Collaborative Proposal: Digitization TCN: Terrestrial Parasite Tracker:

***Overview*.** The Terrestrial-Parasite-Tracker TCN (TPT) will digitize 1.3 million arthropod specimens representing taxa that are important ectoparasites and disease vectors of North American vertebrates. This digitization effort will complement the 15 million vertebrate records, as well as vector and disease monitoring data shared by state and federal agencies. Additionally, the TPT will extend the work of previous TCNs on biotic associations and establish an ontological framework resulting in research data sets that can address a wide-range of host-parasite questions. This is the first TCN to bring together taxonomists from vertebrate and invertebrate collections as well as epidemiologists and ecologists to integrate the vast amount of existing data on vertebrates with the significant new data set on arthropod ectoparasites. The TPT brings together 25 institutions, 12 new to the ADBC program, comprising a cohesive team of curators, data scientists, biodiversity content managers, domain scientists and educators. Finally, by linking contemporary and historic records this project will empower ongoing citizen science and public awareness campaigns to understand distribution changes of arthropod vectors and associated diseases due to climate change and global trade.

Intellectual Merit. Terrestrial arthropod parasites are responsible for significant and pressing issues in wildlife conservation, human health, and livestock productivity, collectively represent an unparalleled spectrum of intimate ecological interactions with vertebrates. Synergies between these areas and an appreciation of the overarching ecological unity of parasites despite their taxonomic diversity is currently hamstrung by a lack of comprehensive specimen level data. Unlocking specimen level knowledge of parasites will facilitate integrated understanding of their ecology and evolution across multiple vital research areas including public and veterinary health, wildlife management, and basic and applied academic research into arthropods and vertebrates. A focus on building links between museums, state and federal institutions, 'hidden' (non-museum) parasite collections, on-going public health monitoring efforts and vertebrate-database records will allow the first ever comprehensive foundation for detecting and predicting responses to accelerated environmental change. Specimens and biological information made available via globally accessible databases provide a pathway to assess the history, relationships and distribution of global diversity. The project will enrich existing vertebrate diversity databases and research by adding ecological interactions layers (parasitic interactions) and promote future interdisciplinary research.

Broader Impacts. Broader impacts emerge from highly cross-disciplinary nature of the TPT network that will foster integrated studies of ecosystem function, zoonotic diseases, and other fields such as conservation genetics and isotope ecology. Digitizing this vast resource significantly strengthens scientific research and instructional infrastructure in the US. Biodiversity information generated by TPT network will offer insights into past events of climate change, contributing to future management and policy decisions with direct impacts on human and livestock health, and conservation biology. Funding will enhance training and experience for US students participating in highly productive collaborations. The TPT Museum collections represent some of the finest resources available to scientists and educators as they tackle emerging challenges for science and society.

I.Urgency and Importance of Digitizing Parasite Collections

Parasitic arthropods inflict an enormous burden on the health of their hosts either directly, or through virulent pathogens that they vector. Indeed, some of these parasites have played key roles in the largest human health crises (World Health Organization website accessed September 2018). For example, fleas, mosquitoes, and ticks vector harmful pathogens responsible for plague, malaria and Zika, and Lyme disease, respectively (**Figure 1**). Similarly, these parasites significantly impact livestock, which threatens agriculture and food security globally (e.g., Pérez de León et al. 2012; Giles et al. 2014). Interactions between humans, animals, and their parasites are constantly changing (Benning et al. 2002; Gratz 2004; Kernif et al. 2016; Eisen et al. 2017) and vectored pathogens transmitted to humans and animals in the U.S. are increasing at an alarming rate. The Center for Disease Control (aka CDC) estimates "*illnesses from mosquito, tick, and flea bites* ***have tripled*** *in the U.S., with more than 640,000 cases reported during the 13 years from 2004 through 2016.*" Human movement, land use, and rapidly changing environments have contributed to both range expansion or distribution changes in many arthropod vector species and the recent surges in the diseases they transmit (e.g., Gratz 2004; Campbell et al. 2015; Kernif et al. 2016; Eisen et al. 2017; Kauffman & Kramer 2017; Sonenshine 2018). Our ability to understand and model the potential risk of parasites is hampered by a lack of baseline information.

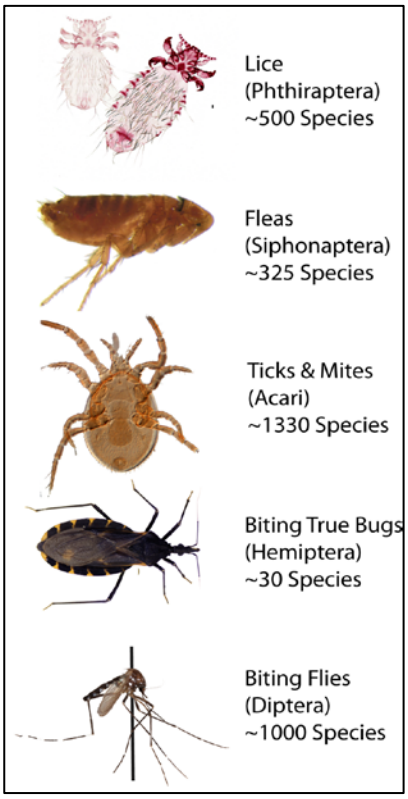

Figure 1: Example of terrestrial arthropod parasites that will be digitized as part of the Terrestrial Arthropod Tracker Network. Images not to scale. Photo credits: S. Hamer, J. Light, H. Lutz, and B. OConnor.

Biodiversity loss is heavily impacted by parasites. Introduced parasites and vectored pathogens devastate populations of many rare and endemic species (Bond 1994; Traveset & Richardson 2006). Island systems are particularly at risk, being susceptible to repeated and ongoing introductions of disease transmitting vectors (Hales et al. 1999; Atkinson & LaPointe 2009; Bataille et al. 2009; Parker et al. 2011; Bonizzoni et al. 2013; Roth et al. 2014). Climate change and globalization continue to provide opportunities for parasites to establish in new regions and on new hosts. Despite these negative interactions, native parasite species may also play a critical role in maintaining ecosystem diversity by thwarting the establishment of invasive parasites. A more thorough understanding of endemic parasite fauna is critical for wildlife management and long-term conservation of biodiversity (Hudson et al. 2006; Hatcher et al. 2012; Cizauskas et al. 2017).

