



Digitization Planning for Large Collections with Specify 7

Grant Fitzsimmons & Theresa Miller



Iriomote cat illustration by Prof. Mitsuru Moriguchi, Okinawa University.
Logo art by Ms. Takako Tomozawa, NPO Osaka Natural History Center.
Used with permission. All rights reserved.

[Digitization Planning for Large Collections with Specify 7](#)
© 2024 by the [Specify Collections Consortium](#) is licensed
under [CC BY 4.0](#)

Introduction

- Specify Supports Research Collections
 - Specify is an intuitive, robust, and highly-customizable software platform for digitizing and managing research collections.
- 25+ Years of Providing Collections Management Solutions
 - Specify has sustained biological research museums and biorepositories with software for managing, integrating, and publishing collections information.
- Engaging with Large Institutions
 - The Specify Collections Consortium has worked with a wide range of institutions, from universities to government agencies, to envision a framework that supports standard workflows and best practices.

Specify 7

Record Set: Fundulus

Cat # 000035591 Accession # 5115 PrevExch #

Cataloger Bentley, Andrew C Cat Date Full Date 03/19/2002

Determinations

Taxon Fundulus escambiae In question

Preferred Taxon Fundulus escambiae Current

Determiner Wiley, Edward O Date Full Date mm/dd/yyyy Type Status

Remarks

Field No. Locality EOW 88-17: 1988-07-16: North America, United States, Alabama, Baldwin: Dyas Creek on Baldwin county road 61: 30.8681

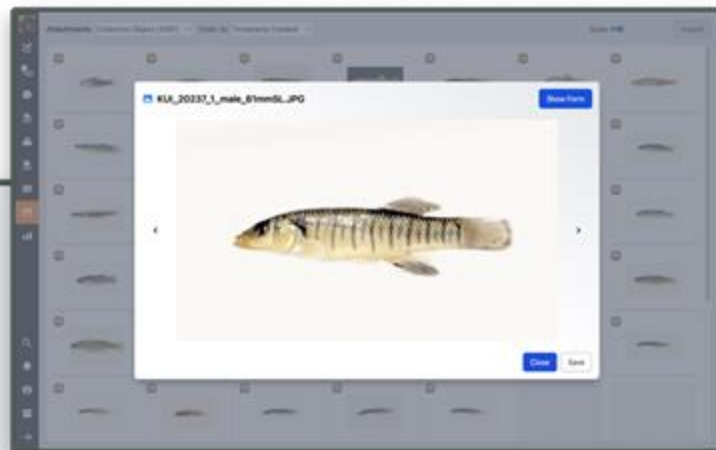
Preparations (1)

Object Type	Count	Actual Count	Is On Loan	Prepared Date	Voucher Storage	Remarks	Preparation Attachments
EOW	3	3		mm/dd/yyyy	Jar		0

Collection Object Collection

- 000006582 KU Fish Tissue Collection
- 000006585 KU Fish Tissue Collection
- 000006580 KU Fish Tissue Collection
- 000006584 KU Fish Tissue Collection

Delete Clone Carry Forward Add Save



Statistics

Last refreshed 2 seconds ago Refresh Download by TSV Edit

Collection	Preparations	Loan
Collection Objects 41,018	EOW 41,018	Items on Loan 1,175
Preparations 44,970	EAS 1,967 / 8,668	Open Loans 58
Type Specimens 335	Shel 912 / 1,496	Overdue Loans 58
	R-Ray 882 / 882	

Locality / Georeference	Type Specimens	Digitization
Locality 9,381	Herbarium 16	Digitized Last 7 Days 0
Georeference Codes 1,889	Paratype 189	Digitized Last Month 55
Georeferenced Localities 8,803	Neotype 1	Digitized Last Year 216
Recent Georeferenced 95.28%	Topotype 1	
	Paratopotype 1	

Other Statistics	Specimens	Geography
Collection Objects with attachments 2,253	Classen 6	Continents 11
Collection Objects with images 2,253	Orders 40	Countries 80
Parent Imaged 5.03%	Families 360	States/Provinces 315
Collecting Events with attachments 3,055	Genera 1,031	Countries 979
	Species 2,651	

OTIS (Outgoing/Threatened)	Classen
OTIS Appendix I 5	Collection Objects cited 5,548
OTIS Appendix II 16	Total citations 840
Endangered 520	
Threatened 1,285	

Specify 7

KU BIODIVERSITY INSTITUTE

Shipping Funding: Ichthyology (2454238 - 898) Date: 2023/07/27
Shipping Address: Packaged by: Bentley, Andrew
Hogue, Gabriela
NC Museum of Natural Sciences, 11 West Jones
Street
Raleigh, NC 27601
USA

Weight: _____
Amount of Postage: \$ _____
Contents: _____ Preserved fish specimens
Value: \$ _____
Check One: _____ No Ethanol
_____ X _____ Under 30ml/500ml total
_____ Over 30 ml/500ml total

Phone No.: (919) 707-8868
Account No.: _____
Insurance Amount: \$ _____
Dimensions: _____ L x _____ W x _____ H

US POSTAL SERVICE DOMESTIC

1st Class _____
Express _____
Priority _____
Parcel Post _____
Media Mail/Book Rate _____

INTERNATIONAL

1st Class _____
Global Priority _____
Air Mail Parcel _____

OPTIONS

Registered _____
Certified _____
Return Receipt _____

UPS: Domestic and International

Ground _____
3 Day Select _____
2nd Day Air _____
Next Day Air (3:00pm delivery) _____
Next Day Air (10:30am delivery) _____
International _____

FedEx: Domestic and International

Priority overnight (delivery by 10:30am) _____
Standard overnight (delivery by 8:00am) _____
1st Overnight _____
2-Day _____
Express Saver (if available) _____
International Priority _____
International Economy _____

Specify 7

Query: Collection Object Sample Query

Basic View Hide Field Mapper Save Query Save All

Data Entry
Trees
Interactions
Queries
Record Sets
Reports
WorkBench
Attachments
Statistics

Collection Object

- Prev Inst
- Previous #
- Remarks
- Cataloger
- Col Obj Attribute
- Collecting Information
- Collection Object Attachments
- Determinations

Determination

- (aggregated)
- Addendum
- Current
- Guid
- Name Usage
- Qualifier
- Remarks
- Type Status

Taxon

- (any rank)
- Kingdom
- Division
- Class
- Order
- Family
- Genus
- Species

Taxon

- Author
- Full Name

Query Builder

- Catalog Number
- Cataloged Date
- Remarks
- Prev Inst
- Determinations
- Determinations
- Determinations
- Determinations
- Determinations

GeoMap - Plotted 1,904 records

Details

Map showing the distribution of records in the Caribbean region, with a focus on the island of Cuba. The map includes a legend and a scale bar.

