Crowdsourcing Specimen Labels
The Crab Shack Experience

Regina Wetzer and N. Dean Pentcheff
rwetzer@nhm.org and *dpentche@nhm.org*

The trouble with wet collections

Wet-preserved collections often have many specimen lots where multiple labels are in the jar with the specimens. Label data cannot be fully read from outside the jar. Therefore, data capture requires opening and disassembling the lot, separating and imaging the labels, then reassembling the lot.

Much of the label information is only apparent from direct inspection. Label format, typography, handwriting, *etc.* are often the clues that allow connection to the correct data sources. Hence, high quality label images are essential for digitization.

The Natural History Museum of Los Angeles County

The wet-preserved marine invertebrate collections at NHM are estimated to have about 633,000 lots, containing about 8.7 million specimens, few of which are digitized.

Because digitizing wet-preserved marine invertebrate lots is time-consuming and complex, it is a daunting challenge. The scale of our problem is encouraging us to explore innovative approaches to digitization.

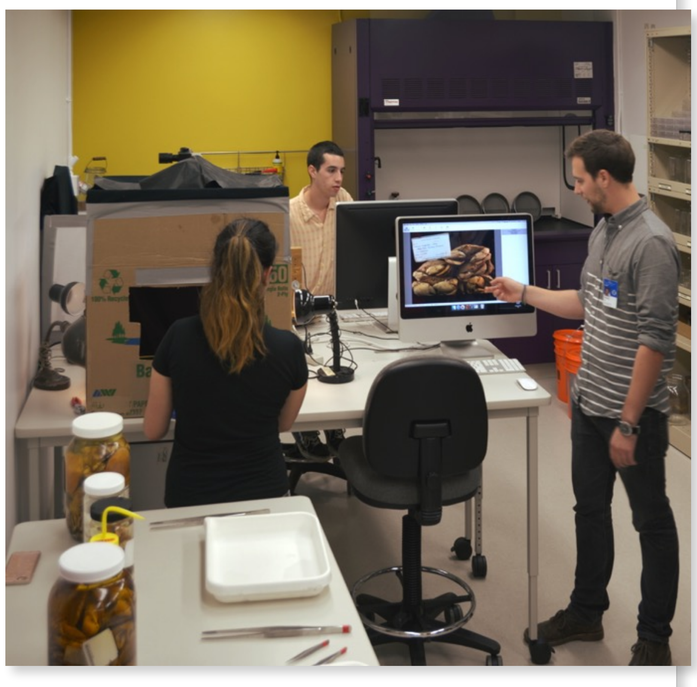
Pilot project: Crabs in the family Cancridae

We picked a single taxon (the family including rock crabs and Dungeness crabs) for a pilot test using about 1,000 lots. Our goal was to develop and assess a pipeline for imaging label data from the jars, ultimately using those images for an experimental crowd-sourced web application for data entry into our collection data fields.

We followed process guidelines developed by *iDigBio* (Nelson, G. et al. *DROID3: Things in spirits in jars* — https://www.idigbio.org/content/workflow-modules-and-task-lists). To maximize efficiency, work was done by small teams of staff and work-study students.

Two separate photo rigs were developed to simultaneously capture images of labels and of the specimen lot itself. For these specimens, research-grade images of specimens are not useful for taxonomic purposes, so the specimen lot images were captured purely for use in collection management.

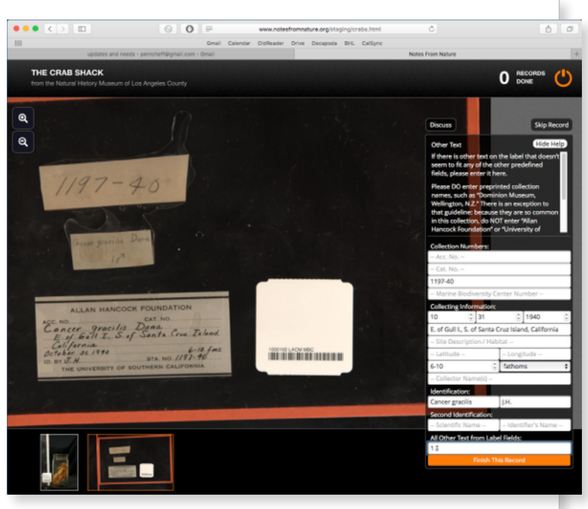
Each lot was labelled with a unique identifier, printed as both a human-readable number and a machine-readable barcode. Using custom software based on the open-source *ZBar* library (http://zbar.sourceforge.net), every image file was renamed automatically by detecting the barcode number in the image.



Crowdsourcing: The Crab Shack

Once label images were gathered, we wanted to allow the public to help us transcribe data from the images to database fields for our specimens.

The *Notes From Nature* team (http://notesfromnature.org) collaborated with us to create a web-based transcription system: *The Crab Shack.*



We launched the transcription project using the *WeDigBio* event in October 2015. To start by using a controlled audience, we invited undergraduate students to an on-site transcribing event on two consecutive days.

Over the next five weeks, Internet visitors transcribed all images. Each set of labels was transcribed four times, resulting in nearly 4,000 transcription records from over 200 visitors.

Lessons learned

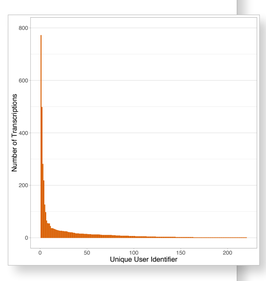
Label image acquisition is time-consuming but inescapable

To digitize collections with multiple labels in wet-preserved lots, there is no alternative to opening the jars, extracting and photographing the labels, then reassembling the lot. It took 10.6 person-minutes to process each lot (171 person-hours for nearly 1,000 lots). Therefore, image acquistion to digitize all 633,000 wet marine invertebrate lots at NHM can be expected to take about 55 person-years.

Complex transcription may not be the best target for crowdsourcing

This transcription task was particularly complex, given the number of data fields available for filling and the ambiguity of assigning label data to database fields. We are still evaluating the efficiency of correcting the amalgamated crowd-sourced transcriptions against direct data entry by trained staff. It is quite possible that untrained transcription of complex data may need so much correction that it is not an efficient approach for data entry.

Most transcription is done by a few participants



Seven of the 220 Crab Shack transcribers (3% of the participants) contributed over half the records. We believe this finding holds great promise for improving crowdsourced data transcription quality: a small effort to train these highly motivated transcribers could result in a disproportionate improvement in overall data quality.

Public outreach value of crowdsourcing transcription is high

Participants at the in-house *WeDigBio* event reported strongly increased interest in museum research specimens (several even expressed interest in volunteering in the museum). Internet participants were not directly surveyed, but spontaneously posted a number of comments indicating that they particularly enjoyed exploring natural history specimens through the *Crab Shack* transcription project. This outreach value of crowdsourcing provides an important potential benefit for digitization projects that is additional to the data acquisition itself.

Acknowledgements

- **WeDigBio** team for Crab Shack interface development (Michael Denslow, Rob Guralnick, Rafe LaFrance) and onsite support (Libby Ellwood).
- **iDigBio** for digitization workflow development.
- **Marine Biodiversity Center** staff for imaging system development, imaging management, and event handling (Kathy Omura, Adam Wall, Jenessa Wall).
- **Kelsey Bailey** for crab specimen photographs.
- **MBC work-study students and volunteers** from the University of Southern California for label and specimen imaging.
- Two hundred and twenty **Crab Shack transcribers**.
- **Time Warner Cable** for inadvertently providing adequate Internet bandwidth for the on-site transcription event.