COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Austin Mast (Florida State University), May 19, 2021

Title: *Rapid Creation of a Data Product for the World's Specimens of Horseshoe Bats and Relatives, a Known Reservoir for Coronaviruses*

Austin Mast CIC Database Profile

NSF Award #: 2033973

YouTube Recording with Slides

May 2021 CIC Webinar Information

Transcript Editor: Macy El Moujabber

Transcript

Austin Mast:

Slide 1
Thanks so much for inviting me. I want to start by recognizing that this was a team effort and we had a great team.

Crises like the pandemic emerge and we find ourselves in urgent need of data. Sometimes the data we need are about biodiversity. In this case, we'd like to know things about bats. In other cases, like oil spills, it might be the entire biota in a particular region about which we need data. Until our work, we didn't have a set of crisis response protocols to rapidly enhance data about an important source of biodiversity information: that is the world's 3 to 4 billion biodiversity specimens.

Slide 2

As we see in our current pandemic, it could be that a narrow subset of those specimens suddenly become critical to crisis response. Specimens have associated information that documents the what was collected, where it was collected, who collected it, and other information. There are also time capsules of potential information, since genomic data can often be derived from the specimen or its disease-causing agents. These are just a few specimens of the species of Horseshoe bat in which the closest relative of SARS-CoV-2 has been found. That

is: *Rhinolophus affinis*.

Slide 3

We targeted a set of three closely-related families, including the family of *Rhinolophus affinis* for specimen data enhancement.

Slide 4

These are the mappings of Horseshoe bat specimens at the two major aggregators of specimen data. I want to emphasize that data coming from collections and served by these aggregators are valuable in their current state. However, the data have some qualities that can be improved through consideration of the data in aggregate and the data have been created over two or more decades, meaning that the data have not all benefited from our current understanding of best practices and the availability of software to improve some steps.

Slide 5

We focused on enhancing the data in these ways and I'll walk through those in bold with you. If you pay attention to the change in slide title you can follow our relatively rapid progression through these activities.

Slide 6

The specimen data coming from the two major aggregators has overlapped but it's not identical. De-duplicating their records produced about 90,000 in scope records.

Slide 7

The records are curated by 118 institutions worldwide. The top 10 of these institutions together share 63% of the records.

Slide 8

We could only assign or assess coordinates for collections - collection locations when those locations are described in the shared data. And about two-thirds of the records had that information. Of those, about two-thirds arrived with pre-assigned coordinates and one third did not. We were able to assess or assign coordinates in 95% of the total possible cases and we modified pre-existing coordinates about half the time. The median amount that a pre-existing coordinate was moved was six kilometers.

Slide 9

Importantly, the relevant metadata fields went from mostly empty to mostly complete with such useful information added as geo-referencing protocol and geo-referencing resources.

Slide 10

In this summary, at the country level, you can see where the greatest number of specimens have been collected by the size of the pie chart and the relative number of new coordinates

added to the specimens from those countries.

Slide 11

Here are the coordinates for collecting locations for each of our focal families.

Slide 12

We compared our coordinates with prior range maps for the species when they were available from the International Union for Conservation of Nature (IUCN). Here's an example range circumscription in red for one species, this is again Rhinolophus affinis, and our coordinates for that species in green. We found that georeferenced specimens suggest range extensions for 153 of the 169 focal bat taxa for which we have these kinds of maps. This is a significant, significant expansion of our understanding of where to find the bats. This is a screenshot of a web-based horseshoe bat data explorer for IUCN map assessors and other stakeholders to look at locality coordinates relative to the current IUCN maps, with links back to complete records in our system.

Slide 13

The records arrived with 2,930 distinct values referencing people who collected or identified the specimens. We were able to assign 803 unique identifiers to a subset of those values. These unique identifiers, or ORCID IDs when the person is living, and Wikidata QIDs when the person is deceased. An additional 437 values representing 359 people are reasonably assigned to persons currently living but who do not yet have an ORCID ID.

Slide 14

To do this we engage 34 people who are mostly bat experts from 13 countries. These experts and our data curators found that they could associate about half of the records to a unique identifier for specimen collector and about two-thirds of those for specimen identifier.

Slide 15

The value of doing this for a crisis response might not be immediately apparent, but it could be among the most important things that we did. We identified 117 living people with ORCID IDs with experience collecting the bats. You might say that it's easy to find the bat experts - just approach their professional societies or do a literature search. However, the bat collectors and those you might find in that way will only partially overlap. The bat collectors turned out to include a considerable diversity of professions including those who are not professional biologists. Here are a few other descriptors - descriptions of collectors with valuable experience doing field work in sometimes remote areas. Remember, excuse me, remember we also identified 359 living bat collectors who don't have ORCID IDs. Together, this is a rolodex of potential contacts for those of you who need to go back into the field to relocate bat populations.

Slide 16

Prior to data enhancement 5.5% of the records had information about associated

sequences. We identified an additional about 1,100 specimens with which we could associate new sequences that we found.

Slide 17

Our versioned data, and importantly our protocols, are shared at Zenodo so that we have now blazed a trail that others can follow for rapid data enhancement of specimens during the next crisis. We expect to share our final versions of everything there very soon. However, I will note that the current version that's up right now is very close to the final version. We're close to submission of a manuscript focused on the work and expect to make the Horseshoe bat data explorer broadly available to those who need it.

Slide 18

The EU recently announced new funding for the creation of records, about 20 - about approximately 20,000 bat specimens, and we expect the foundation that we laid to speed up that work.

Slide 19

I want to thank those who contributed their time and expertise to the people disambiguation and in the first paragraph, but not assigned ORCID IDs there for lack of space. And thank you to the NSF for supporting the work.