InvertNet

Advancing Digitization of Biological Collections

InvertNet: A New Paradigm for Digital Access to Invertebrate Collections

Chris Dietrich

Illinois Natural History Survey

University of Illinois

ASB 2013

Outline

- overview (rationale, scope, goals)
- digitization workflows
- data management, network architecture
- progress report
- future plans

ADBC Goals

- digitize 1 billion specimens in 10 years for $100 million ($0.10/specimen)
- build Thematic Collection Networks (TCNs) to address specific research goals
- link TCNs under national HUB (iDigBio)

InvertNet Rationale

- vast majority of specimens in U.S. collections are invertebrates
  - primarily insects and related arthropods
  - less than 5% available online
  - only label data usually provided
- most invertebrate biodiversity research is specimen-based
  - all knowledge of many species is embodied in collections
- existing digitization methods are inadequate
  - slow and expensive ($1+ per specimen)
  - risk of damage to specimens from handling

InvertNet Goals

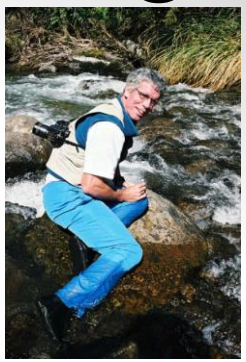
- Digitize all holdings of 22 midwestern arthropod collections (~50 million specimens)
  - Specimen images and metadata (label info)
  - Drawers, vials, slides
  - Advanced imaging (including 3D)
  - Best quality at reasonable cost (~$0.10/specimen)
- Provide access to images and other data via online virtual museum
  - browsable/searchable/zoomable web interface
  - link to other data providers (GBIF,iDigBio etc.)
- Provide platform for research and development of additional tools and resources
  - Data mining and analysis
  - Community building, collaboration, and support
  - Education, outreach, and reference

InvertNet UIUC Team

- Chris Dietrich – Director
  - Systematic Entomologist
- John Hart – CoPI
  - Computer Science - Graphics
- Nahil Sobh – CoPI
  - Computational Multiscale Nanosystems
- Umberto Ravaioli – CoPI
  - Computational Multiscale Nanosystems
- David Raila – Senior Collaborator
  - Computer Science – Sr. Research Programmer
- Others
  - Programmers, research assistants, hourlies

InvertNet Collaborating Curators

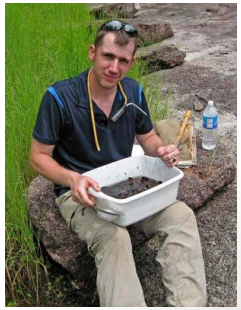
- A. Cognato, MSU
- G. Courtney, J. VanDyk, ISU
- J. Holland, Purdue
- R. Holzenthal, P. Tinerella, Minnesota
- P. Johnson, SDSU
- H. Klompen, M. Daly, OSU
- J. Rawlins, R. Davidson, J. Fetzner, Carnegie Museum
- D. Rider, G. Fauske, NDSU
- A. Short, Kansas
- R. Sites, Missouri
- D. Young, Wisconsin-Madison
- J. Zaspel, Wisconsin-Oshkosh
- G. Zolnerowich, KSU
- D. Rubinoff, U Hawaii
- T. Roberts, U Iowa













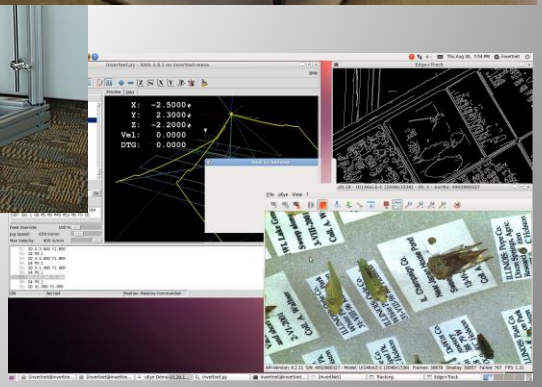
InvertNet
Advancing Digitization of Biological Collections

Other InvertNet Collections

- Eastern Illinois University
- Western Illinois University
- Southern Illinois University
- Illinois State University
- Milwaukee Public Museum
- Northern Michigan University
- U North Dakota
- Valley City State University

Phase 1
(Years 1-2)

- Stage collections for digitization
  - basic housekeeping (drawer and unit tray labels, updating nomenclature, organizing identified material)
  - curator exchanges to upgrade curatorial status of focal taxa
- Develop digitization toolkit/workflow
  - Test variety of capture hardware, software and processes
  - Test and evaluate variety of image processing/reconstruction methods
- Establish web portal at UIUC using HUBzero platform
  - Community development for collaborators
  - Digitization workflow
  - Searchable/browsable web interface for images and label data
- Develop training materials for participants (videos, manuals, wikis, etc.)



