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

The digital collections landscape



N A T U R A L H I S T O R Y M U S E U M



Phascolarctos cinereus	IIIGIIIGAICACIAAIAAICACACCACIACIACIICICCICCIICII
Pseudocheizus pezegzinus	TTTOTATOATCAOTTATAATCACAOCAOTATTATTACTACTATCCCTGCCTGTTCTACCAOCAOCAATTACCATACT
Dendrolagus lumholtzi	TTCOTATOATCCCTAATAATCACAOCAOTCCTTCCTCCTTCATTACCAOTTTACCAOCOOCATTACAATACT
Macropus agilus	MICOTIONICIONNENDENENENENENENENENENENENENENENENENEN
Trichosurus wulpecula	TECOTOLOANCAOTTATAATTACTOCCOTTTACTICTACCACTTCCACTACCAOCAOCAOCAOCAOCAATTACTACTACTACTACTACTACTACTACTACTACTAC
Fetaurus breviceps	TTCOTTION CONSTANTANTACAOCAOTECTACCTICTATCTTTACCAOTATTACCAOCACCACCACTATACTAC
Isoodon obesulus	TTTOTTOATCASTAATCACTOCASTATTATCSCTCCTATCTTCACCASTACTACCASCASTACTACTACCASCASTACTACTACTACTACTACTACTACTACTACTACTACTACT
funitheshynchus anatimas	TECOTIONICAUECTEANERACAUCTECTECTICEATCOLOTICEAUCAUCAUCAUCAUCAUCAUCAUCAUCAUCAUCAUCAUCA
Antechinus swainsonii	TTTOTTTOATCTOTAATAATCACAGCTOTOCTACTACTCTACT
Tursiops truncates	TICOICIONICACICIINOUTACAGEACICIINACINICATINECACENCIACEACEACEACAACIACIAIATIA
Arctcephalus pusillus doriferus	TTCOTOTOXICCOTACTACTACTACACCOCTACTACTACTACCACCACCCTACCACCTOCTACTACTACTACTACTACTACTACTACTACTACTACTAC
Bos taurus	TTCOTATOATCCGTAATTACCGCCGTACTACTACTACTCCCCCCTGTATTACCACCCCGCATCACAATGCT
Capra asgagrus hircus	TEROTORIANCEORCEUSANEACEOCCOSACEACECCEUSCACEECECECECECECECECECECECECECECECECECE
Cebus species	TITOTA CONTROL TO CARTA TACANCANTCO TECHNOLOGIC CANTOC TACCANTACTACTACTACTACTACTACTACTACTACTACTACTACT
Equus caballus	TTCOTTION CONCERTATINGOUCAUTAC TCCTTCTCCCCCCCCCCCCCCCCCCCCCCCCC
Felis catus	TITOTICATCACTCILATCACTOCCCCTCILACTCICCACTCCACCACTTIACCACCOCCATCACCACCCC
Panthera les	TTTOIC GAIGGTTTTAATCACHOCHGTATTGCTACTCCTATGACTGCCAGTTTTAGCAGCAGCATCACHATGCT
Ovis aries	IIIOIAIOAICIGIACIAAIIACIGCCGIACIICICCICICACIICICACIICACAGCAGCIGGACAAIACI





DISCO

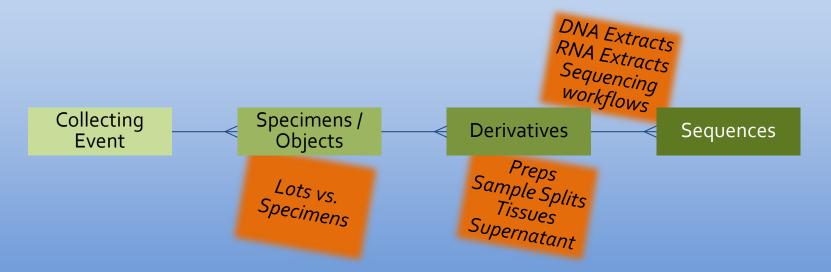
The data flow







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.

