Title
Extending Anthophila research through image and trait digitization (Big-Bee) proposal.

Author
Seltmann, Katja C

Publication Date
2021-08-16

I. Vision:

We propose to catalyze a fundamental shift in how insects are digitized in the ADBC program by establishing Big-Bee, a Thematic Collection Network to provide bee research communities with deep digitization of focal taxa, and a network-specific data portal, The Bee Library. We will focus on outreach to the broader ecological, taxonomic and systematic bee research communities with our involvement in the USDA National Bee Monitoring Network, workforce training, and the creation of 1.2M images of over 572,000 bee specimens to capture global morphological and trait variation. Our aim is to publish openly available image and trait datasets that are easily accessible to expand the research value of the specimens beyond biology and foster partnerships in computer science, industry, and engineering.

II. Introduction and Urgent Need:

Animal pollination accounts for 35% of global food production and 80% of wild plants rely on insect pollination (Klein et al. 2007; Potts et al. 2010). Most of these insect pollinators are bees, which contribute 5% to 8% of total world crop production through pollination services (Klein et al. 2007). In the late 1980s, the value to US crops alone was estimated at $4.6 - $18.9 billion (Michener 2000; O'Toole 1993), which is approximately $14 - $60 billion today, adjusted for inflation. However, where data are available, bee abundance is declining dramatically, especially in agricultural regions, potentially creating deficits in pollination services (National Research Council et al. 2007; Koh et al. 2016). A recently published study by Reilly et al. (2020) also shows that US crops are frequently limited by the lack of pollinators. To avert this looming crisis, researchers need to be able to accurately track pollinator declines. Recent anthropogenic loss of biodiversity and changes in species distributions are already affecting our economy (Calvo-Agudo et al. 2019; Janzen & Hallwachs 2019; Lister & Garcia 2018; Sánchez-Bayo & Wyckhuys 2019). Broadly, these declines are being driven by habitat loss, pesticide use, and disease, among other factors (reviewed in Potts et al. 2010; Goulson et al. 2015). However, the reasons for taxon- or region-specific declines in pollinators are often unknown. Our ability to respond to the crisis of global pollinator declines requires historic and integrated data in order to track how species distributions and flight phenology are changing over time (Cameron et al. 2011; Bartomeus et al. 2013), to monitor the spread of non-native bee species (Gibbs & Sheffield 2009), to elucidate historical trends in abundance over decadal timescales (Biesmeijer et al. 2006; Colla et al. 2012; Bartomeus et al. 2013; Burkle et al. 2013), and to confirm recent bee extinction events (Ollerton et al. 2014).

A recent gap analysis presented by Cobb & Seltmann et al. 2020 at the 4th Annual Digital Data Conference revealed that over 40% of US bee specimens already have digitized label data. This equals approximately 2,171,588 specimens out of the estimated 5,045,313 that exist in US collections. This is a much larger number than other groups of insects (except ants) and is largely due to great interest in bee research, federal pollinator initiatives and prior NSF funding, such as the "Collaborative Research: Collaborative Databasing of North American Bee Collections Within a Global Informatics Network" that produced 369,654 new specimen records now available on iDigBio (Ascher et al. 2016). Many collections already prioritize the digitization of bee label data in projects like the Bumblebee Project (Smithsonian Transcription Center 2020), and ongoing digitization by the USGS and USDA. Yet, the challenge is that even with this data, researchers can only assess the status of a fraction of the 20,000 known global bee species because we lack mechanistic understanding of factors leading from species- to community-scale pollinator decline, driven in part by a paucity of data on important species-level traits related to pollinator response to anthropogenic stresses. Thus, we need to fundamentally change the direction of arthropod specimen digitization to extend specimen digitization beyond the label data and create resources that include researchers early in the process. Historical specimen label data are critical in understanding long-term pollinator declines, but also contain well documented biases, and will not alone answer fundamental questions about insect declines, especially on a global scale, highlighting the need for efforts to capture data beyond what is on the labels (Shirey et al. 2020).

Arguably, molecular methods are one tool for the identification of bees, but the scale at which they can be deployed for widespread monitoring is questionable. Analyses of mitochondrial and nuclear DNA sequence polymorphism have been widely used to discriminate among bee species and subspecies (Francisco et al. 2001; Whitfield et al. 2006; Ramirez et al. 2010; Rasmussen & Cameron, 2010; Theeraapisakkun et al. 2010; Quezada- Euán et al. 2012). These methods require specialized personal knowledge and a well-equipped laboratory (Francoy et al. 2008), or outside services (i.e., Barcode of Life). With these barriers, morphology and identification using local expertise continues to be the default means for identification, and morphometric variation in bee anatomy is sufficient to successfully identify bees, given enough image data (Rattanawannee et al. 2015; Buschbacher 2019), especially when linked with geographic information. In addition, technologies such as geometric morphometrics and computer vision have become increasingly useful in identifying species from images and have sparked interest in non-destructive methods of determination.

III. Intellectual Merit: A Thematic Network Focused on Functional Traits and Images:

Functional traits are phenological, physiological and anatomical characteristics that can be measured at the specimen level, which are not only related to the fitness of an organism, but also its function within the ecosystem (Violle et al. 2007). Bee functional traits are used to understand effects of pesticides (Brittain & Potts 2010), sensitivity of bees to land use pressures (De Palma et al. 2015; Harriston et al. 2018), susceptibility to parasites (Fries et al. 2007), and many other critical questions. These human created pressures are unlikely to affect all species similarly, and the specific effects will be mediated by traits unique to each species (Murray et al. 2009; Roulston & Goodell 2011). For example, species with narrow ecological niches (e.g., restricted geographies, brief adult flight seasons, or narrow floral specializations) are predicted to be more sensitive than species with broader ecological niches (e.g., wide ranging species, with longer flight periods, or more generalized floral associations) (Den Boer 1968; Kassen 2002). However, compiling species-specific ecological traits for researchers is extremely difficult and time consuming, and for many species lacking entirely. This, combined with the known taxonomic difficulties of identifying contemporary specimens (Gonzalez et al. 2013), constitutes a serious impediment to researchers trying to explore ecological mechanisms driving pollinator declines. ***Here, we propose to create a specimen digitization project specifically designed to benefit both bee ecological research and taxonomy by providing open datasets that link ecological and anatomical traits and tackle the problem of specimen identification through traditional and computational methods.***

The digitization of traits from specimens is captured in the concept of the "Extended Specimen," and a growing number of studies are published that explore the relationship between trait diversity, species composition and environmental change in bees (Lendemer et al. 2020; McCravy et al. 2019). Functional diversity of bees in an ecosystem may increase seed set and other measures of pollination success. Blitzer et al. (2016) showed that bee diversity based on nesting habits, sociality, and body size was associated with increased seed set in commercial farms. The commonly used functional traits used in bee research that this project will focus on are hairiness (pilosity), body color, body size, phenology (e.g., seasonality, flight period), nesting biology, sociality, and biotic associations (e.g., floral associates, parasites, pathogens). The ability to score and record other traits, including pollen-collecting structures, surface sculpture and trophic specialization will also be supported by the project. We have enlisted a Research Advisory Board to provide recommendations for potential targets and other traits during the project (see Trait and Character Digitization Targets, below).

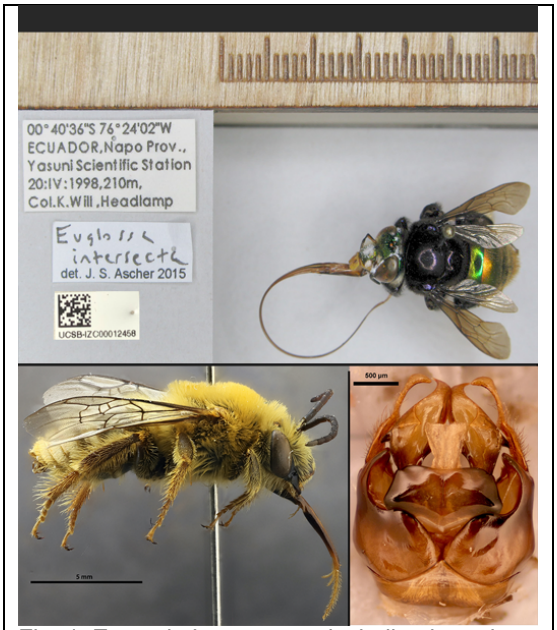

Fig. 1: Example image types including lateral exemplar images of *Peponapis pruinosa*, and male genitalia diagnostic image of the same species.

