

# Fielding the field-to-fields pipeline: From sampling through sequences in practice

N. Dean Pentcheff & Regina Wetzer

*Natural History Museum of Los Angeles County*

NATURAL  
HISTORY  
MUSEUM  
LOS ANGELES COUNTY

DISCO 

Diversity Initiative for the Southern California Ocean

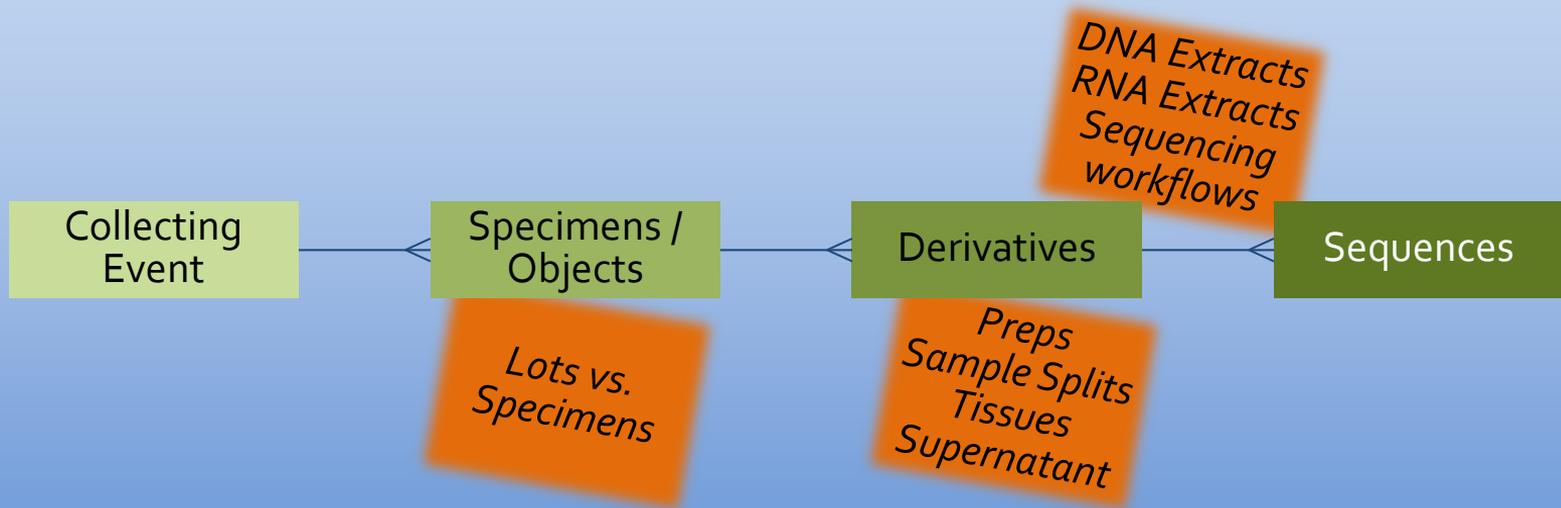
# The digital collections landscape



<i>Phaeocolactes cinereus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Pseudochelirus peregrinus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Tendonegus umbonatus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Macropus agilis</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Trichomurus vulpecula</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Fetaurus bevelceps</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Tasmodon obesus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Urocyon v. marinus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Antechinus swainsonii</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Turilopa truncata</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Acrocephalus pusillus dolifexus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Bos taurus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Capra aegagrus hircus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Canis species</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Equus caballus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Felis catus</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Panthera leo</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA
<i>Ovis aries</i>	TTCTGTTGATGTAATGATGTCAGATGTTACTGATTTCCATTTCTCCGATTTAGGATTAAGATTAATCTATTA



