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

N. Dean Pentcheff & Regina Wetzer
Natural History Museum of Los Angeles County
The digital collections landscape
The data flow

- Collecting Event
- Specimens / Objects
- Derivatives
- Sequences
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
The two cultures

Collections/Specimens          Genetics/Sequences

Collections databases

Collection → Specimens → Derivatives
The two cultures

Collections/Specimens  Genetics/Sequences

Collections databases

Collection → Specimens → Derivatives

Collection Data
The two cultures

Collections/Specimens

Genetics/Sequences

Collections databases

Collection → Specimens → Derivatives

Sequence management

Sequences ← Repository

Collection Data
The two cultures

Collections/Specimens

Collections databases

Collection → Specimens → Derivatives

Collection Data

Genetics/Sequences

Sequence management

Sequences ← Repository

Sequence Data

DISCO

Diversity Initiative for the Southern California Ocean
The two cultures

Collections/Specimens  Genetics/Sequences

Collections databases  

Collection → Specimens → Derivatives  

Sequences ← Repository

Collection Data  Sequence Data

DISCO

Diversity Initiative for the Southern California Ocean
The two cultures

Collections/Specimens

Genetics/Sequences

Collections databases

Collection → Specimens → Derivatives

Sequence management

Sequences ← Repository

Collection Data

Sequence Data

DISCO

Diversity Initiative for the Southern California Ocean
This is why we have friction

Example: Specify
This is why we have friction

Example:

NHM

DISCO
So we build adapters

Collections Database #1
Collections Database #2
Collections Database #3
Repository A
Repository B
Repository C
So we build adapters
So we build adapters
So we build adapters
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:
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?

Collections Database #1

Collections Database #2

Collections Database #3

BOLD

GenBank

GeOME

GBBN

GBIF

iDigBio

...
What is to be done?

Collections Database #1
Collections Database #2
Collections Database #3

BOLD
GenBank
GeOMe
GBNB
GBIF
iDigBio

Move from retail to wholesale
What is to be done?

- Collections Database #1
- Collections Database #2
- Collections Database #3

Move from retail to wholesale

- BOLD
- GenBank
- GeOMe
- GGBN
- GBIF
- iDigBio

DISCO
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
Regina Wetzer  wetzer@nhm.org