IV. Building datasets, identification tools and partnerships in a focused Symbiota Portal:

We will create a Symbiota-based **Bee Library** portal to aggregate bee data and develop novel functionality to manage, search for, and share images and trait data. Functional trait data from specimens is available on labels (e.g., biotic interactions, habitat, floral associates), and some traits can be estimated using label data from many specimens (e.g., flight period, phenology), or scored directly from specimens or specimen images (e.g., hairiness, body size). Additional trait information generally does not exist in collections and come from literature or observational sources (e.g., nesting habit, sociality). In order to assemble a complete resource for bee functional traits, we must obtain information from all of these sources, thus extending the digitization of specimens to include literature, observations and new methods of digitization. Examples of highly successful organism trait-based data resources exist in in plant biology with the Plant Trait Database (TRY; Kattge et al 2011) and entomology with Ant Web (California Academy of Sciences, 2020). Successful trait focused ADBC projects include the California Phenology Project (TCN - Yost et al. 2020).

Training datasets: Big-Bee will publish versioned datasets of the textual trait data and “training datasets” of image data as products. Training datasets play a critical role in computer image analysis, or computer vision applications (Robertson et al. 2019). To achieve species recognition capability in any

machine vision system, training datasets should include multiple images and angles of the same species, including variation based on sex, size and geography. These datasets compile representative images containing the explicit, verifiable identifications of the species they include and are the raw data used in creating models for automated species recognition. Recent advances in machine learning based image recognition have moved attention to data driven approaches, which are done by learning distributions of features present in the underlying data to discriminate species. Convolutional Neural Networks have proven to be particularly successful at classifying and detecting objects in images compared to traditional descriptor-based strategies, and even outperformed humans in terms of classification accuracy (He et al. 2015). They have been successfully applied to species identification tasks for plants (Barre et al. 2017), wild animals in camera-trap images (Norouzzadeh et al. 2018), and birds (Martinsson 2017).

Making images searchable: One present impediment for doing image-based studies with collection data is that it is not presently possible to search and aquire images based on the view of the specimen in the photograph (i.e., head, lateral, label) in any database (SCAN, iDigBio, GBIF). When searching for images of a taxon, label images, or head images, etc. all appear without the ability to filter them based on the desired viewing angle. Big-Bee will improve the ability to pre-filter images by annotating with an "image view" property in Audubon Core. Image searches in Big-Bee will also be based on geographic location, regional taxa (including all bees), and view. Using this functionality, a user could, for example, access all *Bombus* images with labels from Santa Barbara and all lateral view images of any specimen, or a combination thereof. In addition, image views will be tied to ontology terminology in the Hymenoptera Anatomy Ontology (Yoder et al. 2010; Seltmann et al. 2012) so that views are explicitly defined.

To improve consistency in our image products, all collaborating institutions are preparing both 3D and 2D exemplar images using the same imaging system with close coordination and training to create as uniform a dataset as possible. Csaba Molnár from the Biological Research Centre in Szeged, who is computer scientist specializing in biomedical image analysis will serve as a member of our Research Advisory Board (see collaborator letter) and provide expert advice regarding the utility of Big-Bee images for computer vision analysis. In addition, three specific data use pilot projects in computer vision are planned at UNR and UCSB that will use the images to test our data collection methods and their applicability to research.

***Innovations in computer science*:** At UNR, exemplar images will be used in the Notes From Nature (NFN; Hill et al. 2012) platform to annotate hairiness in specific areas of the bee in order to accelerate trait-based annotation of bee images. Hairiness is a novel trait for annotation using citizen science tools and methods to calculate the radius of a circle being placed on the image using the scale bar will be developed using deep learning convolutional network techniques (Sobel 2014; Hough 1959; Duda & Hart 1971). At UCSB, two modes of exploration are tied to Computer Science (CS) student internships that will train both biologist and CS students (see Broader Impacts). The first examines methods for 3D image reconstruction and the second uses convolutional neural networks to advance our ability to identify bees from images. While methods to render images and movies using focus stacked imagery are well developed, capabilities related to 3D modeling are an area of innovation and bees are significant because they are small, hairy, and complex objects that are challenging for traditional photogrammetry (Personal Communication, Tom Duerig, Google Research).

IV. Broader Impacts:

Partnership with the USDA National Bee Monitoring Network RCN: Big-Bee will provide digitization resources for all global bee collections including the ability to digitize specimens through the Big-Bee Symbiota portal, and support from a portal data manager to help train and facilitate upload and data sharing. Part of Big-Bee's success will come from its many partnerships. The recently funded USDA National Bee Monitoring Network RCN (Woodard et al. 2020-2023, see collaborator letter) is a research community of over 100 US bee researchers whose goal is to create a set of national standards and practices for native bee monitoring, and to coordinate and identify national priorities for monitoring and create a training program to support the monitoring workforce, interfacing with the public and policy makers. UCSB and Big-Bee will participate fully with the USDA-RCN by providing a portal and expertise for bee data digitization and specimen data management, including monitoring, image and trait digitization. We will attend RCN meetings and give presentations in order to coordinate fully with the project, including inviting RCN members to join the Big-Bee Research Advisory Board.

Workforce training and development: The PIs, collection managers, staff, interns, students, and volunteers involved in this project are part of the present and future biodiversity workforce. Big-Bee will provide opportunities for all participants to advance skills in biodiversity information science, bee identification, high resolution imaging, museum curation, and outreach. Training will be conducted over Zoom and documented using screen capture for a Big-Bee project YouTube channel. UCSB Data Curator and the ASU Data Manager will develop training workshops, while drawing on expertise from the entire network. Training workshops topics include: 1) 3D and high-resolution imaging provided by Mark Smith,

owner and developer of Macroscopic Solutions, LLC; 2) data curation and workflows using Symbiota portal including georeferencing, data management, trait annotation and Bee Library dataset creation; 3) annotating and defining traits in the Bee Library; and 4) creating a citizen science bee inventory using the Bee Library to create identification tools.

Building the capacity for expanded bee research: We propose to improve our capacity for bee research through simplifying the ability to create reference image datasets; create online and undergraduate focused bee identification training curricula; provide yearlong museum research and digitization internships; and expand interest in insect computer vision topics through undergraduate training. Two online bee morphology and Bee Library workshops will be developed in years two and three of the project. The first will focus on bee male genitalia and it will include how to prepare specimens and use the Bee Library for identification by comparing images to dissected specimens. In year three, a two-day virtual workshop on bee identification and trait determination will train novel bee researchers on how best to use the Bee Library to inform their identification efforts. Participants will use images of focal taxa to learn generic-level identification, and explore approaches to determinations using extracted regional species checklists from the Bee Library and diagnostic image comparison to undetermined specimens. Workshops will be held over Zoom, using digital microscopes to demonstrate dissections. Pre- and post-surveys will be used to evaluate the efficiency of conducting the workshops online, and improvement in bee identification by the participants. Workshops will be promoted through the Entomological Society of America, Ecological Society of America, and Bee RCN. In preparation for the workshops, and for further testing of the Bee Library, PI Carper will test the identification tools as part of a pollinator conservation module in an undergrad/graduate Insect Biology Course and the CU Bee Club. PI Tucker will incorporate the identification tools in projects and outreach activities at the University of Michigan with undergraduates and graduate students participating in independent research.

Undergraduates will also be extensively involved in the imaging and digitization of specimens. The California Academy of Sciences will offer one student a Summer Systematics Institute (SSI) internship, where a student is co-mentored by an entomology department curator and the CAS citizen science department on a project using digitized specimens and data from iNaturalist to look for temporal shifts in distributions or phenology. At the conclusion of the SSI program, this intern will transition into the digitization project for the academic year. In the spring, the intern will be paired with a high school student from the Careers in Science (CiS) program as a peer-mentor. CiS is a year-long paid high school internship program aimed at recruiting local students from underrepresented groups into the STEM-pipeline earlier. All interns will be connected with the Entomological Collections Network and the iDigBio network to ensure that they have a plan for future work or education.

Entomology and computer science: During the first two years of the project, two undergraduate computer science projects will be developed by a Postdoctoral Researcher at UCSB for incorporation into the recently funded NSF “HDR DSC: Collaborative Research: Central Coast Data Science Partnership: Training a New Generation of Data Scientists” initiative. PI Seltmann is a domain specialist for the project who provides real-world problems in data science for students. The first project will explore available methods of computer vision for identifying bees to species. One method is DeepAIBS, which has shown effective in identifying bee species based on wing venation patterns (Buschbacher 2019). The second project will examine several methodologies for 3D reconstruction that may improve on present photogrammetric techniques. An example of a target methodology uses Google Colab, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis (Mildenhall et al. 2020). Both entomology and computer science students will be recruited to work collaboratively on the project.

