Data Cleaning for Analysis and Publication
Using the OpenRefine Software Package

Arctic Data Center, CUAHSI, DataONE, Environmental Data Initiative, GBIF, iDigBio, NEON, Neotoma

Jeanette Clark, Deborah Paul

#datahelpdesk
Ecological Society of America 2020
Career Central
August
The Arctic Data Center
https://arcticdata.io

Data Archive
Portal for data discovery
Tools for data and metadata submission
Support services
Training and Outreach
Data Rescue

5,800 DATA SETS
725K DATA FILES
34 TB DATA STORAGE
990K+ FILE DOWNLOADS

1,500 CREATORS
14,000 USERS

Cumulative Data Objects

720,906 data files
image/jpeg 210,551
image/jpeg 110,151
image/jpeg 129,871
audio/wav

@arcticdatactr
Advancing the Digitization of Biological Collections
iDigBio Hub and Thematic (Museum) Collection Networks

Digitization
Workflows & Protocols
Dissemination

Research Use
Cyberinfrastructure
Tool collaboration
Portal development
ENM workshop
Research focus
Data quality
APIs

Training
Biodiversity informatics
Data skills and literacy
Collections software
Imaging
Project Management

total: 121,428,342

Education Outreach
Citizen Science
K-12 materials
Undergraduate
Fossil Clubs
Mentor teachers

Methods
Workshops
Webinars
Symposia
Conferences
Working Groups
Short Courses
Adobe Connect
Listservs
Publications
Social Media @idigbio
What do we mean by “Clean” Data?

More Data from More Sources =

- Structural Issues
- Inconsistent/unclear missing values
- Mixed data in single columns
- Mixed data types in single columns
- Ambiguous data values
- Data you can’t use
Simple guidelines for data management

• Use a scripted program
• Nonproprietary formats
• Keep a raw version of data
• Descriptive names
• Header line
• Plain ASCII text

Simple Guidelines for Data Management

• Design to add rows, not columns

• Each column should contain only one type of information

• Record a single piece of data only once; separate information collected at different scales into different tables. In other words, create a relational database.