InvertNet
Advancing Digitization of Biological Collections

ASB 2013

Phase 2

(Years 3-4)

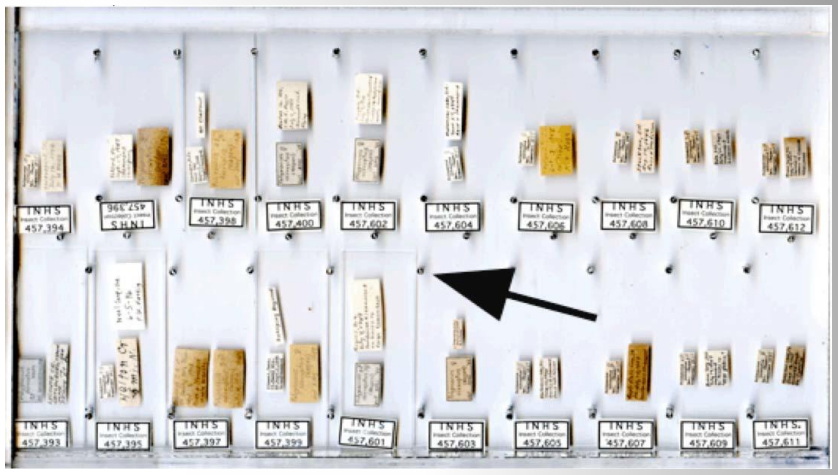
- High-throughput collection digitization
  - capture and provide immediate access to high-quality specimen images
  - crowd source label data capture
- Refine digitization and processing tools
  - further automate workflows
  - image processing/segmentation
  - 3D
- Link to other sites
  - iDigBio, BugGuide
- Incorporate data exploration, analytical, and modeling tools

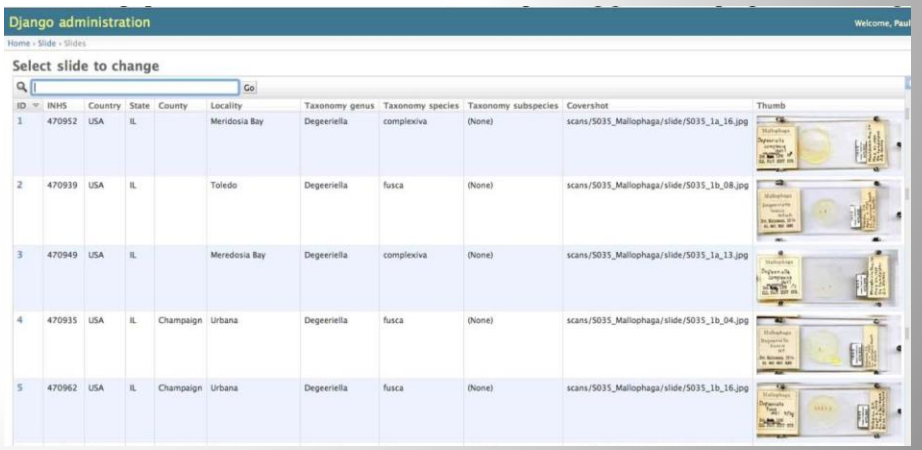
Digitization Workflow: First Pass

- Acquire raw image(s) & metadata for multiple specimens simultaneously
  - entire drawers of pinned specimens: multiple images from different perspectives stitched together for 2D and 3D reconstruction and zoom capability
  - for slides and vials: 2D images of multiple units acquired simultaneously then segmented into individual database containers
- Upload images to centralized repository for further processing
  - includes automated stitching, segmentation of unit trays and specimens
  - semiautomated capture of metadata (taxonomy, labels)
- Advantages:
  - meet cost target of 10 cents/specimen
  - provide rapid access to entire digitized collections