Natural history collections provide documentation of life on Earth over long time periods, often representing the full breadth of organismal diversity across broad geographic regions. Because collections are the primary and permanent repositories for past and present parasite specimens, they have enormous potential to address some of the most significant societal challenges of human and animal health and safety, environmental sustainability, and food security (Dunnum et al. 2017; Schindel & Cook 2018). Collectively, these datasets yield information that can be used to discover host associations, model ecological processes and changes in species distributions, and may even allow researchers to predict the future spread of human and animal disease (Tewksbury et al. 2014; Schindel & Cook 2018).

Arthropod parasite data are underrepresented among digitized specimen data due to a prior focus on fauna and flora exclusive of parasite groups (e.g., plants, herbivore clades, aquatic invertebrates, fossil marine invertebrates) and on parasite hosts (i.e., vertebrates). Thus, even though parasites represent a substantial proportion of organismal diversity (Dobson et al. 2008), their data are not readily accessible (**Figure 2**). This is, in part, because the collections themselves can be difficult to find (Hoberg et al. 2009; Bell et al. 2018a) and are sometimes stored separately from invertebrate collections, in closer proximity to their hosts. These hidden, or 'biotic association' collections tend to be data-rich collections obtained and

curated by vertebrate researchers. Notably, these collections contain specimens and other data that often are completely unknown to the broader community, yet represent irreplaceable knowledge about past organismal habitats, distributions, and parasite-host associations. Because they are often highly represented in these biotic association collections, and given their importance to human and animal health and safety, arthropod parasites are a high priority for integration through digitization. A parasite collections digitization effort would allow us to exponentially increase the number of known biotic associations, transforming multidisciplinary research, which is aligned with NSF's 10 Big Ideas *Growing Convergence Research* initiative.

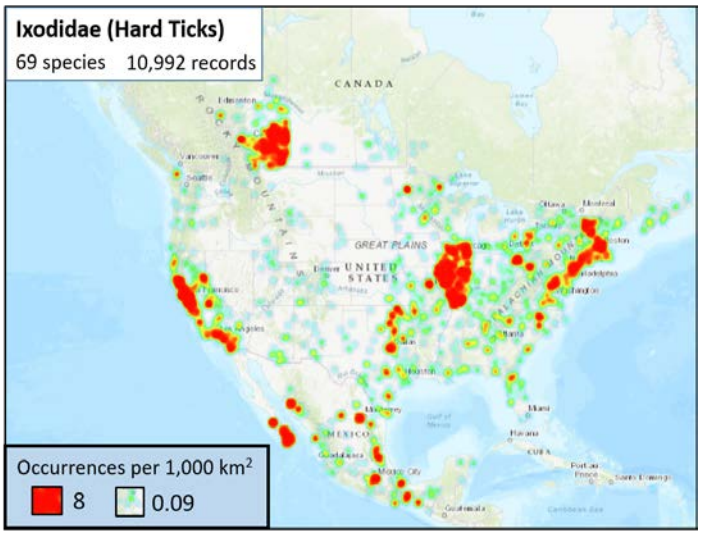

Figure 2. Occurrence records for Ixodidae available on *Symbiota Collections of Arthropods Network* (SCAN). Numbers are representative of all arthropod parasite groups in North America, with low numbers, large gaps and few collections that have contributed data to date. Heat maps are depicting areas with a maximum number of occurrences of 8 (red) and a minimum of 0.09 per 1,000 km^2.

Arthropod parasites represent a massive gap in the context of current digitization efforts in both invertebrate and vertebrate research communities. Given their importance to human and animal health and biological conservation, these collections are in urgent need of digitization. Although institutions with these parasite collections are fully committed to the long-term support of these holdings, an integrated network is needed to lead the digitization and integration of these data.

II. Terrestrial Parasite Tracker (TPT) Network Overview

The Terrestrial Parasite Tracker (TPT) Network will aggregate arthropod parasite collections to build an easily accessible, comprehensive database of parasite-host associations and vector distributions. Our network will provide needed baseline information for research and management of the ecological interactions among parasites, pathogens, and their hosts in North America (including the U.S. & territories). We will work together to digitally provide information on parasite collections by providing research-ready data and images from **1.2+** million parasite specimens, which will be accessible to scientists, educators, wildlife managers, and policy makers worldwide. TPT will significantly increase the visibility, accessibility, and research relevance of these collections using two primary approaches:

TPT Portal and Leveraging Other TCN Developments The TPT data portal will unify data about both parasites and their hosts. The portal will facilitate accessions of rich data produced by our project, and it will act as endpoint from which to serve data. This collective interface will inherently improve the validation, semantics, and utility of digital parasite products (**Figure 3**). Arthropod data will be housed in the Symbiota Collections of Arthropod Network (SCAN) and Darwin core (DwC) archives, which serves as both an online

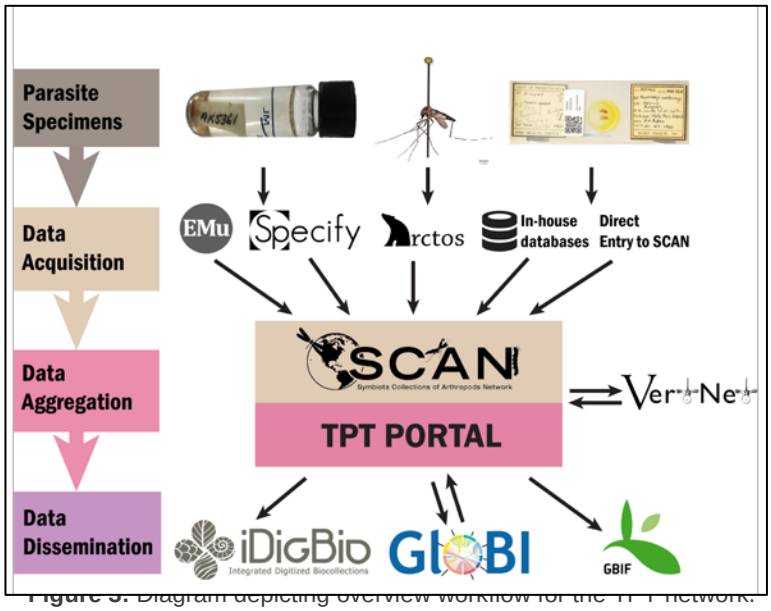

Figure 3. Diagram depicting overview workflow for the TPT network. All resulting data and images will be exported to SCAN and

database and a primary aggregator for North American arthropod data. We will create a dedicated TPT Portal that will selectively filter arthropod parasite data from SCAN, to create a multi-phyla portal that includes occurrence data and images for both arthropod parasites and their vertebrate hosts. With TPT, we will apply existing approaches and infrastructure to digitize, aggregate, and integrate terrestrial arthropod parasite collections and their data with national (i.e., iDigBio) and global (i.e., GBIF) data dissemination initiatives through the establishment of strategic partnerships, resource sharing, and workflow exchanges.