Recognizing untidy data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>species</td>
<td>tree</td>
<td>main trunks</td>
<td>reiterated trunks</td>
<td>limbs</td>
<td>branches</td>
<td>leaves</td>
<td>type</td>
<td>species</td>
<td>main trunk</td>
<td>reiteration</td>
<td>limbs</td>
<td>branch</td>
<td>leaf</td>
<td>TOTAL</td>
<td>% total</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Atlas</td>
<td>255144.9</td>
<td>46020.6</td>
<td>5477.7</td>
<td>13433.2</td>
<td>1101.2</td>
<td>tree</td>
<td>SESE</td>
<td>3569312</td>
<td>213247</td>
<td>53714</td>
<td>230945</td>
<td>17192</td>
<td>4084409</td>
<td>95.3491</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Ballantine</td>
<td>221966.4</td>
<td>7651.6</td>
<td>5922.9</td>
<td>11210.0</td>
<td>1084.8</td>
<td>tree</td>
<td>SESE</td>
<td>135815</td>
<td>0</td>
<td>0</td>
<td>8338</td>
<td>961</td>
<td>45114</td>
<td>3.3876</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Bell</td>
<td>253246.4</td>
<td>5454.3</td>
<td>5792.6</td>
<td>48500.7</td>
<td>1043.4</td>
<td>tree</td>
<td>THSE</td>
<td>31799</td>
<td>0</td>
<td>0</td>
<td>6343</td>
<td>864</td>
<td>39006</td>
<td>0.9106</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Broken Top</td>
<td>130282.9</td>
<td>4805.2</td>
<td>1608.1</td>
<td>5137.4</td>
<td>729.9</td>
<td>tree</td>
<td>ACMA</td>
<td>4444</td>
<td>0</td>
<td>0</td>
<td>925</td>
<td>264</td>
<td>5634</td>
<td>0.1315</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Buena Vista</td>
<td>128833.0</td>
<td>3486.5</td>
<td>0</td>
<td>8552.1</td>
<td>518.4</td>
<td>tree</td>
<td>UMCA</td>
<td>2921</td>
<td>0</td>
<td>0</td>
<td>937</td>
<td>273</td>
<td>4131</td>
<td>0.0964</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Demeter</td>
<td>155896.0</td>
<td>11085.6</td>
<td>3204.3</td>
<td>10054.1</td>
<td>768.7</td>
<td>shrub</td>
<td>RUSP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1974</td>
<td>686</td>
<td>2660</td>
<td>0.0620</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Epimetheus</td>
<td>226987.0</td>
<td>12915.7</td>
<td>1797.2</td>
<td>13885.2</td>
<td>1029.4</td>
<td>shrub</td>
<td>POMU</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1271</td>
<td>1271</td>
<td>0.0296</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Ilu vatari</td>
<td>339586.6</td>
<td>65003.9</td>
<td>12315.6</td>
<td>13987.0</td>
<td>1461.8</td>
<td>shrub</td>
<td>VAOV</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>528</td>
<td>26</td>
<td>552</td>
<td>0.0129</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Kronos</td>
<td>134154.1</td>
<td>12204.4</td>
<td>7232.7</td>
<td>5036.1</td>
<td>597.3</td>
<td>shrub</td>
<td>COCO</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>284</td>
<td>6</td>
<td>289</td>
<td>0.0067</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Pleiades I</td>
<td>182385.2</td>
<td>3735.0</td>
<td>1935.2</td>
<td>10846.6</td>
<td>762.2</td>
<td>shrub</td>
<td>POSC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>107</td>
<td>89</td>
<td>196</td>
<td>0.0045</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Pleiades II</td>
<td>235838.8</td>
<td>11183.4</td>
<td>4306.0</td>
<td>11306.5</td>
<td>877.7</td>
<td>shrub</td>
<td>RPHU</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>18</td>
<td>162</td>
<td>0.0037</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Prometheus</td>
<td>239414.0</td>
<td>25228.9</td>
<td>1612.6</td>
<td>12458.2</td>
<td>1086.0</td>
<td>shrub</td>
<td>XOXI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>112</td>
<td>112</td>
<td>0.0026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Rhea</td>
<td>143710.4</td>
<td>487.8</td>
<td>730.1</td>
<td>5524.2</td>
<td>691.2</td>
<td>shrub</td>
<td>VAPA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>94</td>
<td>4</td>
<td>99</td>
<td>0.0023</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Zeus</td>
<td>243365.7</td>
<td>2885.5</td>
<td>1620.4</td>
<td>19104.7</td>
<td>954.3</td>
<td>shrub</td>
<td>PISI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>3</td>
<td>1781.3</td>
<td>0.0</td>
<td>0</td>
<td>87.6</td>
<td>41.4</td>
<td>shrub</td>
<td>CHLA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>4</td>
<td>6312.0</td>
<td>356.3</td>
<td>73.5</td>
<td>214.1</td>
<td>43.8</td>
<td>shrub</td>
<td>GASH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>5</td>
<td>2068.2</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
<td>2.5</td>
<td>shrub</td>
<td>SAC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>6E</td>
<td>18697.4</td>
<td>0.0</td>
<td>0</td>
<td>1056.2</td>
<td>66.3</td>
<td>shrub</td>
<td>SESE geo</td>
<td>3569312</td>
<td>213247</td>
<td>53714</td>
<td>230945</td>
<td>17192</td>
<td>4084409</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>6W</td>
<td>14651.5</td>
<td>7.7</td>
<td>0</td>
<td>626.3</td>
<td>49.6</td>
<td>shrub</td>
<td>SESE epi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>12</td>
<td>232.1</td>
<td>0.0</td>
<td>0</td>
<td>11.2</td>
<td>10.3</td>
<td>shrub</td>
<td>SESE geo</td>
<td>3569312</td>
<td>213247</td>
<td>53714</td>
<td>230945</td>
<td>17192</td>
<td>4084409</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>18</td>
<td>15632.0</td>
<td>0.0</td>
<td>0</td>
<td>946.3</td>
<td>106.8</td>
<td>shrub</td>
<td>PSME geo</td>
<td>135815</td>
<td>0</td>
<td>0</td>
<td>8338</td>
<td>961</td>
<td>145114</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>19</td>
<td>11805.5</td>
<td>0.0</td>
<td>0</td>
<td>770.1</td>
<td>80.3</td>
<td>shrub</td>
<td>PSME epi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8338</td>
<td>961</td>
<td>145114</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>20</td>
<td>309.5</td>
<td>0.0</td>
<td>0</td>
<td>12.5</td>
<td>5.9</td>
<td>shrub</td>
<td>TSHE geo</td>
<td>31740</td>
<td>0</td>
<td>0</td>
<td>6332</td>
<td>860</td>
<td>38932</td>
<td>0.9900</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>22</td>
<td>25618.3</td>
<td>0.0</td>
<td>0</td>
<td>1504.0</td>
<td>120.2</td>
<td>shrub</td>
<td>TSHE epi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>
Characteristics of Tidy Data

