Integrating Data Quality Feedback

A Data Provider’s Perspective

Mare Nazaire
Rancho Santa Ana Botanic Garden
SPNHC + TDWG, Dunedin 2018
Natural history collections have always contained a wealth of data... a preserved organism and its label are a scientific specimen that has great intrinsic value. ... Before the advent of computers, natural history collections were physical databases from which geographic or ecological analyses and reports could be extracted by human visitation and transcription, usually a laborious and time-consuming task. ... Computerization of label data makes such reports on distribution and ecology of species more readily available to potential users; they add value to the data. Interconnecting the databases brings robustness to the information that natural history collections can provide to policy-making bodies; appreciation of robust data will lead in turn to appreciation of the collections from which those data were taken. Interconnectivity [facilitates] a common goal: the discovery and description of the world's biota.

- Meredith A. Lane
Roles of Natural History Collections
RSA Herbarium

- 1.23 million specimens of vascular plants
- Combined holdings: RSA & POM College
- Curatorial staff: 5.5 FTE
RSA Herbarium

Digitizing efforts
- ~10,000 / yr manual entry
- ~12,000 / yr bulk imports
- ~12,000 / yr images

- 43% databased:
  - California
  - Selected taxa / regions
- 25% georeferenced
- 4% imaged
A Brief History of RSA’s Databasing

- 1987: databasing type specimens, DBase II – limited fields
- 1993: experimentation with FMP - label making
- 1995: databasing southern CA specimens
  - flat files converted to FMP
- 1999: RSA & UCR merge FMP systems
- 2013: FMP crash
A Brief History of RSA’s Databasing

- 2014 – 2017: data migration from FMP to Specify
A Brief History of RSA’s Databasing

- Data migration: ~450,000 records
  - Cleaning, formatting, standardization
  - Combining / separating fields
  - Removing negative accessions, hidden spaces, odd characters
  - Standardizing geography, taxonomy
Publishing Data
Publishing Data
Data Quality Feedback

Sources of feedback

• Individual user (1 – few records)
• Aggregators (large batches)
  - CCH
  - iDigBio
Data Quality Feedback

- Georeferencing errors
  - some can be easy fixes
  - requires re-georeferencing
Data Quality Feedback

- Georeferencing errors
  - some can be easy fixes
  - requires re-georeferencing

<table>
<thead>
<tr>
<th>Country</th>
<th>State</th>
<th>County</th>
<th>Locality</th>
<th>Locality continued</th>
<th>Latitude1</th>
<th>Longitude1</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>California</td>
<td>Los Angeles County</td>
<td>Salt marsh at mouth of Malibu Creek.</td>
<td></td>
<td>34.0000000000</td>
<td>-118.0000000000</td>
</tr>
<tr>
<td>United States</td>
<td>California</td>
<td>Los Angeles County</td>
<td>Salt marsh at mouth of Malibu Creek.</td>
<td></td>
<td>34.0000000000</td>
<td>-118.0000000000</td>
</tr>
<tr>
<td>United States</td>
<td>California</td>
<td>Los Angeles County</td>
<td>Salt marsh at mouth of Malibu Creek.</td>
<td></td>
<td>34.0000000000</td>
<td>-118.0000000000</td>
</tr>
<tr>
<td>United States</td>
<td>California</td>
<td>Los Angeles County</td>
<td>At mouth of Malibu Creek</td>
<td></td>
<td>34.0000000000</td>
<td>-118.0000000000</td>
</tr>
<tr>
<td>United States</td>
<td>California</td>
<td>Los Angeles County</td>
<td>Borders of salt marsh at mouth of Malibu Creek.</td>
<td></td>
<td>34.0000000000</td>
<td>-118.0000000000</td>
</tr>
</tbody>
</table>

Plants of the Santa Monica Mountains, California
(Lat. N. 34°; Long. 118° to 119° W.)

Atriplex lentiformis (Torr.) Wats. ssp. Breweri (Wats.) Hall & Clem.

Dioecious; rounded shrubs up to 1.3 meters tall; borders of salt marsh, mouth of Malibu Creek,
elevation 40 feet, Los Angeles County.