UC 305

Annonaceae

Uvariastrium neglectum

Africa, Angola, Gabon, Macombe, Dingo

Collector # 115 Date: 10/26/1967
Cat
Sheet - 1

University of Canada Biodiversity Institute - Botany

Specify 7

The image displays the Specify 7 software interface, which is used for managing botanical collections. It features a main data entry window and a detailed determination workflow window.

Main Data Entry Window:

- Left Panel:** Contains navigation icons for Data Entry, Trees, Interactions, Queries, Record Sets, Reports, Workbench (highlighted), Attachments, and Statistics.
- Table:** A spreadsheet with columns for Locality Name, Country, State, and other attributes. It lists various botanical specimens with their collection details.
- Bottom Panel:** Includes a search bar, a replace button, and status indicators for Search Results (0/0), Modified Cells (0/0), New Cells (0/0), and Error Cells (0/0).

Determination Workflow Window:

- Top Panel:** Shows the 'Botany Demo Spreadsheet (Collection Object)' with tabs for Basic Table, Clear Mappings, and AutoMap.
- Left Panel:** A tree view showing the hierarchy of collection objects, including Catalog Number, Cataloged Date, SUID, Provenance, Previous #, Remarks, Cataloger, and Col Obj Attribute.
- Main Panel:** A detailed determination workflow for 'Det 1 - Family' and 'Det 2 - Species'. It includes dropdown menus for 'Determinations', 'Taxon', 'Family', 'Genus', 'Species', 'First Name', and 'Last Name'.
- Bottom Panel:** A table showing the results of the determination process, including 'Locality Name', 'Collecting Information', 'Locality', 'Geography', 'Country', and 'State'.

Before You Begin

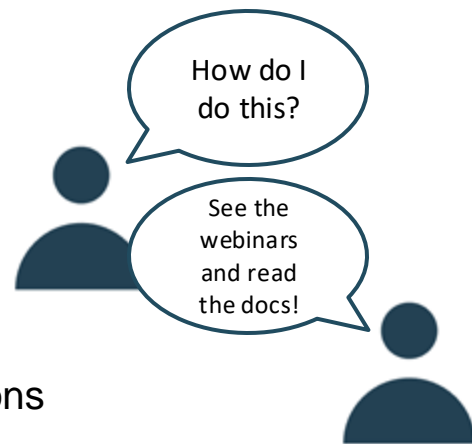
- What do I want in a Collections Management System (CMS)?
 - ✓ Highly customizable data entry forms and bulk data entry via spreadsheets
 - ✓ Transaction management (loans, gifts, accessions, deaccessions, borrows, etc.)
 - ✓ Long-term consistent support, active development, cloud hosted vs. self-hosted
 - ✓ Extensive query system, data visualization
 - ✓ Real-time collection-specific statistics
 - ✓ Public web portal to display my data
 - ✓ Built-in tool for label creation and comprehensive report generation
 - ✓ Ability to publish to GBIF, iDigBio, Symbiota, etc. with a single click

Before You Begin

- Navigating Transitions to Specify
 - The SCC has gathered insights and learned lessons from working with institutions that have successfully transitioned to Specify.
- Sharing Expertise
 - The SCC aims to share its experience to help other large collections anticipate and address key data-related issues when implementing Specify or any collections management system.
 - Whenever possible, collections with a particular focus are put in touch with other collections that have similarly structured data.

Human Factors in Transitioning

- Technical Expertise
 - Transitioning to a new collections management system requires technical expertise to ensure a successful implementation.
- Communication and Collaboration
 - Effective communication and collaboration among collections staff, volunteers, and management are crucial for a smooth transition.
 - Personalized Specify webinars are held with collections staff to hear questions, requests, and opinions.



Human Factors in Transitioning

- Change Management

- Addressing the human factors involved in transitioning to a new system is key to the overall success of the implementation.
- Buy-in from collection staff is essential because it helps ensure that they are engaged and supportive of the change.



Feedback from Zsuzsanna Papp

Natural History Museum of Denmark (NHMD)

DANISH
NATURAL HISTORY
MUSEUMS

- **Management arguments for a specialized CMS:**

- Single system to be supported
- Central repository of all data
- Relational structuring of data
- Ready features for working with data
- Easy sharing of data

- **Advantages of Specify over alternatives:**

- Excellent helpdesk
- Active user community
- Low cost
- Widely used
- Open source

Standardizing before implementation is well worth the effort. Communication about fears and expectations, as well as accommodating individual needs, is essential to minimizing the initial bumps.

Preparing Data for Specify

- **Exporting Data**

- Bringing data out of an existing collections management system (CMS)

- **Standardizing Data**

- Standardizing names of collectors, determiners, catalogers, preparators, etc.
- Cleaning Taxonomy data linked to determinations

