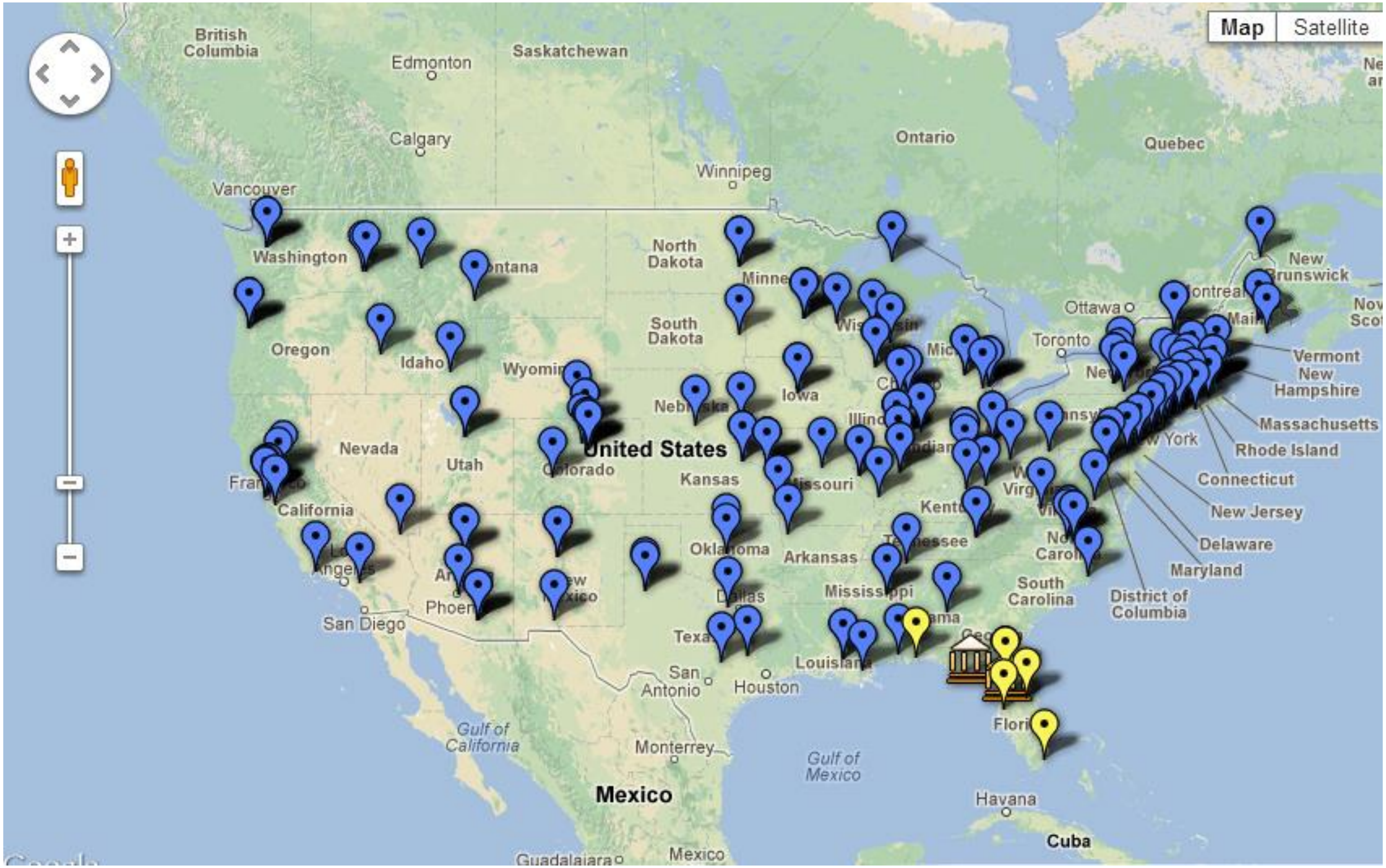
NSF's Grand Challenge

Digitize (text + images) and link one billion specimen records from collections across the U.S.