Observations

● Separate tables for each entity measured
Recognizing untidy data

Table 1

<table>
<thead>
<tr>
<th>species</th>
<th>tree</th>
<th>main trunks kg</th>
<th>reiterated trunks kg</th>
<th>limbs kg</th>
<th>branches kg</th>
<th>leaves kg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SESE</td>
<td>Atlas</td>
<td>256144.9</td>
<td>46020.6</td>
<td>5477.7</td>
<td>13433.2</td>
<td>1101.2</td>
</tr>
<tr>
<td>Ballantine</td>
<td>221966.4</td>
<td>7651.6</td>
<td>5922.9</td>
<td>11210.0</td>
<td>1084.8</td>
<td></td>
</tr>
<tr>
<td>Bell</td>
<td>253246.2</td>
<td>5454.3</td>
<td>5792.6</td>
<td>48500.7</td>
<td>1043.4</td>
<td></td>
</tr>
<tr>
<td>Broken Top</td>
<td>130928.9</td>
<td>4805.2</td>
<td>1608.1</td>
<td>5137.0</td>
<td>729.9</td>
<td></td>
</tr>
<tr>
<td>Buena Vista</td>
<td>128333.0</td>
<td>3486.5</td>
<td>0.0</td>
<td>8552.1</td>
<td>518.4</td>
<td></td>
</tr>
<tr>
<td>Demeter</td>
<td>155896.0</td>
<td>1108.6</td>
<td>3204.3</td>
<td>10054.1</td>
<td>768.7</td>
<td></td>
</tr>
<tr>
<td>Ephemetheus</td>
<td>226987.0</td>
<td>12915.7</td>
<td>1797.2</td>
<td>13585.2</td>
<td>1029.4</td>
<td></td>
</tr>
<tr>
<td>Iluvatar</td>
<td>349586.6</td>
<td>65003.9</td>
<td>12315.6</td>
<td>13987.0</td>
<td>1461.8</td>
<td></td>
</tr>
<tr>
<td>Kronos</td>
<td>134154.1</td>
<td>12204.4</td>
<td>7232.7</td>
<td>5036.1</td>
<td>597.3</td>
<td></td>
</tr>
<tr>
<td>Pleiades I</td>
<td>182385.2</td>
<td>3735.0</td>
<td>1935.2</td>
<td>10846.6</td>
<td>762.2</td>
<td></td>
</tr>
<tr>
<td>Pleiades II</td>
<td>235838.6</td>
<td>1118.4</td>
<td>4306.0</td>
<td>11306.5</td>
<td>877.7</td>
<td></td>
</tr>
<tr>
<td>Prometheus</td>
<td>239414.0</td>
<td>25228.9</td>
<td>1612.6</td>
<td>12458.2</td>
<td>1086.0</td>
<td></td>
</tr>
<tr>
<td>Rhea</td>
<td>5655</td>
<td>487.0</td>
<td>730.1</td>
<td>5522.4</td>
<td>691.2</td>
<td></td>
</tr>
<tr>
<td>Zeus</td>
<td>4177.0</td>
<td>0.0</td>
<td>0.0</td>
<td>87.6</td>
<td>41.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>87.6</td>
<td>41.4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6312.0</td>
<td>356.0</td>
<td>73.5</td>
<td>214.1</td>
<td>43.8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>206.0</td>
<td>0.0</td>
<td>0.0</td>
<td>8.7</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>6E</td>
<td>18697.4</td>
<td>0.0</td>
<td>0.0</td>
<td>1055.2</td>
<td>66.3</td>
<td></td>
</tr>
<tr>
<td>6W</td>
<td>14651.5</td>
<td>7.7</td>
<td>0.0</td>
<td>626.3</td>
<td>49.6</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>614.4</td>
<td>0.0</td>
<td>0.0</td>
<td>28.1</td>
<td>17.0</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>232.1</td>
<td>0.0</td>
<td>0.0</td>
<td>11.2</td>
<td>10.3</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>15632.0</td>
<td>946.3</td>
<td>106.8</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>11805.5</td>
<td>0.0</td>
<td>0.0</td>
<td>770.1</td>
<td>80.3</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>309.5</td>
<td>0.0</td>
<td>0.0</td>
<td>12.5</td>
<td>5.9</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>25618.3</td>
<td>0.0</td>
<td>0.0</td>
<td>1504.0</td>
<td>120.2</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>448.7</td>
