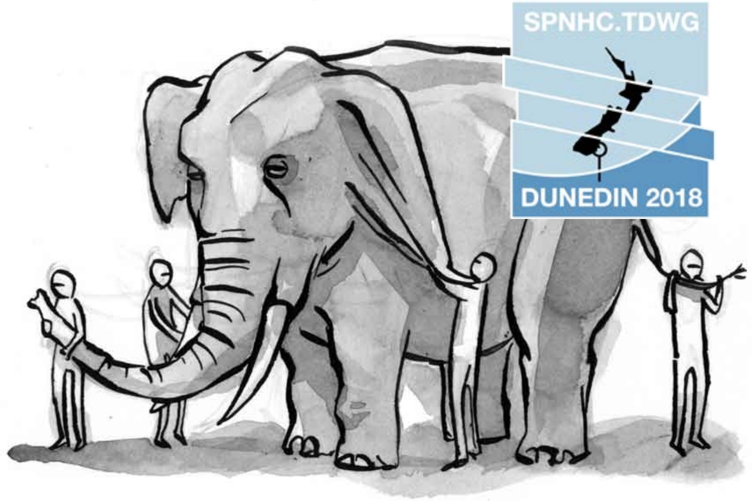
Implementing Collections Data Quality (DQ) Feedback

a survey and your community experience stories to shed some light into the data integration abyss



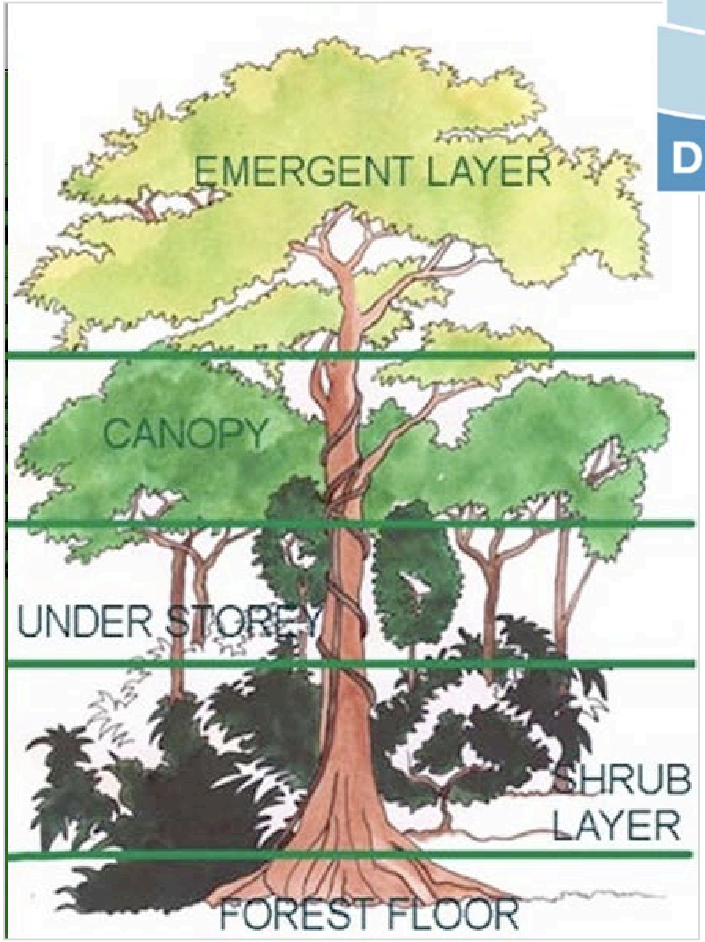
Deborah Paul, iDigBio, Florida State University
Nicole Fisher, CSIRO
@SPNHC-TDWGNZ Thursday 30 August 2018
@idbdeb @fisher_anic http://bit.ly/spnhcdq2018
Collections and Data in an Uncertain World