Seven Thematic Collections Networks (TCNs)

- 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
- 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

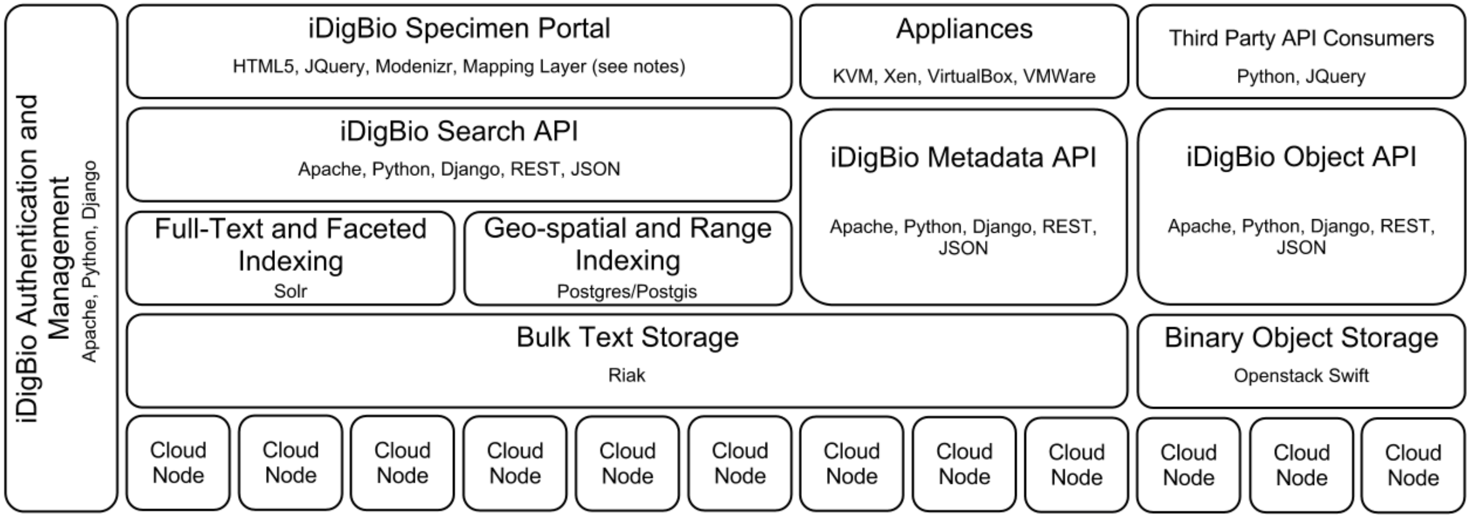
National Resource (iDigBio), Thematic Collection Networks (TCNs), and Collaborators



7 TCNs, 130+ participating institutions, 49 states

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, workshops, person-to-person



iDigBio Authentication and Management
Apache, Python, Django

iDigBio Specimen Portal
HTML5, JQuery, Modenizr, Mapping Layer (see notes)

Appliances
KVM, Xen, VirtualBox, VMWare

Third Party API Consumers
Python, JQuery

iDigBio Search API
Apache, Python, Django, REST, JSON

iDigBio Metadata API
Apache, Python, Django, REST, JSON

iDigBio Object API
Apache, Python, Django, REST, JSON

Full-Text and Faceted Indexing
Solr

Geo-spatial and Range Indexing
Postgres/Postgis

Bulk Text Storage
Riak

Binary Object Storage
Openstack Swift

Cloud Node | Cloud Node | Cloud Node | Cloud Node | Cloud Node | Cloud Node | Cloud Node | Cloud Node | Cloud Node | Cloud Node | Cloud Node | Cloud Node

What Makes iDigBio Unique?

