DigIn Breakout Day 1

How much (what) data to capture and how?

Group members:
Paul Larson
Megan McCuller
Trina Roberts
Libby Ellwood
Mandy Bemis
Henry Choong
Nelson Rios
Kevin Love
Deb Paul
Bill Moser
Ed Gilbert

LABEL, LEDGER, ETC--CORE DATA SOURCES

Paul: stuff that isn't catalogued will have a label without much, and a lot of cross-referencing with ledgers and other sources. Photographing labels won't help much. Mix of typewritten and handwritten. For some, station data exist.

Deb: Tritrophic TCN didn't capture labels. They took data straight into the database.

Trina; LACM size → label capture may be all that's possible

Henry: Capturing the label is necessary for preserving information. If you want to go back later and it hasn't been imaged, you have to go all the way back to the specimen. And old labels are deteriorating, so label images are also a data preservation measure

Ed: People go back to the label all the time, both to add data to transcription and to correct errors. Also: better AI/Machine Learning tools are coming along, and images may be able to be parsed faster later

Libby: We're talking about label images taken quick-and-dirty (Deb holds up cell phone) not studio portraits

Deb: It may not be necessary for all collections to do the same thing

Henry: Sometimes information in multiple places -- jars, labels, lids, fronts and backs… Ed--workflows/tools exist to have multiple images per specimen

Megan--don't keep information that duplicates

Kevin--would we hesitate to digitize stuff without good IDs, etc? Group: NO. Do it--get it out to the community!

WHERE/HOW DO WE START?

Bill--get low-hanging fruit where there is vessel data, or a good data source that can just be keyed in. Do that stuff first. Exemplar data collection, so you know the breadth/scope. A rough inventory, keying in some exemplars to start with.

Deb/Kevin--pre-digitization inventories can be helpful [?]

GEOGRAPHIC-ISH DATA? What do we capture & how do we represent it?

Paul: database issue--three-level terrestrial hierarchy. Political boundaries and marine boundaries don't match up the way county/state/county does.

Megan agrees

Nelson--and these are things without lat/lon?

Some general discussion

Deb--we need to get people to use the standards in DwC--many are not using all those fields as they exist already. There may also be issues with some individual databases representing data from the oceans…

Trina: maybe generate some data entry standards and/or fixes for each of these major database platforms?

Mandy--same problem--replace some of the geography with types like marine/estuarine instead

Paul -- would need to record geographic information somehow
Nelson -- if you know where it is, even with an area of uncertainty, just record whatever the most refined geospatial representation is
Megan--then how do you query when you get a user wanting something like e.g. Florida? Nelson--yes, then you need to be able to query it spatially, rather than with text.
Kevin--and we can then fill in other layers that help query data that are represented geospatially
Nelson--we want to be moving toward having things more spatial, rather than more textual

WHO DOES THE TRANSCRIPTION? CROWDSOURCING?

Paid crowdsourcing/offshoring? --

Henry: your pre-validation has to be pretty good

Libby--yes, it would be (probably) a naive user every time

WHAT DO THE BASE DATA LOOK LIKE?
Is everything in jars? How bad? How slow?
Some discussion of getting labels out of jars

Reporting summary:

- General agreement that for most collections, label photos is a good idea. This is both data preservation and a source of data transcription. The data could be transcribed in multiple ways yet to be determined.

- “Labels” might include jar lids and all kinds of other things. This could be relatively quick, or horribly slow for “condos” or similar arrangements.
- Some collections have key data that is not on labels--esp ledgers, field notes, etc. This needs to be taken into account.
- Some problems exist for representing marine data in existing databases. We could fix some of this with existing fields. Or new standards?
- Crowdsourcing: if we do it, attention to front-end data curation is important. With crowdsourcing, messy data in → messy transcription out.
- Flexibility in workflow procedures between collections, or between projects within a collection because of variability in the arrangement and location of pertinent data.