Improving Natural Science Collections Data through Quality Control for Research using Workflows Embedded in a FilteredPush network

James A. Macklin

Agriculture and Agri-Food Canada

Agriculture et Agroalimentaire Canada

James Hanken, Maureen Kelly, Sven Koehler, David Lowery, Bertram Ludaescher, Paul J. Morris, Robert A. Morris, Tianhong Song

Museum of Comparative Zoology

Supported by NSF DBI-0960535
Outline

• FilteredPush architecture
• Quality control issues in specimen and observation data
• Kurator
  – Issues and solutions
  – Technology
Detecting QC Issues
Taxonomic QC

- Does scientific name exist?
- Is composition and spelling correct?
- Is the author string present and/or correct?
- Is it an accepted name or synonym according to an authoritative resource?
- Is the name ambiguous i.e., represents more than one organism (homonyms)?
Geographic QC

- If a georeference is not present, can we assert one?
- Is the latitude and longitude bounded by the record's country and/or state/province data?
- Do the coordinates have a terrestrial organism mapped to an aquatic environment?
- Are the coordinates bound by the known distribution of the organism?
Date/Time QC

- Do the date/times conform to a standard (ISO 8601)?
- Does the verbatim date agree with the structured date?
- Was the collector alive when the collecting event took place?
- Was the organism at a given life-stage when the collecting event occurred (phenological)?
- Does the collection number correspond to an acceptable time range for the collector?
Plan → Do → Check → Act

Classical QC
Shewhart, 1939

Define → Measure → Analyze → Improve

Total Data Quality Management
Wang, 1998

Continuous Quality Control
# Kepler Kurator

## Input Data Set

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TaxonID</td>
<td>ScientificName</td>
<td>HigherTaxon</td>
<td>TaxonRank</td>
<td>TaxonAccNumber</td>
<td>NamePublicName</td>
<td>TaxonomicName</td>
</tr>
<tr>
<td>2</td>
<td>105838</td>
<td>Pontoporia</td>
<td>105837</td>
<td>Species</td>
<td>valid</td>
<td>Hershkovitz, P., Cat...</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>105830</td>
<td>Inia d'Orb...</td>
<td>113648</td>
<td>Genus</td>
<td>valid</td>
<td>Hershkovitz, P., Cat...</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>105831</td>
<td>Inia geoffr...</td>
<td>105830</td>
<td>Species</td>
<td>valid</td>
<td>Hershkovitz, P., Cat...</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>105839</td>
<td>Lipotes vex...</td>
<td>105832</td>
<td>Species</td>
<td>valid</td>
<td>Hershkovitz, P., Cat...</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>113518</td>
<td>Susu Less</td>
<td>113648</td>
<td>Genus</td>
<td>valid</td>
<td>Hershkovitz, P., Cat...</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>155044</td>
<td>Susu</td>
<td>113518</td>
<td>Species</td>
<td>valid</td>
<td>Hershkovitz, P., Cat...</td>
<td></td>
</tr>
</tbody>
</table>

Keywords: SPNHC 2011, Filtered Push, Kepler, Specimen Curation

[http://www.youtube.com/watch?v=DEkPbvLsud0](http://www.youtube.com/watch?v=DEkPbvLsud0)
Embedded Workflow

- **ComadDirector**
  - DataCacheDir: property("KURATION_SPECIMENCURATION_DIR") + "cache"
  - FuncDir: property("KURATION_SPECIMENCURATION_DIR") + "function"
  - VisualizationFile: property("KURATION_SPECIMENCURATION_DIR") + "function/visua..."

Diagram:
- MongoDBCollectionReader
- ScientificNameValidator
- GEORefValidator
- FPAnnotationInserter
- MongoDBCollectionWriter
- MongoCurationSummaryWriter
Launch Quality Control Workflow on Harvested Data

**Quality Control My Data:**

**Analysis:** Specify query parameters, select a workflow, and run.

**Query Parameters:**
- Property: Institution Code
- Value: NAU

**Workflow:**
- SpecimenCurationTaxonGEO-1

**Run**
Quality Controlled Result Set

Workflow http://f1.acis.ufl.edu:8080/fedora/objects/changeme:6529/datastreams/XML/content has been run!
Message id: 7714e55-ad61-4397-812d-4ad03afcb319

Quality Control My Data:

Analysis: Specify query parameters, select a workflow, and run.