- Ingest all contributed data with emphasis on GUIDs, not only a restricted set of selected data elements
- Maintain persistent datasets and versioning, allowing new and edited records to be uploaded as needed
- Ingest textual specimen records, associated still images, video, audio, and other media
- 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
- Facilitate sharing and integration of data relevant to biodiversity research
- Provide computational services for biodiversity research

Recent and Ongoing Activities

- Assessment of common and effective practices (paper in *ZooKeys*)
- Minimum information for scientific collections working group
- Collaborative georeferencing pilot project at Godfrey Herbarium
- Digitization workflows working groups
- Biodiversity Informatics Manager working group
- Public Participation in Digitization of Biodiversity Specimens workshop
- Georeferencing working group & train-the-trainers workshop
- OCR/natural language processing working group & Hackathon
- ASB symposium and workshop
- Series of digitization training workshops
- Call for appliances
- Specimen data portal implementation
- Server hosting

iDigBio
Integrated Digitized Biocollections

iDigBio

Integrated Digitized Biocollections

Getting Started with Digitization

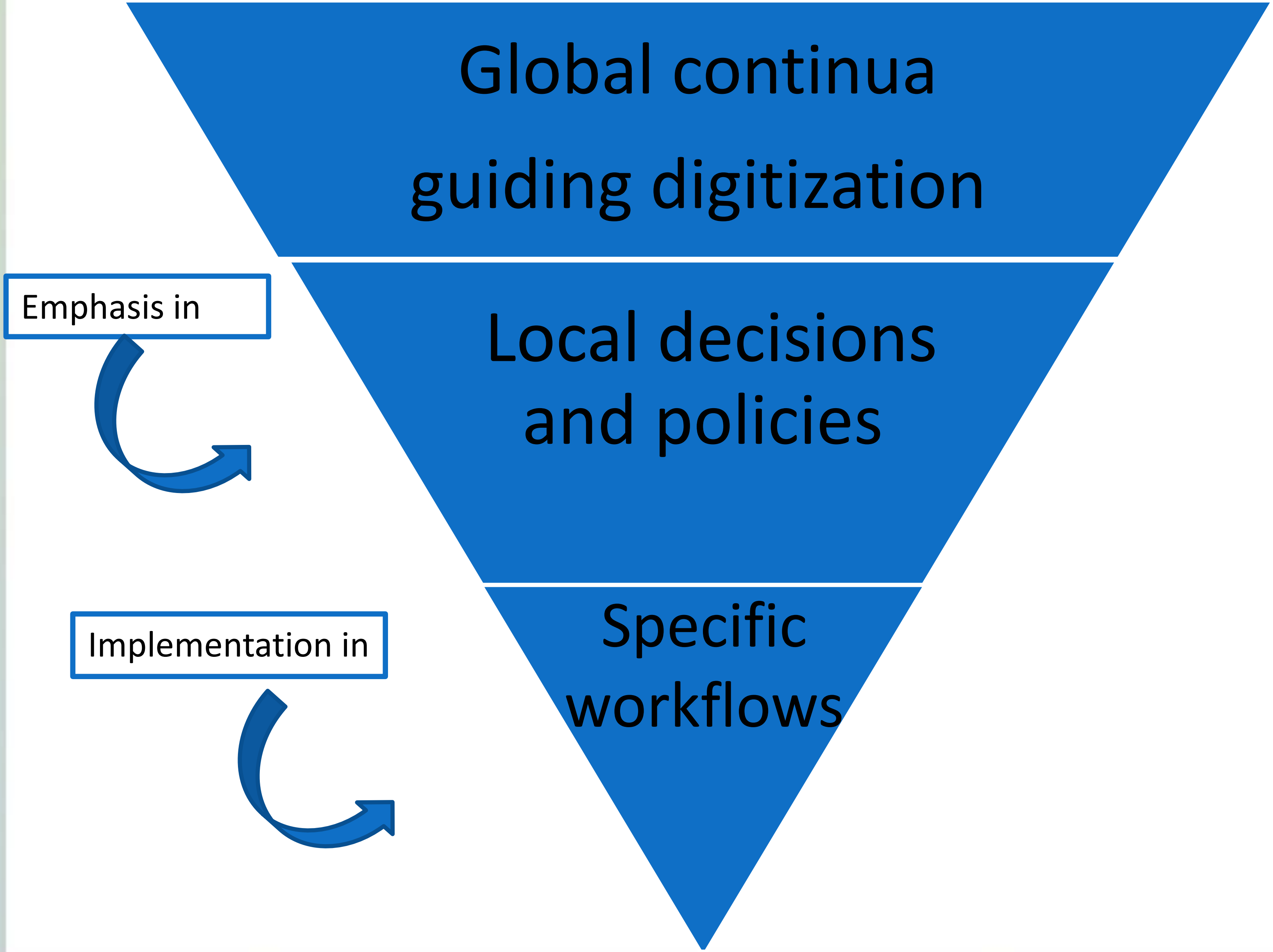
Gil Nelson
Florida State University

Ultimate Goals of Biological Collections Digitization

Output level: An abundance of scientifically useful and accessible data.

