UC SANTA BARBARA

THE Current

October 22, 2019
Harrison Tasoff

Creative Curation

To some people, collection management may seem little more than dusty bookkeeping — organizing, cataloguing, processing specimen loans. But modern collections are actually quite sophisticated. Such resources are beginning to rival the big data operations in industry and tech.

UC Santa Barbara hosts dozens of collections, many of which fall under the purview of the Cheadle Center for Biodiversity and Ecological Restoration (CCBER). The center's director, Katja Seltmann, is applying techniques from computer and information science to revolutionize our use of the immense datasets that research collections hold. She's incorporating these methods into her work on a $4.3 million National Science Foundation initiative investigating terrestrial parasites.

Seltmann is leading the biodiversity informatics component of NSF's new Terrestrial Parasite Tracker project, which involves 27 different research institutions. Arthropods are major carriers of human disease worldwide, she explained, but scientists don't know how they'll respond to the changing environment.

She and her colleagues are working to structure this information using ontologies: sets of concepts and categories that characterize the properties and relationships within a particular field. The system of scientific names for organisms and the classes we sort them into is one example of an ontology.

"We want to very formally create structured links between our statements and the terms in those statements," Seltmann said. This structure will enable researchers to

take advantage of powerful statistical techniques and natural language processing.



Calyptra minuticornis is one of the species of vampire moths known to consume both fruit and blood.

Photo Credit: LUCINDA GIBSON & KEN WALKER, MUSEUM VICTORIA CC BY 3.0 AU

For example, some vampire moths feed on fruit by piercing its skin, but they also occasionally suck blood. So, this observation is broken down into phrases, such as “vampire moth:eats:blood,” and “vampire moth:eats:fruit.” All of the terms in the phrase have formal definitions in the online database. The items have additional tags such as “in nature” and “under experimental conditions.”

Seltmann plans to use ontologies on Ontobee, an online data server designed for ontologies. These are developed for many kinds of projects and are used extensively for annotating genomes and understanding model organisms. The system already

has a wealth of terms and relationships to take advantage of.

“Terms like ‘host of’ or ‘biotically interacts with’ already exist,” she said, “however, the ability to annotate and share complex ideas — like species x interacts with species y on some body part — is a harder problem.” Global Biotic Interactions is one tool that helps with the process, along with other databases for managing natural history collection information.

The Terrestrial Parasite Tracker project spans 1.3 million specimens of parasitic arthropods. “Collections are pivotal to working with arthropods,” Seltmann said, “because there’s just so many of them.”



Entomology graduate student Rachel Behm selects insects from the cases in the preparatory room. Another room across the hall holds the bulk of the collection.

Photo Credit: HARRISON TASOFF

What's more, a lot of the important information about these samples is qualitative, especially the relationships between parasite and host.

The ontology and informatics she's working on will open these collections to new methodologies and allow institutions to easily link their resources for large-scale studies.

"What we're talking about with the Terrestrial Parasite Tracker project is the next generation for biodiversity information science and how it can revolutionize the way we think about studying biodiversity," Seltmann said.

The ways in which we use data have evolved over the past few decades, requiring more advanced methods to search and share information. Researchers are beginning to use analyses of information across many collections, rather than simply looking at individual specimens.

Some of this information is text, but a great deal is less concrete, which makes it more difficult to incorporate into a searchable database. For instance, the time of year a specimen was collected, its size relative to other individuals, and its association with other organisms and features in its habitat.

Researchers often will include qualitative information on a specimen's record, but differences in word usage between individuals means this generally defies analysis using conventional big-data tools. Seltmann's goal is to devise ways to make this sort of information accessible to computers as well as humans. In fact, the Biodiversity Collections Network recently released a report detailing how networking specimen data is the next evolution of natural history collections.

"It's a wholistic way of thinking about natural history collections," Seltmann explained. "Each specimen is a little piece of information with certain properties, but if we put them all together then we can ask really big questions."

About UC Santa Barbara

The University of California, Santa Barbara is a leading research institution that also provides a comprehensive liberal arts learning experience. Our academic community of faculty, students, and staff is characterized by a culture of interdisciplinary collaboration that is responsive to the needs of our multicultural and global society. All of this takes place within a living and learning environment like no other, as we draw inspiration from the beauty and resources of our extraordinary location at the edge of the Pacific Ocean.