<td>0.0</td>
<td>0.0</td>
<td>18.5</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>type</th>
<th>species</th>
<th>main trunk</th>
<th>reiterated</th>
<th>dry masses (kg)</th>
<th>leaf</th>
<th>TOTAL</th>
<th>% total</th>
</tr>
</thead>
<tbody>
<tr>
<td>tree</td>
<td>SESE</td>
<td>3569312</td>
<td>213247</td>
<td>53714</td>
<td>239095</td>
<td>17192</td>
<td>4084409</td>
</tr>
<tr>
<td>tree</td>
<td>PSME</td>
<td>135815</td>
<td>0</td>
<td>8338</td>
<td>961</td>
<td>145114</td>
<td></td>
</tr>
<tr>
<td>tree</td>
<td>TSHE</td>
<td>31799</td>
<td>0</td>
<td>6334</td>
<td>864</td>
<td>390066</td>
<td></td>
</tr>
<tr>
<td>tree</td>
<td>ACMA</td>
<td>4444</td>
<td>0</td>
<td>925</td>
<td>264</td>
<td>5634</td>
<td></td>
</tr>
<tr>
<td>tree</td>
<td>UNCA</td>
<td>2921</td>
<td>0</td>
<td>937</td>
<td>273</td>
<td>4131</td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>RUSP</td>
<td>0</td>
<td>1974</td>
<td>666</td>
<td>2660</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>POMU</td>
<td>0</td>
<td>1271</td>
<td>1271</td>
<td>0.0296</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>VA0V</td>
<td>0</td>
<td>552</td>
<td>552</td>
<td>0.0129</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>VA0V</td>
<td>0</td>
<td>289</td>
<td>289</td>
<td>0.0067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>POMU</td>
<td>0</td>
<td>107</td>
<td>196</td>
<td>0.0045</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>VAPA</td>
<td>0</td>
<td>44</td>
<td>44</td>
<td>0.0037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>VAPA</td>
<td>0</td>
<td>94</td>
<td>94</td>
<td>0.0023</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>CHIA</td>
<td>0</td>
<td>11</td>
<td>11</td>
<td>0.0027</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>GASH</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0.0009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shrub</td>
<td>SACA</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.0002</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>species</th>
<th>geo</th>
<th>epi</th>
</tr>
</thead>
<tbody>
<tr>
<td>SESE</td>
<td>3569312</td>
<td>213247</td>
</tr>
<tr>
<td>SESE</td>
<td>3569312</td>
<td>213247</td>
</tr>
<tr>
<td>PSME</td>
<td>135815</td>
<td>0</td>
</tr>
<tr>
<td>PSME</td>
<td>135815</td>
<td>0</td>
</tr>
<tr>
<td>TSHE</td>
<td>31799</td>
<td>0</td>
</tr>
<tr>
<td>TSHE</td>
<td>31799</td>
<td>0</td>
</tr>
<tr>
<td>ACMA</td>
<td>4444</td>
<td>0</td>
</tr>
<tr>
<td>UNCA</td>
<td>2921</td>
<td>0</td>
</tr>
<tr>
<td>RUSP</td>
<td>0</td>
<td>1974</td>
</tr>
<tr>
<td>POMU</td>
<td>0</td>
<td>1271</td>
</tr>
<tr>
<td>VA0V</td>
<td>0</td>
<td>552</td>
</tr>
<tr>
<td>VA0V</td>
<td>0</td>
<td>289</td>
</tr>
<tr>
<td>POMU</td>
<td>0</td>
<td>107</td>
</tr>
<tr>
<td>VAPA</td>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>VAPA</td>
<td>0</td>
<td>94</td>
</tr>
<tr>
<td>CHIA</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>GASH</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>SACA</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Characteristics of Tidy Data

Observations

- Separate tables for each entity measured
- Each row represents a single observed entity
Recognizing untidy data