Collections/Specimens

Genetics/Sequences





Collections/Specimens

Genetics/Sequences

Collections databases

Collection -> Specimens -> Derivatives





Collections/Specimens

Genetics/Sequences

Collections databases

Collection -> Specimens -> Derivatives

Collection Data





Collections/Specimens

Genetics/Sequences

Collections databases

Collection -> Specimens -> Derivatives

Sequences <-- Repository

Collection Data





Collections/Specimens Genetics/Sequences

Collections databases		Sequence management	
Collection -> Specimens -> Derivatives		Sequences < Repository	

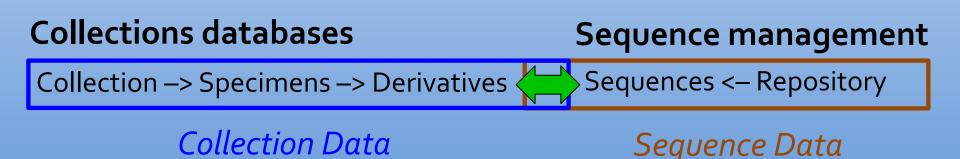
Collection Data

Sequence Data





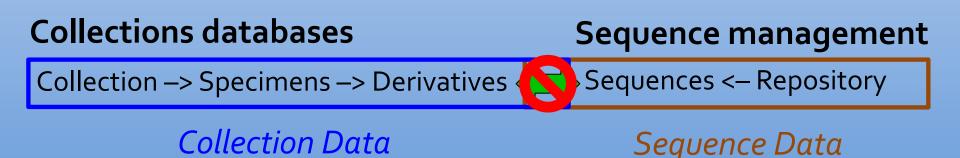
Collections/Specimens Genetics/Sequences







Collections/Specimens Genetics/Sequences







This is why we have friction

Example: Specify

NATURA	L
HISTORY	
MUSEUN LOS ANGELES COUNT	

DnaSequenceID: int(11) TimestampCreated: datetime(0) TimestampModified: datetime(0) Version: int(11) onectionMemberID: int(11) Ambiguous Residues: int(11) BOLDBarco elD: varchar(32) BOLDLastUp lateDate: date BOLDSampleD: varchar(32) BOLDTrans_tionMatrix: varchar(64) CompA: i (11) . int(11) CompG: int(11) compT: int(11) GenBankAccessionNumber: varchar(32) GeneSequence: text MoleculeType: varchar(32) Number1: float(20, 10) Number2: float(20, 10) Number3: float(20, 10) Remarks: text TargetMarker: varchar(32) Text1: varchar(32) Text2: varchar(32) Text3: varchar(64) _ ._ ..

dnasequence

	<u>*</u>	_		
	materialsample			
	MaterialSampleID: int(11)	٦.		
	TimestampCreated: datetime(0)			
	TimestampModified: datetime(0)	>6		
	Version: int(11)			
	CollectionMemberID: int(11)			
	5GBNA sorbanceRatio260_230: float(20, 10)			
	GGBNAb prbanceRatio260_280: float(20, 10)			
ə	GGBNRAb orbanceRatioMethod: varchar(64)			
	GGBNConcentration: float(20, 10)			
	GGBNConcentrationUnit: varchar(64)			
	GGBNMateralSampleType: varchar(64)			
	GGBNMediun: varchar(64)			
	GGBNPurific tionMethod: varchar(64)			
	GGBNQuality varchar(64)			
	GGBNQuality CheckDate: date			
	GGBNQuality Remarks: text			
	GGBNSample Designation: varchar(128)			
	GGBNSampl <mark>Size</mark> : float(20, 10)			
	GGBNVolume: float(20, 10)			
	GGBNVolun eUnit: varchar(64)			
	GGBNWeig t: float(20, 10)			
	GGBNWeinhtMethod: varchar(64)			
	GGBNWeghtUnit: varchar(64)			
	GUID: varchar(128)			
	hteer 1: int(11)			
	Integer2: int(11)			
	Number1: float(20, 10)			