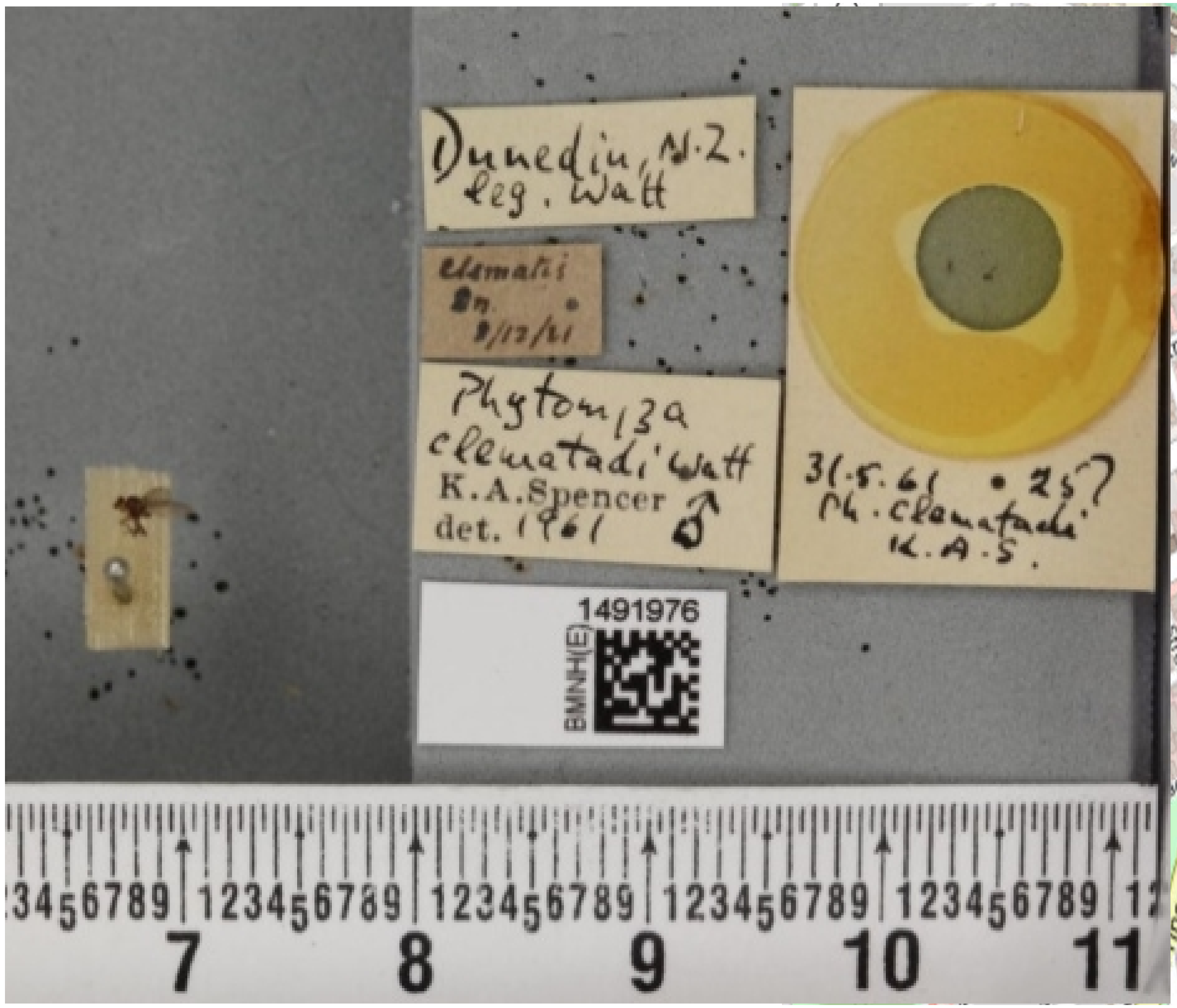
iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All images used with permission or are free from copyright.

SPNHC.TDWG DUNEDIN 2018

Topics

- What we did and why
  - Survey, SIG, Report
- Highlights of Some Results
- What next?
  - Our speakers
  - SIG tomorrow 11-1230
    - your abyss?
  - Survey report *in progress*
  - Darwin Core Hour *follow-up*





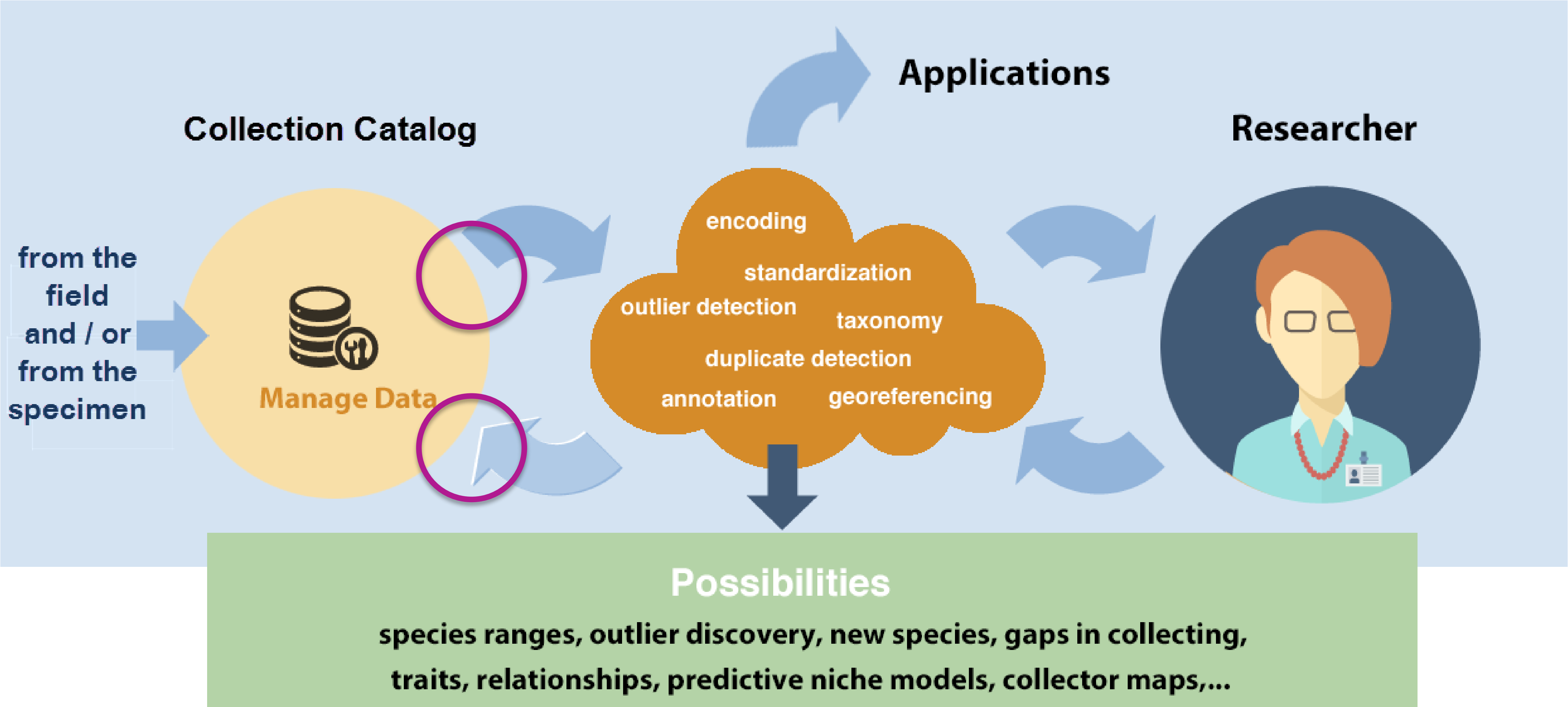
Tennyson Street
Dowling Street
Burlington Street
High Street
The Exchange
Thomas B
Princes Street
Bond Street
Crawford Street
Vogel Street
Cumberland Street
Main South Line
Wharf Street
Roberts Street
Buller Street
SH1
Leaflet | Ma