- **Normalizing Data**

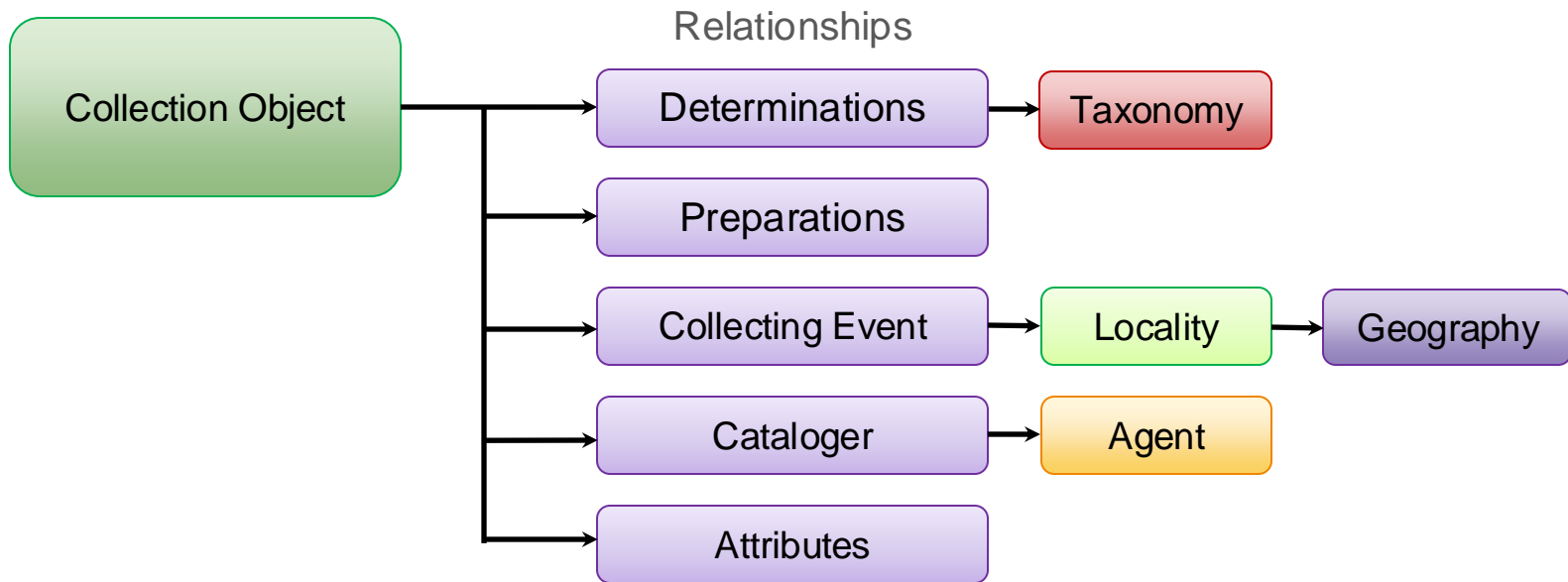
- Parsing agent names into individual columns (e.g., first name, last name, etc.)
- Separating location information into 'Collecting Events' and 'Localities'
- Separating taxon full names into individual columns (e.g., genus, species, etc.)

- **Error Checking**

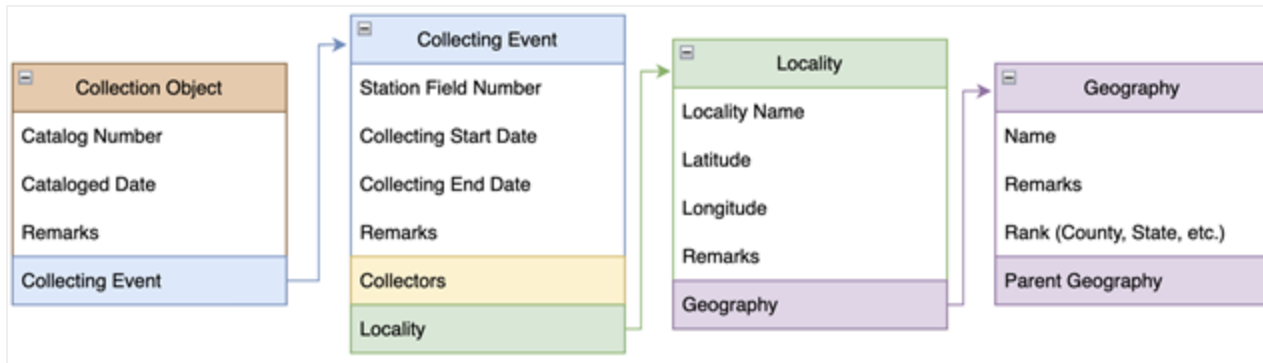
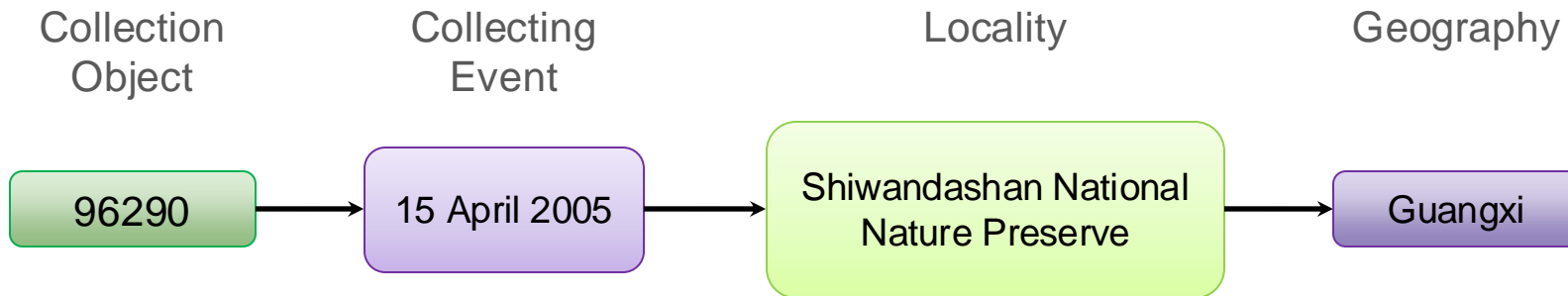
- Verifying the accuracy and completeness of the existing data.

This takes place at every step in the process!

