Overview: The CalBug Project, housed out of the The Essig Museum of Entomology, is a collaborative initiative between nine California institutions with a goal to digitize over a million specimens. Digitization involves imaging both specimens and their labels as well as database their collection info. The CalBug project also is attempting to georeference, or locate the latitude and longitude coordinates, for these million specimens (some dating back to the 18th century) so that they can be better used for research. The project uses many student workers, graduate students and volunteers to capture the images and data. Over the past few years, it has participated in the Notes from Nature project, which helps connect citizen scientists to scientific research. Through the images generated of the specimen labels by the team at the Essig Museum, citizen scientists digitally transcribe the data that can be read from the image. The Essig, after each label is transcribed by 2-4 citizen scientists, runs an R algorithm to find the most accurate transcription and transfer it into the Essig’s database. These combined efforts have accumulated in over 209,000 specimen records and over 400,000 images and counting. This project has a large scope and an ever-increasing scale. My proposed artifact is the redesigned web interface for the CalBug Project (Figures 1-4) and the search crosswalk to the database fields (Figure 5). This supports the discussion surrounding the data categorization for different users.

Key Concepts: Chapter 7, Classification, Search and Retrieval, Taxonomy

What is being organized? The insect specimens in the CalBug project are digitized on an individual level, with unique identifying numbers, and new specimen records and their associated data are continually being added to the digital collection. Both the specimens and their data are being organized. Existing groups of specimens are prioritized for digitization and new physical specimens are accessioned into the collection and are databased upon arrival.

Why is it being organized? An individual specimen’s associated data can be highly variable, however, as long as a specimen has the time and place it was collected (no matter how vague) associated with it, it is valuable research material. This 202 project focuses on the multiple levels of organization that happen with the specimens and their information. Each level of organization supports a different type of user. The physical specimens are organized to facilitate the collection manager’s use of the collection. When physical specimens need to be borrowed, they must be efficiently found, packaged, and sent out on loan, so fastidious organization is key when locating thousands of specimens. The digital organization of the collection also facilitates the museum staff and collection manager by allowing for expanded interaction with the collection by using the database. The digital collection’s web interface affords to make the collection accessible for researchers and novices alike, as well as to foster data sharing to other data repositories. Since the specimen data follows digital curatorial standards, a web interface that allows these fields to be easily searchable and navigable can add to the use of the collection for a broader audience.
**How is it being organized?** The data of the CalBug Project is organized according to Darwin Core (DwC), a standard “designed to facilitate the exchange of information about the geographic occurrence of species and the existence of specimens in collections.”

Certain specimen attributes have concrete institutional parameters, such as unique identifying numbers and taxonomic identification, however other attributes have less strict parameters (e.g. a precise location of where a specimen is found), although specific DwC fields. Although there are institutional taxonomies in place for information associated with a specimen’s collection and identification, the CalBug search interface redesign allows for an outward-facing re-organization of the existing fields.

**When is it being organized and by whom?** The categorization and organization happens at multiple times for one specimen. If identified, the specimen is already inserted into the taxonomic classification scheme -- the hierarchy of how species are related. This scientific warrant is inherited and replicated in the physical curation of the collection and specimens are further sorted (within a taxon) by geographic region. Aligning with taxonomic categories provides a clear hierarchy for sorting and locating physical specimens and, with changes in taxonomy having to be published, makes collection maintenance fairly consistent.

The specimens are organized a second time when they are databased, either by interns or through Notes from Nature. The data is stored in a MySQL database that uses mostly DwC fields, an institutional taxonomy for specimen data. The digitization of specimens, through utilizing DwC institutional semantics, makes collection maintenance, governance and interaction easier as the collection manager can search in a multifaceted manner, better understand the holdings of the museum, and track specimens for loans. The unique specimen numbers allow for individual tracking and the other DwC fields provide multiple areas for accurate search and retrieval.

For the CalBug search interface design, the specimens retain their classification hierarchy within the database, however, the outward-facing search fields aim to serve a broader audience, not just the collection manager and museum staff. As the redesign focuses on researchers and students, the classification appears to focus more on taskonomy instead of the institution taxonomy (see Fig 1). The 20 search fields provided in the search interface, while actually searching through the ~100 fields in the database, facilitate precise information retrieval. Although less search fields might indicate lower accuracy, user testing has proven that the new search design improves accuracy since the use doesn’t need to know exactly what DwC field in which to enter search parameters. The search is further expanded by having a ‘Search any field’ box, which literally looks in every DwC field for a term, as well as “Common Name” field, to support novice searches, such as ‘beetle’ and ‘butterfly’ instead of ‘coleoptera’ and ‘lepidoptera’. The intrinsic properties of the specimens lend the results to simple (alphabetic and numeric) sorting as well as searching (through the ‘Refine’ option) on the list view of the results pages (Figure 2) and additional views of results (Figures 3,4) help users locate desired specimens and reorganize as needed to suit their needs.

---

Figure 1: CalBug Search Interface
Figure 2: Search Results: List View

<table>
<thead>
<tr>
<th>Cat. No.</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
<th>Species</th>
<th>Year Collected</th>
<th>State/Province</th>
<th>Country</th>
<th>Collector</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMEC200281</td>
<td>Coleoptera</td>
<td>Carabidae</td>
<td>Cicindela</td>
<td>sp.</td>
<td>2002</td>
<td>California</td>
<td>United States</td>
<td>Kipling Will</td>
</tr>
<tr>
<td>EMEC200282</td>
<td>Coleoptera</td>
<td>Carabidae</td>
<td>Cicindela</td>
<td>sp.</td>
<td>2002</td>
<td>California</td>
<td>United States</td>
<td>Kipling Will</td>
</tr>
<tr>
<td>EMEC200283</td>
<td>Coleoptera</td>
<td>Carabidae</td>
<td>Cicindela</td>
<td>sp.</td>
<td>2004</td>
<td>California</td>
<td>United States</td>
<td>Kipling Will</td>
</tr>
<tr>
<td>EMEC200284</td>
<td>Coleoptera</td>
<td>Carabidae</td>
<td>Cicindela</td>
<td>sp.</td>
<td>2004</td>
<td>California</td>
<td>United States</td>
<td>Kipling Will</td>
</tr>
<tr>
<td>EMEC200285</td>
<td>Coleoptera</td>
<td>Carabidae</td>
<td>Cicindela</td>
<td>sp.</td>
<td>2004</td>
<td>Iowa</td>
<td>United States</td>
<td>Kipling Will</td>
</tr>
</tbody>
</table>

Figure 3: Search Results: Image View

EMEC212767
Cicindela limballis Identified by: [tray label] in null Country: Canada Year Collected: 1967

EMEC212768
Cicindela limballis Identified by: [tray label] in null Country: Canada Year Collected: 1978

EMEC212769
Cicindela limballis Identified by: [tray label] in null Country: United States Year Collected: 1986

EMEC212771

EMEC212772

EMEC212773
Figure 4: Search Results: Map View

Search results

Map View of search results showing locations in the western United States with markers indicating specific points of interest.
Figure 5: CalBug Database Search Field Crosswalk