This project will leverage digitization workflows established by other invertebrate-focused Thematic Collections Networks (TCNs) such as NSF-InvertNet (slide and vial imaging workflows), NSF-SCAN (Symbiota Portal), and NSF-LepNet (georeferencing protocols and research advisory board implementation), along with other NSF-funded efforts to engage citizen scientists in transcription label data through the national crowd-sourcing of museum label data project 'Notes from Nature'. The TPT Network also offers a unique opportunity to build on the work of the Tri-Trophic TCN to understand the importance of biotic associations as a key process in driving biodiversity. There is a substantial and growing database of vertebrates, but still inadequate data on their arthropod vector parasites (SCAN currently holds 18 million records for North American arthropods but only contains 214,000 records for all parasitic arthropod taxa; retrieved from SCAN portal September 2018). For comparison there are over 11 million vertebrate specimen records and 330 million observation records for North American vertebrates (GBIF, 2018) Precise occurrence data, high-resolution images, and a newly developed associations linkage for arthropod parasites are expected to activate biological research in ways that have not been possible before, cutting across multiple disciplines.

Goals of Terrestrial Parasite Tracker

Specimen Digitization:

1. Transcribe and georeference label data from **1.2+** million arthropod parasite specimens from 22 collections across North America (U.S. and territories) including ~55,000 specimens from biotic-association collections.

2. Produce **600,000+** High Throughput (HTP) scans of fluid-stored slide-mounted parasite specimens.

3. Produce **220,000+** research-grade images for exemplar parasite species, including over **500 type specimens**.

Data Mobilization:

4. Produce a dedicated Terrestrial Parasite Tracker (TPT) Symbiota portal to serve as the primary aggregator for both live and snapshot collections within the TPT Network.

5. Digitize 500,000 biotic association records, a substantial increase over the 7,528 current records in SCAN . These will be integrated with a growing literature database maintained by Global Biotic Interactions (GLoBI).

6. Extend previously existing automated identification tools (i.e., Fieldguide) to develop rapid-identification applications for parasite images generated by TPT. Images curated through the Fieldguide library will also be available for trait analyses.

Broader Impacts:

7. Build interdisciplinary bridges between convergent research communities through facilitation of training workshops, sharing of workflows and data between academic and non-academic institutions, and during scientific meetings (e.g., Entomological Society of America, Ecological Society of America, American Society of Parasitologists, etc.).

8. Engage the public in the importance of natural history collections and educate broad audiences of their use in addressing societal issues.

9. With a focus on terrestrial arthropod parasites, enhance curriculum for educators at multiple levels and partner with citizen science projects and other initiatives.

III. Intellectual Merits

Specimen digitization often leads to novel research (e.g., Foley et al. 2008; Heidorn 2008; Soltis 2017; Stephens et al. 2009). TPT will provide research-ready baseline data that will catalyze new research and education initiatives, thus having immediate and long-lasting benefits for our understanding of organismal associations, biodiversity, and beyond. We have identified four core areas of research that can directly utilize TPT data to answer a variety of questions centered on North America and island systems (Pacific and Caribbean).

***Biological Associations*.** Of the ~214,000 vertebrate parasite records that have biotic associations on SCAN, only 7,528 have host associations. TPT will harness critical recent historical data to significantly increase the availability of biotic-association data from parasites for downstream research. We estimate TPT will produce, 500,000 biotic-association records, increasing the number of available records by more than 66-fold. In linking parasite records to vertebrate databases, TPT will facilitate the creation of lists of parasite-host biotic associations (including documentation of novel associations), resulting in a better understanding of the role of parasites in ecosystem functions. In this way, TPT will advance research on parasite-host coevolution and the evolution of specialist and generalist parasites (e.g., Tripet & Richner 1997; Vázquez et al. 2005; Appelgren et al. 2018; Matthee et al. 2018; Stokke et al. 2018), which will drastically influence our understanding of ecology, evolution, and community interactions (Clayton et al. 2015).

Contributions to Disease Ecology. A major aim of disease ecology is to understand the effects of ecological interactions on the spread of pathogens and the intensity of the diseases they cause (Collinge & Ray 2006). Vector-borne diseases emerging in human and animal populations worldwide have a high impact on global economy, animal production, and public health (e.g., Gubler 1998; Jongejan & Uilenberg 2004; Tatem et al. 2006; Weissenböck et al. 2010; Dantas-Torres et al. 2012). TPT will facilitate research in disease ecology with the digitization of terrestrial arthropods such as mosquitoes, ticks, kissing bugs, fleas, and lice (**Figure 1**) that vector parasites and pathogens that cause disease in humans and animals, including viruses (e.g., Zika, West Nile, chikungunya, etc.), bacteria (e.g., Lyme disease, typhus, tularemia, plague), protozoans (e.g., malaria, babesiosis) and roundworms (e.g., filariasis). TPT will also address other needs in organismal health by identifying specimens significant for tracking diseases over time (e.g., George 1987; Ávila-Arcos et al. 2013; Brooks et al. 2014; DiEuliis et al. 2016; Fournier et al. 2016; Dunnum et al. 2017).

***Changing Species Distributions*.** TPT will provide precisely georeferenced localities of parasite specimens, which will allow human health and safety officials to produce accurate distribution maps for terrestrial arthropod parasites species, most notably those that vector disease-causing pathogens. In addition to predictable impacts on human health, invasive arthropod vectors will likely have damaging effects as they migrate to new regions (e.g., Benning et al. 2002; Gratz 2004; Traveset & Richardson 2006; Kernif et al. 2016; Eisen et al. 2017; Chalkowski et al. 2018; Sonenshine 2018). From a conservation perspective, these parasites and vectored diseases can threaten native ecosystems (Bond 1994; Holmes 1996). With precise georeferenced specimen data, TPT will track arthropod vectors as their distributions shift, alerting public health officials and conservation managers to the potential of emerging infectious diseases. Finally, coextinctions of parasites with their hosts are a major loss of biodiversity (Carlson et al. 2017; Cizauskas et al. 2017) and much of this diversity will be lost before it is documented. This loss could be devastating, especially considering the critical role parasites play in maintaining ecosystems (Poulin 1999; Hudson et al. 2006; Hatcher et al. 2012; Bell et al. 2018b; Frainer et al. 2018). TPT will help identify under-represented parasite groups in urgent need of sampling and threatened parasite diversity in need of conservation.