From the Collection of: NHM K

From the Collection of: NHMUK	
Collector: Watt	
Phytomyza clematadi Watt	
Found on: *Clematis*	
Date Collected: 1921-12-09	
Date Due	Borrower's Name
1961	det. K. A. Spencer ♂

Australasia; New Zealand; Otago; Dunedin, -45.8788, 170.5028

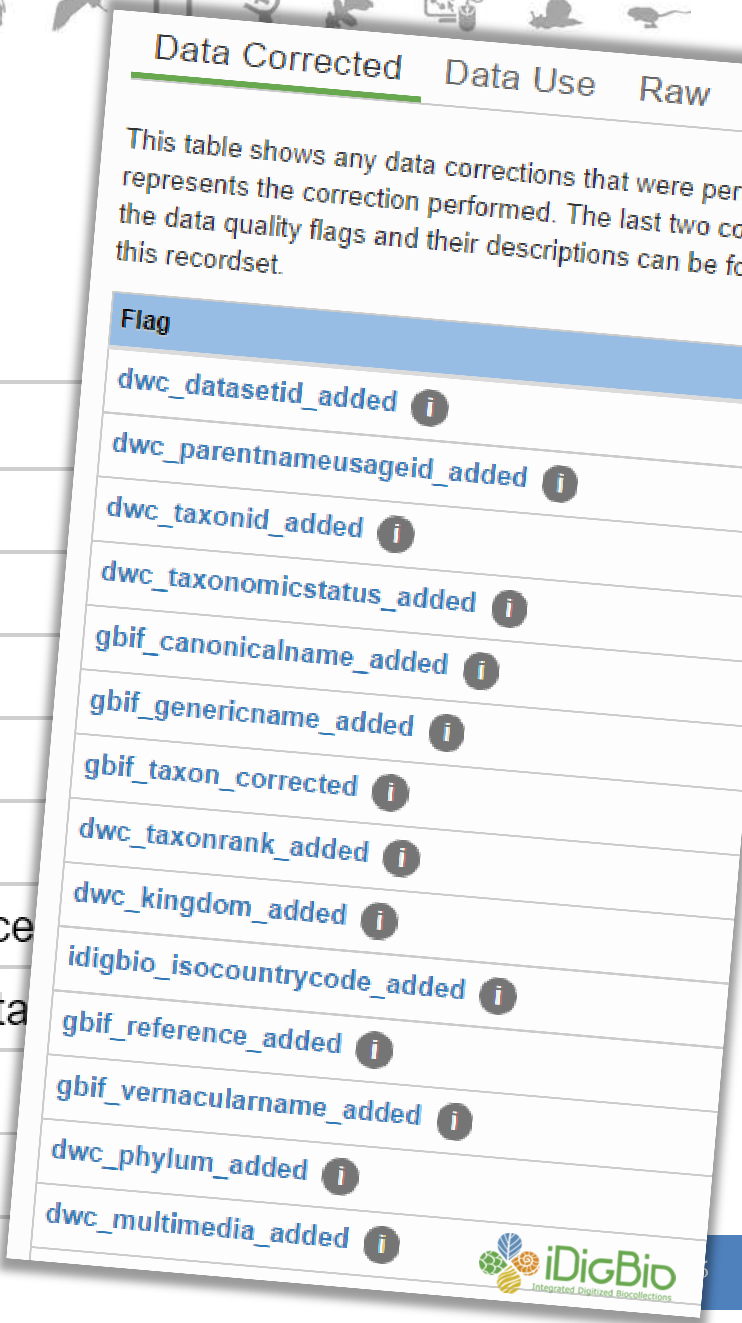
Biodiversity Data Pipeline: where are (some of) the bottlenecks (now)? What enables use of data quality feedback and supports data integration?



species ranges, outlier discovery, new species, gaps in collecting, traits, relationships, predictive niche models, collector maps,...