Peter H. Raven
Henry J. Thompson.

September 6, 1959.
Data Quality Feedback

• Georeferencing errors
  - requires locating physical specimen
Data Quality Feedback

- Duplicate accessions
Data Quality Feedback

- Duplicate accessions
  - replace barcode as identifier

These are the records that the person has switched to using the barcode instead of the original RSA/POM accession:

logging: ACC: Barcode field not NULL, using barcode for ID B178841==RSA0036212<
logging: ACC: Barcode field not NULL, using barcode for ID POM100829B==RSA0007636<
logging: ACC: Barcode field not NULL, using barcode for ID POM100810A==RSA0003956<
logging: ACC: Barcode field not NULL, using barcode for ID POM100810B==RSA000396<
logging: ACC: Barcode field not NULL, using barcode for ID POM1008160==RSA0003976<
logging: ACC: Barcode field not NULL, using barcode for ID POM100315==RSA0000402<
logging: ACC: Barcode field not NULL, using barcode for ID POM100316==RSA0000491<
logging: ACC: Barcode field not NULL, using barcode for ID POM10034==RSA0001600<
logging: ACC: Barcode field not NULL, using barcode for ID POM10047==RSA0006513<
logging: ACC: Barcode field not NULL, using barcode for ID POM10048==RSA0012486<
logging: ACC: Barcode field not NULL, using barcode for ID POM100674==RSA0002986<
logging: ACC: Barcode field not NULL, using barcode for ID POM100675==RSA00029860<
logging: ACC: Barcode field not NULL, using barcode for ID POM10067==RSA0001587<
logging: ACC: Barcode field not NULL, using barcode for ID POM100691==RSA0000946<
logging: ACC: Barcode field not NULL, using barcode for ID POM100070==RSA0114414<
logging: ACC: Barcode field not NULL, using barcode for ID POM100719==RSA0102439<
logging: ACC: Barcode field not NULL, using barcode for ID POM100874==RSA016428<
logging: ACC: Barcode field not NULL, using barcode for ID POM100816==RSA0013865<
logging: ACC: Barcode field not NULL, using barcode for ID POM100817==RSA0009438<
logging: ACC: Barcode field not NULL, using barcode for ID POM100819==RSA0009431<
logging: ACC: Barcode field not NULL, using barcode for ID POM100820==RSA0009447<
logging: ACC: Barcode field not NULL, using barcode for ID POM100821==RSA0009448<
logging: ACC: Barcode field not NULL, using barcode for ID POM100822==RSA0009446<
logging: ACC: Barcode field not NULL, using barcode for ID POM10084==RSA016907<
logging: ACC: Barcode field not NULL, using barcode for ID POM10090==RSA0076245<
logging: ACC: Barcode field not NULL, using barcode for ID POM10091A==RSA0076246<
logging: ACC: Barcode field not NULL, using barcode for ID POM10093==RSA0082380<
logging: ACC: Barcode field not NULL, using barcode for ID POM10099==RSA0012487<
logging: ACC: Barcode field not NULL, using barcode for ID POM100==RSA0002933<
logging: ACC: Barcode field not NULL, using barcode for ID POM01105==RSA0012763<
logging: ACC: Barcode field not NULL, using barcode for ID POM01115==RSA001276<
logging: ACC: Barcode field not NULL, using barcode for ID POM01183==RSA00979632<
logging: ACC: Barcode field not NULL, using barcode for ID POM01244==RSA0041947<
logging: ACC: Barcode field not NULL, using barcode for ID POM01311==RSA007880<
logging: ACC: Barcode field not NULL, using barcode for ID POM01426==RSA00450886<
logging: ACC: Barcode field not NULL, using barcode for ID POM01427==RSA0057703<
logging: ACC: Barcode field not NULL, using barcode for ID POM01430==RSA0057701<
logging: ACC: Barcode field not NULL, using barcode for ID POM01444==RSA0045087<
logging: ACC: Barcode field not NULL, using barcode for ID POM01448==RSA0057639<
Data Quality Feedback