Relational Database Structure

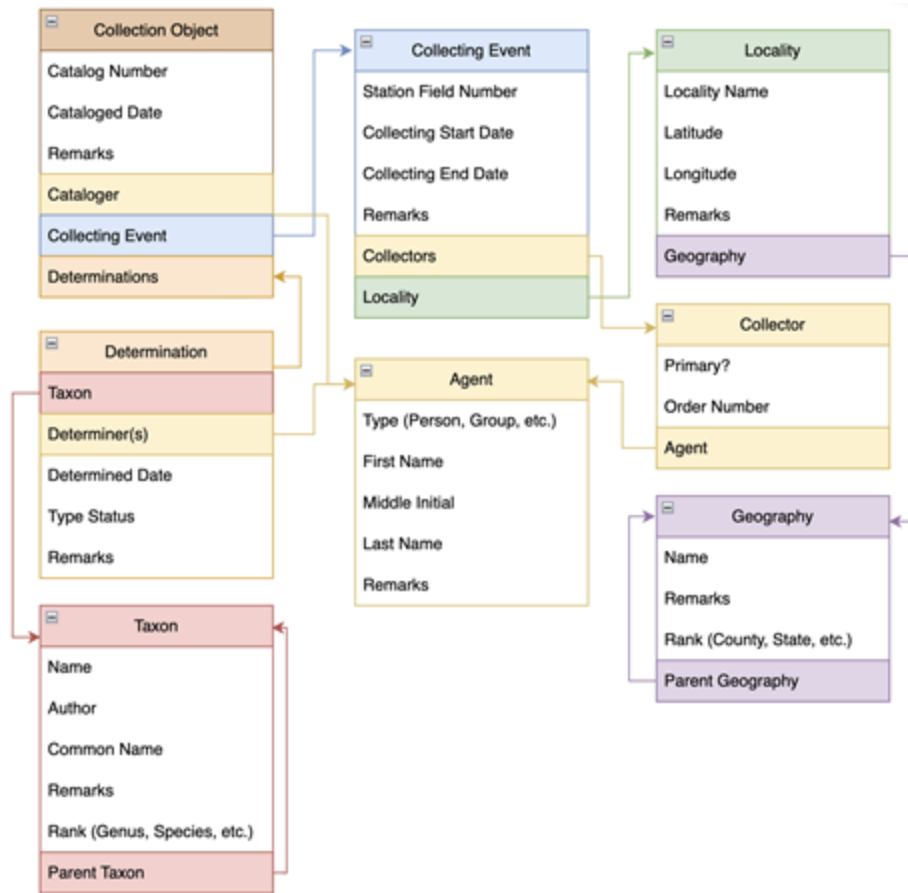


Making a Collection Object



Preparing Data for Specify

Relationships



Normalizing Data

- When we talk about ‘normalizing’ data, we refer to the process of making records shareable and consistent between records.
 - This involves organizing the data to reduce redundancy and improve data integrity, ensuring that it can be easily accessed, analyzed, and compared.

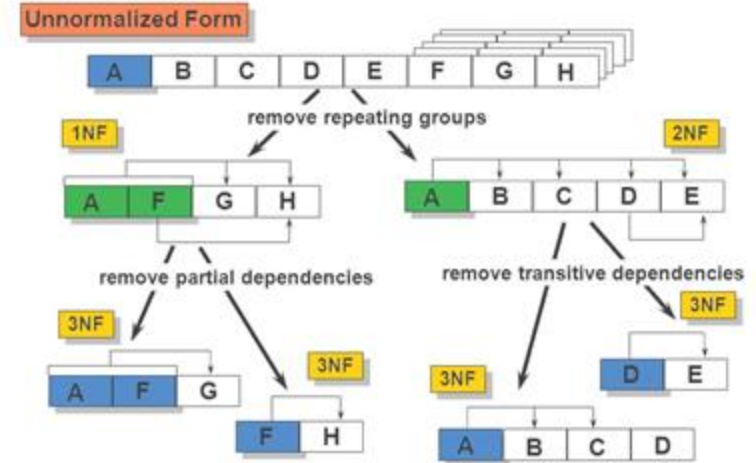
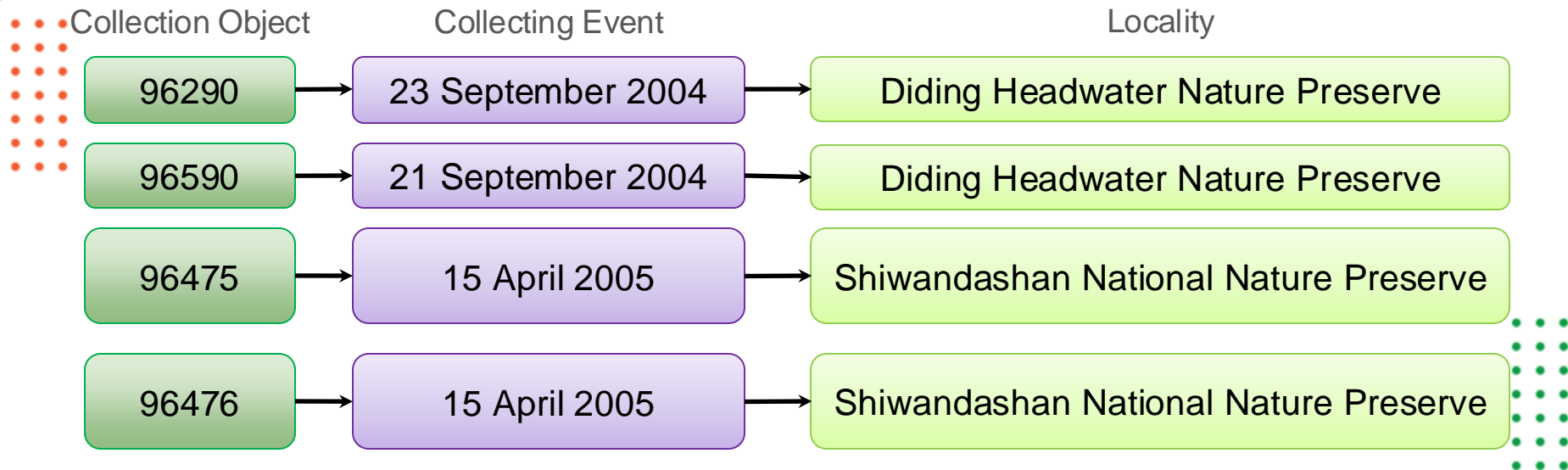