Parasite Systematics, Taxonomy, and Species Trait Analyses. Parasites are estimated to represent 40% of all animal species (Dobson et al. 2008; Weinstein & Kuris 2016), yet much of their true diversity is unknown. This is, in part, because major parasite clades are outside the most heavily studied arthropod groups (i.e., the five megadiverse insect orders, macroscopic arachnids, etc.). Compilation of specimen-level data, especially from biotic association and other non-museum collections, will facilitate comprehensive systematic approaches and alpha-taxonomic studies of these parasite groups. TPT will enable next-generation molecular studies through discovery of ethanol-stored museum specimens allowing researchers to test phylogeographic, phylogenetic, and biogeographic hypotheses relating to parasite biodiversity (e.g., Short et al. 2018; Wood et al. 2018). Moreover, linking of large numbers of parasite-host associations will facilitate the study of ecological patterns repeated across independent taxonomic groups (Clayton et al. 2015). TPT will make available online research-grade images of thousands of specimens which will enable non-experts to identify economically and medically important

species, including invasive species. TPT images will facilitate morphological revisions and the development of diagnostic tools (e.g., Tsangaras & Greenwood 2012; Kocher et al. 2017; Lopes et al. 2017) and production of taxonomic keys. Resulting TPT images will also prove to be invaluable for studies examining how morphological and phenological traits respond to global change over time (Kissling et al. 2018).

IV. TPT Network Development - parasite-tracker.org and Project Management

Participating Institutions. The TPT network is comprised of participants from 27 institutions (including 22 research collections, **Table 1**) across the U.S. (****Institutions with PIs new to the ADBC program*****)**: Purdue University and Milwaukee Public Museum (PU, MPM - lead Institutions), *Academy of Natural Sciences of Drexel University (ANS), *Bernice Pauahi Bishop Museum (BPBM), *Brigham Young University (BYU), *California Academy of Sciences (CAS), Clemson University (CU), Field Museum of Natural History (FMNH), Illinois Natural History Survey (INHS), Michigan State University (MSU), Ohio State University (OSU), *Pennsylvania State University (PSU), *Texas A&M University (TAMU), University of Hawaii (UH), University of Michigan (UM), University of Minnesota (UMSP), University of Nebraska (UNL), *University of New Hampshire (UNH), *University of New Mexico (UMN), *University of Utah (UU), University of Wisconsin Madison (UWM), and *University of Wisconsin Stevens Point (UWSP).

INSTITUTION	Imaged Pinned Specimens	Scanned Vials	Scanned Slides	Total Transcribed	Legacy Data Import	Data Crowdsourced
ANS	30	21,200	7,289	**28,519**	20,000	
BPBM	4,500	35,000	10,000	**70,000**		55,000
BYU			79,200	**79,200**		
CAS	5,241	6,289		**99,840**		6,289
CU	200	9,370		**14,640**		
FMNH	809	6,000	189,000	**198,000**		
INHS	400	9,906	13,111	**73,258**	8,214	
MSU	1000	530	2100	**36,000**		
MPM	200	1500	1000	**2,700**		
OSU			6,000	**6,000**		
PSU	200	8,301	5,000	**38,301**	images	
PU	300	300		**27,600**		
TAMU	500	13,595	23,452	**165,000**		13,595
UH	4,600		6,400	**11,316**		
UM			180,000	**180,000**		30,000
UMSP	48		56,204	**56,252**		
UNL			10,000	**10,000**	20,000	
UNH	4,500	1000	1000	**30,000**		
UNM		18,500	2,000	**43,000**	2,500	
UU		8,300	39,700	**48,000**		
UWM		6,426	13,053	**19,985**		
UWSP			9,280	**9,280**		
TOTAL	**22,528**	**146,217**	**653,789**	**1,246,891**	**50,714**	**104,884**

Table 1. Digitization plan for participating TPT collections. Institutions with significant biotic-association holdings are: ANS, FMNH, TAMU, and UNM (~55,000 specimens).

Due to the intrinsic complexity of parasite collections, including biotic-association collections, we have enlisted key individuals at five institutions to assist in the development of informatics tools that will allow us to more effectively aggregate, integrate, and publish collections data online for downstream research: University of California Santa Barbara (UCSB), University of California Berkeley (UCB), Northern Arizona University (NAU), University of Florida (UFL), and University of Nevada, Reno (UNR). We will also collaborate with other collections not supported by this TCN and partner with parasite-related initiatives in the U.S. (i.e., UW Madison's Midwest Center of Excellence for Vector-borne Disease, University of Notre Dame's VectorBase, Walter Reed Biosystematics Unit, and NOLA Mosquito and Rodent Pest Control Board). Each of these institutions has expressed a need for these parasite data to support their institutional and project goals.

Taxonomic and Geographic Focus. Of the ~12 arthropod orders of medical-veterinary interest (Mullen & Durden 2018), we will focus on 'ectoparasite' and/or 'vector' groups, defined by either by a) feeding on the host, or b) living their entire life on the host. With this approach, we anticipate comprehensive coverage of species across all major parasite groups for North American (including U.S. & territories) within the following orders (**Figure 1**): Ixodida (~100 species), Mesostigmata (~230), Trombidiformes (~500), Sarcoptiformes (~500), Diptera (including mosquitoes, sand flies, biting midges, black flies, horse and deer flies, bot flies, louse flies; ~1000), Hemiptera (~30), Phthiraptera (~500), and Siphonaptera (~325). It will not be possible for TPT institutions to digitize all holdings of all parasite groups within the three-year project period; however, we aim to digitally capture the vast majority (80%) of North American parasite species diversity (~3,100 species) in these museums. Further, we aim to capture specimens that are historically important and critical for research goals. The TPT Network has developed a 'subsampling strategy' to meet these goals based on the following criteria: **1)** Specimens digitized represent a major strength of each participating institution's holdings (e.g., historically significant parasite collections, like the pre-1950's Atyeo collection of bird parasites at UM), **2)** Specimens digitized represent a comprehensive sampling (geographically and/or species diversity) of an arthropod parasite lineage, **3)** Specimens digitized represent a hidden and/or biotic collection not previously associated with other arthropod holdings or collections, or **4)** Specimens which are a target for a particular research project(s), as directed by the Research Advisory Board (see below). In addition to arthropod data we will also ingest snapshots of all vertebrate specimen data for our study area (>11 million records) and select observational records from the 330 million GBIF records (see Biotic Associations - GloBI below).

Taxonomy and Terms. In addition to the systematic expertise available within our team of experts, TPT will interact with other federally-supported parasite data aggregators such as VectorMap, VectorBase, and the Upper Midwest Center of Excellence for Vector-Borne Disease (UW-Madison) to ensure effective use of shared vocabularies, taxonomies, nomenclature, terms related to collection records, host association data (if present), and other ecological information.