Creative community engagement: We will engage the broader community via a science-themed radio program, Unknown Territories (Cohen 2016), produced by science communication interns, and disseminated (live, online, and podcast) via the college community radio station, KCSB 91.9 FM, whose broadcasts reach a potential audience of hundreds of thousands across central California. Seltmann has produced over 400 episodes of the Unknown Territories over the three years. The project will produce a short, “sciencey” news segment called The Buzz for broadcast on KCSB News and during the weekly Unknown Territories show. The Buzz will be promoted for syndication at other college radio stations and the content will highlight natural history, conservation and research in a creative fashion. Undergraduate college radio and communications internships will be provided to produce the segments and the radio broadcasts, which will require research into the subject by the students, recording interviews with researchers within and outside of the project, audio editing and pre-and post- production training.

Diversity, accessibility, inclusivity, and training: Because these activities depend on remotely-accessible image data, we will have a great deal of flexibility to conduct training workshops regardless of on-campus work restrictions that may occur due to COVID-19. Collections developing undergraduate course materials that involve bee identification, computer science or use of data derived from specimens, will contribute lesson plans to BLUE (Biodiversity Literacy for Undergraduate Education - Ellwood et al. 2019). All network partners will continue their efforts to recruit students from underrepresented groups

using the resources available at participating institutions including the Society for the Advancement of Chicano and Native American Scientists, and the California Alliance for Minority Participation. CAS will recruit students San Francisco State University (SFSU) and the City College of San Francisco (CCSF). SFSU has an undergraduate student population of Hispanic biology majors (43%) exceeding the proportion of Hispanic residents in the state of California (38%), and CCSF has among the highest proportions of African American students of any college in the Bay Area (7%). UCSB is a Hispanic Serving University and we promote our internships with the UCSB Office of Education Partnerships, whose mission is to increase college-going rates of low-income students and those who will be the first in their families to pursue higher education. The UMMZI will continue to recruit and encourage student participation from the Doris Duke Conservation Scholars program (DDCSP) and the Undergraduate Research Opportunity Program (UROP). UROP provides paid research opportunities throughout the year allowing students who may come from less financially stable communities the ability to conduct research without having to choose between an internship and paying bills. The DDCSP provides summer research opportunities specifically for students that are at risk or underrepresented in science. The UCMC will also recruit students through UROP and will seek guidance from the Center for Inclusion and Social Change to increase recruiting efforts to under-represented groups.

V. Project Plan:

Participating Institutions: Our network is composed of entomologists - ecologists, morphologists, systematists, taxonomists, data scientists and evolutionary biologists. We are invested in the idea of creating a global network focused on the dissemination and discoverability of bee information, providing resources to improve bee trait data dissemination and specimen identification, and providing resources for natural history collections to digitize historic bee specimens. Twelve collections (Fig. 2) are participating in the formation of the network across the US, three large collections new to the ADBC program (UCB, FSCA, LACM), seven mid-sized to large collections involved with the ADBC program (CAS, SEMC, UNH, FSCA, UCMC, ASUHIC and UMMZI), and two smaller collections (UCSB, SDMC). Six collections have expertise in hymenopterology, bees, bee ecology and bee traits. Combined, our collections provide a rich historical record from across the country including T.D.A Cockerell's collection (1905-1930s; UCMC), Charles Michener's collection (1934-2015; SEMC), U.N Lanham (1930s, 1960s-1988, UCMC), Francis C. Evans collection (1948-1997; UMMZI), and L. Walter Macior Bumble bees (1967-2007; UCMC). More contemporary records from across the US include specimens collected by Michael Engel and Victor Gonzalez (KU), Virginia Scott and Adrian Carper (UCMC), and Gordon Fitch et al. and Erika Tucker et al. (UMMZI). Specimens from extensive ecologically focused and urban impact studies include: restoration ecology surveys (UCSB), Los Angeles urban collections from the BioSCAN project (LACM), numerous Colorado Native bee research projects (UCMC), Michigan Edwin S. George Reserve ecological surveys and southern Michigan urbanization research studies (UMMZI), UC Reserve surveys and the California Insect Survey (ESSIG), Kremen's agriculture and native ecosystem interface (ESSIG), and Gordon Frankie's lifelong urban bee ecology works (ESSIG). Additionally, our network of institutions house a number of important specialty collections such as the Type collection (MCZ), the Euglossini and *Osmia* associated herbaria host plant collection (FSCA), the Michelbacher squash bee surveys collection and Howell Daly collection (ESSIG), a vast bee dissections and wings collection (SEMC), Peggy Byron's Colorado Bumble bees collection (UCMC), the Roy Snelling collections from Mexico, Kenya, Tanzania, and Irian Jaya (LACM), an extensive collection of fly parasitoids and commensals of bees with hosts (LACM), and a large collection of bee mite associates and hosts (UMMZI).

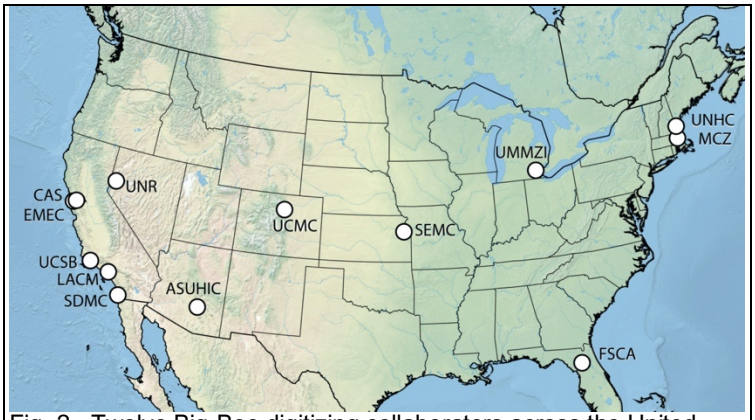

Fig. 2 - Twelve Big-Bee digitizing collaborators across the United States. Additional informatics and management centers are at University of Nevada, Reno (UNR), UCSB, and ASU, and Data Providers at USGS Patuxent Wildlife Research Center (Laurel, MD) and Karlsruhe Institute of Technology (Karlsruhe, Germany).

Project partners, including federal collections (USGS), account for additional bee expertise, data and approximately 4000 additional high-resolution images (see Droege collaboration letter). Thomas van de Kamp at Karlsruhe Institute of Technology (KIT) Institute for Photon Science and Synchrotron Radiation (IPS) has also agreed to provide CT scans of bee male genitalia to the project, with the aim to scan

~2000 bees depending on the availability of beam time during the project years. Five collections are in California as the state is a hotspot for bee diversity in both wild and urban ecosystems (Frankie 2009; Meiners 2019), and the digitization of pollinators in the state is synergistic with the funded California Phenology Project (CAP TCN) whose aim is to digitize California's herbaria. Their digitization efforts are creating a rich dataset tracking the distribution and phenology of the floral resources for California bees.

Digitizing institution	Digitized specimens	Focus-stacked images	3D image suites	Additional Images
Arizona State University (ASUHIC)	10,000	0	0	
California Academy of Sciences (CAS)	142,325	5,474	500	
Florida State Collection of Arthropods (FSCA)	90,495	5,294	500	
Harvard University (MCZ)	52,324	6,492	500	
Karlsruhe Institute of Technology (KIT)	2000	0	0	2,000 (microCT)
Natural History Museum of Los Angeles County (LACM)	72,681	6,543	500	
San Diego Natural History Museum (SDMC)	14,224	797	3,000	
University of California-Berkeley (EMEC)	69,297	9,892	100	
University of California, Santa Barbara (UCSB)	8,697	5,092	4,000	
University of Colorado (UCMC)	62,677	10,160	100	
University of Kansas (SEMC)	9,495	28,185	100	
University of Michigan Museum of Zoology (UMMZI)	27,815	15,198	500	
University of New Hampshire Collection of Insects and Arthropods (UNHC)	6,783	12850	500	1,000 (CLSM)
USGS Native Bee Inventory and Monitoring Lab (USGS BIML)	4000	4,000	0	
Total:	**572,813**	**109,977**	**10,300**	**3,000**

Table 1: Digitization goals including publishing 572K specimens digitized using multiple imaging modalities; 109K high-resolution, focal stacked exemplar and diagnostic images; 10,300 3D image suites consisting of ~64 focal stacked images per suite; and 3,000 CT or CLSM images. Grey highlighted rows are non-funded partners.

