Dasyllepis insignis

Type

Labrador

12838

Liguus fasciatus austiniannus Guitart
1. Loma de la Cruz, Sancti Spiritus, Cuba.
2. Camino del Ceney, Sancti Spiritus, Cuba.

PARATYPES.

No. 153420
Type

Liguus fasciatus austiniannus Guitart
Camino del Ceney, Sancti Spiritus, Cuba.

R. Guitart / Agyja

Guitart !

Sphaerochinus grumulans (Hagen)
Naples Station, 1878.

Red. May 11, 1878

Loew Coll.
<table>
<thead>
<tr>
<th>Catalogue Number</th>
<th>Original Number</th>
<th>Name</th>
<th>Locality</th>
</tr>
</thead>
<tbody>
<tr>
<td>5255</td>
<td>1</td>
<td>Tudora megacheilis randecklepusis</td>
<td>Rende Klip, Curacao, Dutch Antilles</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>Hate Ridge</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N.B.E.
Continuous Quality Control

Classical QC
Shewhart, 1939

Total Data Quality Management
Wang, 1998

With tests for Fitness for Purpose
Fit for what purpose?
We **Guess** (at user communities and their data quality needs)

- **Alpha taxonomists**
  - Rich transcriptions of locality, citation, type status, identification history.

- **Global change researchers**
  - Current identification, georeference, date collected

- **Collection management staff**
  - Parts/Preparations, storage medium, loans, accessions, permits, higher taxonomy.
Managing MCZ data as a long term asset

Labels

Handwritten Ledgers

Data Capture

Data Cleanup

Data Enhancement

MCZbase

Agents
Journal Names
Higher Geography
Higher Taxonomy
Scientific Names

Type Citations
Georeferences
Media Objects

DataShot (O-I-D)

Spreadsheet Bulkloads

New Collections

Collections without ledgers

Muse-Fish
Muse-Herp
Muse-IP
FileMaker-Mala

FileMaker-IZ
Mantis-Ent
CollectionManager-Orn
CollectionManager-Mam
Welcome to the Museum of Comparative Zoology’s Wikipedia page for using the museum-wide database, MCZbase. Following pages include basic guides, tools, and tips to using and navigating the most common functions in MCZbase and encouraged to create additional documents as necessary listing unique department-specific protocols to be used with this resource.

MCZbase primarily holds specimen data from the ten research collections but also has the ability to database other data. For publication, all MCZbase specimen records should be cited using the institution, department, prefix and catalog number (e.g., MCZ:Ich:4693, MCZ:Mamm:Bangs-8114, MCZ:HerpOBS:2).

The site includes four sections:

- **Field Definitions**: A glossary of concepts and field headings used in MCZbase.
- **Field Definitions by Screen**: A listing of field definitions organized by the screen to which they pertain. This section is meant to serve as a field-by-field reference for staff as they carry out different management actions in MCZbase (creating loans, entering data, tracking permits, etc.).
- **How To's**: Step-by-step instructions on how to use key aspects of the database.
- **Tools**: Various tools and functionalities available within the MCZbase system.
Shared Approach to Data Quality
Reduce errors, Assert Uncertainties
Fully represent known facts

Plan
Do
Check
Act

New Data Capture
Legacy Data Capture
Georeference

Fix issues found in data
Improve Process
Training: Improve process documentation
Improved Authority Files
Improved User Interfaces

Internal
Sample
Look for Outliers
Completeness
Clean Data With Data
External Authorities
Patterns
Outliers
Measure
Data Quality Problems are easy to find.

Data Quality Problem: Out of Range

Latitude: Greater than 90 degrees.
Easy to Find Georeference Errors

• Compare coordinates (decimal latitude, decimal longitude, datum), with GIS polygon for country (or other geopolitical hierarchy units, or ocean for marine data).