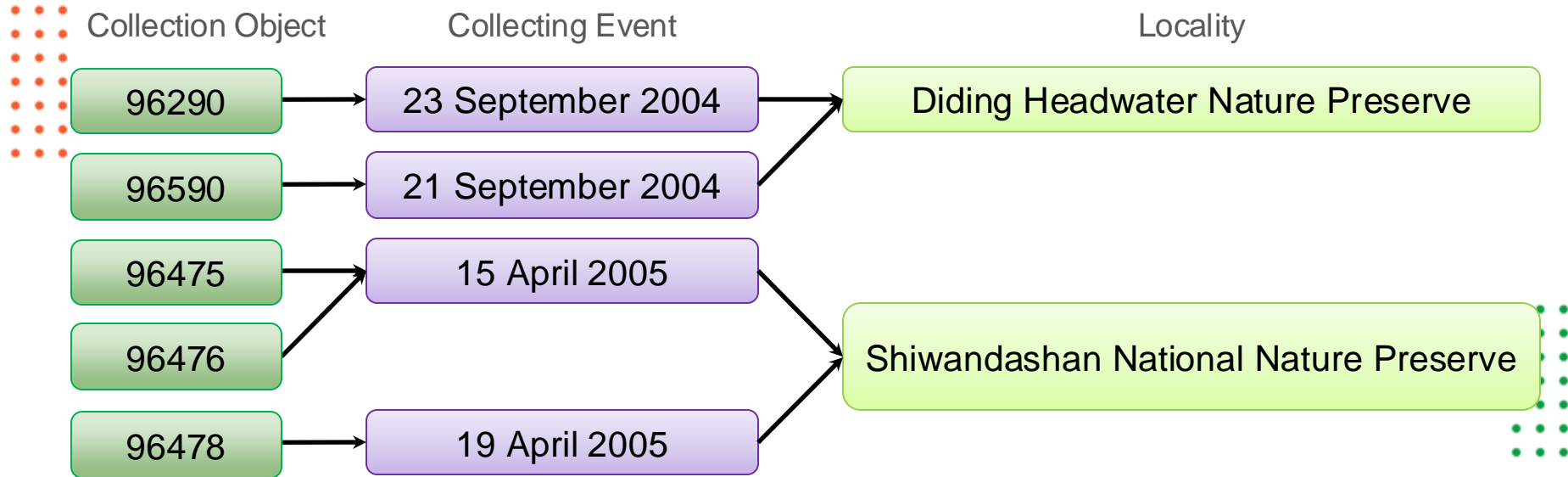
Figure 1: Data Mining Normalization. (n.d.). Galaktikasoftware. <https://galaktikasoftware.com/blog/data-mining-normalization.html>

De-normalized Data



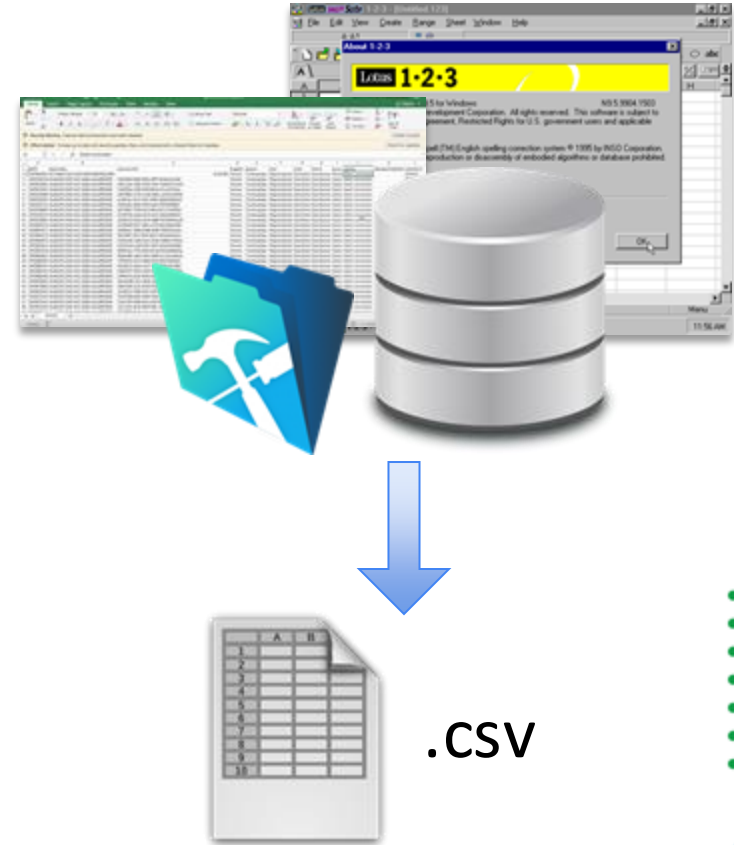
Normalized Data

Data is shared through **relationships**, linked behind-the-scenes in the database



Exporting Data

- Data is often captured in one or more other systems before being ingested to Specify.
 - In most cases, data must be accessible in spreadsheet (CSV, Excel, etc.) form
 - The mechanism for exporting depends on the source of the data
- Once exported, data is frequently disorganized or lacking standardization.



Standardizing & Cleaning Data

- Collections staff can use software like OpenRefine, Excel, Python, or other tools to clean and organize data.



Preparing Data for Specify

Standardizing & Cleaning Data (OpenRefine)



Transform Text

Transform menu options:

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- Replace smart quotes with ASCII
- To titlecase
- To uppercase
- To lowercase
- To number
- To date

Resulting data table:

Collectors - Last Name 1 1 1	Collectors - Last Name 1 1 2	Collectors - Last Name 1 1 3
SPRAGUE		
SPRAGUE		
TEIMER	O	
FINDLEY	J	M
FINDLEY	J	M
FINDLEY	J	W
FINDLEY	J	S
FINDLEY	J	S

Find Misspellings

OpenRefine: Collecting Event Cleaning for Specify

Facet / Filter: Collectors - Last Name

10 choices Sort by: name count

Collecting Information - Remarks	Locality	Collectors - First
MAYFIELD GOLF COURSE		
MAYFIELD GOLF COURSE		
MAYFIELD GOLF COURSE		
MAYFIELD GOLF COURSE		
MAYFIELD GOLF COURSE		
MAYFIELD GOLF COURSE		
MAYFIELD GOLF COURSE		
MAYFIELD GOLF COURSE		
MAYFIELD GOLF COURSE		
MAYFIELD GOLF COURSE		

Parse Names

Geography - Full Name

Nicaragua, Chinandega
Nicaragua, Chinandega
Nicaragua, Esteli
Nicaragua, Esteli
Nicaragua, Esteli
Nicaragua, Esteli
Nicaragua, Esteli
Nicaragua, Esteli
Nicaragua, Esteli
Nicaragua, Esteli
Nicaragua, Matagalpa

Geography - Full Name 1

Nicaragua
Nicaragua
Nicaragua
Nicaragua
Nicaragua
Nicaragua
Nicaragua
Nicaragua
Nicaragua
Nicaragua
Nicaragua

Geography - Full Name 2

Chinandega
Chinandega
Esteli
Esteli
Esteli
Esteli
Esteli
Esteli
Esteli
Esteli
Matagalpa

Preparing Data for Specify

Standardizing & Cleaning Data (OpenRefine)



Standardize Dates

The screenshot shows the OpenRefine interface with a table titled "Collecting Information - CollectionDate". A dropdown menu is open, displaying a list of dates: 01/01/1915, 01/01/1915, 01/01/1932, 01/01/1937, 01/01/1937, and 01/01/1939. A blue arrow points from the table to the dropdown menu.

