# Digitization Planning for Large Collections with Specify 7

Grant Fitzsimmons & Theresa Miller

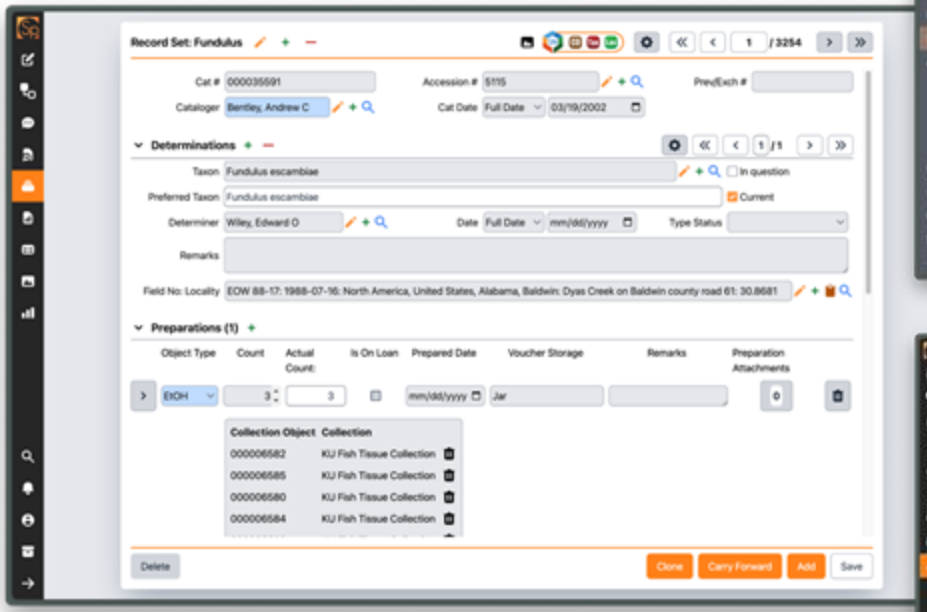
SPNHC TDWG 2024 沖縄 JAPAN

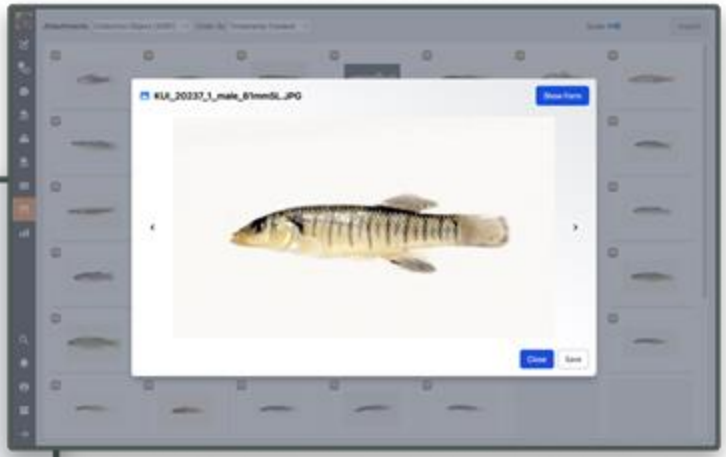
# Introduction

- 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.

KU Biodiversity Institute & Natural History Museum, The University of Kansas
Muséum genève
CSIRO
Florida Museum
CSIC Consejo Superior de Investigaciones Científicas
Danish Natural History Museums
LSA College of Literature, Science, and the Arts, University of Michigan