• If coordinate isn't inside country, but some transformation (swap latitude/longitude, change signs of latitude/longitude) of it is, propose the transformation as a correction.
**Scelotes gronovii**

**Dasser Island, Cape Colony**

Africa: South Africa: Cape

[no verbatim date data] (1700-01-01 - 2100-01-01)

---

**South Africa is not on the Mediterranean**
Finding one problem may lead to another

Specific Locality: **Dasser** Island, Cape Colony

Coordinates: 33.424206° 18.083° (WGS84), Error: 3.9 km
CSD (MCZ); 2006-09-11; Fuzzy Gazetteer Coordinate Remarks: used **Dassen** Island ISL, extent estimated to be 1.5 km. Map scale 1: 1 500 000.
FilteredPush → Kurator

Pl: James Hanken
David B. Lowery
Paul J. Morris
Robert A. Morris
John Wieczorek

NSF: DBI #1356438 & #1356751 (Kurator)
NSF: DBI #0960535 & #0646266 (FilteredPush)

PI: Bertram Ludäscher
Timothy McPhillips
James A. Macklin

Former Project Participants
Lei Dou
Tianhong Song
Chinua Iloabachie
Sven Koehler
Maureen A. Kelly
Chuck McCallum
Donna Tremonte
Zhimin Wang

Data Quality Workflow

1. Load Data
2. Check scientific name
3. Check basisOfRecord
4. Check date collected
5. Check lat/long
6. Geolocate
7. Write out results

Repositories:
- IPNI, IF, WoRMS, COL, GBIF
- GNI
- GBIF
- HUH Botanists
- SCAN Entomo.
- Geolocate

Tools:
- ParseOptions
- CSVReader
- ScientificNameValidator
- nameValidatedRecords
- BasisOfRecordValidator
- bcorValidatedRecords
- InternalDateValidator
- dateValidatedRecords
- GEORefValidator
- geoRefValidatedRecords
- MongoSummaryWriter
- outputFilename
- nameAuthority
- inputFilename
- inputSpecimenRecords
- outputFile
- (outputFilename)
This spreadsheet represents the output of data quality control software run on a copy of harvested occurrence record data. The software runs a workflow composed of a set of "actors" each of which examines a facet of the data (one for checking scientific names, one for checking georeferences, and one for checking data collected). These actors assert that problems exist in the data (which in some cases are real problems, and in other cases are not), and can each propose a change to correct the problem.

<table>
<thead>
<tr>
<th>Date Collected</th>
<th>Determiner</th>
<th>Scientific Name</th>
<th>Scientific Name Authorship</th>
<th>Taxon Id</th>
<th>Family</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1920-11-02</td>
<td>Catalog</td>
<td>Hemidactylum scutatum</td>
<td>(Temminck &amp; Schlegel)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1959-04-26</td>
<td>Catalog</td>
<td>Ambystoma tigrinum</td>
<td>(Shaw, 1802)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1965-05-27</td>
<td>Catalog</td>
<td>Lithobates clamitans melanota</td>
<td>(Rafines.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1904-08-16</td>
<td>Catalog</td>
<td>Plethodon cinereus</td>
<td>(Green, 1818)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1987-03-01</td>
<td>Catalog</td>
<td>Plethodon glutinosus complex</td>
<td>(Green, 1818)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1954-01-01</td>
<td>[no agent data]</td>
<td>Gamma narida</td>
<td>[Green, 1818]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date Collected</th>
<th>Scientific Name</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1873-01-01/1924-08-16</td>
<td>Placopecten magellanicus</td>
<td>we have proposed this change</td>
</tr>
</tbody>
</table>

**DateValidator:**
- no change needed; looks good to us
- there seems to be a problem, but we don't know how to solve it
- no change needed; looks good to us
- there seems to be a problem, but we don't know how to solve it
- there seems to be a problem, but we don't know how to solve it
- there seems to be a problem, but we don't know how to solve it
- there seems to be a problem, but we don't know how to solve it
- there seems to be a problem, but we don't know how to solve it
- we have proposed this change

**GeoRefValidator:**
- don't know
- no change needed; looks good to us
- no change needed; looks good to us
- no change needed; looks good to us
- no change needed; looks good to us
- no change needed; looks good to us
- no change needed; looks good to us
- no change needed; looks good to us
- there seems to be a problem, but we don't know how to solve it
- there seems to be a problem, but we don't know how to solve it
- no change needed; looks good to us
- we have proposed this change
Placopecten magellanicus  Gmelin, 1791

Found accepted name Placopecten magellanicus
Source: Catalog of Life.
Authorship: Differs only in Parentheses
Authorship Similarity: 0.833

Where is the Defect: Implications

● Database?
  – Fix It. (data curator); Change database application to prevent this class of error? (developer); Develop policy and training to prevent the defect.

● Migration of Data from Legacy Database?
  – Find the other records with this problem (DBA); Review the other records with this problem (data curator)

● Mapping to Darwin Core for Aggregation?
  – Fix the mapping (DBA).

