Abstract

Diverse datasets of observations and experiments are now available for reuse, integration, and synthesis. However, the largest time investment is still in cleaning and combining data until all primary data are completely understood and in a similar format. That process can be simplified by creating a common, flexible data pattern for synthesis projects, which also highlights the most important dataset features for those who anticipate their data being used in synthesis. Here, we describes EDI’s process for harmonizing datasets, with an example “design pattern” for ecological community survey data.

Process

1. Identify active synthesis projects to leverage data-use practices and target (L0) data
2. Identify important attributes of L0 data (common and rare)
3. Examine related work in the domain; consider how these are currently used
4. Develop a candidate design pattern (DP, L1) to hold important attributes in related tables
5. Reformate L0 data into the L1 design pattern
6. Examine metrics & identify issues, adapt the format as necessary
7. Build tools to create metadata (EML), and use data in the L1 format

Harmonizing Ecological Community Survey Data to Facilitate Synthesis and Reuse
Margaret O’ Brien1, Colin Smith2, Corinna Gries2
with ecocomDP workshop participants
1University of California Santa Barbara, 2University of Wisconsin Madison

Model Features Compared

<table>
<thead>
<tr>
<th>Feature</th>
<th>ecocomDP</th>
<th>Popperl</th>
<th>DC Archive</th>
<th>BioTIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decentralized</td>
<td>yes</td>
<td>no</td>
<td>yes, qualified</td>
<td>no</td>
</tr>
<tr>
<td>Files stored as ASCII</td>
<td>yes</td>
<td>no</td>
<td>yes, qualified</td>
<td>no</td>
</tr>
<tr>
<td>Sampling site nesting</td>
<td>unlimited</td>
<td>&lt;5 levels</td>
<td>in some arrangements</td>
<td>some</td>
</tr>
<tr>
<td>Ancillary data accommodated</td>
<td>any</td>
<td>some</td>
<td>some</td>
<td></td>
</tr>
<tr>
<td>Variables mapped to external vocabularies</td>
<td>yes</td>
<td>no</td>
<td>suggested, not explicit</td>
<td>uses a local vocabulary</td>
</tr>
</tbody>
</table>

For Creators - Step 1
- Guides, FAQs
- Data table cleaning functions
- Table validation and quality control
- Scripts create EML metadata
- Upload to EDI

For Users - Step 2
- Find all ecocomDP
- Explore metadata
- Aggregate ecocomDP
- Filter and subset

Data available in the ecocomDP format

Maintenance:
- In progress: Automated updates for ongoing time-series

Takeaways

- **Scientists:** Talk to EDI if you ...
  - Use community survey data
  - Know of data appropriate for ecocomDP
  - Have synthesis projects in other scientific domains
  - Need to save time cleaning data

- **Data Managers:**
  - Note important L0 features identified by ecocomDP
  - Sampling site nesting can be understood
  - Locations are complete (with latitude, longitude)
  - Taxa can be resolved (e.g., species binomials)
  - Work with EDI to build robust measurement vocabularies

Contact: margaret.obrien@ucsb.edu
GitHub: github.com/EDIorg/ecocomDP

Acknowledgements: This material is based upon work supported by the National Science Foundation under grants #1565103 and #1629233. Any opinions, findings, conclusions, or recommendations expressed in the material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.