Research Advisory Board (RAB). We will assemble a research advisory board following the model developed by the NSF-funded LepNet TCN. The TPT-RAB will establish a similar process that maximizes efficiency of digitization for the TCN and opportunities for research collaboration. The TPT-RAB will be comprised of researchers from both academic and non-academic institutions. They will help define project goals and outcomes, develop conditions of the collaboration (e.g., how to handle data for student projects, sensitive data, meeting deadlines for grant proposals requesting data, publications, etc.). The TPT-RAB also will determine digitization priorities and moderate authorship discussions.

Macropod Imaging Coordinators. Eight TPT institutions with significant parasite collections that are near numerous smaller parasite collections will serve as coordinators for imaging specimens using Macropod systems (ANS, BMCAS, UNH, MPM, PSU, TAMU, UM, and UNM). The Macropod is a photography system that was developed specifically for rapid processing of arthropod vector images and is currently utilized by the Army Public Health Center (APHC: Website accessed 28 September 2018); equipment specification details and workflow are provided in the "research-grade imaging" section (IV Digitization and Image Processing). Each center will also receive imaging hardware systems (e.g., iMac Pro: see Data Management Plan). Combined, these institutions will coordinate imaging of ~20,000 pinned specimens and ~200,000 slide-mounted mites and other non-mite parasite specimens less than 5 mm over the course of three years (including 500+ types). Specimens stored in vials and larger slide-mounted material will be scanned using standard work stations consisting of custom-built trays and flatbed scanners following protocols established by other TCNs (e.g., InvertNet, LepNet); all participants will install basic scanning work stations in their primary workspace.

Recruitment. TPT will strive to build diverse and inclusive applicant pools from undergraduates and recent graduates in the biological sciences for our technician positions. We will advertise positions broadly across participating campuses and affiliated universities. The positions will be targeted to those who have an interest in pursuing a career in biodiversity informatics, collection management, or entomology and parasitology research with the intention to provide the experience of participating in a national-scale biodiversity project.

Supervision and Training. At each institution, a point person will be assigned to supervision and training of technicians and students, as they will work in the same space and serve as the point person in regard to transcription of label data, standards, and georeferencing. Over the course of the project, the technicians will receive training beyond data entry, and will be mentored by PIs. As part of training, technicians will be encouraged to attend iDigBio's webinars to keep abreast of current developments in digitization and collection management.

Milestones and Project Tracking. Digitization progress will be tracked by an automated service built into Symbiota that reports the number of images and records for each institution on a monthly basis, including status of crowd-source processing of images. Reporting can also be provided for specific taxa for each collection, monthly, through regular reporting by PIs both in individual institutions and across the TPT network. It will be particularly important to monitor progress in the first months of the project, because workflows for data entry are already in place in some institutions and data entry should flow smoothly at those institutions. Other institutions will require assistance by established digitizers, and a "mentor/mentee" partnerships will be established between institutions with digitization experience and those who are new to digitization. Milestones will be established and milestone evaluation will be conducted quarterly, in accordance with research goals for the network. Both MPM and FMNH have developed detailed project management workflows and other digitization resources which will be made readily available to the TPT network.

V. Digitization and Image Processing

Cost-Benefit Analysis. The composition of parasite specimen holdings and progress in digitization at each of the 22 TPT institutions varies significantly. Therefore, we have implemented cost-analysis standards to be used across all participating collections. We have set a network-wide rate of $1.25 per specimen (direct cost) to capture specimen label data and georeference records with select, research-grade images for high research-value material. Because of the diversity of storage methods for parasite specimens, this TCN incorporates a wide variety of data acquisition techniques and products. **1)** Label data will be transcribed for 1.2+ million specimens. We estimate transcription to take an average of 30 specimens per hour (based on MPM rate). **2)** Slide-mounted material will be batch scanned for label data and low-resolution specimen images using flatbed scanners and trays at the rate of 240 slides per hour, if pre-cleaning of slides is not needed. **3)** Vials will be scanned the same workflow as slides and the rate is estimated at 100 samples per hour. **4)** High-resolution images of high research-value pinned and slide material will be produced at the rate of 20 samples per hour. Our cost rate also includes all project-specific basic curation and taxonomy updates. Our estimates are based on data from three existing TCNs (SCAN, LepNet, InvertEbase) with over three years of experience digitizing all arthropod taxa, and from assessment data collected at PU, MPM, and FMNH for hundreds of thousands of specimens digitized. This estimate includes extra time required for transcription of host data or 'cloning' of host records into parasite databases (i.e., biotic-association collections). A detailed breakdown of the digitization goals for each institution is provided in Table 1.

Label Imaging and Data Transcription. Collectively, we will capture label data from 1.2+ million specimens across 22 research collections, using trained students and technicians, collection staff, and supervised volunteers. Eight participating collections will oversee research-grade imaging of specimens and label data for over 20,000 pinned specimens and 200,000+ mites, fleas and lice. Through development of the TPT portal, we will integrate images and transcribed data from previously digitized specimens from the following institutions: ANS (~20,000 specimens including host data for 95% of records), UNL (~20,000 specimens including host data for 95% of records), INHS (~8,000 specimens), PSU (~15,000 images), and UNM (~2,500 specimens including host data for 95% of records); SI-National Museum of Natural History will contribute available U.S. data during the project period (~10,000 records).

***Fluid-stored specimens*.** We will capture label data and images from 150,000 ethanol-preserved specimens. Ethanol-stored material will be scanned using trays and standard flatbed scanners following previously developed workflows (i.e., for InvertNet and LepNet TCNs). Vials stored in bale-top jars can be

scanned in batches, using a flatbed scanner with a flat jig for placement; however, labels will be removed from vials and placed on trays next to the specimens (fixed with a clean microscope slide).

***Slides*.** We will capture label data and images from 650,000 slide-preserved specimens. We will scan slides with larger specimens using the same batch scanning protocol as fluid-stored specimens (see **Figure 4** for sample workflow). Select high-research value specimens will be imaged using a Macropod workflow developed for HTP imaging of slide-mounted specimens.