Query Parameters:
Property: Institution Code
Value: NAU
Add

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>institutionCode</td>
<td>NAU</td>
<td>Delete</td>
</tr>
</tbody>
</table>

Workflow:
SpecimenCurationTaxonGEO-1

Run

Summary of analysis (defaults to the last run or you may select a previous result from the list below):

<table>
<thead>
<tr>
<th>Institution Code</th>
<th>Collection Code</th>
<th>Catalog Number</th>
<th>Family</th>
<th>Scientific Name</th>
<th>Scientific Name Authorship</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAU</td>
<td>NAUF</td>
<td>NAU4F A0010139</td>
<td>Curculionidae</td>
<td>Trigonoscuta yumaensis Pierce</td>
<td></td>
<td>32.72528</td>
<td>-114.485613</td>
<td>United States</td>
</tr>
<tr>
<td>NAU</td>
<td>NAUF</td>
<td>NAU4FA0010142</td>
<td>Curculionidae</td>
<td>Ericydeus lautus (LeConte, 1856)</td>
<td></td>
<td>-109.153673</td>
<td>33.895516</td>
<td>USA</td>
</tr>
<tr>
<td>NAU</td>
<td>NAUF</td>
<td>NAU4FA0010140</td>
<td>Curculionidae</td>
<td>Trigonoscuta yumaensis Pierce</td>
<td></td>
<td>32.72528</td>
<td>-114.62361</td>
<td>USA</td>
</tr>
<tr>
<td>NAU</td>
<td>NAUF</td>
<td>NAU4FA0010141</td>
<td>Curculionidae</td>
<td>Ericydeus lautus (LeConte, 1856)</td>
<td></td>
<td>31.895516</td>
<td>-109.153673</td>
<td>USA</td>
</tr>
</tbody>
</table>

Export as CSV  Export as XLS
With Workflow Metadata

### Quality Control My Data:

**Analysis:** Specify query parameters, select a workflow, and run.

**Query Parameters:**
- **Property:** Institution Code
- **Value:** NAU

**Workflow:**
- SpecimenCurationTaxonGEO-1

**Run**

Summary of analysis (defaults to the last run or you may select a previous result from the list below):

| Institution Code | Collection Code | Catalog Number | Family               | Scientific Name               | Scientific Name Authorship | Latitude      | Longitude     | Country      | State/Province |
|------------------|-----------------|----------------|----------------------|-------------------------------|----------------------------|---------------|---------------|--------------|----------------|                |
| NAU              | NAUF            | NAU4F A00100139 | CurculionidaeTrigonoscuta yumaensisPierce |                             |                            | 32.725204     | -114.485613   | United States | Arizona        |
| NAU              | NAUF            | NAU4FA0010142  | CurculionidaeEricydeus lautus (LeConte, 1856) | 109.1636733 13.896316 | USA Arizona Sunny F |

**Actor Name**
- ScientificNameValidator
- GEORefValidator

**Actor Run**
- ?

| Actor Name | Actor Run | Actor Result | Institution Code | Collection Code | Catalog Number | Family               | Scientific Name               | Scientific Name Authorship | Latitude | Longitude | Country | State/Province |
|------------|-----------|--------------|------------------|-----------------|---------------|----------------------|-------------------------------|---------------------------|----------|------------|---------|----------------|                |
| NAU        | NAUF      |              | NAU4F A00100139 | CurculionidaeTrigonoscuta yumaensisPierce |                             | 32.725204     | -114.485613   | United States | Arizona        |
| NAU        | NAUF      |              | NAU4FA0010140  | CurculionidaeEricydeus lautus (LeConte, 1856) | 109.1636733 13.896316 | USA Arizona Sunny F |