Digitization workflow: slides

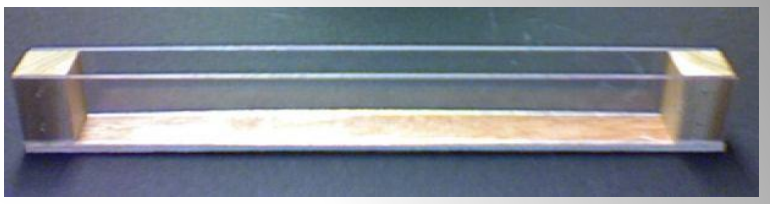
1. place 20 slides face down on clear tray
2. scan image of tray
3. segment image using pixel map
4. individual slide images automatically placed in separate database containers
5. semiautomated capture of label data



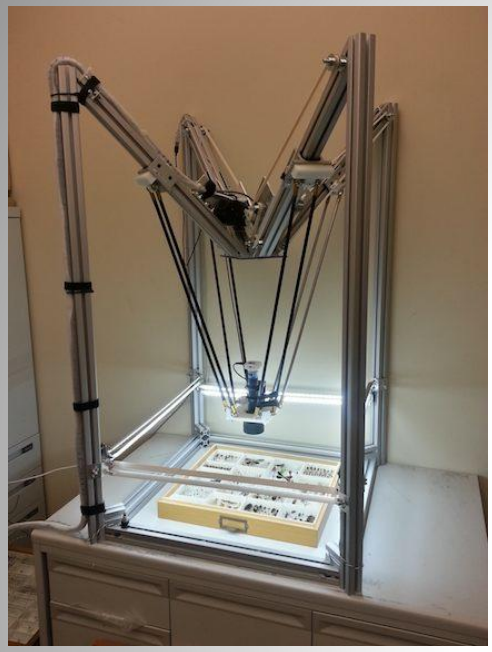


Digitization workflow: vials

1. place 48 vials in 3 custom racks and rotate so labels oriented consistently
2. Place racks on scanner bed
3. Scan at 600 dpi
4. Flip racks over and scan opposite side
5. Segment images as for slides



Drawer Imaging



- Delta Robot, digital camera, telecentric lens captures grid of single, close-up images at 40-60 x/y coordinates and 5 perspectives