● Defect in the DQ Software?
  – Fix the bug (developer).

● Error in the authority consulted?
  – Notify the authority.
Bartonia tenella

Incorrect correction proposed: Bartonia tenella Muhl. ex Willd.

The International Plant Names Index

Plant Name Details

Gentianaceae Bartonia tenella Willd.

Neue Schriften Ges. Naturf. Freunde Berlin iii. 445. 1801

Id: 366463-1 Version: 1.1.2.3.1.2.2.1 Hide Record history

Record history:
- root: Initial value made on 2003-07-02 00:00:00.0
- matt: (A) Auto fixing ex made on 2003-12-04 15:37:39.0
- Challis: (A) Auto fix. changing Willd. to Willd. - Carl Ludwig von Willdenow 1765-1
- Barker: (A) batch fix to remove title variants prior to standardization made on 200
- Barker: (A) Auto fix. changing Muhl. to Muhl. - Gotthilf Henry Ernest Muhlenberg 
- Barker: (A) Auto fix changing in Ges. Naturf. Fr. Neue Schr. to Neue Schriften Gr
- Gandhi: (A) GCI Authorship revised made on 2016-01-27 23:21:11.0

Gandhi: (A) GCI Authorship revised made on 2016-01-27
<table>
<thead>
<tr>
<th>Decimal Latitude</th>
<th>Decimal Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>42.45409393</td>
<td>-71.10149384</td>
</tr>
<tr>
<td>41.7625</td>
<td>-70.7222</td>
</tr>
<tr>
<td>42.45409393</td>
<td>WAS: 71.10149384; CHANGED TO: -71.10149384</td>
</tr>
<tr>
<td>41.7625</td>
<td>WAS: 71.10149384; CHANGED TO: -71.10149384</td>
</tr>
<tr>
<td>55.565922</td>
<td>-62.0727539</td>
</tr>
<tr>
<td>WAS: -42.45409393; CHANGED TO: 42.45409393</td>
<td>WAS: 71.10149384; CHANGED TO: -71.10149384</td>
</tr>
<tr>
<td>41.7625</td>
<td>-70.7222</td>
</tr>
</tbody>
</table>

**Black Box Correction**

decimalLatitude: Was: -42.45409 Changed to: 42.45409
decimalLongitude: Was 71.10149 Changed to: -71.10149

FP-Akka GeoreferenceValidator
Internals of a Data Quality mechanism can be complex.
decimalLatitude: Was: -42.45409 Changed to: 42.45409
decimalLongitude: Was 71.10149 Changed to: -71.10149

| coordinates are on the Earth's surface. | Coordinates not inside country. | transposed/sign changed coordinates to place inside the provided Country UNITED STATES | Transposed/sign changed coordinates are near (within 200.0 km) georeference of locality from the Geolocate service.

(1) coordinates are on the Earth's surface.

Criterion in context: Latitude and longitude form a valid coordinate.
Specification: decimalLatitude >= -90 and <= 90;
                decimalLongitude >= -180 and <= 180
Mechanism: FP-Akka georeferenceValidator
Assertion: Compliant
(2) Coordinates not inside country

**Criterion in context:** decimalLatitude and decimalLongitude fall within the boundaries of the country.

**Specification:** The coordinate specified by decimalLatitude, decimalLongitude and datum falls within a polygon for the country as found in the Country vector data from Natural Earth, if a country is provided and if no ocean is provided.

**Mechanism:** FP-Akka georeference validator, using Natural Earth country vector GIS layer.

**Assertion:** Not Compliant.
transposed/sign changed coordinates to place inside the provided Country UNITED STATES

Criterion in context: If test (2) failed, does any combination of transpositions or sign changes of decimalLatitude and decimalLongitude place the coordinate inside the provided country.

Specification: Create a list of all of the transpositions and sign changes of decimalLatitude and decimalLongitude. Exclude any transposition that is not a valid coordinate. Return a list of each transposition and sign change that falls within the polygon for the country as found in the Country vector data from Natural Earth, if a country is provided and if no ocean is provided.

Mechanism: FP-Akka georeference validator, using Natural Earth country vector GIS layer.