Constituency level: High quality exposure of the content and value of scientific collections.

Improvement level: Collaboration and workflow sharing across the collections community.



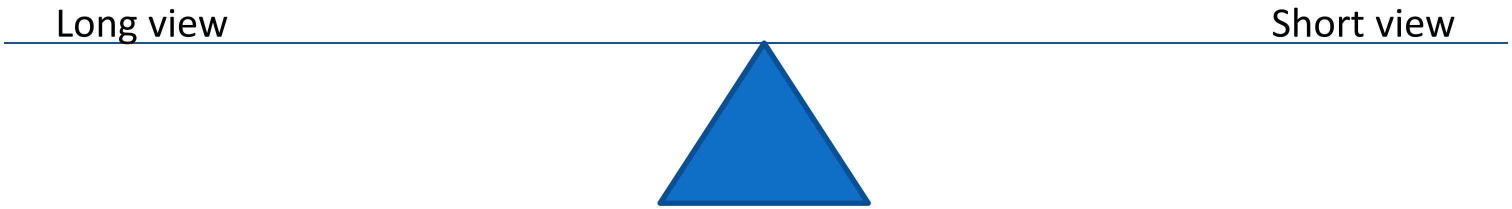
Global continua guiding digitization

Local decisions and policies

Specific workflows

Emphasis in

Implementation in



Long view

Short view

- **Taking the long view means developing doable, effective, and sustainable strategies for robust digitization NOW.**
- **Taking the short view means balancing long term goals with short term constraints, including a commitment to implementing future enhancements.**

Pressures mitigating the long view

So much data, so little time.

Our collections are not getting smaller.

The funding agencies have high output expectations.

We only have 3 years to get this done.

All of our data and all of our specimens are important.

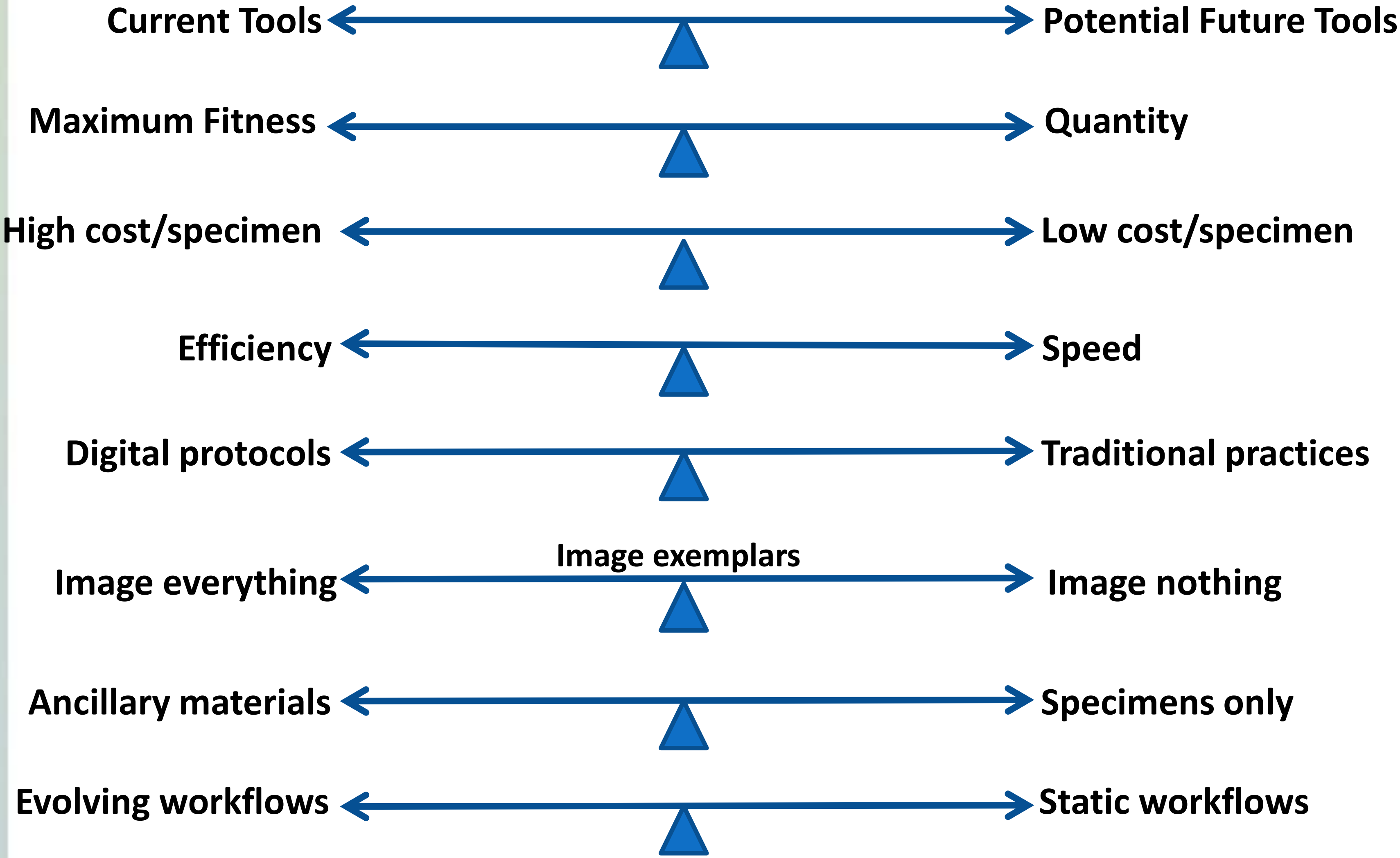
Let’s just use the images!