- Single images stitched to yield Gigapixel images from multiple viewpoints

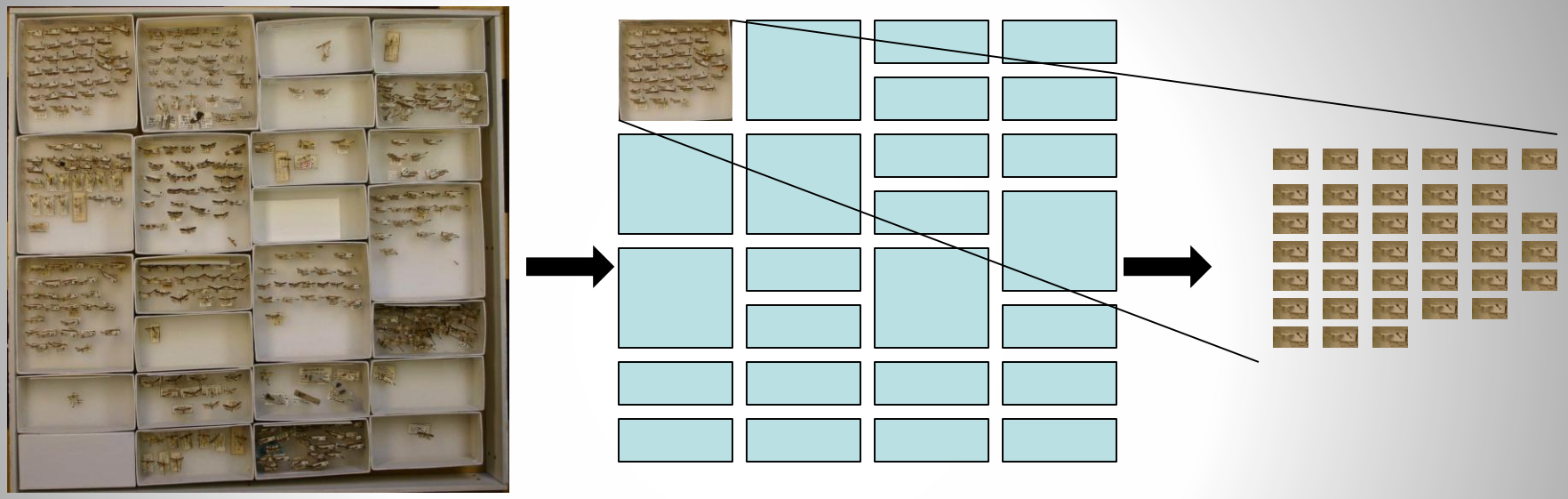
Top-down view

Angled view

- Enables virtual tilting

InvertNet
Advancing Digitization of Biological Collections

Image segmentation/annotation



1. capture image of drawer + metadata (location, contents)
2. segment unit trays (image analysis software)
3. segment specimens
4. capture label data

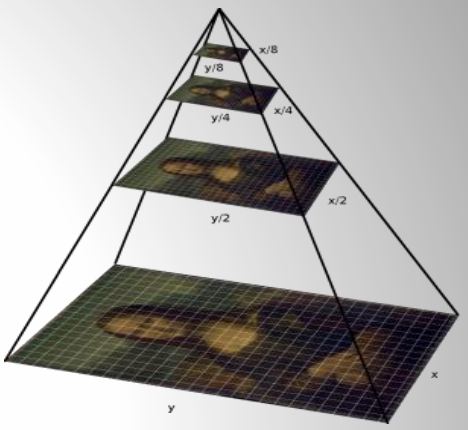
InvertNet Web Infrastructure

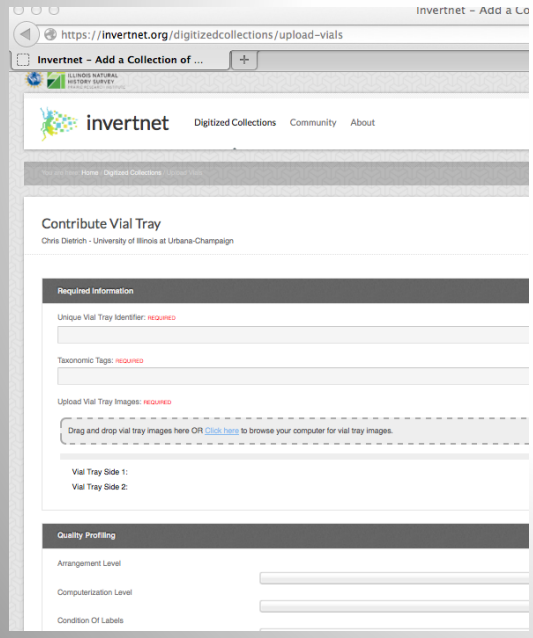
- HUBzero Cyberinfrastructure
  - Dynamic web 2.0 platform for scientific research and educational activities ("CMS on steroids")
    - Browser-based access to databases/semantic repositories
    - Extensible backend supports highly interactive tools
      - Image processing, searching, analytics, etc.
      - Integration with high-performance computing resources
    - Integration with FEDORA preservation and archiving
- InvertNet.org
  - Digitization workflows
  - Image processing/rendering
  - Databases
  - Community building/interaction/collaboration
    - wikis/blogs/groups
    - polls/wish lists
    - links to social networking sites
  - Analytical tools
  - Developer tools (hardware environments, virtual machines, testbeds)
  - Education/Outreach tools

InvertNet Data Management

Current ingest pages for slides and vials:

- drag and drop chunked uploading
- tagging, profiling, batch submission
- CoL taxonomic tree- and tag-based site search
- zoomable viewer supporting Tiled Pyramidal TIFF image stacks