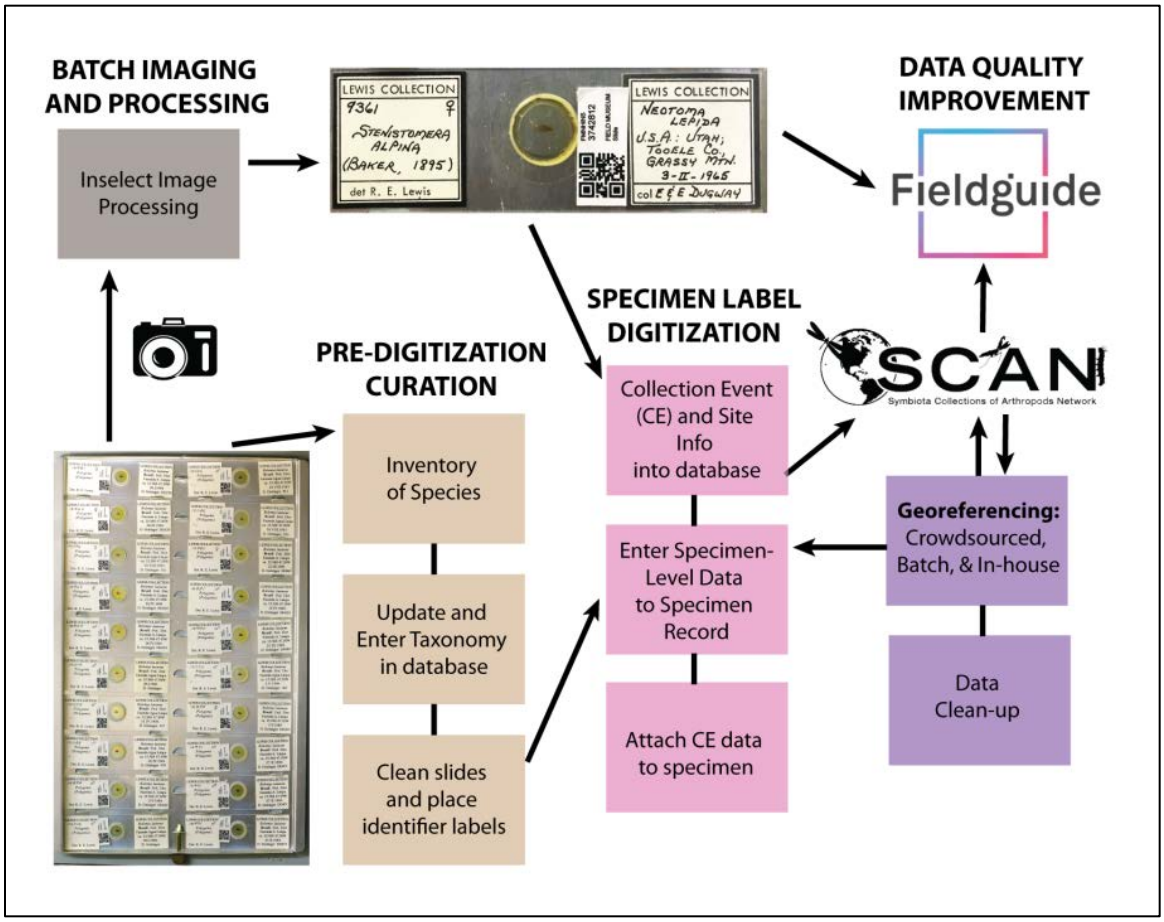

Figure 4. Low-resolution slide-scanning workflow showing process from specimen positioning to image processing.

Image Processing. Resulting batch scans of vials and slides will be segmented into individual images of single specimens using the software package Inselect (**Figure 4;** Hudson et al. 2015). Inselect assigns bounding boxes around objects of interest in the scan (using the cookie cutter function) and then parses that segment of the parent image or scan. IrfanView (freeware software offered by Irfan Skiljan) will be used to batch rename image files based on scanned barcode and species name. We have developed detailed electronic user guides for Inselect and will provide all TPT institutions with these resources. Image files will be linked with specimen data and unique identification numbers and loaded into the TPT portal via upload features of Symbiota. Resulting slide images will be used as training images for our rapid ID tool (see Broader Impacts).

Crowdsourcing. Label data from estimated 100,000 records will be transcribed through crowdsourcing using 'Notes from Nature', an NSF funded project run through Zooniverse, where museums upload batches of images into 'Expeditions'. Label data from those images is transcribed by three independent volunteers, and the data from those three transcriptions is then 'reconciled' and cleaned data is returned to providers. Notes from Nature has successfully transcribed over 2 million

specimens from 20 institutions by more than 10,000 volunteers. We have offset transcription costs for some of the larger TPT collections through pre-planned CrowdSourcing estimates.

Georeferencing. Data harvested from TPT collaborators will be georeferenced **1)** through Geolocate (either the portal or as part of a batch backend submission to Geolocate directly), or **2)** 'in-house' prior to IPT harvest cycles. For institutions that choose to complete georeferencing 'in-house', we will provide the necessary resources such as will the *Guide to Best Practices for Georeferencing* (Chapman & Wieczorek 2006); specimens with less descriptive locality data will be georeferenced using the point-radius method for calculating uncertainty of coordinates (Wieczorek et al. 2004).

Research-grade Imaging. Eight participating collections will oversee research-grade imaging of specimens for over 20,000 pinned specimens representing 1,200+ species and 200,000 slide-mounted ticks and mites representing 1,000+ species. Standards for images will be modified based on criteria developed as part of the LepNet TCN. Our imaging workflows for pinned specimens will use the Macroscopic Solutions, LLC Macropod imaging system. This technology performs a variety of functions including high-resolution macro/micro imaging, 3D scanning/modeling, and high speed video (**Figure 5)**.