• Misspelled taxon names
Data Quality Feedback

- Discrepancies in elevation
  - requires locating the physical specimen
Data Quality Feedback

iDigBio

- formatting
### Data Quality Feedback

**iDigBio**
- formatting
- may require Specify tech
Data Quality Feedback

iDigBio

- formatting
- may require Specify tech
- still need explanation
Data Quality Feedback

iDigBio

- formatting
- may require Specify tech
- still need explanation
- worth correcting?
Data Quality Feedback

When the data quality feedback you’re receiving is wrong:

• Example – rev_geocode_flip_both_sign
Data Quality Feedback

When the data quality feedback you’re receiving is wrong:

- Example – rev_geocode_flip_both_sign

**Specimen Record**

**Barleria lancifolia**  T. Anderson

*From Rancho Santa Ana Botanic Garden Herbarium*

- **Continent**: Africa
- **Country**: South Africa
- **Latitude**: -29
- **Longitude**: 22

**Map**

Lat: -29, Long: 22

South Africa. The Transvaal. Soutpansberg District, North of the Soutpansberg.
Data Quality Feedback

When the data quality feedback you’re receiving is wrong:

- Example – rev_geocode_flip_both_sign

<table>
<thead>
<tr>
<th>Family</th>
<th>Scientific Name</th>
<th>Date Collected</th>
<th>Country</th>
<th>Institution Code</th>
<th>Basis of Record</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acanthaceae</td>
<td>Barleria lancifolia</td>
<td>1991-05-08</td>
<td>South Africa</td>
<td>RSA</td>
<td>PreservedSpecimen</td>
</tr>
<tr>
<td>Convolvulaceae</td>
<td>Dichondra occidentalis</td>
<td>1990-03-10</td>
<td>Mexico</td>
<td>RSA</td>
<td>PreservedSpecimen</td>
</tr>
<tr>
<td>Acanthaceae</td>
<td>Barleria heterothica</td>
<td>1991-05-08</td>
<td>South Africa</td>
<td>RSA</td>
<td>PreservedSpecimen</td>
</tr>
<tr>
<td>Acanthaceae</td>
<td>Barleria geipini</td>
<td>1995-04-29</td>
<td>South Africa</td>
<td>RSA</td>
<td>PreservedSpecimen</td>
</tr>
<tr>
<td>Acanthaceae</td>
<td>Barleria bremekampi</td>
<td>1995-04-29</td>
<td>South Africa</td>
<td>RSA</td>
<td>PreservedSpecimen</td>
</tr>
</tbody>
</table>

- Only 1 of 5 records flagged indicated an actual error
Other Issues with Data Quality

CCH

- georeferencing and re-ingesting coordinates back into home institution database
Data Quality Feedback

Steps involved to correct errors (often with individual records)

- Review specimen record in database
- Verify the source of the error: Is it a data entry error or an error on label?
  - Often entails pulling the specimen from the collection
- Check feedback against other sources
  - Georeferencing (GEOLocate, Google Earth)
  - Taxonomy (Tropicos, TNRS, IPNI)
- Make the correction **IF** the feedback is correct

Total time involved: 5 – 30 minutes / specimen
Challenges & Issues

• Lack of resources:
  - time
  - funds
  - staff

• Determinations: require expertise

• Specimen is missing!
Solutions to Improving Data Quality

Minimize errors:
• Detailed data entry protocols
• Data entry training sessions
• QC checks for first few weeks
• Monthly QC checks on all data entry

DQ Feedback:
• Group errors into type
• Designated time to address errors
Solutions to Improving Data Quality

Barcoding Bonanza with Cal Poly Pomona students
Solutions to Improving Data Quality
Acknowledgements

RSA curatorial staff: Erin Berkowitz, Sierra Sutton, Lauren Tucker, Christina Varnava, Grace Rice, LeRoy Gross, Carrie Baldwin, Brandy Watts
Deb Paul, iDigBio
Joanna McCaffrey, iDigBio
Gil Nelson, iDigBio
Theresa Miller, Specify
Ed Gilbert, Symbiota