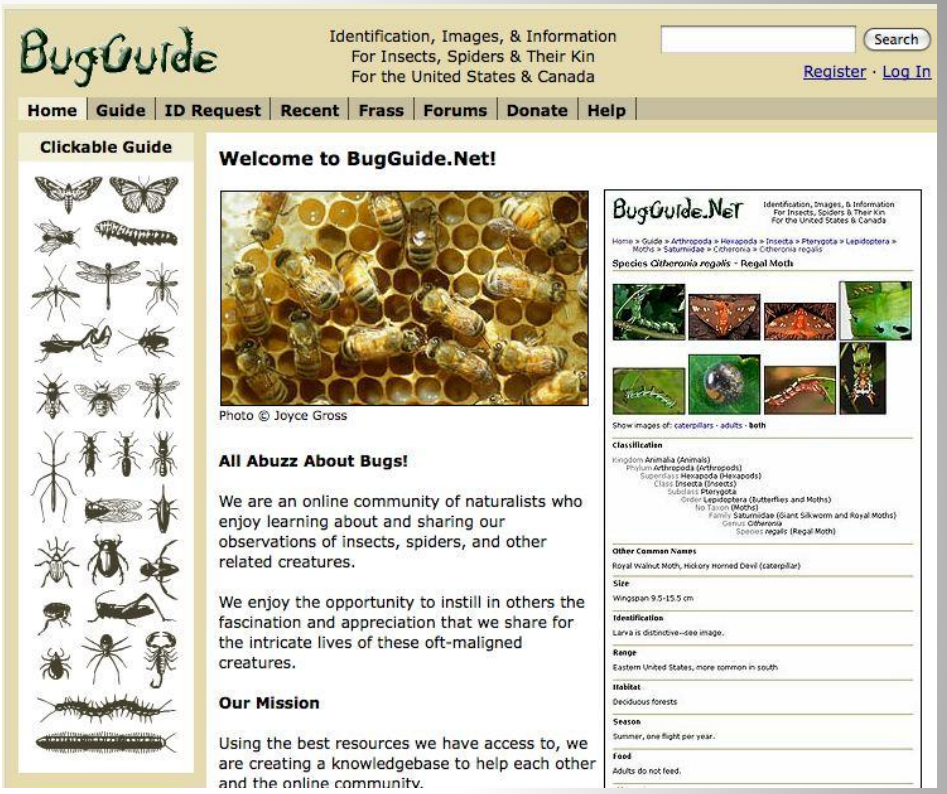


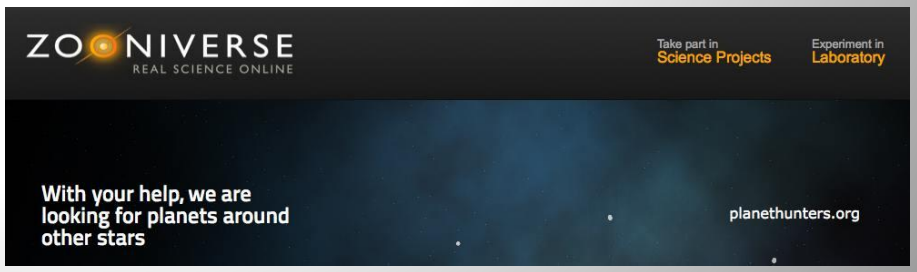
TCN Themes

- Environmental change
  - changes in biota over time reflect changes in climate, landscape use, etc.
- Species discovery
  - high-res images of specimens, including unsorted/unidentified materials, become accessible to expert taxonomists at remote locations
- Species identification
  - replicate images of identified species used for morphometric analysis and improved identification accuracy/automated identification

Outreach

- link to BugGuide
  - users compare photos of live bugs to images of identified specimens
- crowd-sourcing label data capture (Zooniverse)