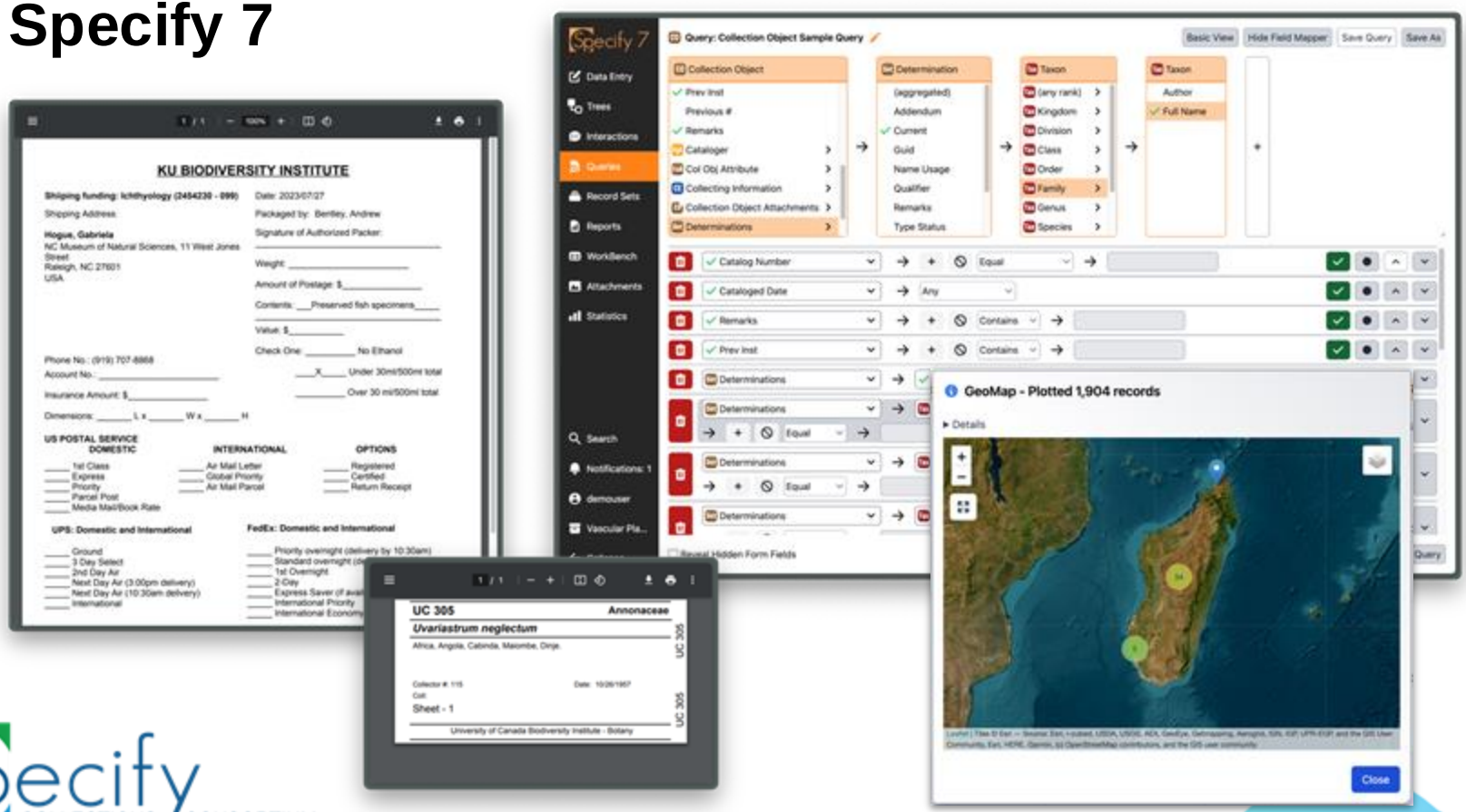
# Specify 7



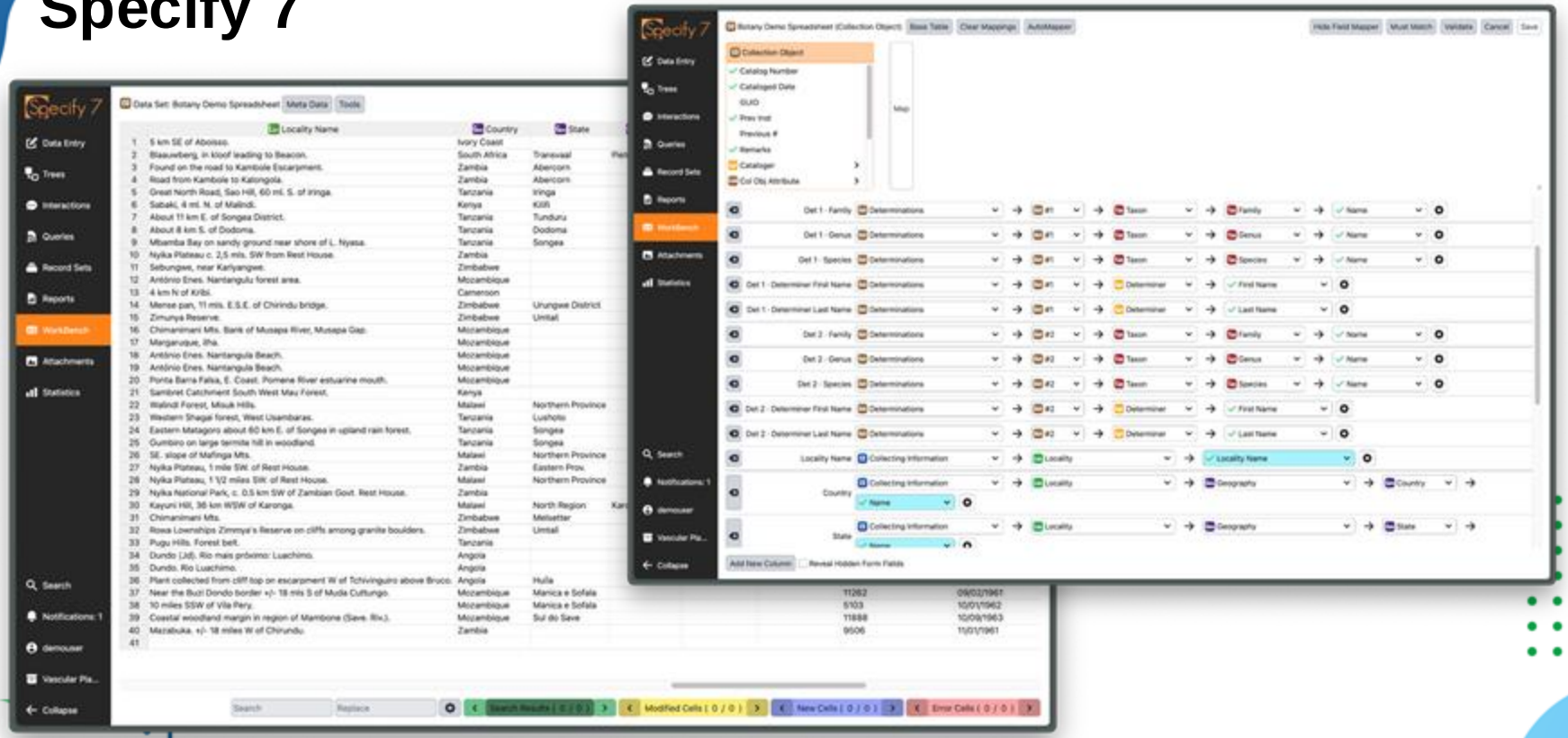

Specify COLLECTIONS CONSORTIUM

# Specify 7




Specify COLLECTIONS CONSORTIUM

# Specify 7




Specify COLLECTIONS CONSORTIUM

# 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

Specify COLLECTIONS CONSORTIUM

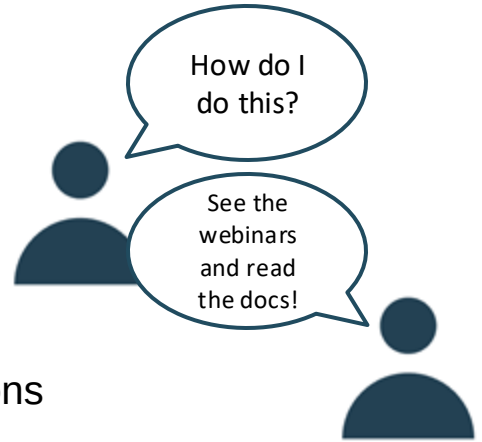
# Before You Begin

- 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.