Survey to evaluate community experience *integrating DQ Feedback*

Left	Term	Right
believe that using the ALA's	data	quality assertions is an invaluable
for validating and improving our	data	. We would like to prioritise
actually deciding which of the	data	quality assertions the ALA adds
and gave us feedback on	data	flow and quality of coordinates
do a thorough job of	data	evaluation and cleaning Some the
different ways to standardize the	data	, of which there are many
emails in order to clean	data	. These messages are minimal. Once
be more of a priority.	data	quality improvement posterior to data
Data quality improvement posterior to	data	entry has not been a
taxonomies). I have found some	data	entry errors such as GPS
the benefit of cleaning their	data	or do not know to



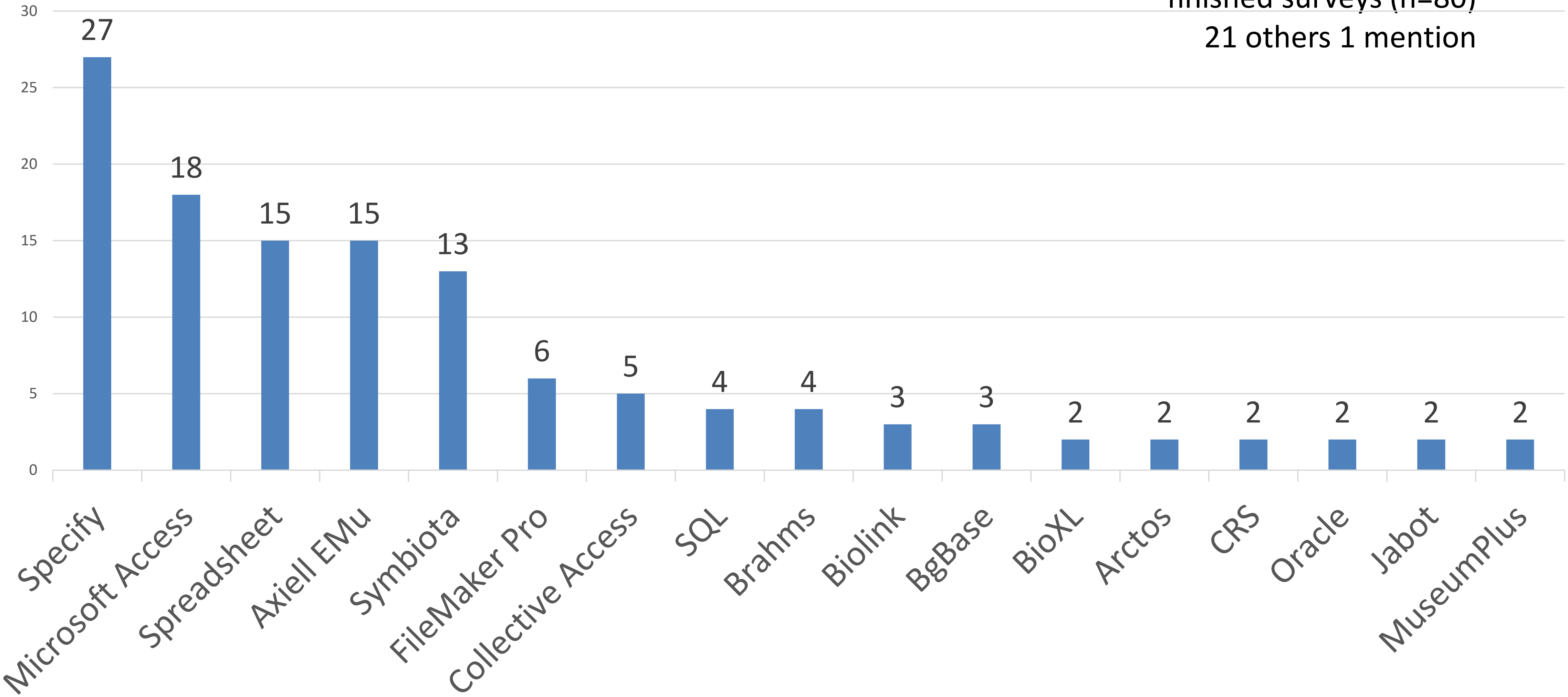
Data Corrected Data Use Raw

This table shows any data corrections that were per
represents the correction performed. The last two co
the data quality flags and their descriptions can be fo
this recordset.

Flag
dwc_datasetid_added ⓘ
dwc_parentnameusageid_added ⓘ
dwc_taxonid_added ⓘ
dwc_taxonomicstatus_added ⓘ
gbif_canonicalname_added ⓘ
gbif_genericname_added ⓘ
gbif_taxon_corrected ⓘ
dwc_taxonrank_added ⓘ
dwc_kingdom_added ⓘ
idigbio_isocountrycode_added ⓘ
gbif_reference_added ⓘ
gbif_vernacularname_added ⓘ
dwc_phylum_added ⓘ
dwc_multimedia_added ⓘ

