Database schemas for natural history collections

Laura Brenskelle[1] and Andrea Thomer[2]

[1]The University of Texas at Austin

[2]University of Illinois at Urbana-Champaign

Introduction

- How do the structures of different natural history collections database schemas vary, and how well do these schemas match the information of the science they are mapping?

- Compared two databases and their schemas
  - Specify & Arctos

- Used Darwin Core as our measure of "the information of the science"

Darwin Core (DwC)

- Community-developed metadata schema for natural history collections
  - Based on Dublin Core

- Provides standardized way to describe NHC data (specimens and observations)

- Makes data aggregation & use simpler

- Does not define relationships of terms

- Many different ways to convey metadata
  - XML

Database Schemas

- How tables in a database relate to each other
- Different systems have different database schemas
  - Specify, Arctos, KE EMu, Microsoft Access, etc.
- May contain some but not all DwC terms

Specify vs. Arctos

- How do different databases represent Darwin Core in their schemas?

- Occurrence DwC terms in Specify & Arctos

Collaborative Collection Management System

Occurrence elements

- occurrenceID
- catalogNumber
- recordedBy
- recordNumber
- individualID
- sex
- lifeStage
- establishmentMeans
- occurrenceStatus
- associatedSequences

occurrenceID

- Definition: global unique identifier
- Arctos – Catalogued_Item → Collection_Object_ID
- Specify – CollectionObject → GUID

catalogNumber

- Definition: institutional specimen/observation identifier
- Arctos – Cataloged_Item → Cat_Num
- Specify – CollectionObject → catalogNumber

recordedBy

- Definition: collector/observer
- Arctos – Collector:Agent_ID → Agent:Agent_ID → Person:First_Name, Last_Name, Middle_Name
- Specify – Collector → collectorID → Agent → First Name, Last Name, Middle Initial

recordNumber

- Definition: field number
- Arctos – unclear; Coll_Obj_Other_ID_Num table (includes Other_ID_Type, Other_ID_Number, etc.)
- Specify – CollectionObject → fieldNumber

individualID

- Definition: identifier for single individual that may have been resampled
- Arctos – likely handled through Specimen_Part table (fields Part_name, Sample_from_Obj_ID, Derived_from_Cat_Item)
- Specify – no specific field for this; would depend on how you interpret “Collection Object”

sex

- Controlled vocab: unknowable, undetermined, male, female, hermaphrodite, gynandromorph
- Arctos – no specific field for this; Attributes table?
- Specify – Morph Bank view; 32 bit string

lifeStage

- Controlled vocab: zygote, embryo, larva, juvenile, adult, sporophyte, gametophyte, spore, gamete, pupa
- Not a defined field in Specify or Arctos?
- Arctos – could be covered by Attributes table

establishmentMeans modern

- Controlled vocab: native, introduced, naturalised, invasive, managed, uncertain
- Not in Specify or Arctos
- Arctos – could be covered by Attributes or Specimen_Event → habitat

occurrenceStatus modern

- Controlled vocab: present, absent, common, irregular, rare, doubtful
- Not in Specify or Arctos
- Arctos – possibly Attributes or Specimen_Event → habitat

associatedSequences

- Definition: list of identifiers of genetic sequence information
- Arctos – nothing specific; possibly Attributes table or Specimen_Part or Specimen_part_attributes
- Specify – DNA Sequence → genbankAccessionNumber

Other observations...

- The biogeography-specific fields do not seem to fit in either schema

- Arctos – specimen information is spread through a few tables

- Unclear if controlled vocabulary is present
  - Enforce DwC controlled vocabularies through picklists when possible

- Labeling fields from a particular schema
  - Arctos – not done
  - Specify – certain fields labeled as ABCD schema

Co-opt at your own risk!

Co-opting fields does not change them in the backend!

This could make future portability and data-sharing difficult.

Acknowledgements

- Gil Nelson, iDigBio

iDigBio
Integrated Digitized Biocollections

- Matt Brown
- Angie Thompson, Chris Sagebiel, Ann Molineux, Unmil Karadkar, Tim Rowe, Chris Bell
- You for listening!

Questions?