Assertion: The sign changed coordinate 42.454 -71.101 falls within the polygon for the United States.
Criterion in Context: If test (2) failed, and test (3) returned one or more results, does any combination of transpositions or sign changes of decimalLatitude and decimalLongitude place the coordinate within a configurable (20 km default) distance of the coordinate returned from a query on the GeoLocate web service for the country, stateProvince, county and specificLocality?

Specification: Query the GeoLocate webservice with the provided country, county, stateProvince, and specificLocality. Compare this result with each valid transposition of decimalLatitude and decimalLongitude that lies within the provided country. If any transposition lies within the configurable distance from a high confidence GeolocateResult, return that transposition of the original coordinates as a proposed correction.

Mechanism: FP-Akka georeference validator, using the GeoLocate web service.

Assertion: The sign changed coordinate 42.45409 -71.10149 falls within 20 km of a Geolocate georeference for the the provided locality.
Structured data about data quality lets us produce targeted reports on data quality issues, and gives the consumer of the report a chance to drill into the details (for the cases where the details are important)
Tools can have differing opinions

<table>
<thead>
<tr>
<th>State/Province</th>
<th>Locality</th>
<th>Decimal Latitude</th>
<th>Decimal Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sabah</td>
<td>8 km S Sapulut</td>
<td>4.62851</td>
<td>116.47158</td>
</tr>
<tr>
<td>Galápagos</td>
<td>Galápagos Islands</td>
<td>-0.76951</td>
<td>-85.91155</td>
</tr>
<tr>
<td>Galápagos</td>
<td>Galápagos Islands</td>
<td>-0.76969</td>
<td>-85.91183</td>
</tr>
<tr>
<td>Galápagos</td>
<td>Galápagos Islands</td>
<td>-0.76965</td>
<td>-85.91153</td>
</tr>
<tr>
<td></td>
<td>Approx. 105 mi. N. of Cabo, Orange</td>
<td>WAS: -5.475; CHANGED TO: 5.475</td>
<td>-51.35833</td>
</tr>
<tr>
<td></td>
<td>Approx. 105 mi. N. of Cabo, Orange</td>
<td>WAS: -5.475; CHANGED TO: 5.475</td>
<td>-51.35833</td>
</tr>
<tr>
<td>Ogooué-Maritime</td>
<td>Moukalaba-Doudoug National Park Buffer Zone</td>
<td>-3.08247</td>
<td>10.43732</td>
</tr>
<tr>
<td></td>
<td>about 30 miles SSE of Joao Pessoa</td>
<td>-7.15833</td>
<td>-34.425</td>
</tr>
<tr>
<td></td>
<td>Approx. 105 mi. N. of Cabo, Orange</td>
<td>WAS: -5.475; CHANGED TO: 5.475</td>
<td>-51.35833</td>
</tr>
<tr>
<td></td>
<td>Approx. 105 mi. N. of Cabo, Orange</td>
<td>WAS: -5.475; CHANGED TO: 5.475</td>
<td>-51.35833</td>
</tr>
<tr>
<td>Galápagos</td>
<td>Galápagos Islands</td>
<td>-0.76951</td>
<td>-85.91155</td>
</tr>
<tr>
<td></td>
<td>Approx. 105 mi. N. of Cabo, Orange</td>
<td>WAS: -5.475; CHANGED TO: 5.475</td>
<td>-51.35833</td>
</tr>
<tr>
<td></td>
<td>Approx. 105 mi. N. of Cabo, Orange</td>
<td>WAS: -5.475; CHANGED TO: 5.475</td>
<td>-51.35833</td>
</tr>
<tr>
<td>Cochaiba</td>
<td>Cocapata</td>
<td>-16.95</td>
<td>-66.71667</td>
</tr>
<tr>
<td>Mazaruni-Potaro</td>
<td>Potaro River at Tukeit Fall</td>
<td>WAS: 5.21667; CHANGED TO: -5.21667</td>
<td>-59.41667</td>
</tr>
<tr>
<td>Nord-Kivu</td>
<td>Forest, Lulenga -River Road</td>
<td>WAS: 2.41306; CHANGED TO: -2.41306</td>
<td>29.37306</td>
</tr>
<tr>
<td></td>
<td>Approx. 105 mi. N. of Cabo, Orange</td>
<td>WAS: -5.475; CHANGED TO: 5.475</td>
<td>-51.35833</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Napo</td>
<td>Rio Pucuno</td>
<td>-0.8</td>
<td></td>
</tr>
<tr>
<td>Sabah</td>
<td>Batu Punggul, 16 km ESE Sapulut</td>
<td>4.64045</td>
<td>116</td>
</tr>
</tbody>
</table>