# The data flow



*Collectively how are we doing  
on this data landscape?*

# The landscape in practice

It's a bit sparse



# The landscape in practice

Few institutions have the full pipeline



*Photo of L.A. Water flowing above the Owens Valley.*

# The two cultures

Collections/Specimens

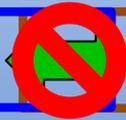
Genetics/Sequences

**Collections databases**

**Sequence management**

Collection → Specimens → Derivatives

Sequences ← Repository



*Collection Data*

*Sequence Data*

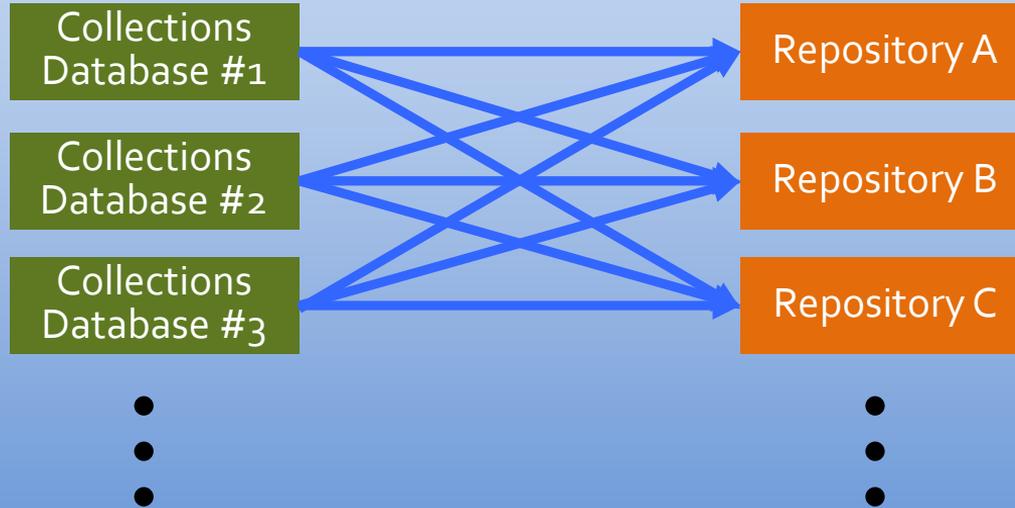
# This is why we have friction

Example:

~~Specify~~  
DISCO

Field Name	Type	Options / Comments (Click to toggle)
* specimen ID	Number	Indexed, Auto-enter Serial, Always Validate, Required
* parent	Number	Indexed, Always Validate, By Value List, Numeric Only
* collection ID	Text	Indexed, Auto-enter Calculation replaces existing value
* taxon ID	Number	Indexed, By Value List, Message
* determined by	Text	Indexed
* determination date	Date	Indexed
* determination reliable	Text	Indexed, Auto-enter Data, Always Validate, By Value List
* determination remarks	Text	Indexed
* determination timestamp	Timestamp	
* determination history	Text	Indexed
* genus species description	Text	Indexed
* taxon description	Text	Indexed
* molecule usefulness	Text	Indexed
* physical location	Text	Indexed
* specimen count description	Text	Indexed
* specimen exists	Text	Indexed, Auto-enter Data, Always Validate, By Value List
* BOLD voucher status	Text	Indexed, Auto-enter Data, Always Validate, By Value List
* remarks	Text	Indexed
* export marker	Text	Indexed
* creation	Timestamp	Indexed, Creation Timestamp (Date and Time)
* change	Timestamp	Indexed, Modification Timestamp (Date and Time), Creation
* BOLD Project	Text	Indexed
* BOLD Collection Code	Text	Indexed, Auto-enter Data
* BOLD Institution Storing	Text	Indexed, Auto-enter Data
* BOLD Identifier Institution	Text	Indexed, Auto-enter Data
* BOLD Extra Info	Text	Indexed
* BOLD Museum ID OVERRIDE	Text	Indexed
* BOLD Museum ID	Calculation	Indexed, from Specimen Database, = If ( IsEmpty ( Tri
* BOLD Process ID	Text	Indexed
* specimen label cache	Text	Indexed
* specimen label	Calculation	Unstored, from Specimen Database, = /* if no taxonic
* islisted	Calculation	Unstored, from Specimen Database, = If ( Taxonomy:)

# So we build adapters



# Adapters are horrid time sinks



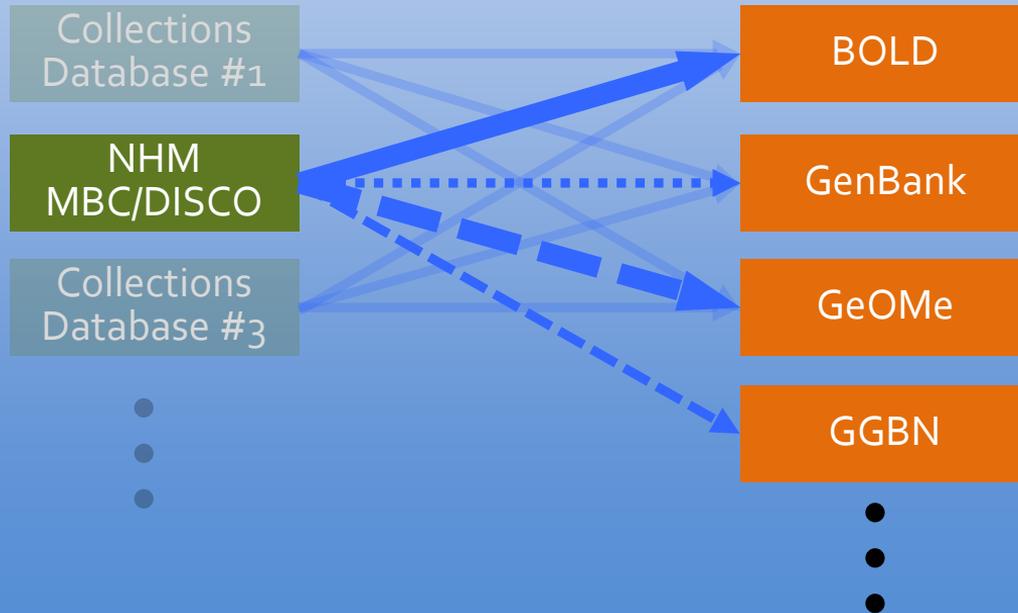
Recast identifiers,  
combine fields,  
convert units,  
fix filename limits,  
massage taxonomy,  
convert date format,  
*etc.*