Specify COLLECTIONS CONSORTIUM

# 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.




Specify COLLECTIONS CONSORTIUM

# 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.


Specify COLLECTIONS CONSORTIUM

# 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.*

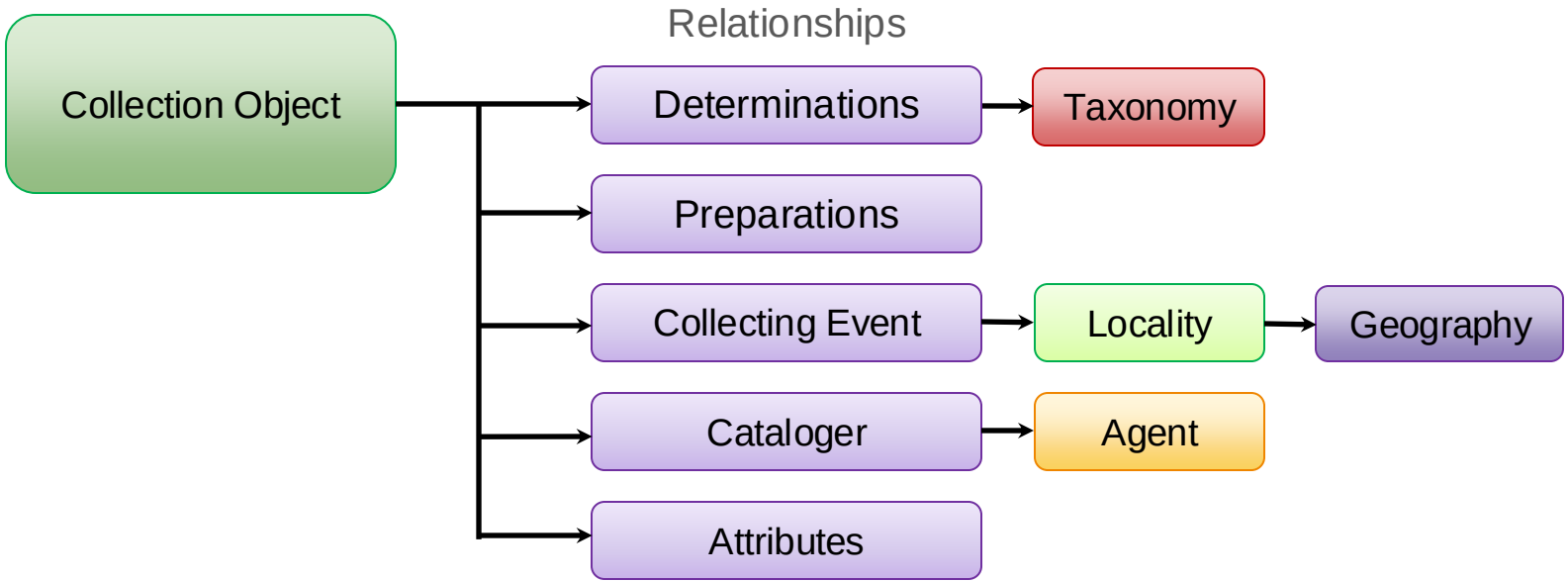
Specify COLLECTIONS CONSORTIUM

# Preparing Data for Specify

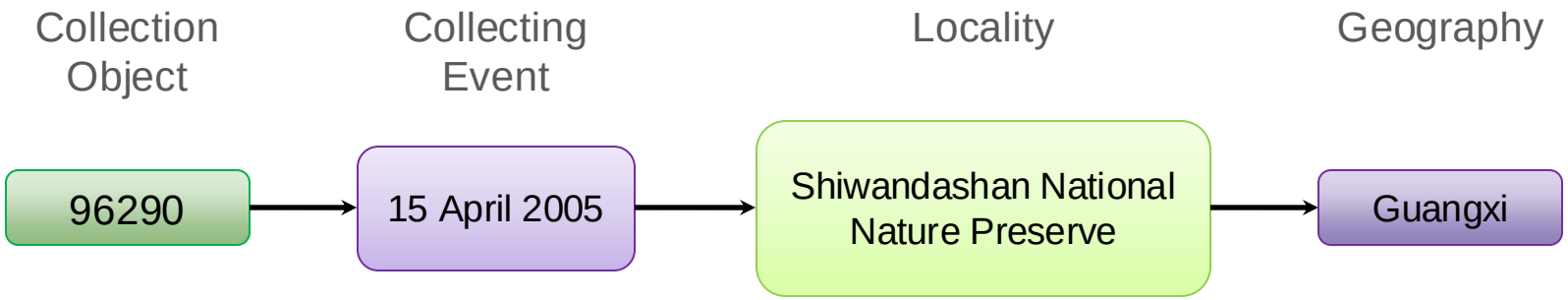
- **Exporting Data**
  - Bringing data out of an existing collections management system (CMS)
- **Standardizing Data**
  - Standardizing names of collectors, determiners, catalogers, preparators, etc.
  - Cleaning Taxonomy data linked to determinations
- **Normalizing Data**
  - Parsing agent names into 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!