Digitization Deliverables: In total, Big-Bee will create over 1.2M images of 572K bee specimens, including over 519K specimens with labels for transcription of labels from images. To date, digitization of arthropods has largely been focused on label data, often without imaging the label itself. The Symbiota Collection of Arthropods Network (SCAN) database, which aggregates globally shared arthropod data, contains 19,179,675 arthropod specimen records (August 18, 2020). Of these records, 18% have an image associated with the specimen record. Specimen label images are important for data quality control and specimen images are needed for identification, trait and taxonomic discovery. In addition, the sex of bee specimens, important for understanding flight phenology and floral host resources, will be recorded for all Big-Bee specimens. As of September 10, 2020, only 34% of bees have sex recorded in online data.

In this project, we have a series of detailed digitization targets (Fig. 1; Table 1) intended to create complete data products for the focal taxa that emphasize research in bee functional traits and image analysis. Each target provides the ability to record specific functional traits and record other life history characteristics from the specimens or labels. An individual specimen maybe imaged using multiple modalities. These are:

⇒ Over 519,000 newly digitized specimens with **Label with Specimen Images,** focused on complete digitization in network collections for the 5,509 target taxa. Label with Specimen Images will be used to transcribe label data, score morphological traits (Body Color, Hair Color, Body Size and presence of mites and other attached arthropod Biotic Associates), update taxonomic identifications and determine specimen sex;

⇒ 31,000 high resolution multi-stacked **Diagnostic Images including Male Genitalia Images** that visually describe focal features for species and genus level identification;

⇒ 27,000 high resolution multi-stacked images of bee parasites;

⇒ 51,977 **Exemplar Images** of approximately 1/25 of all target taxa specimens (~8%) across network collections to capture morphological variation across taxa, inform computer vision applications, scoring traits (Hairiness, Body Color, Hair Color, Body Size and presence of mites and other attached arthropod Biotic Associates), update identifications and sex;

⇒ 10,300 **3D Image Suites** consisting of 64 images shot in 360 degrees around a specimen for 3D image reconstruction. The image suites will be available as datasets to inform computer vision applications and will result in over 659,000 high resolution multi-stacked images of bees; and

⇒ **Life History Traits** from literature data for all focal taxa including Phenology, Nesting Biology, Sociality and Biotic Associations Including Floral Specialization, Parasites and Pathogens.

VI. Focal Taxa:

There are approximately 20,000 species of bees worldwide and we will capture labels, images and traits for over ¼ of the global species (~5,500 species). The goal of this network is to ultimately assemble label, trait, and image data for all bee species. However, choosing focal taxa will lay the foundation for a depth of bee data not currently available elsewhere, that can be applied to all taxa and related biotic associates. The focal taxa are: *Bombus*, *Peponapis*, *Xenoglossa*, and *Xylocopa* (Apidae); Osmiini and *Megachile* (Megachilidae); *Andrena* (Andrenidae); *Colletes* (Colletidae); and all genera of Melittidae (Table 2).

The taxa have been selected for a number of reasons. First, and foremost, the majority are of significant bio-economic importance because of their utility as alternative pollinators to honeybees in agricultural settings. All Apidae genera (*Bombus*, *Peponapis*, *Xenoglossa*, *Xylocopa*) are economically important in agricultural systems and *Osmia*, *Bombus* and *Megachile* are highly managed for agriculture in the United States (Wilson et al. 2016; López-Uribe et al. 2016; Pitts-Singer et al. 2016). The other taxonomic groups were selected for ecological and evolutionary importance. The *Andrena*, *Megachile*, and *Colletes* are known to exhibit wide variation in specificity to floral associates. There is also great variation in nesting biology and sociality in the selected taxa, from sociality in *Bombus*, to aggregate nesters in *Colletes* and Melittidae, to solitary nesters in many other genera. Some species are of conservation importance as their populations have declined (Cameron, et al. 2011). As one of the most commonly studied genera, some *Bombus* have official conservation status with 91 species evaluated by IUCN that lists five species as critically endangered, eight as endangered, four near threatened, and ten species as vulnerable. In contrast, at least two of the genera (*Osmia*, *Megachile*) included in our selection contain introduced species. Of evolutionary significance, Melittidae are important to the study of bee evolution as they are basal to the rest of bees (Danforth et. al, 2006), and contain many pollen specialists. Additionally, Osmiini is a tribe that has a worldwide distribution with many evolved life histories strategies including parasitism and are especially diversified in Mediterranean and desert climates (Michener, 2000).

Family	Tribe	Genus	Species
Apidae		*Bombus*, *Peponapis*, *Xenoglossa*, *Xylocopa*	678
Megachilidae	Osmiini	*Osmia, Afroheriades, Ashmeadiella, Atoposmia, Chelostoma, Haetosmia, Heriades, Hofferia, Hoplitis, Hoplosmia, Noteriades, Ochreriades, Othinosmia, Protosmia, Pseudoheriades, Stenoheriades, Stenosmia, Wainia, Xeroheriades*	1087
	Megachilini	*Megachile*	1522
Melittidae		*Dasypoda, Samba, Capicola, Eremaphanta, Hesperapis, Ceratomonia, Meganomia, Pseudophilanthus, Uromonia, Afrodasypoda, Macropis, Promelitta, Melitta, Rediviva, Redivivoides*	206
Andrenidae		*Andrena*	1556
Colletidae		*Colletes*	460
		Total species	**5509**

Table 2: Focal taxa for Big-Bee, representing over 1/4 of global bee species and many significant species of bio-economic importance.

VII. Trait and Character Digitization Targets:

Body color, hair color, body size: Body Size is one of the most common morphological traits used to explore functional response of bee communities to anthropogenic disturbance. It is easily estimated via quantifiable features, such as intertegular distance (ITD), the distance between the wing-coverings (Cane 1987) and is directly related to foraging behavior (e.g., Greenleaf et al. 2007; Spathe & Weidenmuller 2002), which has implications for species-specific response to disturbance. For example, using museum records of 438 bee species in the northeastern US, Bartomeus et al. (2013) demonstrated that large-bodied bees, namely *Bombus* spp., were particularly vulnerable to increasing anthropogenic land use, which other studies have also confirmed and attributed to potentially increased food-limitation for large species (Scheper et al. 2014). Given its importance, we lack good estimates of body size for many species. Big-Bee will produce standardized measures from dorsal views, providing inter-specific body sizes estimates, as well as intra-specific and regional variation in body size using multiple specimens from collaborating institutions. Insect color can come from a diversity of traits and serves many biological and ecological functions that have broad evolutionary implications (Badejo et al. 2020). In bees, Body Color and Hair Color can be quite different, with the integument sometimes having metallic reflections, or

iridescence. Quantifying these color traits would not only provide important intra- and inter-specific variation for taxonomists, but could help researchers interested in exploring ecological or evolutionary drivers of bee color.

Hairiness (pilosity): We will also record pubescence, or hairiness, as a salient trait of pollinators that has been linked to their effectiveness as pollinators in agricultural crops (Woodcock et al. 2019; Phillips et al. 2018), is related to thermoregulation (Stiles 1979) and climate-driven geographic distributions (Peat et al. 2005), which could potentially be predictive of species-specific response to climate change. We will measure hair density according to Roquer-Beni et al. (2019) using citizen science volunteers to count hairs from an image in a given area (see Data and Bioinformatic Workflows below)

Life history traits: *Biotic Associations Including Floral Specialization, Parasites and Pathogens:* Museum specimen data has suggested that declines in host-plants can be a major driver of bee declines (Scheper 2014), though results have been limited in scope. While male bees forage primarily for nectar, female bees forage for both nectar and pollen to provision nests and likely contribute more to pollination services than males This difference in utilization of floral resources is important in understanding the floral specialization of bees (Danforth et al. 2019), thus recording the sex of the bee is critical during the digitization process. This project will leverage prior work from the Terrestrial Parasite Tracker TCN to use Global Biotic Interactions (GloBI, Poelen et al. 2014) ability to index biotic interactions from existing collection management systems (e.g., Arctos, Symbiota, EMu, Excel, MS Access, Specify), thus avoiding the need to build new cyber-infrastructures or disrupt associated institutionalized workflows. In addition, GloBI can index bee associations from the literature. As of September 2020, digitized bee specimens recorded in Global Biotic Interactions contained 302,926 bee floral associations with 8423 unique plants, 25,992 interactions with non-arthropod parasites, and 6277 arthropod parasites.