Find and Resolve Duplicates

The screenshot shows the OpenRefine interface with a table titled "Cluster and edit column 'Last Name'". The table displays three clusters of names: Richards, Atkins, and Al Issai. The "Method" is set to "Nearest neighbor", "Distance function" is "PPM", "Radius" is "1.0", and "Block chars" is "6". The "Auto-update" checkbox is checked. The "3 clusters found" status is displayed. The table has columns for "Cluster size", "Row count", "Values in cluster", "Merge?", and "New cell value".

Cluster size	Row count	Values in cluster	Merge?	New cell value
2	2	Richards Richardson	<input type="checkbox"/>	Richards
2	3	Atkins (2 rows) Atkinson	<input type="checkbox"/>	Atkins
2	3	Al Issai (2 rows) Al Issay	<input type="checkbox"/>	Al Issai

Additional settings on the right:

- # Rows in cluster: 2 — 3
- Average length of choices: 7 — 9
- Length variance of choices: 0 — 1

Preparing Data for Specify

Institutional Hierarchy

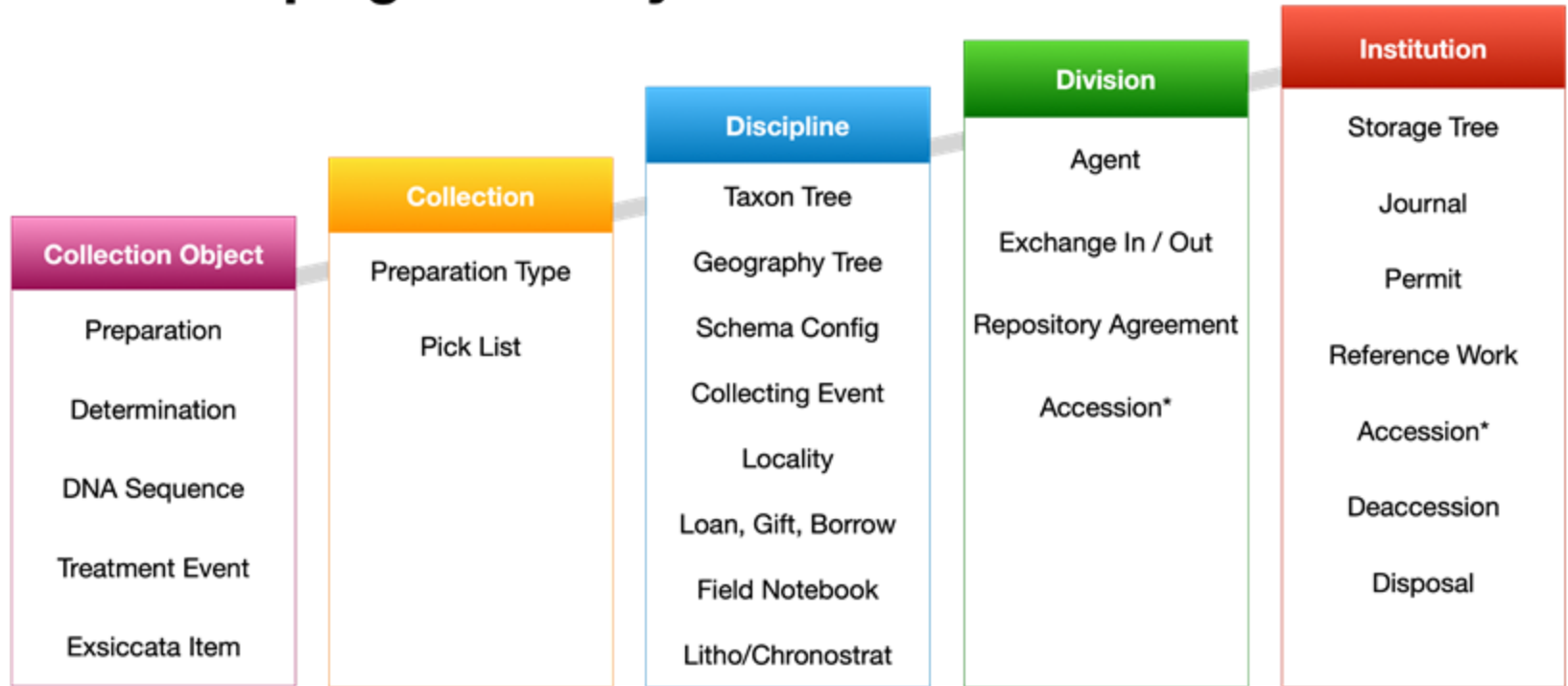
Institution	Natural History Museum							
Division	Vertebrate Zoology				Botany			
Discipline	Ichthyology		Herpetology		Non-Vascular Plants		Vascular Plants	
Collection	Wet	Dry	Amphibians	Reptiles	Mosses	Lichens	Herbarium	Pollen

The decisions made regarding data sharing and scoping can have a significant impact on the success of a Specify implementation.