Collection Management Software

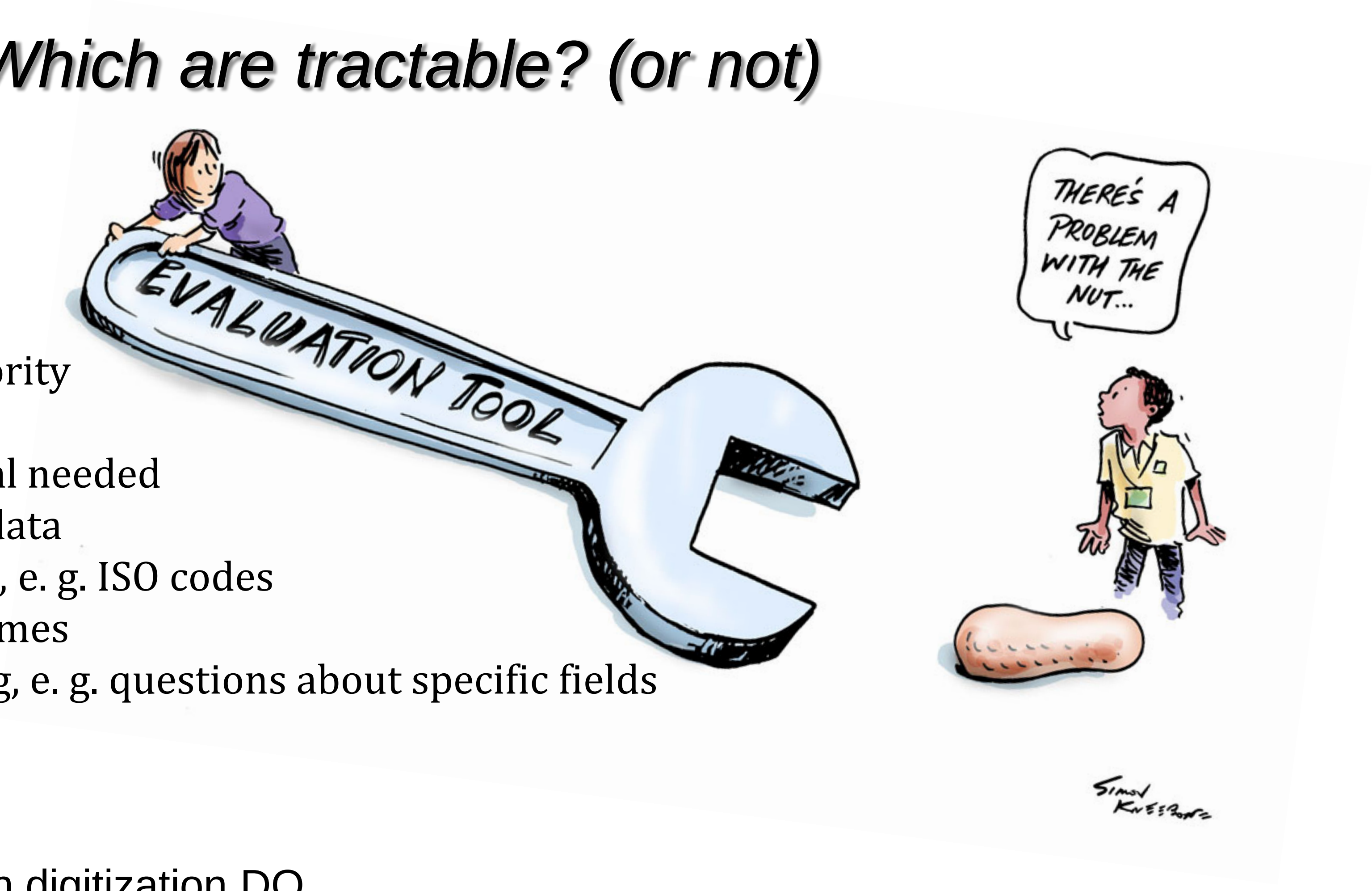
n=104 responders
software mentioned > 1 time
finished surveys (n=80)
21 others 1 mention



Not (yet) using dq feedback.
Why? Which are tractable? (or not)

- top selections were
  - lack of resources (time, staff, funds),
  - *not aware of feedback,*
  - *software challenges,*
  - *job is too massive,*
  - *not knowing where to find this feedback, and*
  - *need more biodiversity informatics skills*

Using dq feedback.
What challenges noted? Which are tractable? (or not)



- lack of resources (time, staff)
- data quality feedback
  - organization priority differences
  - massive scope
    - have to prioritize or not a priority
  - workflow issues
    - assigning / tracking / approval needed
    - no one to manage geospatial data
    - can’t make changes requested, e. g. ISO codes
  - erroneous feedback, e. g. taxon names
  - data standards knowledge missing, e. g. questions about specific fields
  - difficult to interpret
- Skills missing, or software impeded
  - changing by hand
- Prefer to work on curation rather than digitization DQ
  - An opportunity here

DQ workflows to reveal skills, literacy, software

Data and Collections Literacy

- collection knowledge
- data entry
- formatting
- import functions
- knowledge of related biodiversity data
- knowing which records need attention
- parsing
- querying
- scripting
- taxonomic, geography, geology skills
- track collection / collector
- track down correct date
- understanding relational databases
- understanding feedback

Software and skills

- Emu
- how “the Atlas” works
- advanced spreadsheet skills
- database software
- postgreSQL
- able to use SQL / MySQL
- FileMaker
- Specify
- Symbiota
- Open Refine

DQ feedback – benefits and changes

- dates
- distribution data
  - understanding scope of DQ issues
- fixing inaccuracies
- georeferences
- misspellings
- taxonomic name “insights”
- using dq feedback for prioritization
  - example: biosecurity and trade

> “…crowdsourcing data error detection is a better way to go rather than hiding data until you are sure it is all correct.”