All the same observation? No.
Characteristics of Tidy Data

Observations

- Separate tables for each entity measured
- Each row represents a single observed entity

Variables

- All values in a column are of the same type
Recognizing untidy data

<table>
<thead>
<tr>
<th>species</th>
<th>tree</th>
<th>main trunk</th>
<th>reiterated trunk</th>
<th>limbs</th>
<th>branches</th>
<th>leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>SESE</td>
<td>Atlas</td>
<td>25644.9</td>
<td>4802.0</td>
<td>547.7</td>
<td>13433.2</td>
<td>1101.2</td>
</tr>
<tr>
<td>SESE</td>
<td>Ballantine</td>
<td>221966.4</td>
<td>7651.6</td>
<td>5922.9</td>
<td>11210.0</td>
<td>1084.8</td>
</tr>
<tr>
<td>SESE</td>
<td>Bell</td>
<td>253246.4</td>
<td>5454.3</td>
<td>5792.6</td>
<td>48500.7</td>
<td>1043.4</td>
</tr>
<tr>
<td>SESE</td>
<td>Broken Top</td>
<td>130928.9</td>
<td>4805.2</td>
<td>1608.1</td>
<td>5137.4</td>
<td>729.9</td>
</tr>
<tr>
<td>SESE</td>
<td>Buena Vista</td>
<td>128833.0</td>
<td>3486.5</td>
<td>0</td>
<td>8552.1</td>
<td>518.4</td>
</tr>
<tr>
<td>SESE</td>
<td>Demeter</td>
<td>155896.0</td>
<td>11085.6</td>
<td>3204.3</td>
<td>10054.1</td>
<td>768.7</td>
</tr>
<tr>
<td>SESE</td>
<td>Epimetheus</td>
<td>226987.0</td>
<td>12915.7</td>
<td>1797.2</td>
<td>13585.2</td>
<td>1029.4</td>
</tr>
<tr>
<td>SESE</td>
<td>Iliuvatar</td>
<td>349586.6</td>
<td>65003.9</td>
<td>12315.6</td>
<td>13987.0</td>
<td>1461.8</td>
</tr>
<tr>
<td>SESE</td>
<td>Kronos</td>
<td>134154.1</td>
<td>12204.6</td>
<td>7232.7</td>
<td>5038.9</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Pleiades I</td>
<td>182385.2</td>
<td>3735.0</td>
<td>1935.2</td>
<td>1084.3</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Pleiades II</td>
<td>235838.8</td>
<td>1118.4</td>
<td>4306.0</td>
<td>1139.0</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Prometheus</td>
<td>239414.0</td>
<td>25228.9</td>
<td>1612.6</td>
<td>1245.8</td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Rhine</td>
<td>14710.4</td>
<td>487.8</td>
<td>730.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>Zeus</td>
<td>243365.7</td>
<td>2885.5</td>
<td>1620.4</td>
<td>1910.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1761.3</td>
<td>4632.0</td>
<td>206.0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3632.0</td>
<td>356.0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>206.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>6E</td>
<td>18697.4</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>6W</td>
<td>14651.5</td>
<td>7.7</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>614.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>232.1</td>
<td>0</td>
<td>11.2</td>
<td>10.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>15632.0</td>
<td>0</td>
<td>946.3</td>
<td>106.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>11805.5</td>
<td>0</td>
<td>770.1</td>
<td>80.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>309.5</td>
<td>0</td>
<td>12.5</td>
<td>5.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>25618.3</td>
<td>0</td>
<td>1504.0</td>
<td>120.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>geo</td>
<td>3569312</td>
<td>2123274</td>
<td>53714</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SESE</td>
<td>epi</td>
<td>135815</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSME geo</td>
<td>135815</td>
<td>0</td>
<td>8338</td>
<td>961</td>
<td>145114</td>
<td>1.0</td>
</tr>
<tr>
<td>PSME epi</td>
<td>135815</td>
<td>0</td>
<td>8338</td>
<td>961</td>
<td>145114</td>
<td>1.0</td>
</tr>
<tr>
<td>TSHE geo</td>
<td>31740</td>
<td>0</td>
<td>6332</td>
<td>860</td>
<td>38932</td>
<td>0.99</td>
</tr>
</tbody>
</table>

All the same variable? No.
Characteristics of Tidy Data

Observations

- Separate tables for each entity measured
- Each row represents a single observed entity
- Observations (rows) are all unique