Preparing Data for Specify

Table Scoping Hierarchy

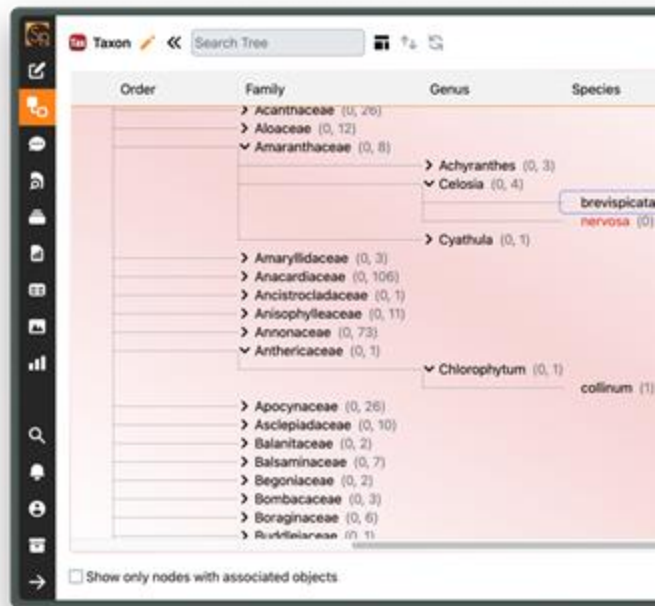
**Accession can be either at the Institution or Division level*



Visibility

Determining Data Structure

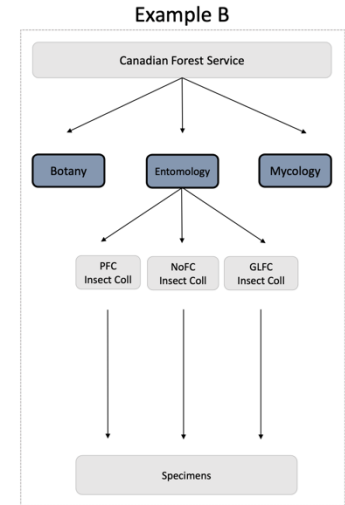
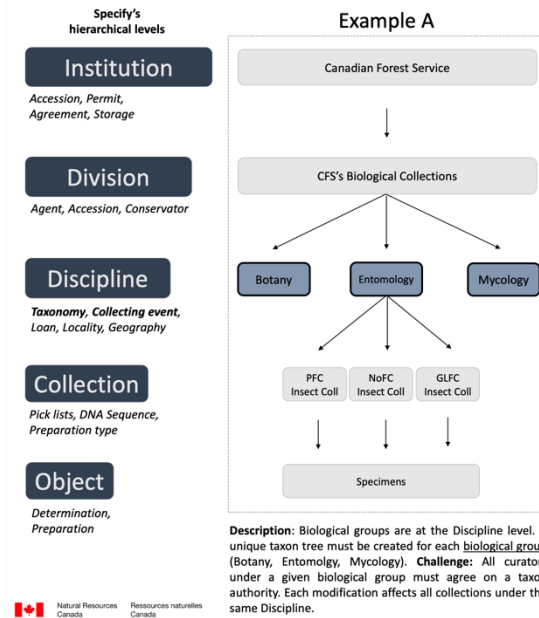
- Centralized vs. Decentralized Approach
 - Institutions must decide what collection data should be shared between collections versus what should be managed separately.
 - Is it desirable to share:
 - Taxonomy
 - Occurrence data
 - Collecting Events and Localities
 - Agents (People, Groups, or Organizations)
 - Data entry forms & specimen labels



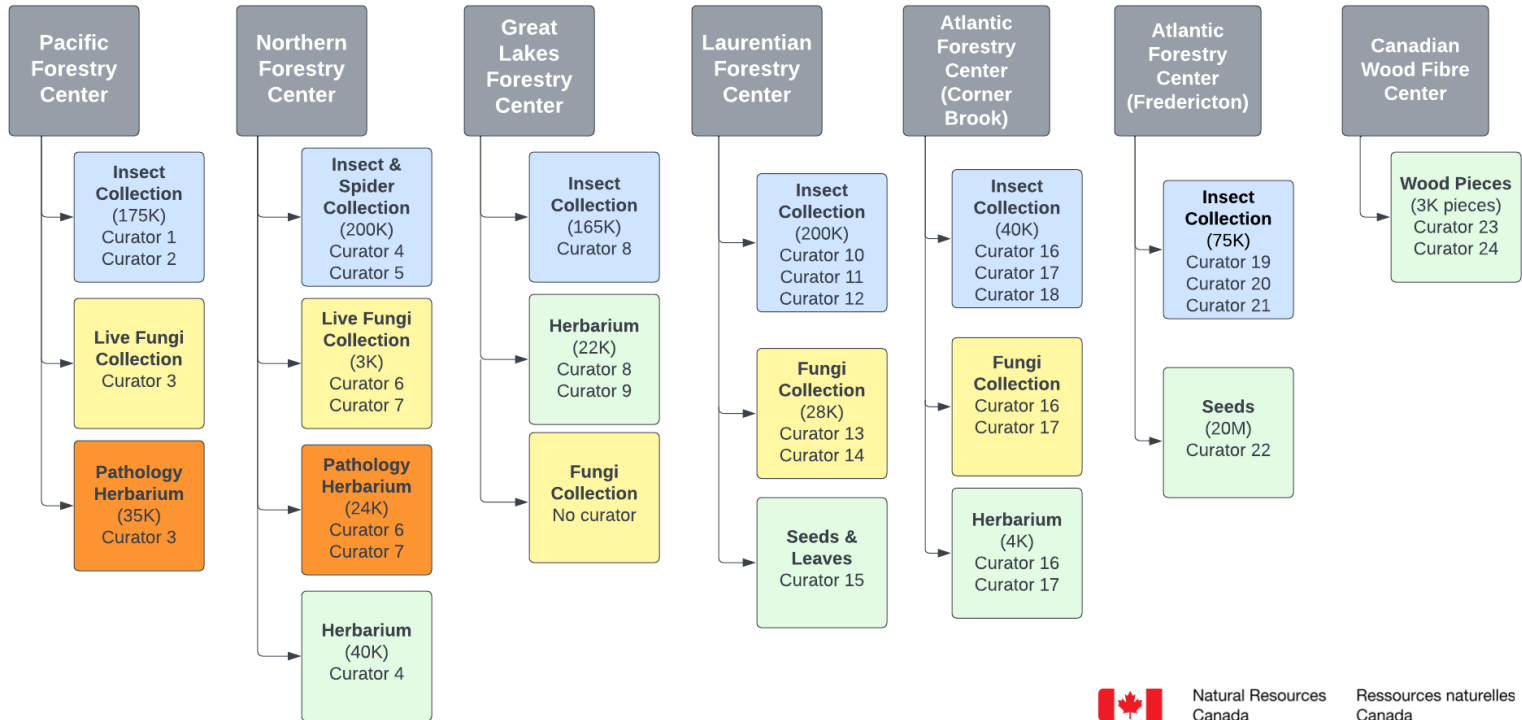
Sharing Data Across Physical Locations

Natural Resources Canada – Canadian Forestry Service

- Navigating the Balance
 - If you have 3 physical centers, do all entomology collection managers manage taxonomy together?
 - Should all managers have access to each others collections?