**MCZ data in GBIF**
Flagged as: Latitude inverted
FP-Akka disagrees
Interoperaibility problem
Herpetology A-125892

previous number: MCZ-A-35781-83

Anelides aeneus (Cope and Packard, 1881)

Country: United States
State/Province: North Carolina
County: Macon

Specific Locality: Macon near Highlands

Collecting Source: wild caught

Remarks: Original entry had #MCZ-A-35781-83, but these belong to Eleutherodactylus abboti, so recatalogued x .95, LAT

Accession: 5139

Media Details
MCZ Herpetology
Amphibia 125001-129950,
pg. 36

[Report Bad Data]
The Bernardo Assertion

:Evidence_0 a oad:Evidence ;
  cnt:chars "These are clearly incorrect records. [This species does not occur in the Dominican Republic]"@en .

:Expectation_0 a oad:Expectation_Solve_With_More_Data .
I have identified a problem in your data.
I can't tell what the solution should be.

It will take research on the recipient's part to identify the cause of the problem and to resolve the problem.

Effort is Assymetrical: Easy to find, hard to resolve.
Data Quality problems are easy to find in large numbers.

There can be too many issues

- 1,780,781 MCZ occurrence records
- 533,342 ScientificNameValidator: Curated
- 7,176 DateValidator: Curated
MCZ Malacology taxon name cleaning: Units of Work

- Database Application
- CSV list of taxa
- FP-Akka
- QC Report CSV File
- QC Report
- SQL Update Queries
- Manual edits by data curator
- Select subsets of the report
Exported data for input to FP-Akka

dbpk,scientificName,scientificNameAuthorship
38422,"Tectarius muricatus","(Linnaeus, 1758)"
56443,"Pteria colymbus","(Roding, 1798)"
38512,"Tellina radiata","Linnaeus, 1758"
38427,"Engina turbinella","(Kiener, 1835)"
38523,"Volachlamys tranquvebaria","(Gmelin, 1791)"
77101,"Conus ximenes","Gray, 1839"
FP-Akka output

- "dbpk","scientificName","authorship","guid","status","sciNameWas","sciNameAuthorshipWas","provenance"

- "77101","Conus ximenes","Gray, 1839","urn:lsid:marinespecies.org:taxname:428413","Valid","","","Found **exact match in WoRMS.** | Authorship: Exact Match Similarity: 1.0"

- "56443","Pteria colymbus","(Röding, 1798)","urn:lsid:marinespecies.org:taxname:420735","Curated","","","Found **plausible match in WoRMS:** Similar Author, Year Exact | Authorship: Similar Author, Year Exact Similarity: 0.917"
MCZ Malacology Taxon Name Checking

- Taxonomy table – authority file for scientific names.
- Task: Clean up scientific names in Malacology after migration of legacy data into MCZbase.
- **First try:** Run All Probably marine extant mollusk names in MCZbase against WoRMS, give a list of ~15,000 proposed changes to the collection staff.
- Two years later, issue is still open in issue tracking system….
MCZ Malacology Taxon Name Checking

- **Second try**: Run All Probably marine extant mollusk names in MCZbase against WoRMS.
- Identified specific sets of similar problems with proposed corrections, a few hundred records each.
- Gave first set: We had a Sowerby as an author, WoRMS asserted which of the 4 Sowerbys that was.
- Two weeks later, first batch corrected by the collection, gave them second batch.
Finding units of work through rich QC assertions

- "39116", "Vexillum rubrum", "(Broderip, 1836)", "urn:lsid:marinespecies.org:taxname:751919", "Curated", "", "Broderip, 1836", "| can't construct sciName from atomic fields | Found plausible match in WoRMS: Differ only in Parentheses | Authorship: Similarity: 0.8571428571428571

grep "Differ only in Paren" output.csv > output_paren.csv
Structured Data about Data Quality issues allows us to break a large set of issues into trackable units of work, each forming a set of similar problems.
Fit for what purpose?
We Guess (at user communities and their data quality needs)

• Alpha taxonomists
  – Rich transcriptions of locality, citation, type status, identification history.

• Global change researchers
  – **Current identification**, georeference, date collected

• Collection management staff
  – Parts/Preparations, storage medium, loans, accessions, permits, higher taxonomy.
IDs by experts