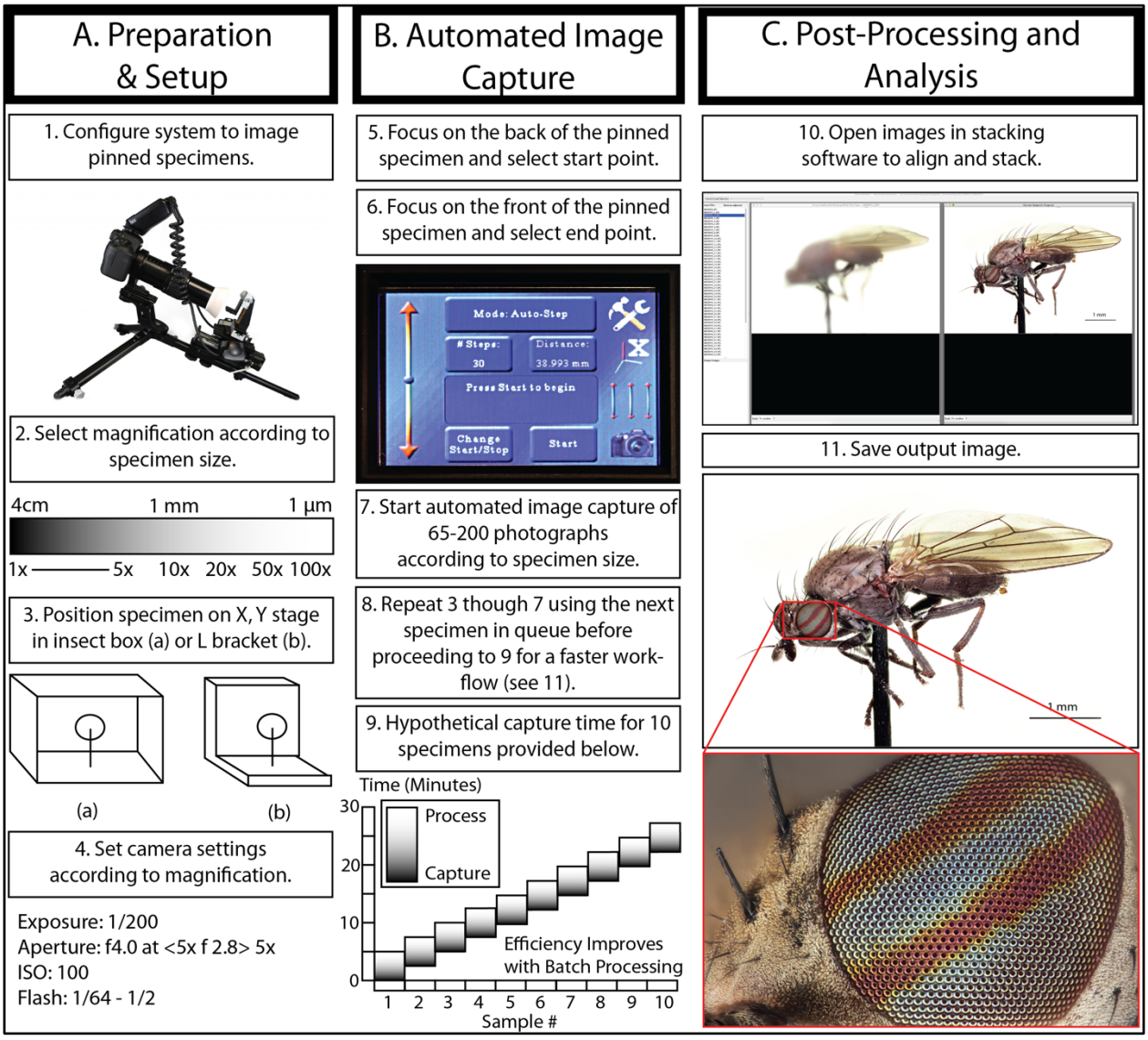

Figure 5. Macropod imaging workflow from specimen positioning to image processing.

Macropod systems are fully portable and automated devices capable of capturing detailed, high resolution images for samples ranging from 1 micron to infinity (¥). The Macropod rotates a specimen

360° around an axis to produce 3D movies and point clouds that can be used for 3D printing, volumetric analyses, and as visual aids for research and educational purposes. For pinned insects and slide-mounted microparasites, we will use the 1-5x stereoscopic lens. No special infrastructure is needed for a Macropod setup other than any solid surface (i.e., table). Minimal sample preparation is required and imaging methods are non-destructive (see **Figure 5** for workflow).

Rapid, high-throughput species-level ID through Fieldguide. Images generated by all TPT institutions will be used for the development of a rapid parasite ID tool. Fieldguide (Fg) is an API, website, and suite of smartphone applications that provide image recognition services to the research community as well as the general public (**Figure 6**). Fieldguide convolutional neural net processes provide automated species identification via its apps and API-as-a-service to third-parties including a deep integration with SCAN, which was developed through funding to LepNet. Most automated identification services (e.g., iNaturalist) focus on field images, whereas Fg has concentrated on creating specimen image libraries from collections. Fieldguide allows data providers to process their images for three primary purposes; **1)** provide species-level suggestions from images that are not identified to species, **2)** screen existing human identifications, and **3)** allow rough sorting of specimens in the initial curation phase into taxonomic groupings above the species level. Fieldguide will harvest TPT parasite slide images and associated data (locality, host, date) from SCAN to train a convolutional neural network (Fig. 6) . Fieldguide will provide API endpoints available to TPT and launch an Android and iPhone app for parasite identification by modifying the existing apps built for the LepNet TCN. Due to phenomenal advances in smartphone lens technology, we foresee these Apps utilized by vector-monitoring labs, veterinarians as well as museum personnel.

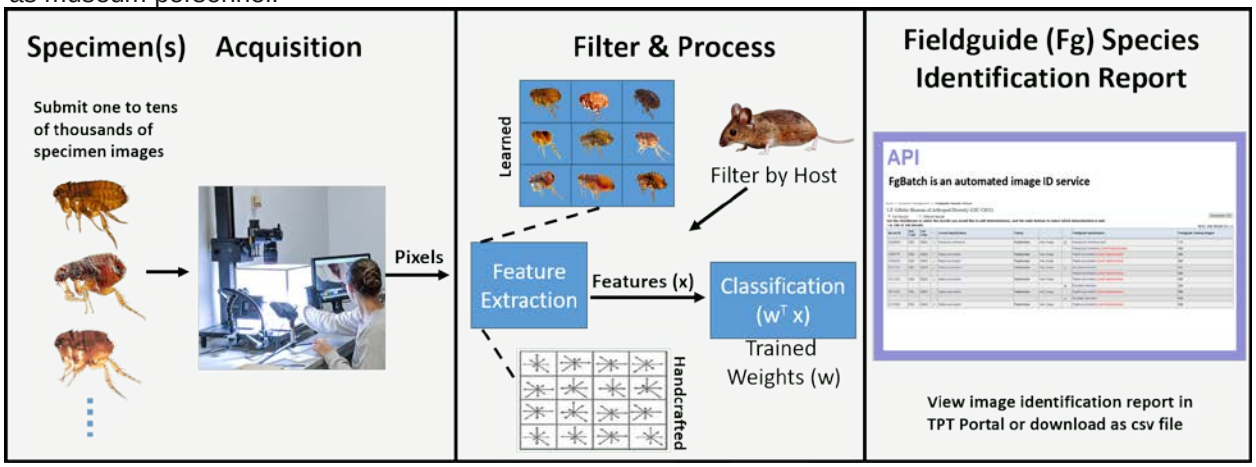

Figure 6. Fg workflow from specimen(s) through the identification report that allows image providers to select identification from list and apply; a determination history is maintained to document changes to identification specimens (e.g., provenance).

VI. Data Portal

The TPT Network offers a unique opportunity to build on the ongoing work of the SCAN and Symbiota, which provides the underlying data integration platform for SCAN. Symbiota is presently the most widely used software to organize and annotate data for the ADBC program, and provides automated data feeds to iDigBio and GBIF. The 30+ Symbiota data portals cover the full range of metazoan diversity, so we are well positioned to share synthetic data on arthropod parasites and their vertebrate hosts.