Fig. 3: Bee mites (all of the brown bumps) can often be seen on specimen. They will be recorded during digitization.

Parasites: One of the causal agents in decline of bees are arthropod parasites and the pathogens they may vector. Big-Bee will target two large bee parasite collections (phorid flies, mites) and expose other bee-parasite, bee-pathogen relationships from the literature including Meloidae beetles, other hymenopteran and dipteran parasitoids, brood parasitoids, kleptoparasites and Strepsiptera (Whitfield & Cameron 1993). We will digitize the LACM collection of bee-associated phorid flies (5500 specimens) and about 200 species of phorid flies that are known to be associated with bees. We will also digitize 14,600 mite specimens at the UMMZI and train digitizers in partner collections to recognize mites on bees during the digitization process, entering them into the database as present on specimens when found (Fig. 3).

Phenology (e.g., seasonality, flight period): The timing of seasonal life-history events is critical to understanding how biotic interactions change through time and in response to changing climates. Major changes in pollination networks through time have been attributed to shifting phenology of both plants and their pollinators (Burkle 2013), creating concern over potential phenological mismatches in pollination. Moreover, asymmetrical phenological responses of plants and pollinators can also be mediated by different human disturbances such as land-use change (Fisogni et al. 2020), suggesting that interactions between climate change and other human stresses may exacerbate the effects of already susceptible pollination networks. While museum specimen data has been used to explore the phenological traits of bees and their host plants (Bartomeus et al. 2011), increasing access to phenological records, and in relation to other life history traits (see *Nesting Biology and Sociality* below), could help elucidate additional ecological drivers of phenological change.

Nesting biology & sociality: Nesting biology is a dominant factor driving bee-response to agricultural land-use (Forrest 2015). For example, bees nesting above ground have been shown to be more impacted by habitat loss and fragmentation, compared to below ground nesting species which are impacted by agricultural practices such as tilling (Williams et al. 2010). Such trait-specific responses within bee communities can impact pollination services to crops (Hoehn et al. 2008), and may be reinforced by surrounding human disturbance (Martins et al. 2015). Moreover, species-specific differences in bees' response to climate change is likely a result of both ecological and evolutionary factors, though few studies have explored ecological drivers of flight phenology in bees. A recent study of 67 bee species in the Colorado Rocky Mountains, found that while adult bee emergence was mostly driven by the onset of snowmelt, a climate variable, the peak of activity and date of senescence were more driven by life-history traits, namely where species nest and in what stage of development they overwinter (Stemkovski et al. 2020).

Focal diagnostic features including male genitalia images: Bee identification is still primarily done using keys and morphological characters. The availability of diagnostic character images, or images that illustrate focal characters used to define bee species, are difficult to find (located in the literature) or unavailable. Big-Bee will image these characters for bee species and genera following recent revisionary work and the most widely used texts for identification including *Bees of the World* by Charles Michener, *The Bees of the Eastern United States* by Theodore Mitchell and *The Bee Genera of North and Central America* by Charles Michener, Ronald J. McGinley, and Bryan N. Danforth.

Male terminalia (apical most abdominal sternites and male genitalia) are arguably the most important diagnostic character in insect systematics and have been the focus of Apoidea research since Audoin (1821). Almost all critical keys for species identification of bees use male genitalia (Mitchell 1960; Milliron 1970; Kuhlmann 2003) and images of male genitalia are necessary for descriptions of new species (Ayala & Griswold 2012; Onuferko 2019; Hall et al. 2016). Male genitalic characters can also help resolve higher level phylogenetic relationships (Roig-Alsina 1993; Cameron & Williams 2003; Williams et al. 2008). Major texts, including Michener's book Bees of the World (Michener 2000), rely on male genitalic characteristics. If searched by "male genitalia bee identification" in Google Scholar, over 7,380 papers have been published since 2016. Despite the fact that these characters are the most reliable and sometimes the sole sources of accurate identification of bee species (Hall et al. 2016), the understanding of these characters has commonly been lost on many bee observers and ecologists. We believe this is largely due to the lack of informative illustrations. We understand that many morphologists may not share images widely, however, we will provide the infrastructure for them to do just this, making an important character system transparent and accessible. In addition, Big-Bee will image every species of bee in the target taxa, providing the largest and most complete dataset for this character system.

VIII. The Bee Library:

Central to our network is the creation of the Bee Library Symbiota portal. This portal will provide the infrastructure necessary to manage, curate and serve images generated by participating institutions and innovate functionality to support rich trait, biotic interaction and image data. Symbiota is presently the most widely used software to organize and annotate data in the ADBC program. The 30+ different Symbiota data portals span the full range of metazoan diversity, so Big-Bee is well positioned to share data on bees, bee parasites, and bee floral hosts with existing portals. Symbiota functions as both an online database that allows users to perform easy, web-based data entry and a biodiversity content management system that allows for seamless ingestion of data from providers that use both Symbiota and non-Symbiota databases. The strength of Symbiota lies in the ease of packaging data from disparate datasets into a single portal, where the data can be refined, annotated and evaluated for quality together. These datasets can come from anywhere and the system is flexible to manage both specimen and observation-based data, thus not all records in Symbiota need to be tied to a museum specimen, and Symbiota can aggregate data from literature and observation sources (i.e., iNaturalist, BOLD, GloBI).

Building on the efforts of other TCNs: The Bee Library expands on innovations from existing Symbiota portals, including ADBC California Phenology Thematic Collections Network (CAP TCN) development of phenology trait annotation and the Terrestrial Parasite Tracker (TPT TCN) integration with GloBI and publication of Zenodo project reports. Big-Bee will also further develop the Occurrence Trait Management Tool from the newly funded Building a global consortium of bryophytes and lichens: keystones of cryptobiotic communities (GloBaL TCN). Presently, several versions of the Symbiota software are available to TCNs. We have chosen to extend the version that is maintained at ASU's Biodiversity Knowledge Integration Center (BioKIC) because of the advancements in trait and image functionality already integrated into this version that does not exist elsewhere. In addition, Big-Bee is benefiting and contributing to functionality being developed for the NEON biorepository project at ASU's BioKIC, which uses the same version of Symbiota as planned in Big-Bee.

This project aims to contribute developments toward a decentralized but globally coordinated system of biodiversity portals, thus allowing taxon or topic specific portals to exist and build community around shared interests (Sterner et al. 2020). Symbiota can already harvest taxonomic information via taxonomic authorities, APIs and taxonomic information can be batch loaded. The Documenting Marine Biodiversity through Digitization of Invertebrate Collections (DigIn TCN) funded this year includes Symbiota development that can be used to synchronize a taxonomic thesaurus against another Symbiota taxonomic thesaurus, or a CSV import file. These developments will assist in coordinating Hymenoptera taxonomy across SCAN, Ecdysis (Symbiota portal for insects at ASU), and the Bee Library. We will expand on the development, creating an API driven solution to automatically synchronize one portal's taxonomy against another.

Building on the Symbiota ecosystem of tools: The Bee Library will support discovering and structuring biotic interactions from collection data by **1)** expanding the ability for Symbiota to link external resources including the ability to link to externally managed literature (via Weblink DOI), observations

(e.g., iNaturalist), other image datasets (e.g., CT, CLSM and 3D models held outside of ASU), and other related specimens; **2)** investigating options for sharing Image View (e.g., dorsal, ventral) as a property in Audubon Core; **3)** extending the Occurrence Trait Management Tools to create an administrative user interface that will be used to establish new trait definitions, adjust existing traits and manage permissions; **4)** develop an Attribute Mining Tool to assist in standardizing trait and biotic association data typically defined within various fields (e.g., associatedTaxa, occurrenceRemarks, habitat, substrate, OCR output, etc.). This tool will function similar to the mining tool used by the CAP TCN, currently being used to extract verbatim phenology traits; **5)** Adjust the Dataset Management tools to enable sort, filter, and augment downloadable datasets based on image properties (e.g., dorsal, ventral, head). This will extend the Dataset Management Tool functions, whose core developments funded via NEON; and **6)** use built-in OCR functionality on images of insect labels. This functionality is already built-in Symbiota, but the utility of the Salix tool for insect labels is untested, and we will create improved libraries for insect labels.

IX. Data and Bioinformatic Workflows:

The Big-Bee digitization workflow (Fig. 4) follows established best practices for insect specimen digitization and innovates in trait capture and dissemination methods.