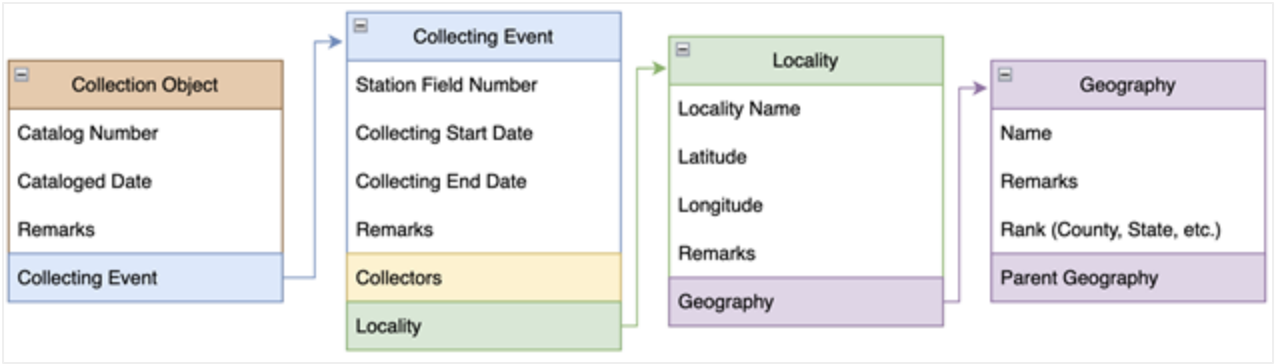
Specify COLLECTIONS CONSORTIUM

# Relational Database Structure



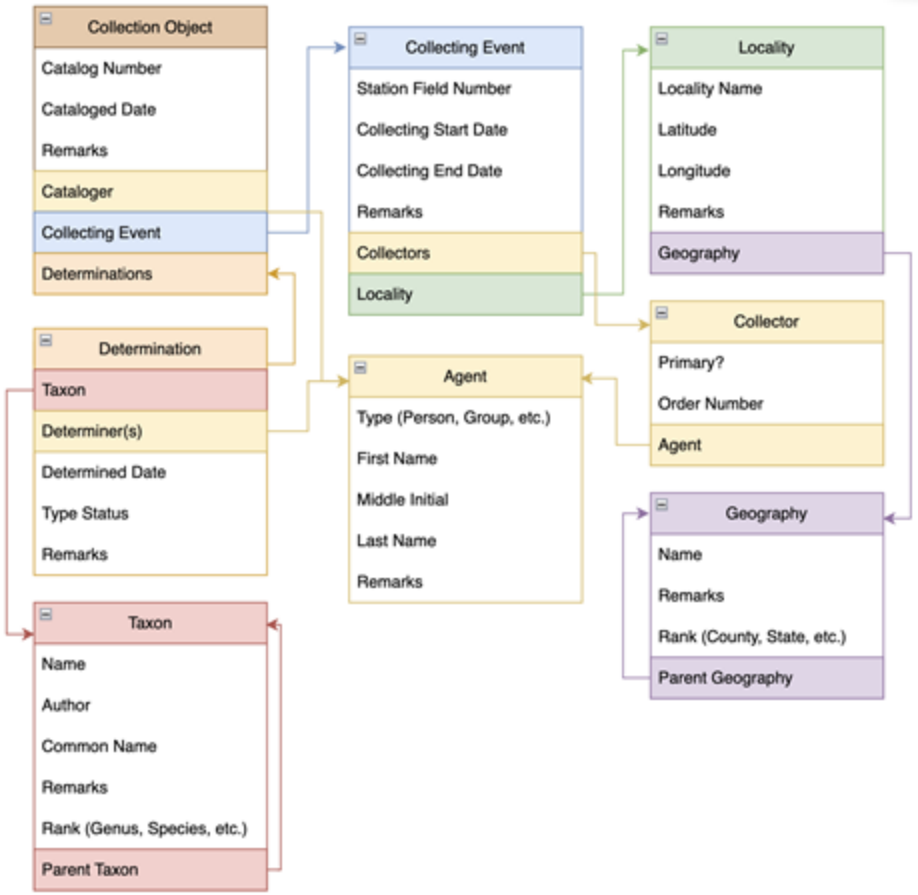
# Making a Collection Object



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.

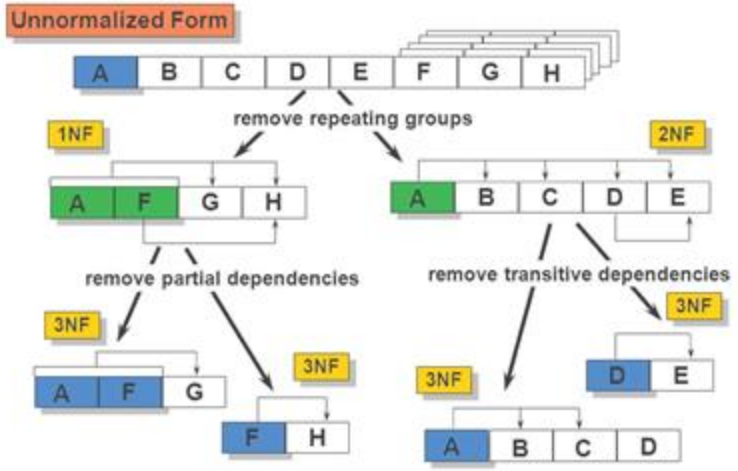
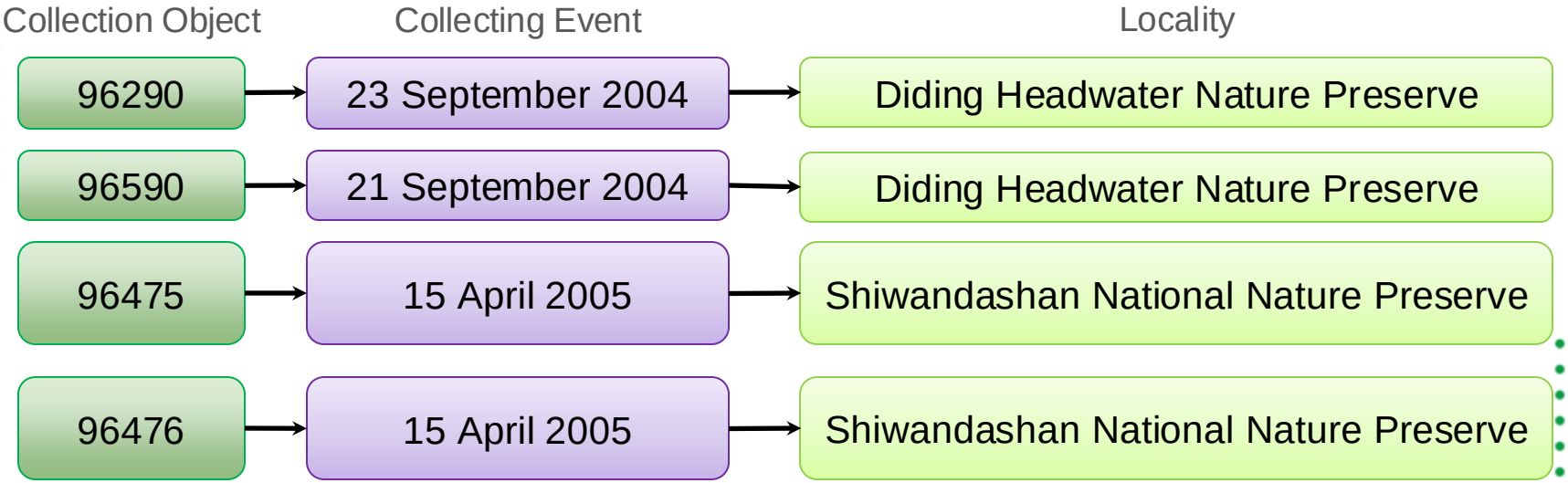



Figure 1: Data Mining Normalization. (n.d.). Galaktikasoft. https://galaktika-soft.com/blog/data-mining-normalization.html