or the Southern California Ocean

This is why we have friction

Example: NHM DISCO

NATURAL HISTORY M U S E U M LOS ANGELES COUNTY

	Fie	eld Name	Туре	Options / Comments (Click to toggle)
	٠	specimen ID	Number	Indexed, Auto-enter Serial, Always Validate, Required
	٠	parent	Number	Indexed, Always Validate, By Value List, Numeric Only
	٠	collection ID	Text	Indexed, Auto-enter Calculation replaces existing value
	٠	taxon ID	Number	Indexed, By Value List, Message
	٠	determined by	Text	Indexed
	٠	determination date	Date	Indexed
	٠	determination reliable	Text	Indexed, Auto-enter Data, Always Validate, By Value L
	٠	determination remarks	Text	Indexed
	٠	determination timestamp	Timestamp	
	٠	determination history	Text	Indexed
	٠	genus species description	Text	Indexed
	٠	taxon description	Text	Indexed
	٠	molecule usefulness	Text	Indexed
	٠	physical location	Text	Indexed
	٠	specimen count description	Text	Indexed
	•	specil en exists	Text	Indexed, Auto-enter Data, Always Validate, By Value L
		BOLD vucher status	Text	Indexed, Auto-enter Data, Always Validate, By Value L
		remarks	Text	Indexed
	٠	export marker	Text	Indexed
	٠	creation	Timestamp	Indexed, Creation Timestamp (Date and Time)
	٠	change	Timestamp	Indexed, Modification Timestamp (Date and Time), Ca
	٠	BOLD Project	Text	Indexed
I	٠	BOLD Collection Code	Text	Indexed, Auto-enter Data
	٠	BOLD Institution Storing	Text	Indexed, Auto-enter Data
	٠	BOLD Ider ifier Institution	Text	Indexed, Auto-enter Data
	٠	BOLD Extended Info	Text	Indexed
	٠	BOLD Muteum ID OVERRIDE	Text	Indexed
		BOLD Museum ID	Calculation	Indexed, from Specimen Database, = If (IsEmpty (Tri
		BOLD Flocess ID	Text	Indexed
	•	specialen label cache	Text	Indexed
	٠	specimen label	Calculation	Unstored, from Specimen Database, = /* if no taxonic
	٠	islisted	Calculation	Unstored, from Specimen Database, = If (Taxonomy:

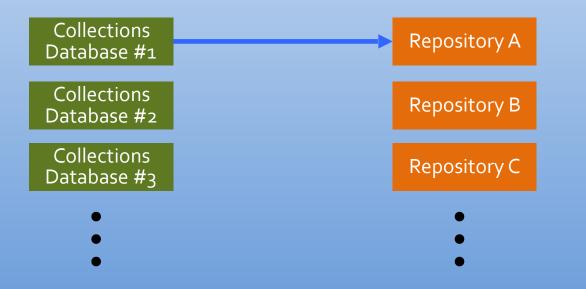
Options / Comments (Click to toggle)
Indexed, Auto-enter Serial, Always Validate, Required
Indexed, Always Validate, By Value List, Numeric Only
Indexed, Auto-enter Calculation replaces existing valu
Indexed, By Value List, Message
Indexed
Indexed
Indexed, Auto-enter Data, Always Validate, By Value L
Indexed
Indexed
Indexed, Auto-enter Data, Always Validate, By Value L
Indexed, Auto-enter Data, Always Validate, By Value L
Indexed
Indexed
Indexed, Creation Timestamp (Date and Time)
Indexed, Modification Timestamp (Date and Time), Ca
Indexed
Indexed, Auto-enter Data
Indexed, Auto-enter Data
Indexed, Auto-enter Data
Indexed
Indexed
Indexed, from Specimen Database, = If (IsEmpty (Tri
Indexed
Indexed
Unstored, from Specimen Database, = /* if no taxonic