# We do it under duress

In our case:



Diversity Initiative for the Southern California Ocean



Diversity Initiative for the Southern California Ocean

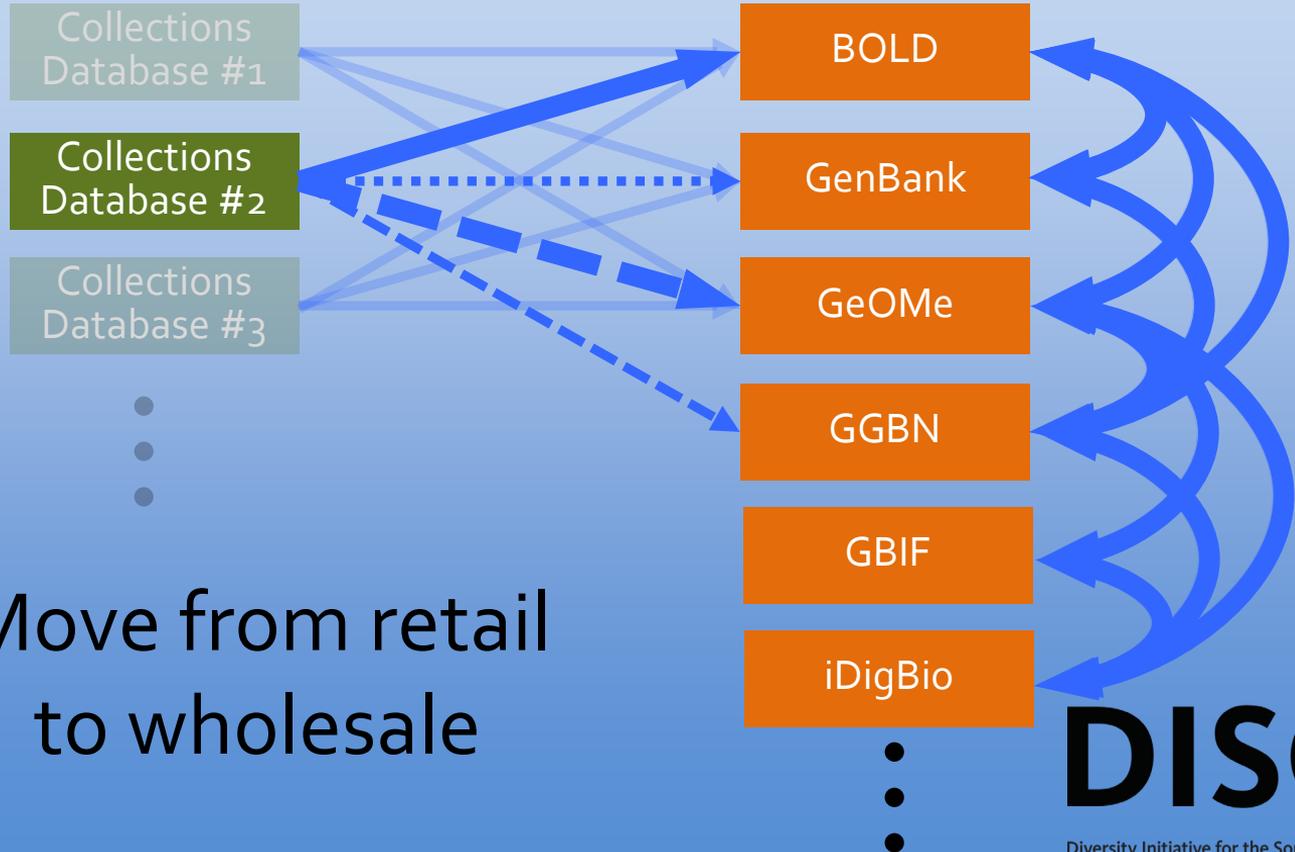
# The landscape in practice

Few institutions have the full pipeline



*Photo of L.A. Water flowing above the Owens Valley.*

# What is to be done?



Move from retail  
to wholesale

*May our data flow as smoothly  
as waves to the shore...*

Dean Pentcheff pentcheff@nhm.org  
Regina Wetzer wetzer@nhm.org



Diversity Initiative for the  
Southern California Ocean

*May our data flow as smoothly  
as waves to the shore...*

*Thanks to:*

*GGBN*

*NMNH LAB*

Dean Pentcheff [pentcheff@nhm.org](mailto:pentcheff@nhm.org)

Regina Wetzer [wetzer@nhm.org](mailto:wetzer@nhm.org)



Diversity Initiative for the  
Southern California Ocean