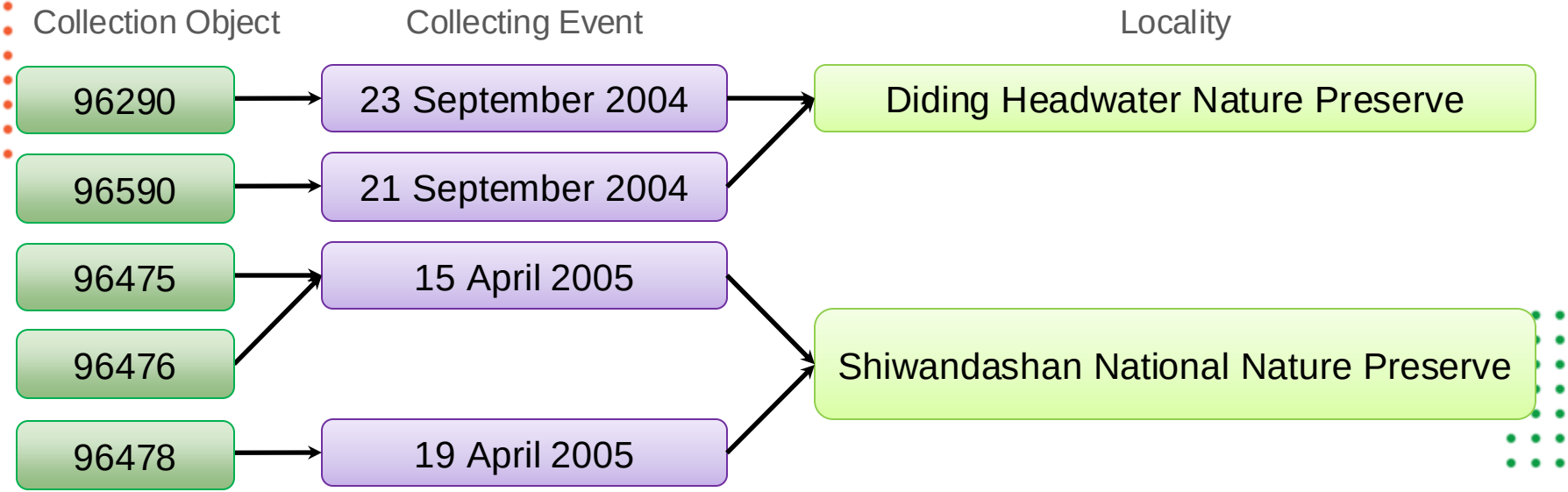
Specify COLLECTIONS CONSORTIUM

# De-normalized Data



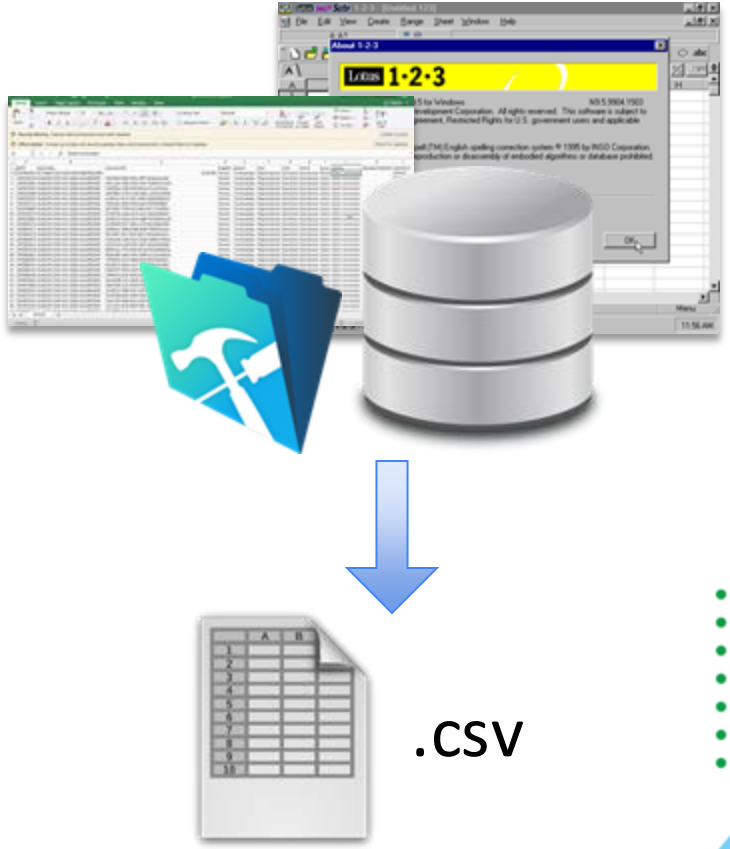
# Normalized Data

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.




Specify COLLECTIONS CONSORTIUM

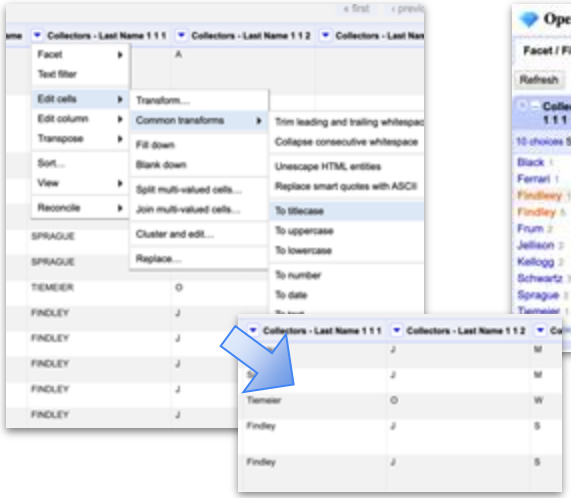
# Standardizing & Cleaning Data

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

Specify COLLECTIONS CONSORTIUM

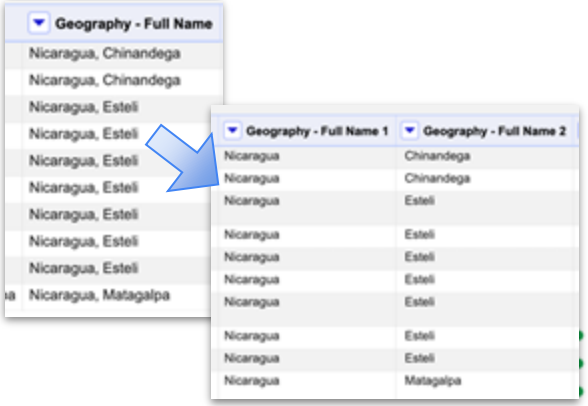
# Standardizing & Cleaning Data (OpenRefine)