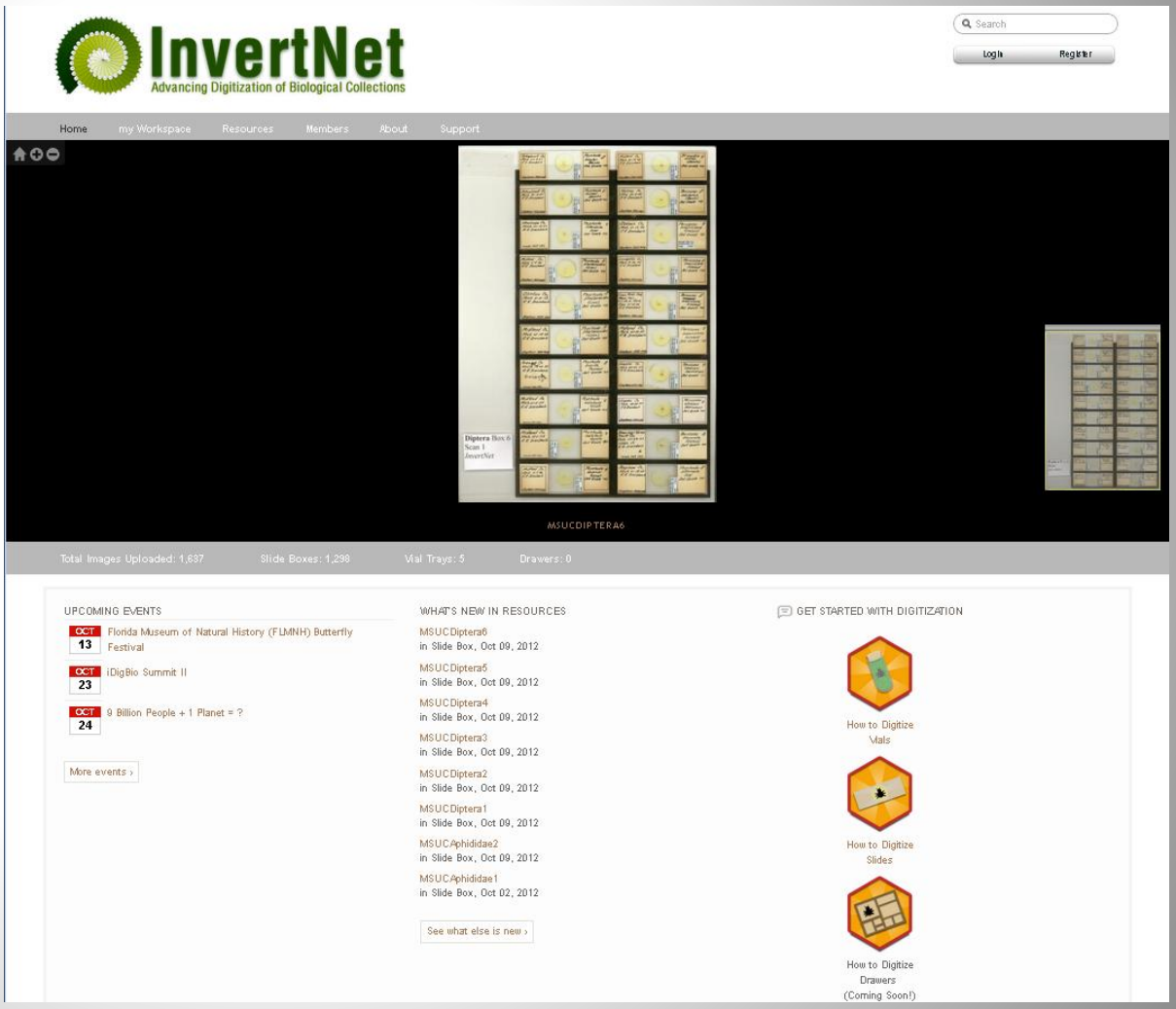
InvertNet
Advancing Digitization of Biological Collections

Summary

- Short-term goals (4 years):
  - digitize 50 million specimens from 22 collections
  - provide access via virtual museum
  - provide tools supporting theme-related research, education and outreach
- Long-term goals
  - incorporate federal and non-US collections
  - include all invertebrates worldwide

Website

- InvertNet.org
- registration is open to all and available now; please join us!



Acknowledgements

Collaborators: J. Hart, N. Sobh, O. Sobh, D. Raila, U. Ravaioli, C. Taylor, A. Cognato, G. Courtney, J. Holland, R. Holzenthal, P. Tinerella, P. Johnson, H. Klompen, M. Daly, J. Rawlins, R. Davidson, J. Fetzner, D. Rider, G. Fauske, A. Short, R. Sites, D. Young, J. Zaspel, G. Zolnerowich

Funding: NSF ADBC program

InvertNet
Advancing Digitization of Biological Collections