(1) Specimen Imaging: Specimen imaging will occur at each institution and specimens will be pushed through one of four workflows, based on the research-value of the specimen and institutional strengths. A single specimen may be imaged by one or several modalities, but all specimens that pass through the pipeline will have at least one image. UCSB and SDNHM will borrow material for 3D imaging from larger collections. Since SEMC has an extensive collection with most of the global bee species, SEMC will loan specimens to UCSB and UNHC for imaging of diagnostic features.

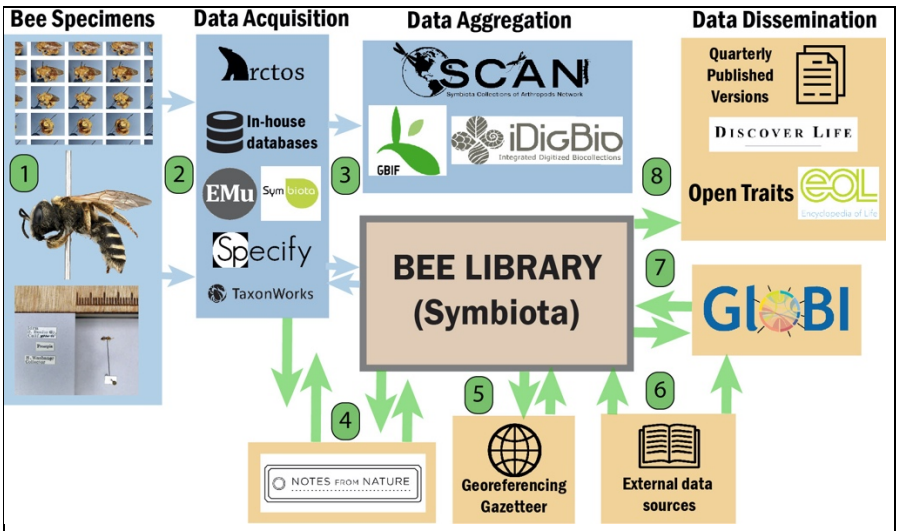

Fig. 4: Generalized bioinformatic workflow for the Big-Bee project. A workflow that follows existing best practices and establishes new systems for data dissemination and trait digitization.

(2) Institutional Databases: Specimen occurrence data and images will be entered into diverse institutional databases that collections presently use to digitize and manage data and images (e.g., EMu, Specify, TaxonWorks, Symbiota Monarch, Symbiota UCSB Collection Network and Arctos). High resolution JPG images will be imported into institutional systems for sharing.

(3): DwC-A and IPT: The Integrated Publishing Toolkit [IPT] or Darwin Core [DwC] archives will be generated by institutions to share data and images with the Bee Library, GloBI and other downstream data aggregators and users (e.g., iDigBio, GloBI, GBIF and SCAN). A collection will only need to produce one IPT or DwC-A to share data with all of these resources.

(4) Notes From Nature (NFN): Big-Bee will utilize NFN in three ways: **(a)** digitize specimen labels; **(b)** measure Body Size; and **(c)** annotate images for Hairiness. Expeditions (series of images) will be created on NFN for both label transcription and trait measurement. NFN is an online transcription tool for museum specimen data that, through the volunteer effort from the general public, has transcribed information from over 2.5 million specimens with over 10,000 volunteers. NFN currently has 5 active TCNs and has helped many institutions move specimens through digitization pipelines. Big-Bee will create focused expeditions, with clear instructions to volunteers in order to generate excitement over the trait annotation. Some institutions, like EMEC, have a long history of using NFN for all of their label transcription, so in some cases, expeditions for digitizing specimen labels may come directly from an individual institution. Expeditions to measure Body Size and Hairiness will come directly from the Bee Library because the Bee Library innovations to manage and share trait data. The project will repurpose the Zooniverse measurement tool, which will be used in an ADBC digitization project for the first time. In separate expeditions, volunteers will be asked to annotate images for trait data. Volunteers will record the

distance between the wing-coverings (ITD) using an in-photo scale bar in the first expedition. In the second, volunteers will be asked to estimate the hairiness of the bee by counting the number of hairs in a circle. Hairiness is a novel trait for annotation using citizen science tools, and the methods by which the circle is placed on the image will be developed using deep learning convolutional network techniques (Sobel 2014; Hough 1959; Duda & Hart 1971). Measurements are performed by multiple users and the mean score will be recorded in the Bee Library.

(5) ***Collaborative Georeferencing:*** To increase efficiency, georeferencing will be done collaboratively in the Bee Portal or using the Geolocate Collaborative Georeferencing Tool. If funded, Big-Bee will leverage the Georeferencing Gazetteer - Biodiversity Enhanced Locality Service (BELS) project that was recently recommended for funding through NSF (PI Allen at UCR is also a PI on BELS). Already in prototype, BELS is a georeferencing gazetteer built from all GBIF and iDigBio georeferences. BELS can be run through the Bee Library or through the Geolocate Collaborative Portal. BELS is a data service that will determine how many of the localities submitted to the service already have georeferences. In addition, BELS will expedite georeferencing by sending the remaining localities through GEOlocate. Overall, BELS will reduce the total number of georeferences that need to be manually curated. Data quality tools are integrated into the process and georeferenced coordinates will be returned to individual institutions via a Darwin Core Archive.

(6) ***External Data Sources:*** Big-Bee will integrate additional bee data from network collections, monitoring programs and federal collections (e.g., USGS, USDA, Smithsonian) via established IPTs or spreadsheet uploads. 1.5M records of bee data will come from network collections alone. The additional bee data are necessary to create a complete picture of bee occurrence records in the Bee Library. As our network grows, Bee Library will support live collections, and is flexible for data providers as to where and how their data are stored while ensuring that there are clear pathways for providing data to aggregators. Bee observations will be added from iNaturalist, BugGuide and from the literature. Literature records will go in the portal as a Literature Observation. All records will be linked to the originating source of the data through an expanded linked resources tab within the Symbiota occurrence editor, which includes the ability to link to both internally and externally managed literature (via Weblink DOI), observations (e.g., iNaturalist), and specimens in other databases.

(7) Global Biotic Interactions: GloBI is the largest global resource of interaction data, providing open access to over 5 million species interaction records, covering over 100,000 taxa sourced from hundreds of existing datasets, and citing more than 100,000 references. Big-Bee will leverage work in biotic interaction modeling initiated by the Terrestrial Parasite Tracker TCN (TPT TCN). The TPT TCN uses GloBI to facilitate access to, and monitor the availability of, 500,000 vertebrate arthropod interaction records. GloBI uses open source software and is actively cited in papers and review articles (e.g., Hortal et al. 2015; Nielsen et al. 2018). GloBI consists of three components: **(a)** an automated crawler that continuously discovers, integrates, links, and indexes existing interaction data and associated annotations from heterogeneous openly available data sources; **(b)** data services that provides access to query aggregated and linked interactions online via a web API, R package (rglobi), and offline-enabled access via Elton software (Poelen et al. 2020); and **(c)** published data archives (Poelen 2020) of aggregated and linked species interaction data in various formats (e.g., Darwin Core Archive, RDF/N-Quads, Neo4j, data dump, and CSV/TSV files). GloBI integrates with a multitude of existing biodiversity and bioinformatics projects such as NCBI taxonomy (Federhen 2012), GBIF Backbone Taxonomy (GBIF 2020), Encyclopedia of Life (Pafilis et al. 2015; Parr et al. 2014), Global Names Architecture (Pyle 2016), and CrossRef (Pentz 2001) to provide taxonomic resolution and citation DOI lookup. Vocabularies are tied to ontology (UBERON, Relations Ontology, EnvO; Mungall et al. 2012; Buttigieg et al. 2013; Seltmann et al. 2013; Pafilis et al. 2015; Simons & Poelen 2017), but data sharing is not dependent on resolution of a specific data sharing model or consensus over terminology.