## Transform Text




## Find Misspellings

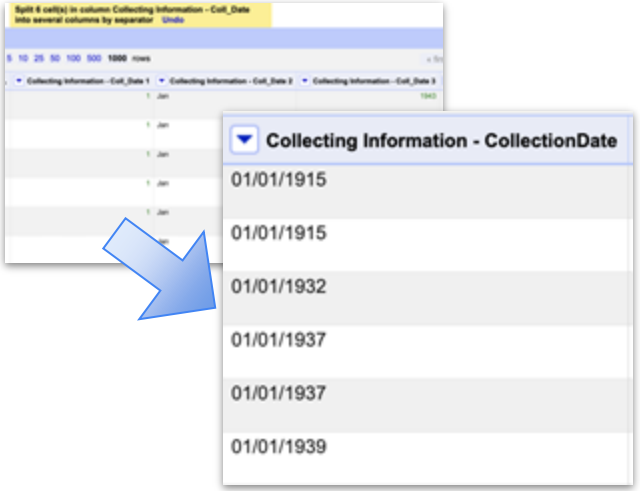

## Parse Names



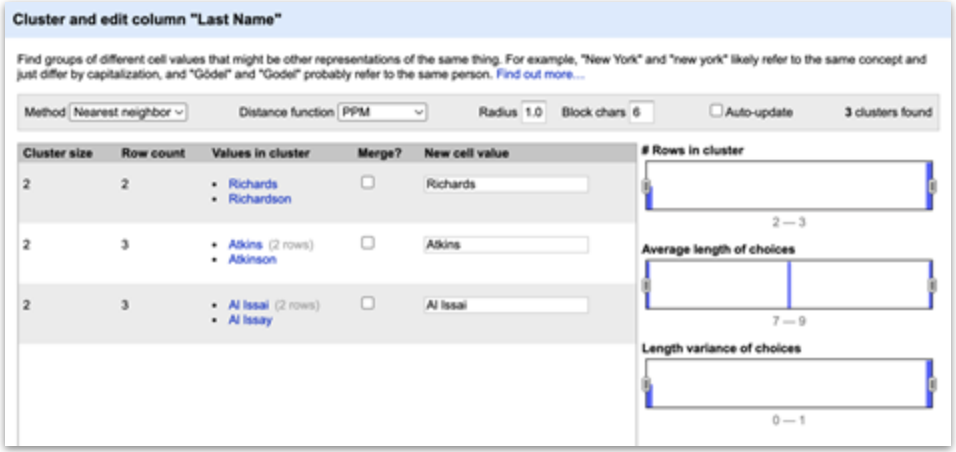

Specify COLLECTIONS CONSORTIUM

# Standardizing & Cleaning Data (OpenRefine)

## Standardize Dates




## Find and Resolve Duplicates




Specify COLLECTIONS CONSORTIUM

# Institutional Hierarchy

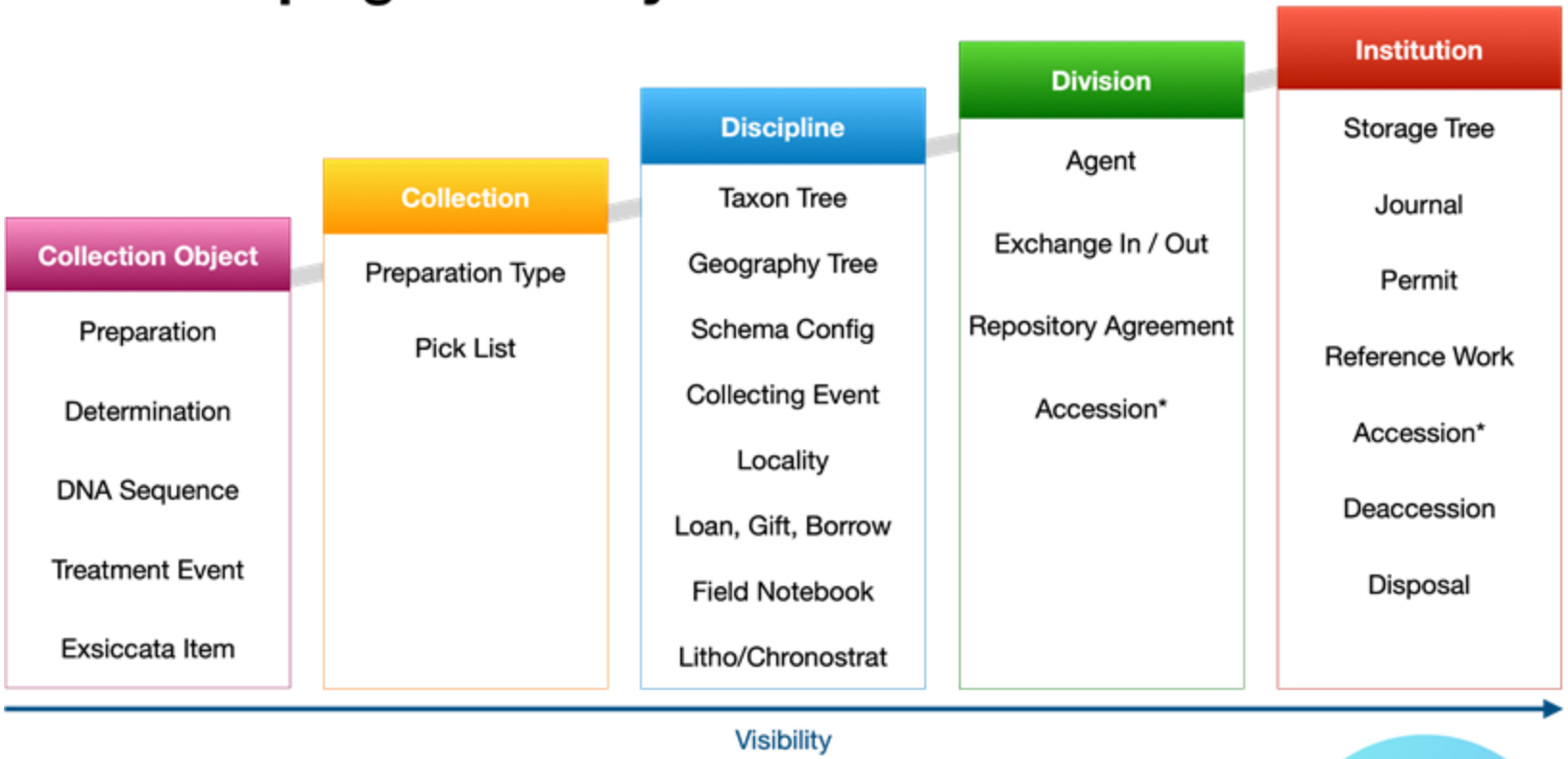
| Institution | Natural History Museum | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Division | Vertebrate Zoology | | | | Botany | | | |
| Discipline | Ichthyology | | Herpetology | | Non-Vascular Plants | | Vascular Plants | |
| Collection | Wet | Dry | Amphibians | Reptiles | Mosses | Lichens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