Comments and requests

- “county” boundary checks
- embed DQ tests in collections software?
- pest - host data quality checks
- community-proofing
- re-format of dq feedback (data downloads)
- metrics tracking
- more workshops, webinars, (data mgmt., open refine, …)

“Making changes that are suggested is a big task and it is scary to make large batch changes.”

Key points or *why it's important to shed light on the abysses*

- awareness of DQ issues

[now we are] "aware of how many different terms are used for the same things … I hope as a community we can fix this issue soon."

https://bit.ly/niceuglydata

Key points

- asking for and getting help
  - do you know what to do? is it working?
- need for transparency in DQ processes
  - manage expectations

Key points

- software bottlenecks
- skills bottlenecks
  - … need for biodiversity data mobilization skills
    - source and development of skills?
    - changing roles in collections?
    - expectations for the future?

"… huge need for the community to proof on-line data of all kinds. Extra eyes regularly find things that have [been] missed …"

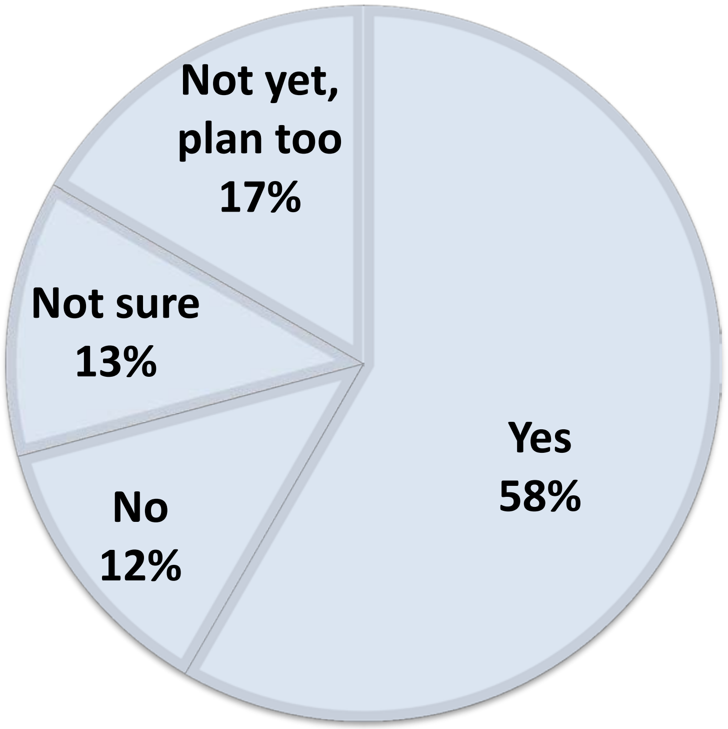
What's next for data integration at the DQ feedback step?

- Algorithms can't do it all
- AI can't do it all
- Data standards can't do it all

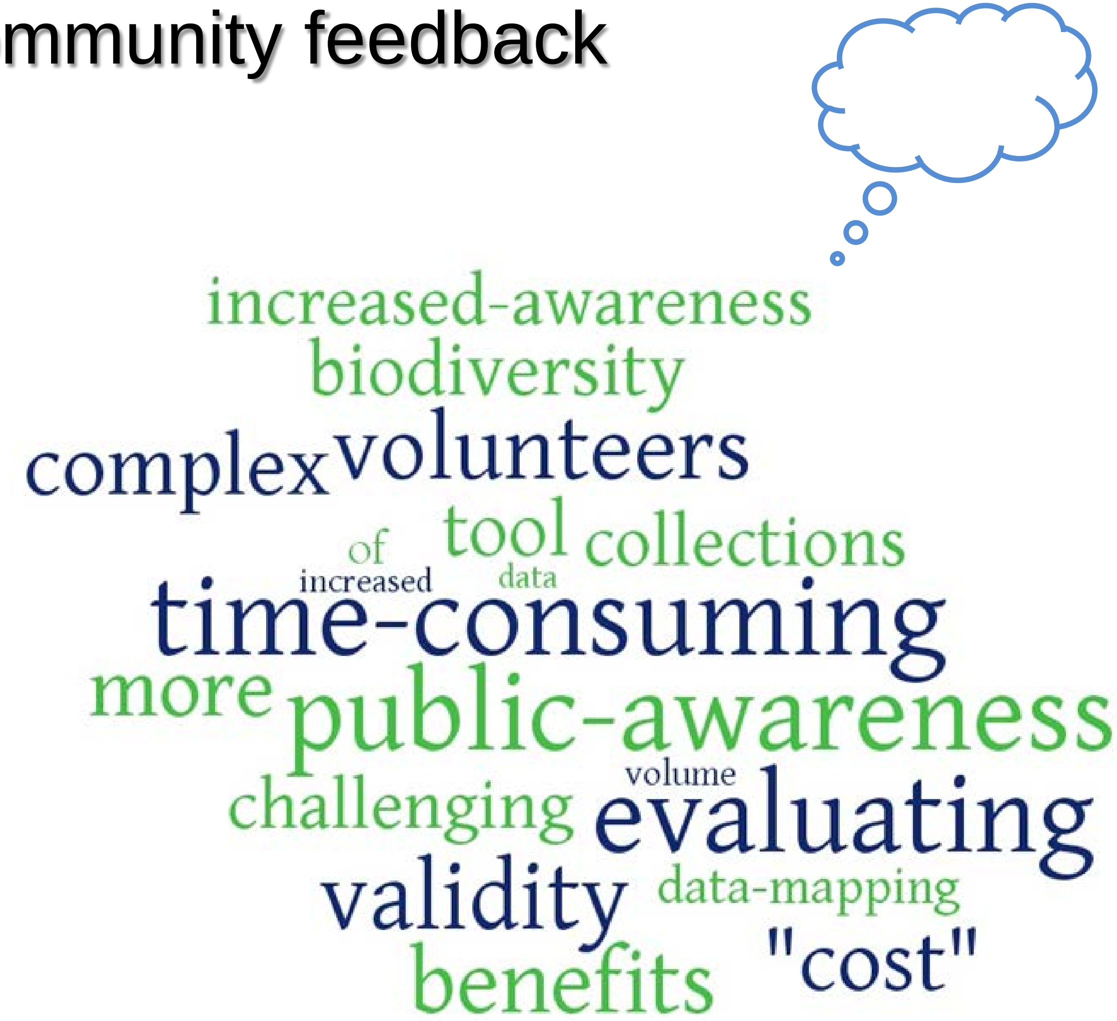
- What's your role?
- What's your abyss?