Do the minimum now and enhance it later.

Future Tools Favoring the Short View

- OCR, NLP, and ICR (handwriting analysis) improvements.
- Automated image analysis for data extraction.
- Data mining of labels.
- Robotic technologies, conveyor belts, etc.
- Improvements in discovery/capture/use of duplicates.
- Improvements in voice recognition and other data entry technologies.
- Post-digitization tools for curation and quality control.
- Field data capture.

Digitization Decision Continua



Current Tools ↔ Potential Future Tools

Maximum Fitness ↔ Quantity

High cost/specimen ↔ Low cost/specimen

Efficiency ↔ Speed

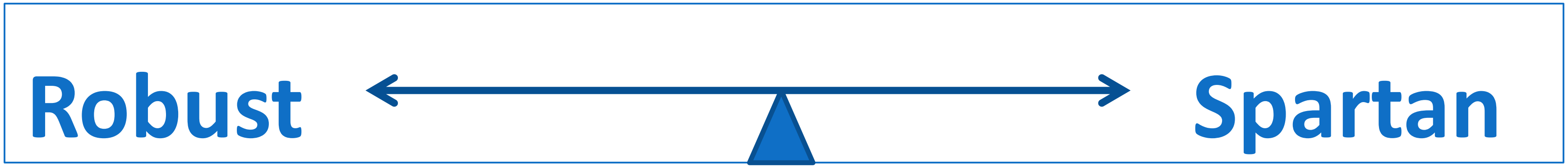
Digital protocols ↔ Traditional practices

Image everything ↔ Image nothing

Image exemplars

Ancillary materials ↔ Specimens only

Evolving workflows ↔ Static workflows



Robust

Spartan

Facilitators	Facilitators
Emphasize immediate fitness for use	Emphasize output
Robust datasets	Skeletal datasets
Data validation/cleaning	Defer validation/cleaning
Integrated quality control	Deferred quality control
Integrated georeferencing	Deferred georeferencing
Intensive physical curation	Deferred digital curation
Record historical annotations	Record current determination
Staff specialization	Staff generalization
Small collection	Large collection
Emphasize images	Emphasize data
High quality images	Low quality images

iDigBio

Integrated Digitized Biocollections

Thank you!