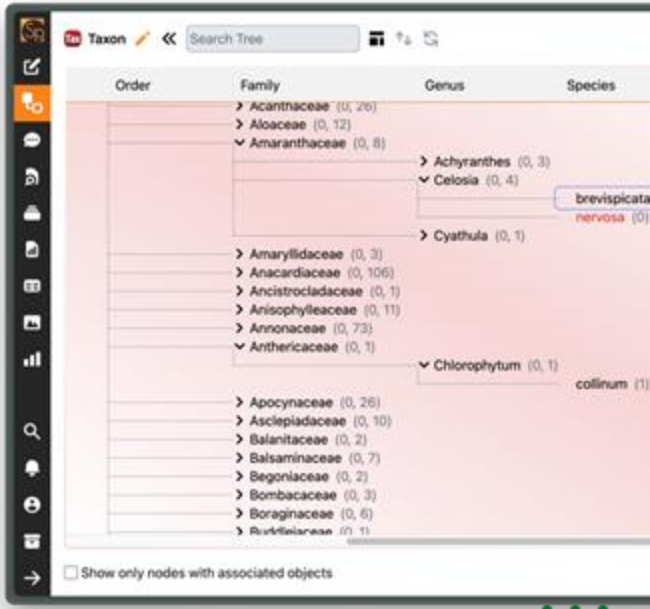
# Table Scoping Hierarchy

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



# 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

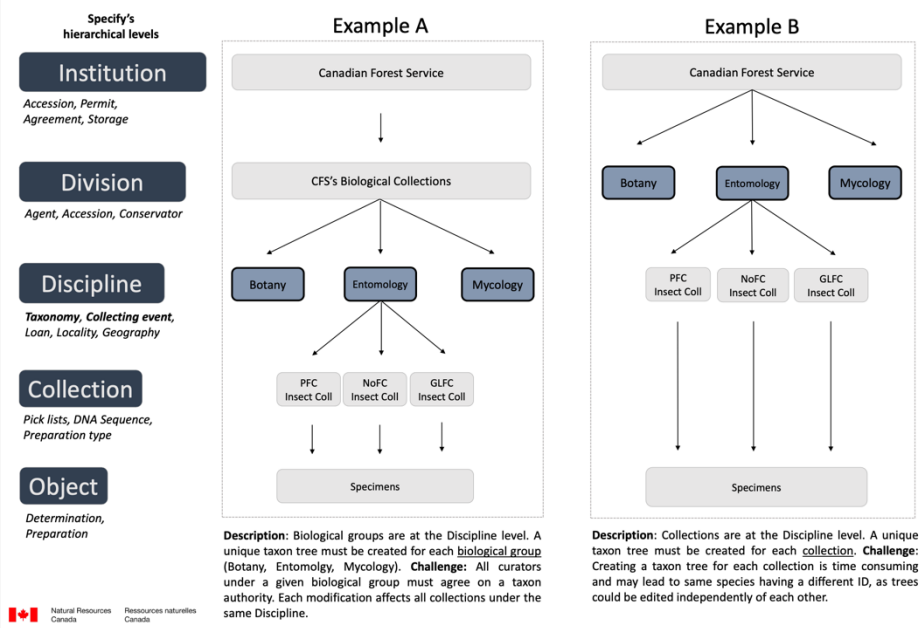



Specify COLLECTIONS CONSORTIUM

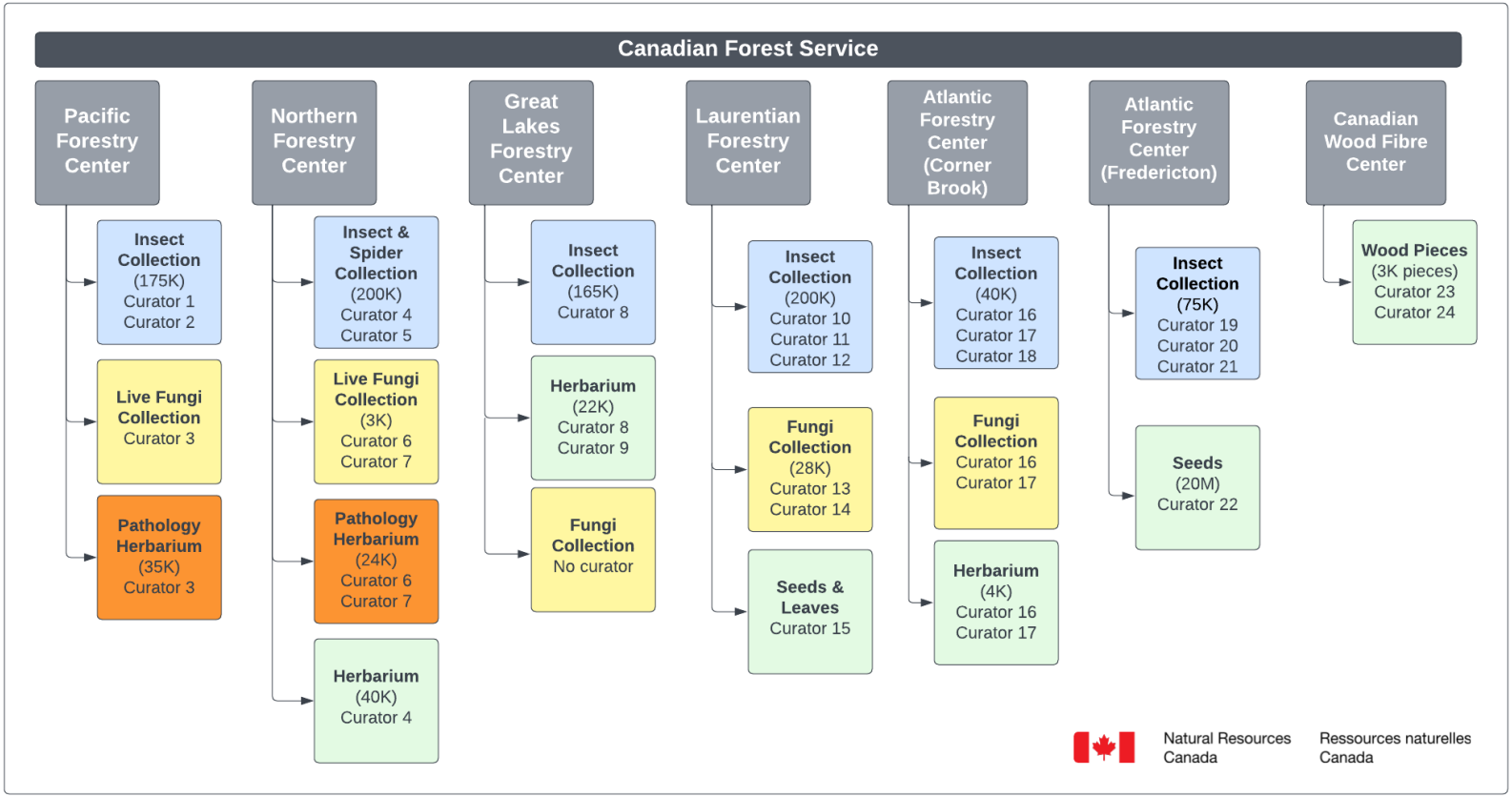
# Sharing Data Across Physical Locations