Integrating Crowdsourced Data - community feedback

- 64 responses
- 54.69% use crowdsourced data

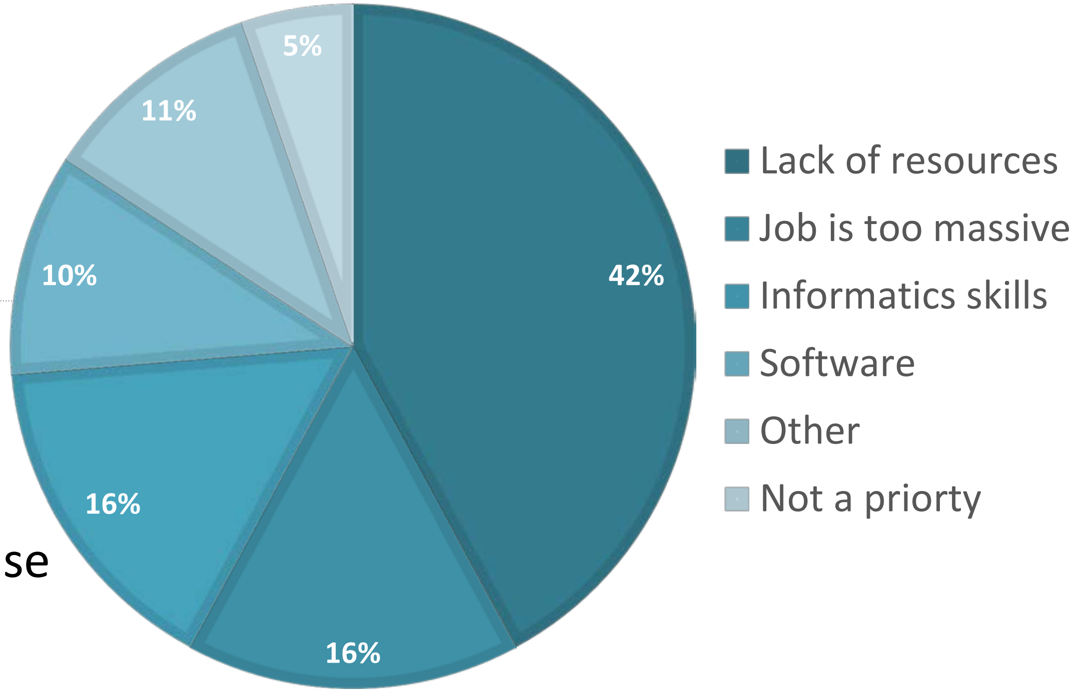


Integrate crowdsourced data



Challenges - integrating crowdsourced transcription data?

- 42% responses - “Lack of resources”
- 16% responses – “Job is too massive”



Other Challenges listed ……

- Complex data-mapping (not standardised DC)
- Data needs formatting
- Data validation
- More staff for validation and integration in database
- Procedure challenges

Integrating crowdsourced transcriptions – what would help?

more focus on the issue from the "top”

broader ability to rate "trustworthiness" of transcribers

crowdsourcing portals integrated with local database
*note – Symbiota and @NfromN

standardization (process & data)

more expert validators

Summary & future thoughts …

- Does crowdsourcing actually save time?
- Many other benefits … increase in awareness of collection

- ICEDIG / DiSSCo : evaluating costs/benefits of different transcription choices.
  https://icedig.eu/content/deliverables

ICEDIG.EU

• D4.2	Data quality in transcription	January 2019
• D4.3	Data standards in transcription	February 2019
• D4.4	Interoperability with institutional collection management systems	April 2019
• D4.5	Cost analysis of transcription methods	December 2019
• D5.1	Recommendations for volunteer transcription systems and a source repository	April 2019