DISCO



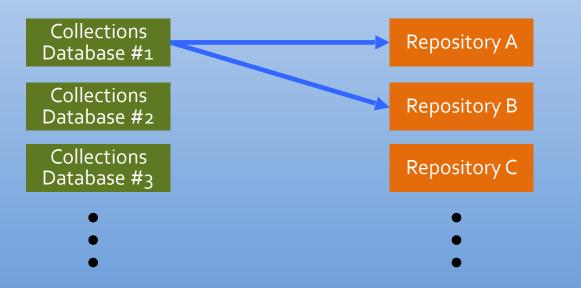






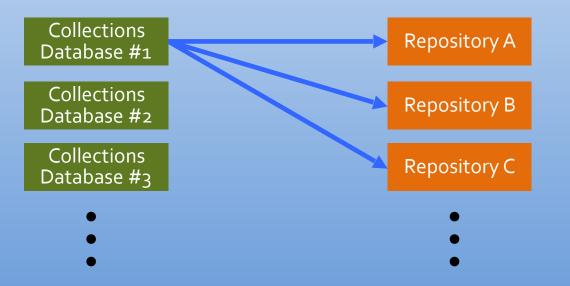






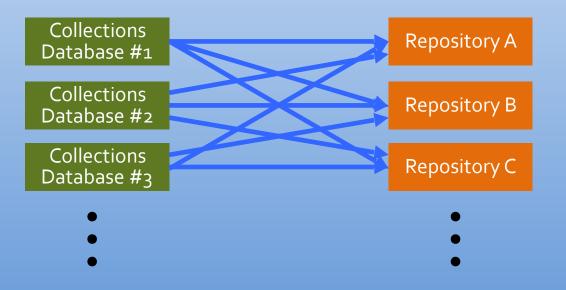








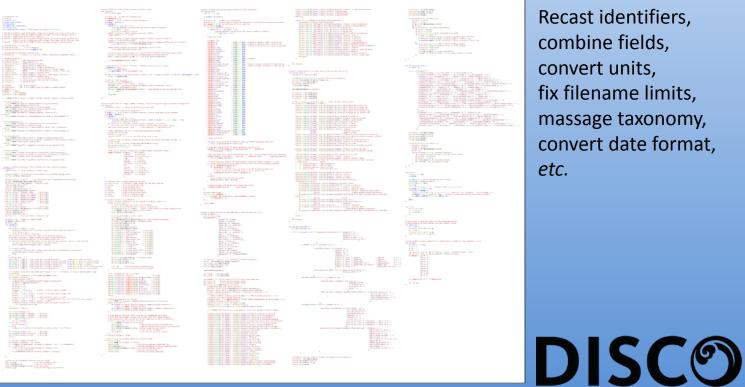








Adapters are horrid time sinks



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

Diversity Initiative for the Southern California Ocean

NATURAL HISTORY MUSEUN LOS ANGELES COUNTY

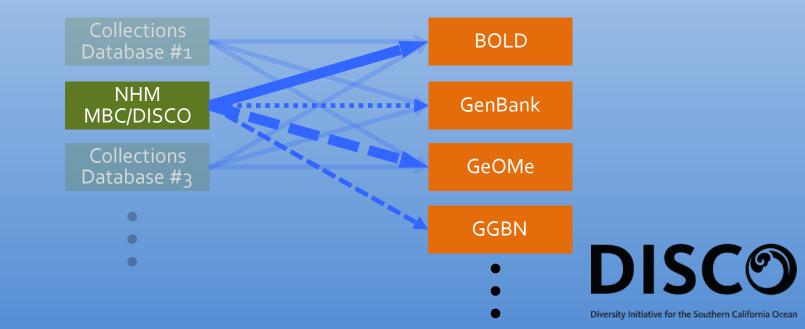
We do it under duress

In our case:





Diversity Initiative for the Southern California Ocean



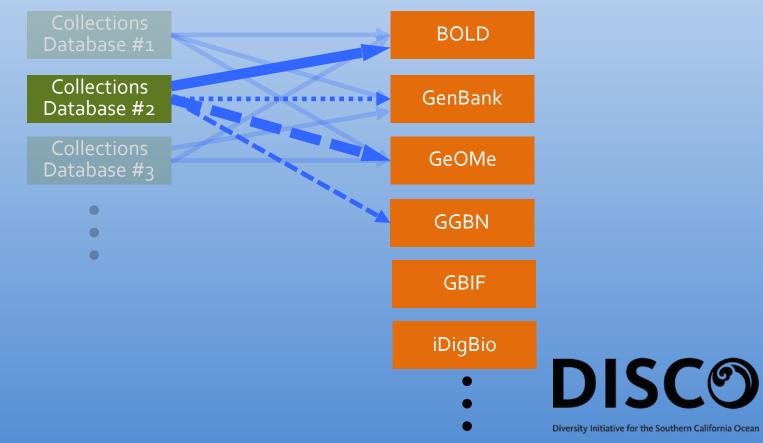
NATURAL HISTORY MUSEUM LOS ANGELES COUNTY

The landscape in practice Few institutions have the full pipeline



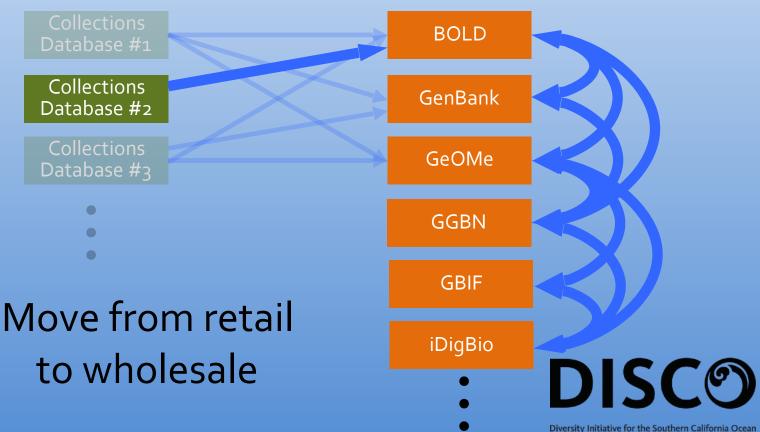
Photo of L.A. Water flowing above the Owens Valley.

What is to be done?



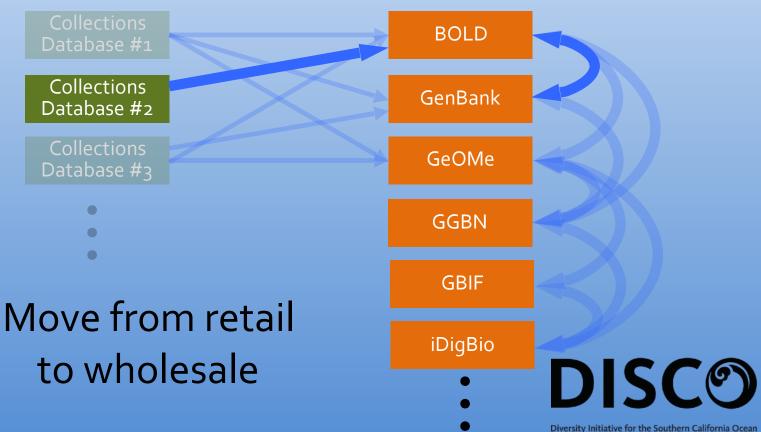
NATURAL HISTORY MUSEUM LOS ANGELES COUNTY

What is to be done?



N A T U R A L H I S T O R Y M U S E U M LOS ANGELES COUNTY

What is to be done?



N A T U R A L H I S T O R Y M U S E U M LOS ANGELES COUNTY

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