(8) Data Dissemination: Our results will also be disseminated through published datasets as data papers, or versions of the data produced from the Bee Library using Zenodo and other semantically enabled publication sources. These data papers will facilitate detailed project tracking over time. Data papers that include a DOI allow for easier citation and research repeatability (Chavan & Peney 2016). These data papers will be shared through participation in the OpenTraits Network (Open Traits Network 2020), a global community of researchers working towards standardizing and integrating trait data across all organisms, Encyclopedia of Life TraitBank, provided to Discover Life through collaborator Sam Droge, and discoverable through Google and other Web search capabilities. Published datasets Big-Bee will provide include: **(a)** quarterly versions of entire database data for tracking overall project results; **(b)** yearly species level functional trait synopses detailing the trait results for each species (*hairiness, body color, body size, phenology, nesting biology, sociality, and biotic associations);* **(c)** and image datasets that contain resolvable links to the images for each 3D image suite. Reliable citation of images will be maintained using hash technology, which will allow citation and cloning of images of known provenance (Poelen 2020). Hashes are a unique signature for each image and will provide us with the ability to know

if images found in two locations (i.e., Big-Bee, iDigBio) are exactly the same. Thus, copies of images can be retained in any discoverable location and be compared for exact similarity. In addition, the Symbiota functionality of the **Dataset Management Tools** we are extending will allow any user of the Bee Library to create publishable datasets. For example, a researcher or educator interested in starting a survey program for bees in San Francisco can create a dataset of specimens found in their area of interest, based on taxa, or another filter. This dataset can then be downloaded as a DwC-A file, including links to the available images on the Bee Library.

X. Digitization Methods, Imaging Modalities and Innovations:

The target image types are ***Label with Specimen Images, Exemplar Images, Focal Diagnostic Features including Male Genitalia Images, and 3D Image Suites***. To increase quality and efficiency in digitization, the Big-Bee project will innovate in many areas, both in digitization methods and imaging workflows. The Label with Specimen data capture requires inexpensive equipment that are readily available and can be easily acquired by collections. Research grade multi-stacked imaging (exemplar, diagnostic, and 3D) will be generated using a single, compact system, the Macropod imaging system, the same imaging system successfully being used for the TPT TCN.

Fig. 5: One 3D model is created from ~64 high resolution multi-stacked images. These images are packaged together as an image dataset, useful for innovation in 3D reconstruction and computer vision identification methods.

Digitization of labels with Specimens: Every new specimen digitized will have a standardized dorsal habitus image with labels, on a standardized imaging stage provided to all institutions. The specimen will be imaged with a grey background and ruler. Specimens will be sexed during image label capture, and these images will be sent to Notes from Nature for label transcription, sexing from images, and trait measurements including Body Size (see Digitization Goals).

Exemplar and Diagnostic Images: Exemplar images will include head (front), dorsal, and lateral views of all bee specimens. These images will be generated using the focus stacking techniques and hardware solutions (Macropod Pro Imaging System) by Macroscopic Solutions in a standardized way across Big-Bee participating collections. Macroscopic Solutions technology is used to non-destructively produce 2D imagery and 3D models that are completely in focus, color accurate and high resolution for wet, dry and pinned specimens larger than 1.0 microns. Exemplar specimens of slide mounted wings, detailed diagnostic images, and male genitalia images will also be taken for all bee target taxa using the Macropod Pro. All images will include a scale bar. Macropod systems are fully portable and automated devices capable of capturing detailed, high resolution images for samples ranging from 1.0 to infinity. Male genitalia will also be imaged using microCT and CLSM.

3D Images: Different methods for capturing 3D models of small sized subjects presently include CT scanning, SEM/TEM and laser scanning, however, photogrammetry is the only method capable of rendering color in the 3D model. While color profiles are significant, the structural quality and resolution of models generated with photogrammetry are gradually inferior as samples reduce in size. Specific challenges for samples smaller than 2-3 cm include inconsistencies in brightness, color tone, chromatic aberration, background color and seamless motion along 6 axes. Recent advancements in software, image quality, lighting and backdrops are allowing photogrammetry to be used in combination with focus stacking to generate high quality 3D models for materials >500 microns. This is achieved through the creation of a microscopic universal stage that seamlessly maneuvers small specimens along 6 axes of translation/rotation within a working distance of only a few centimeters. Custom backdrops and lighting workflows are being generated to standardize image capture and create image data with very little deviation in color or brightness. Variations that may occur are capable of being automatically corrected using image correction software prior to 3D modeling and a novel stereo point cloud generation method is being used to more efficiently create 3D models. The Macropod captures focus stacked imagery in 5°

intervals per 360° around 2 distinctive axes that intersect at 90 degrees. Each 3D image takes about 2 hours to create the 64-image focal stacked images for one 3D reconstruction (Fig. 5). The images are isolated per stack, batch processed in Zerene Stacker, script processed in Photoshop CS6 and combined in Agisoft Metashape. The models generated by Agisoft Metashape record textural, structural and volumetric information used for analysis and virtual education purposes. In partnership with Macropod Solutions LLC, several technological innovations will result from this proposal including **Seamless motion along 6 Axes.** Currently, the Macropod allows seamless translation along the x, y and z axes and rotation around the y axes. Macroscopic Solutions is creating a low-cost universal stage that will permit seamless motion along 6 axes within a very short working distance. **Consistent Lighting.** Backdrops for 3D modeling vary depending on strobe configuration and color. Hotspots, glare or bright areas prohibit 3D modeling software from analyzing structural details necessary for point cloud generation. Macroscopic Solutions is developing a round dual-purpose shroud that encapsulates the specimen and universal stage while directing light towards a specialized lens mounted diffuser designed to retroreflect light towards the specimen regardless of a changing specimen orientation. This completely eliminates artifacts in the final images and allows for seamless point cloud generation. The shroud is painted neutral grey to standardize color profiles. **Automatic Image Editing in Scripts.** Scripts profiles will be shared to clean up focus stacked imagery. The clean imagery reduces build time and creates a higher quality model after a texture has been applied. **Stereographic Mosaic.** Macroscopic Solutions has shared data with Agisoft Metashape to create a stereographic generation method that is both efficient and accurate. It requires masking 1-2 images of a ~150 image dataset, low density image alignment and final model generation. This method is particularly good for materials with uneven textures such as hair or complex wing structures.

XI. Timeline and Management:

Day to day: Each collection PI or designated lead personnel in the collection will be responsible for the daily management of their imaging efforts. Collection leads and PIs will meet regularly over Zoom to coordinate and share digitization best practices. UNR will support the creation of new expeditions for collections in NFN. To support collections using volunteers and interns to digitize specimens, UCSB will set up a weekly Label Transcription and Georeferencing Working Group via Zoom to work collaboratively on georeferencing and transcribing labels. Big-Bee will reuse the CAP TCN online Georeferencing Course and other materials created during the project, including creating a YouTube Channel of how-to videos, which was successfully implemented in both the CAP and Tri-Tropic TCNs (Yost 2020). Volunteers, students and collection staff will be invited to attend within and outside of the network to include other collections and researchers digitizing bees.

Activity	Institutions	Y1	Y1	Y2	Y2	Y3	Y3
Bee Library Installation	ASU	X					
Bee Library Development	UCSB, ASU	X	X	X	X		
Trait Definitions	all, Research Advisory Board	X	X				
Imaging	all	X	X	X	X	X	X
Notes From Nature Expeditions	UNR, all		X	X	X	X	X
Label Transcription and Georeferencing	all		X	X	X	X	X
Score Traits	all	X	X	X	X	X	X
Workforce Training	UCSB, ASU, MCZ	X	X	X	X	X	X
Identification Workshops	UCMC, LACM, UNHC, UMMZI, UCSB				X	X	
Computer Science Internships	UCSB			X	X	X	X
Undergraduate Internships	UMMZI, UCMC, CAS			X	X		

Table 3: Project deliverables, responsibilities and training are distributed throughout the network.

Timeline: Year 1 will be devoted to portal development and aggregation of bee data into the Bee Library. In addition, training and imaging will start year 1. Imaging will continue through years 1-3, and trait scoring will be conducted throughout the term of the project.

Data management: Overall project management is centered at UCSB. The Bee Library will be hosted at ASU and ASU will provide data management support for uploading data to the Bee Library and helping collections in portal management activities. UCSB and ASU will collaborate for innovations in the Bee Library, aggregation of all bee data records, data cleaning, and quality control. UCSB will provide quarterly feedback to collections about data formatting, especially important to improve overall georeferencing, and formatting of trait and biotic interaction data. Individual institutions are responsible for

maintaining original copies of their own images (e.g., TIF, DNG) and JPG versions will be uploaded to the Bee Library, hosted through ASU. JPG storage requirements for the 1,500,000 total images has been estimated as a total of 6.3TB (average 4.2 MB each). ASU will additionally provide backup for those JPG images. Additional storage is provided by UCSB for CT, CLSM and 3D images and models. These will be shared through the portal as representative JPG images on the ASU server, linked to full download archives via UCSB.