Export as CSV Export as XLS
# Quality Control Result

## Workflow Actor

<table>
<thead>
<tr>
<th>Institution Code</th>
<th>Collection Code</th>
<th>Catalog Number</th>
<th>Family</th>
<th>Scientific Name</th>
<th>Scientific Name Authorship</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAUF</td>
<td>NAU4F A0010139</td>
<td></td>
<td></td>
<td>Curculionidae</td>
<td>Trigonomus yumaensis</td>
<td>Pierce</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAUF</td>
<td>NAUFA0010140</td>
<td></td>
<td></td>
<td>Curculionidae</td>
<td>Ericydeus laurus</td>
<td></td>
<td>-109.163673</td>
<td>31.896516</td>
</tr>
<tr>
<td>NAUF</td>
<td>NAU4F 0010141</td>
<td></td>
<td></td>
<td>Curculionidae</td>
<td>Trigonomus yumaensis</td>
<td>Pierce</td>
<td>32.72528</td>
<td>-114.82361</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Actor Name</th>
<th>Actor Run</th>
<th>Actor Result</th>
<th>Institution Code</th>
<th>Collection Code</th>
<th>Catalog Number</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScientificNameValidator</td>
<td>?</td>
<td></td>
<td>NAUF</td>
<td>NAU4F A0010139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEORefValidator</td>
<td>X</td>
<td></td>
<td>NAUF</td>
<td>NAUFA0010140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEORefValidator</td>
<td>X</td>
<td></td>
<td>NAUF</td>
<td>NAU4F 0010141</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Export as CSV](#) [Export as XLS](#)
<table>
<thead>
<tr>
<th>Institution Code</th>
<th>Collection Code</th>
<th>Catalog Number</th>
<th>Family</th>
<th>Scientific Name</th>
<th>Scientific Name Authorship</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAU</td>
<td>NAUF</td>
<td>NAU4F 0010139</td>
<td>Curculionidae</td>
<td>Trigonoscuta yumaensis</td>
<td>Pierce</td>
<td>32.725204</td>
<td>-114.485613</td>
<td>United States</td>
</tr>
</tbody>
</table>

The georeference has been corrected for a transposition error. It is inconsistent with the text-based locality data and needs human review for validation.
Annotations with corrections available for ingest

### Annotations

<table>
<thead>
<tr>
<th>Inbox (1)</th>
<th>Trash</th>
<th>NAU (1)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Curator</th>
<th>Summary</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kepler Workflow System Solve With More Data: NAU4FA00101422013-07-03T17:12:22EDT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kepler Workflow System New Georeference: NAU4F A0010139 2013-07-03T17:12:20EDT</td>
<td></td>
</tr>
</tbody>
</table>

---

**Museum of Comparative Zoology**

**Harvard University**

**Harvard Herbaria**

**UMass Boston**

**NSF**

**NSF DBI #0646266 #0960535**

**Agriculture and Agri-Food Canada**

**Agriculture et Agroalimentaire Canada**
Annotation: Corrected Georeference
List of services

• External:
  – Global Names Index (GNI)
  – GBIF checklist bank (IPNI, WORMS, Index Fungorum, etc.)
  – GeoLocate

• Internal:
  – GBIF Cache
  – Land-based polygons (Natural Earth); political polygons (Geocommunity)
  – Harvard Index of Botanists
  – Flora of North America phenology data
Issues and Solutions

- Response time of services
  (varies and we cannot control this!)
  - Caching
  - Parallelization

- Available services (not enough!)
  - Internalize

- Curation requiring human intervention
  - Asynchronous processing
  - Syndication based on “interests”
Technology

- **Kepler** scientific workflow platform
  - Run headless (no UI)
    - **COMAD** (collection-oriented model of computation)
      - “Conveyer-belt” like pipeline of data objects
    - **AKKA**
      - Workflow parallelization engine
  - **Provenance**
    - Provenance browser
    - Spreadsheet visualization
Acknowledgements and Resources

Funding: NSF DBI-0960535

Project Wiki: http://wiki.filteredpush.org/

Code Repository: http://filteredpush.sourceforge.net/