Variables

- All values in a column are of the same type
- All columns pertain to the same observation (row)
- Each column represents either an identifying or measured variable
A not-so-reproducible workflow
Building a reproducible workflow

Raw, messy data

Clean, raw data

Merged/summarized derived data

Figures, tables, maps
Where data come from matters! (a sample)

- **Excel**
  - Automatic conversion of gene names to dates or floating point numbers*
  - Date values can be converted when transferring data between operating systems and applications

- **Text (e.g. CSV) & Excel**
  - Free-form structure - lack of enforcement of column-row structure, type consistency

- **Text (e.g. CSV)**
  - Inconsistent structure - quotes, commas, missing values, spaces

- **Database**
  - Enforced structure - tables, column typing
  - Specialized methods for interaction (pros and cons to this)

---
The ESAUSSEE Data Help Desk

who we are and how to find us

Amber Budden, @aebudden, @DataONE_org, aebudden@epscor.unm.edu

Deborah Paul, @idbdeb, @idigbio, dpaul@fsu.edu

Dmitry Schigel, @dschigel, @GBIF, dschigel@gbif.org

Karl Benedict, @kbene, kbene@unm.edu, president@esipfed.org

Kristen Vanderbilt, @vanderbik, @EDIgotdata, krvander@fiu.edu

Kyle Copas, @kylecopas, @GBIF, kcopas@gbif.org

Laura Brenskelle, @lbreorsk, @idigbio, lbrensk@ufl.edu

Margaret O'Brien @, @EDlgotdata, margaret.obrien@ucsb.edu

Megan Jones, @MeganAHJones, @NEON_sci, mjones01@battelleecology.org

Rebekah Wallace, www.eddmaps.org, bekahwal@uga.edu
Messy data? Repetitive data tasks?
Increase Reproducibility and Productivity using tools like Open Refine
Data lessons compiled - inspired by workshop

Georeferencing for Research Use of Museum Collections Data

- Data mapped to standards
  - supports use and re-use (e.g. Darwin Core DwC, Ecological Metadata Language EML)
  - standards help with data validation and cleaning

- Data have issues
  - what are some you have experienced
  - need to be addressed before applying research methods
  - keep raw data raw
  - track your changes

- Data visualization is key
  - QGIS lessons
  - Open Refine
  - R, etc.

Carabidae (beetles) of California
(Fun!) features and functions in Open Refine

- runs on your computer (not in the cloud)
- data formats supported
- raw data
- column manipulation
- text facet
- routine cleaning (white space)
- clustering
- step-wise editable task script
- APIs
- regular expressions
- export
- share project files
Open Refine - getting started is quick and easy

- download and install
- launch
- import your data
- your raw data is NOT touched
- supported data formats
- subset data
Open Refine - managing columns

- reorganize columns easily
Open refine - text facet

*lists and counts the distinct values in a column*
Open Refine - the magic of clustering algorithms
or how to find issues that abc sort won’t
and fix them all at once - no hunting
Open Refine - manages pesky white spaces

<table>
<thead>
<tr>
<th>specificEpithet</th>
<th>scientificName</th>
<th>weight</th>
<th>length</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>californicus</td>
<td>Facet</td>
<td>30.5</td>
<td>165</td>
<td>male</td>
</tr>
<tr>
<td></td>
<td>Text filter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>californicus</td>
<td>Edit cells</td>
<td></td>
<td></td>
<td>Transform...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text
- Blank out cells

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>23.5</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>176</td>
</tr>
</tbody>
</table>
Open Refine - add data to your data using APIs

that's application programming interface
Open Refine - saves your steps

- supports reproducibility
- tracks your work for you
- easy to go back to earlier steps with confidence
Open Refine - export your data, share project files

select the **format**
export subsets too
and **project files**
Open Refine - make some friends

- share this tool with students, friends, families, colleagues
- imagine future tools, think beyond spreadsheets

Increase Reproducibility and Productivity using tools like Open Refine

Magic is here. Ask for it, plan for it.
Looking for next steps now?

*R, Open Refine, and Data Management resources*

- The #datahelpdesk is ready to offer data assistance!
- #CareerCentral Q and A: Wednesday, August 5th, 9:30-10:30 PDT (12:30-1:00 EDT)
- Data Carpentry lessons
- on Twitter #ESA2020 #datahelpdesk