ASU Developer Gilbert will work with the ASU Portal Manager to provide infrastructure support, and the ASU Portal Manager will work with the UCSB Data Curator to jointly provide the data management support. ASU Data Manager will (a) support and maintenance of the hosting environment including OS, web servers, and databases; (b) installation of regular updates for software; (c) monitor and maintain server and portal security (e.g., installation of security patches, SSL certificates, etc.); (d) maintain an archive of nightly data dumps of the MySQL database; (e) facilitate a central image server along with the workflows and processing scripts needed for batch image ingestion; and (f) perform SQL queries for custom data extractions, transformations, and data cleaning. Data management support includes creation of new collection profiles, performing regular data refreshes of snapshot datasets, setting up new ingestion profiles for batch submission and storage of specimen JPG images and derivatives, monitoring and cleaning specimen datasets, building and maintaining taxonomic thesaurus and other authority support tables, publishing specimen records to iDigBio and GBIF, and troubleshooting intermittent workflow and data issues. User assistance and help desk assistance will be provided via web and video tutorials, training sessions, and one-on-one consultations. Training workshops will be created by the UCSB Data Curator for creating datasets using the Bee Library, including implementing these in a local citizen science project. Data publications will be overseen by UCSB, in collaboration with the Research Advisory Board, UCSB Postdoctoral Researcher and GloBI innovator Jorrit Poelen.

Research Advisory Board: A research advisory board following the model developed by the TPT TCN will be developed to guide the project. The board will establish processes that maximize efficiency of digitization and opportunities for research collaboration. It will be composed of researchers across disciplines and from academic, agency, and industry partners. They will help define project goals and outcomes, develop conditions of the collaboration (e.g., defining traits, sensitive data, meeting deadlines for grant proposals requesting data, publications, etc.), and evaluating data publications from the Bee Library. The Board also helps determine digitization priorities and moderate authorship discussions.

Project assessment/data tracking: UCSB will be responsible for tracking the rates at which records are added to the Bee Library. The use of data will be tracked via Symbiota's built in tools that track searches as well as number of downloads. Surveys with workshop participants (see Broader Impacts) will be conducted pre-and-post workshops to evaluate the effectiveness of the images and Bee Library for improving bee identification and bee research. The data in the Bee Library will be published on Zenodo (Peters et al. 2017), and citation statistics will be compiled from Zenodo and GBIF.

Sustainability: The Macropod imaging stations will continue to be used by collections after the completion of this project. Each institution will receive training in imaging and they will have all the tools necessary to continue digitization. The complete datasets for bee species provide a template to continue digitization for all global bee species and 3D imaging in insects. The PIs will actively pursue future industry and university partnerships in data science, engineering and computer science with the promotion of easily accessible large image datasets. There is a strong expectation of continued baseline services for the duration of the NEON project for the Big-Bee portal as ASU is providing Symbiota portal hosting and management services to over 30 portal communities, including the NEON biorepository.

XII. Prior NSF (IM-Intellectual Merit; BI-Broader Impacts):
("¥" symbol in Literature Cited section indicates paper in Prior NSF)

The Hymenoptera Ontology: part of a transformation in systematic and genome science (HAO). **Seltmann & Miko** DBI-0850223 (2009-2013; $1,411,508). IM: Developed and made available the first multi-taxa insect ontology. 10 papers published. BI: 2 postdocs, 2 graduate students, and 2 undergraduate students trained. National and international outreach to the scientific community via 4 workshops and conference presentations.

Digitization TCN: Collaborative Research: Capturing California's Flowers: using digital images to investigate phenological change in a biodiversity hotspot. Maser & **Seltmann** DBI-1802181 (2018-2022; $255,844). IM: Digitize 40,000 plant specimens and develop phenology trait scoring. Two papers published. BI: 1 graduate student, and 28 undergraduate students trained. National and international outreach to the scientific community via 2 workshops and conference presentations.

Digitization TCN: Digitizing collections to trace parasite-host associations and predict the spread of vector-borne disease. **Seltmann** DBI-1901926 (2019-2022; $29,023). Create pipelines for biotic

interaction digitization using GloBI. 3 software and data publications BI: Two workshops. **Tucker** DBI-1902113 (2019-2022; $367,004). IM: Database specimens of 180,000 parasite specimens. BI: 6 undergraduate students trained. One workshop and 1 conference presentation.

Digitization PEN: Field Museum of Natural History Partnership with the Southwest Collection of Arthropods Network. **Maier** - transfer of support to Sierwald DBI-1802353 (2018-2021; $163,485) IM: Digitized 100,000 beetle specimens in 860 species. No publications were produced under this award.BI: This project will support informal educational opportunities focused on ground-dwelling arthropods.

Digitization PEN: Integration of data from the San Diego Natural History Museum with the Lepidoptera of North America Network. **Wall & Horsley** DBI-1903299 (2019-2021; $159,770). IM: This project will digitize over 150,000 Lepidoptera specimens. No publications were produced under this award. BI: Over 6 undergraduates and many volunteers will be trained in digitization techniques and collections-based research. Data from this project will add invertebrates to conservation risk analysis of BCP being conducted by CONABIO (Comisión Nacional para el Conocimiento y Uso de la Biodiversidad).

All Diptera Biodiversity Inventory of a Tropical Mid-Elevation Site. **Brown** and Borkent DBI-1145890 (2012-2014; $900,000). IM: Inventory all flies collected at a single 4 ha cloud forest site in Costa Rica yielding 4332 species of Diptera and providing the first verifiable diversity estimate of a major group of insects at a single site in the tropics. 6 papers published BI: Web sites and blogs documented the findings and Costa Rican biologists were trained and employed during the project.

Digitization TCN: Collaborative Research: Building a global consortium of bryophytes and lichens: keystones of cryptobiotic communities. Bungartz and **Gilbert** DBI-2001394 (2020-2023; $344,384.00). IM:Twenty-five US herbaria will digitize 1.2 million bryophyte and lichen specimens. Deep-learning approaches provide powerful new investigative tools. BI: K-12 object-based learning outreach programs, online lesson plans, Science Clubs and educator workshops, and videos using Learning Glass.

ABI Development: Notes from Nature: Advancing a Next Generation Citizen Science Platform For Biocollection Transcription. **Allen** ABI-1458527 (2014-2020**;** $594,428) IM*:* Provide a scalable means to increase the rate of mobilized collection data. 2 papers published BI: Connecting the public with scientists and collections specialists and further significant reach via the NFN blog and social media.

Digitization TCN: Fossil insect collaborative: A deep-time perspective to studying diversification and response to environmental change. **Engel** DBI-1304957 (2013–2018; $174,151). IM: Digitized seven major collections of fossil insects. No publications were produced. BI: Data served via the web.

IOS Collaborative Proposal: Scents of Self: How Trade-offs Shape Self/Non-self Recognition Cues in a Supercolonial Insect. **Tsutsuil** & Fisher OS-1557934 (2016-2020; $605,479).IM: Investigate determinants of desiccation resistance in the invasive Argentine ant. 3 papers published BI: 3 graduate students, one post-doctoral researcher, and ~20 undergraduates trained. Created the Backyard Biodiversity Project, guiding citizen scientists through the collection of > 31,000 arthropods from backyard swimming pools.

Understanding the Evolution of Diet Breadth through Ecoimmunology. Bowers & **Carper** (Postdoctoral Scholar) DEB-1456354 (2015-2019; $275,236). IM: Understand the role of immune response in mediating diet breadth in herbivorous insects. 4 papers published BI: 2 postdocs, 2 graduate students, and 11 undergraduate students were trained at the University of Colorado. 15 conference presentations, created museum exhibit, *Becoming Butterflies*.

National Ecological Observatory Network (NEON) Biorepository. **Franz & Gilbert** DBI 1724433(2018-2020; $4,216,393 subcontract from Battelle Memorial Institute). IM: Establish the NEON Biorepository - the primary repository for NEON-collected samples (ca. 110,000 samples annually) and publish NEON sample data through a dedicated, Symbiota-based biodiversity data portal. 3 papers published. BI: 8 undergraduates, and provided loan-related services to 60 external research groups.

Digitization TCN: Lepidoptera of North America Network: Documenting Diversity in the Largest Clade of Herbivores. **Pierce** DBI-1601124 (2016 -2020; $179,204) IM: Digitization and imaging effort involving 27 partnering natural history collections to mobilize biodiversity data from 3 million specimens of butterflies and moths. To date, the MCZ has generated 152,864 records and 41,179 high-resolution images, with 98% of specimens identified to species and 66% georeferenced. BI: Images and data generated from this project will be available online for education, identification and research on herbivores world-wide.

Nothing to report: **Talamas**, **Grinter**, **Oboyski**, **Gonazalez-Betancourt**