Canadian Forest Service

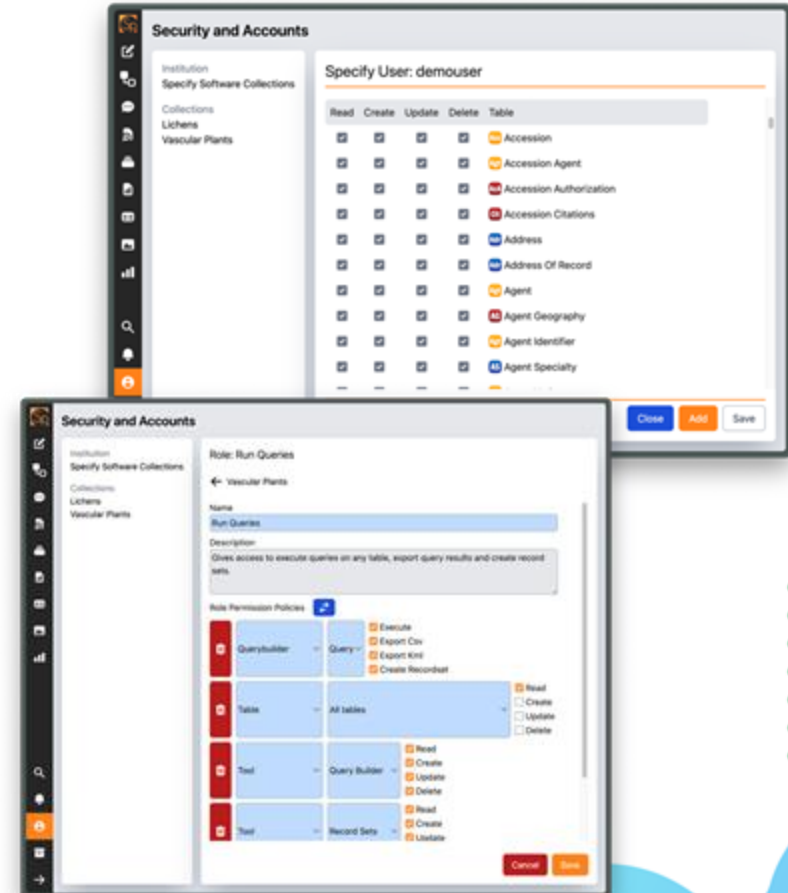


Natural Resources
Canada

Ressources naturelles
Canada

Security & Accounts

- User Access
 - Separate users into 'groups' (guest, student, manager, administrator)?
 - Give specific permissions to individual users
- Data Confidentiality
 - Enable read-only access by the public? Specific records?



Preparing for Data Publishing

- Data Publishing
 - Specify 7 can function as an Integrated Publishing Toolkit (IPT) to publish collections data to GBIF and other data aggregators.
 - RSS feed
 - ZIP archive
 - Collections work with our team and others to map to the appropriate terms.
 - GBIF extensions are supported and exports can be one or many queries.



Conclusion

- **Clarify Collection Requirements:** Gain a deep understanding of your collection's needs and objectives.
- **Seek Assistance:** Our team is dedicated to guiding you through this transition.
- **Address Human Factors:** Factor in human elements like technical proficiency, communication, and teamwork during the transition process.
- **Thorough Data Preparation:** Ensure meticulous planning and preparation for data import, including standardization, normalization, and error validation.
- **Consider Data Visualization and Sharing:** Evaluate data visualization needs, public web portal requirements, GBIF publishing, and data sharing implications carefully.

Special Thanks

Project Support: All Specify Collections Consortium Member Institutions, SCC Staff

Previously supported by several US NSF grants, 1986-2017

Founding Partners



Special Thanks

Full Members

Cornell University Museum of Vertebrates	Queensland Dept. of Environment and Science	Western Australian Herbarium
Natural Science Collections Facility, SANBI	Paleontological Research Institute	Laurentian Forestry Centre
Santa Barbara Museum of Natural History	Royal Botanic Garden Edinburgh	
Swedish Museum of Natural History	Royal Botanic Gardens Victoria	

Solution and Associate Members

Agriculture and Agri-Food Canada	Earlham College	New Brunswick Museum	University of Colorado
Ateneo de Naga University, Philippines	Emory University	New Mexico State Univ. Herbarium	University of Iowa
Auburn University	Estación Biológica de Doñana	National Institute of Water and Atmospheric Research	University of Massachusetts
Bailey-Matthews National Shell Museum	Father Saturnino Urios University, Philippines	North Carolina Museum of Natural Sciences	University of Minnesota
Bernice Pauahi Bishop Museum	Florida Fish and Wildlife Conservation Commission	North Carolina State University	University of Nebraska Herbarium
Brigham Young University	Fisheries and Oceans Canada	Ohio State University Orton Geological Museum	University of New South Wales
Brown University	Gothenburg Natural History Museum	Ohio State University Museum of Biological Diversity	University of Otago
California Academy of Sciences	Hebrew University of Jerusalem	Oranm Academic College of Education	University of Oregon
Calvert Marine Museum	Illinois State University	Oregon State University	University of Rochester
Canadian Food Inspection Agency	Indiana University	Pennsylvania State University	University of Texas at Austin
City of Orange County, Parks and Recreation	Institut de Recherche pour le Développement	Pioneer Trails Regional Museum	University of Washington
Cleveland Museum of Natural History	Kent State University	Raymond M. Alf Museum of Paleontology	University of Wyoming
College of Idaho	Lauer Foundation	Royal Botanic Garden Madrid	University of Zurich
College of William and Mary	Louisiana State University	San Diego Natural History Museum	Virginia Institute of Marine Science
Cornell University Entomology	Michigan State University	South Dakota School of Mines and Technology	Wesleyan University
Cornell University Plant Pathology	Montreal Insectarium	Southeastern Louisiana University	West Virginia University
Cranbrook Institute of Science	Museum of the Rockies	Unitec Institute of Technology	Yugra State University
Delaware Museum of Natural History	Museu de Ciències Naturals de Barcelona		
Duke University Herbarium	National Museum of Costa Rica		
Duke University Lemur Center	Natural History Museum Basel		