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?

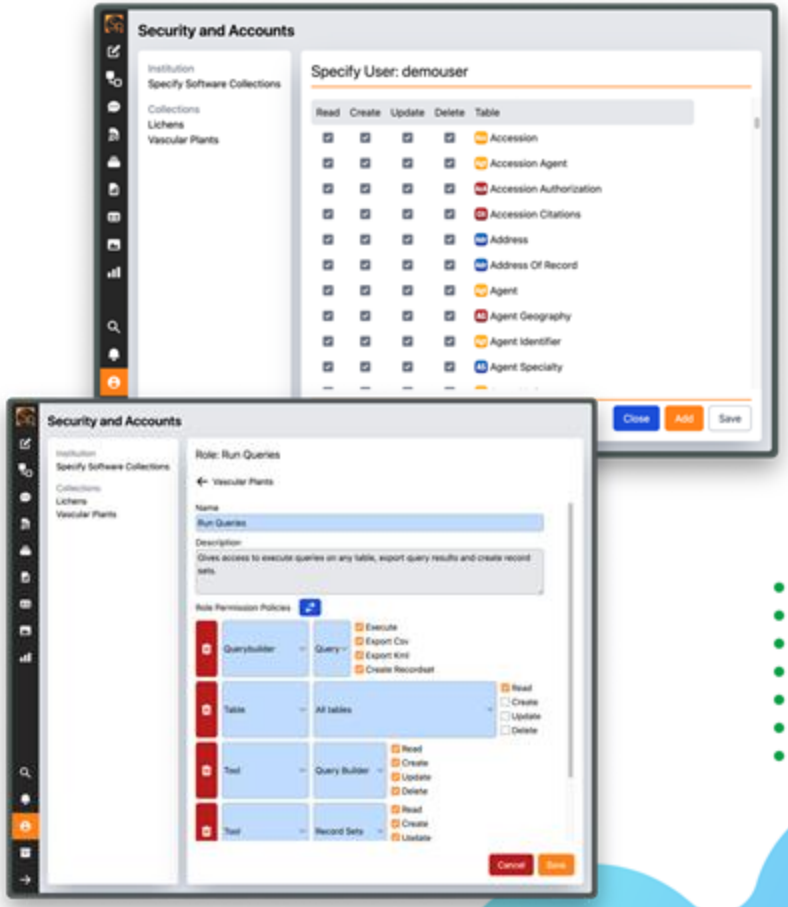



Specify COLLECTIONS CONSORTIUM


Canadian Forest Service
Pacific Forestry Center
Insect Collection (175K) Curator 1 Curator 2
Live Fungi Collection Curator 3
Pathology Herbarium (35K) Curator 3
Northern Forestry Center
Insect & Spider Collection (200K) Curator 4 Curator 5
Live Fungi Collection (3K) Curator 6 Curator 7
Pathology Herbarium (24K) Curator 6 Curator 7
Herbarium (40K) Curator 4
Great Lakes Forestry Center
Insect Collection (165K) Curator 8
Herbarium (22K) Curator 8 Curator 9
Fungi Collection No curator
Laurentian Forestry Center
Insect Collection (200K) Curator 10 Curator 11 Curator 12
Fungi Collection (28K) Curator 13 Curator 14
Seeds & Leaves Curator 15
Atlantic Forestry Center (Corner Brook)
Insect Collection (40K) Curator 16 Curator 17 Curator 18
Fungi Collection Curator 16 Curator 17
Herbarium (4K) Curator 16 Curator 17
Atlantic Forestry Center (Fredericton)
Insect Collection (75K) Curator 19 Curator 20 Curator 21
Seeds (20M) Curator 22
Canadian Wood Fibre Center
Wood Pieces (3K pieces) Curator 23 Curator 24
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?




Specify COLLECTIONS CONSORTIUM

# 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.

Biodiversity Information Standards TDWG

GBIF

iDigBio Integrated Digitized Biocollections

GGBN

Symbiota

Specify COLLECTIONS CONSORTIUM

# 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.

Specify COLLECTIONS CONSORTIUM

# Special Thanks

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

*Previously supported by several US NSF grants, 1986-2017*

## Founding Partners

KU Biodiversity Institute & Natural History Museum, The University of Kansas

Muséum Genève

CSIRO

Florida Museum

CSIC Consejo Superior de Investigaciones Científicas

Danish Natural History Museums

LSA College of Literature, Science, and the Arts, University of Michigan

# Special Thanks

## Full Members

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

## Solution and Associate Members

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