Publications and Activities

- Belbin L, Daly J, Hirsch T, Hobern D, Salle JL. A specialist’s audit of aggregated occurrence records: An “aggregator”s’ perspective. *ZooKeys*. 2013;(305):67-76. doi:10.3897/zookeys.305.5438.
- Mesibov R. An audit of some processing effects in aggregated occurrence records. ZooKeys. 2018;(751):129-146. doi:10.3897/zookeys.751.24791.
- SPNHC_TDWGNZ W08 - Standardizing data to Darwin Core using R: A hands-on workshop with lessons learned from the TrIAS project. (2 – 3.30pm, Thursday)
- Project Paleo: Citizen Curation and Community Science at the Natural History Museum of Los Angeles County. - Elizabeth R Ellwood (4 – 4.20pm, Tuesday)

"… now on to the stories!"

Challenges For Implementing Collections Data Quality Feedback: Synthesizing the community experience.

2.20 – 2.40pm :	**Arthur Chapman**	Data Quality – Whose responsibility is it?
2.40 – 3.00pm :	**Mare Nazaire**	Integrating Data Quality Feedback: a Data Provider's Perspective.
3.00 – 3.20pm :	**Robert Cubey**	Label Transcript is Done – Now what do we do with that Data?
3.30 – 4.00pm	**Coffee Break**	
4.00 – 4.20pm :	**Andrew Bentley**	Practical use of aggregator data quality metrics in a collection scenario.
4.20 – 4.40pm :	**Teresa Mayfield**	Who Has Time for Biological Collections Data Quality Feedback? Maybe a Community Can Help.
4.40 – 5.00pm :	**Sharon Grant**	Repatriation of Augmented Information to an Institutional Database.

From our speakers' data integration stories to yours

- SPNHC #SIG on DQ Feedback is tomorrow for your part of the *#biodiversity #dataIntegration story*

iDigBio
Integrated Digitized Biocollections

Collaboration is key!

SPNHC.TDWG
DUNEDIN 2018

Florida State University 1851

SC net
Small Collections Network

Cretaceous World

FLORIDA MUSEUM

OPEN Tree of Life

Spark

ECN

THE CARPENTRIES

VertNet

ATLAS OF LIVING AUSTRALIA

myFOSSIL

AXIELL

EMu

software carpentry

PENSOFT

UF
UNIVERSITY of FLORIDA

DATA CARPENTRY
BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

ePANDDA

EPICC

ROpenSci

GUODA
http://guoda.bio/

ICEDIG.EU

DiSSCo
Distributed System of Scientific Collections

ARCTOS

Symbiota
Mobilizing Biodiversity Data

GBIF

GREAT LAKES INVASIVES NETWORK

SYNTHESYS
Synthesis of systematic resources

XXXII CONGRESSO BRASILEIRO DE ZOOLOGIA
FOZ DO IGUAÇU 2018

Smithsonian Institution

Specify
Software Project

BIODIVERSITY COLLECTIONS NETWORK

iDigBio
Integrated Digitized Biocollections

VACS
Duration: 00:01:33.910000

iDigFossils

oVert

SWANS
Collections

LUOMUS
LUONNONTIETEELLINEN KESKUSMUSEO

CETAF
CONSORTIUM OF EUROPEAN TAXONOMIC FACILITIES

eol
Encyclopedia of Life

NatureServe

LBCC

SCAN
Symbiota Collections of Arthropods Network

USGS
science for a changing world

Biodiversity Information Standards
TDWG

NSF

SYNTHESYS
Synthesis of systematic resources

ARBOR
revolutionary workflows

THE MACROALGAL HERBARIUM CONSORTIUM

MAM

fossil insect COLLABORATIVE

ECOLOGICAL SOCIETY OF AMERICA
esa

BRISTOL CITY COUNCIL

culture team

Invert-E-Base

BLUE
Biodiversity Literacy in Undergraduate Education

GEOLocate

SERNEC
Southeast Regional Network of Expertise and Collections

LepNet

NSC ALLIANCE

DWC hour

North American Network of Small Herbaria

THE JOHN ELLERMAN FOUNDATION

4B757261746F72
K

SoRo

WeDigBio
Worldwide Engagement for Digitizing Biocollections

Mobilizing New England Vascular Plant Specimen Data
to Track Environmental Changes

MiCC
Microfungi Collections Consortium

MYCOLOGY COLLECTIONS PORTAL

SPNHC
ADVANCING COLLECTIONS CARE

Lifemapper

NATURAL HISTORY MUSEUM

Notes from Nature

Effechecka

DISCOVER LIFE

Scratchpads
biodiversity online

BIOSPEX

BLUE
Biodiversity Literacy in Undergraduate Education

American Institute of Biological Sciences

Kia ora

from Nicole Fisher and Deborah Paul

see you tomorrow too at the SPNHC #SIG Share your data integration stories – successes and snafus too!

Special thanks to Shari Ellis, iDigBio Project Evaluator, for her guidance when developing our ideas for this work. Very kind thanks to all our speakers for being ready, willing, and yes, even eager to share their data stories – including the juicy bits.

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/idigbio

idigbio.org/rss-feed.xml

webcal://www.idigbio.org/events-calendar/export.ics

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All images used with permission or are free from copyright.