Symbiota functions as both an online 'database' that allows users to perform easy, web-based data entry (i.e., live collections) and a biodiversity content management system that allows for seamless ingestion of data from providers that use a non-Symbiota database (i.e., snapshot collections). The most common databases used by data providers that contribute to snapshot collections are Arctos, EMu, Specify, and institutional databases. Each collection retains autonomy of their data while contributing to a community database and tools to provide comprehensive data to researchers and educators. Symbiota offers a broad spectrum of downstream research applications that are continually being developed and available to end users including, but not limited to, the generation of multi-species distribution maps,

batch georeferencing, computer-aided identification, and datasets for phylogenetic and ecological analyses (**Figure 7**).

SCAN is the largest Symbiota portal, aggregating 18 million specimen records and 2 million images from over 120 North American arthropod collections for **all** arthropods. SCAN is also the primary repository for occurrence data produced by two other continuing TCNs (LepNet, InvertEbase). We will create a dedicated TPT Symbiota multi-phyla portal "Terrestrial Parasite Tracker Network Portal" (TPT Portal) that will integrate with SCAN. Arthropod collections will either be 'live' in TPT Portal or snapshots from SCAN or another data provider source (via Integrated Publishing Toolkit [IPT] or Darwin Core [DwC] archive). The system will provide complete flexibility for data providers as to where and how their data is stored while ensuring that there are clear pathways for providing data to aggregators (e.g., iDigBio and GBIF) and preventing duplicate records.

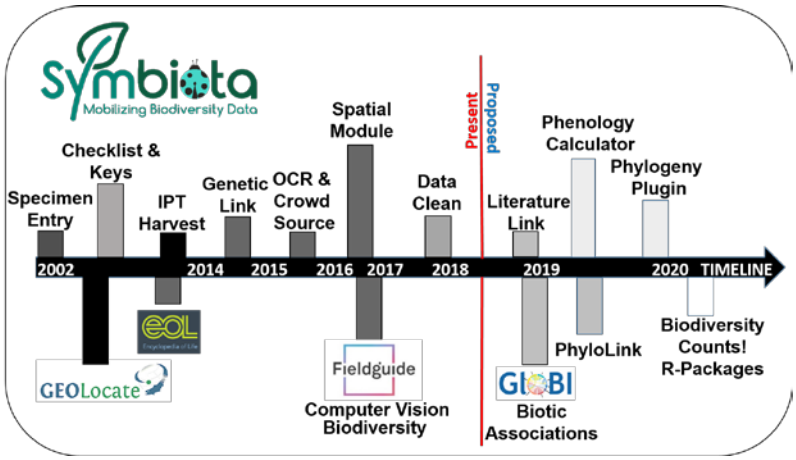

Figure 7. Timeline showing initiation specific functions built into Symbiota (above Timeline bar) or third-party modules that link to Symbiota data (below Timeline bar). Size of vertical bars indicates complexity in function; shading indicates degree of completion.

TPT will integrate additional parasite specimen data from vetted monitoring programs and federal collections (e.g., VectorBase, Walter Reed). Likewise, vertebrate data will be ingested from existing publicly available datasets via IPTs, such as those hosted by VertNet, which aggregates most of the vertebrate occurrence data in North America. Currently, there 11 million vertebrate specimen records and 330 million observations for North America. We will not generate new vertebrate data as part of the TPT project. The same will be true of disease data, which will be harvested from disease aggregators (e.g., VectorMap, ArboNet). Available disease data will likely be more coarse, but there is enough data to establish a baseline for some vector-borne diseases.. Hosting vertebrate and disease data on the TPT Portal will allow end users to map host, parasite and disease data simultaneously and download integrated parasite-host .csv files for analysis outside the portal (e.g., Map of Life).

GLoBI: Harnessing the Data Revolution Through Biotic Interaction Data Sharing. Despite the importance of species interaction data to biological research, there are not enough records and few standardized methods exist to systematically collect, record, share, and integrate biotic interaction or biotic-association datasets. This is true for digitized natural history records, there are no established standards for encoding and sharing organismal interaction information. This leads to the loss of valuable species interaction information, which is difficult to find in the literature and challenging to integrate into datasets for analysis due to differences in terminology and standards. Overall, this lack of standardized methods for working with biotic-associated data hinders the ability to share or evaluate these data for future research. The collaboration between TPT and GloBI will make significant progress in producing a large data set that is structured and will inform other projects and provide a model for future collection practices (cite the ? and Cook Nextgen 2018 paper)…….

It is difficult to estimate the magnitude of this information loss, whether in terms of lost species interaction data, or research costs. For natural history specimens, such losses may be estimated by examining the database exports of the major data aggregators of specimen data from U.S. institutions, the GBIF, and iDigBio. In these exports, species interaction data are commonly stored as free text in *associatedOccurrences* and *associatedTaxa* fields as defined by the Darwin Core Standard (Wieczorek et al. 2012). Using those fields as a proxy for how much species interaction data might be discoverable, GBIF has 553,000 records and iDigBio 1,006,685 records with those fields populated with possible

species interactions (GBIF 2017; Page et al. 2015). Many other interactions are expected to be found in specimen records as notes (e.g., Darwin Core: *occurrenceRemarks*).

TPT will contribute 500,000 biotic interaction records to **Global Biotic Interactions** (GLoBI; Poelen 2014 et al., 2017), and in turn GloBI will organize, standardize and integrate our records into existing species interaction datasets accessible via web tools (**Figure 8**) and machines (e.g., via data archives, APIs, software libraries). TPT will also integrate interaction data that GloBI harvests form other sources (e.g., publications). Using a collaborative approach, GloBI has quickly grown to be the largest global resource to provide open access to about 3.2 million species interaction records, covering approximately 100,000 taxa over 100 data sources and citing more than 100,000 references. GloBI exclusively 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: **1)** an automated crawler that continuously discovers, integrates, links, and indexes existing interaction data (e.g., TPT portal) and associated annotations from heterogeneous openly available data sources; **2)** a data service that provides access to query aggregated and linked interactions online via a web API, R package (rglobi), and offline-enabled access via Elton; and **3)** data archives 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 2012, 2017), EOL API (Pafilis et al. 2015; Parr et al. 2014), Global Names Architecture (Pyle 2016), and CrossRef (Pentz 2001) to provide complex features such as taxonomic resolution and citation DOI lookup. Vocabularies are tied to ontology (UBERON, Relations Ontology,

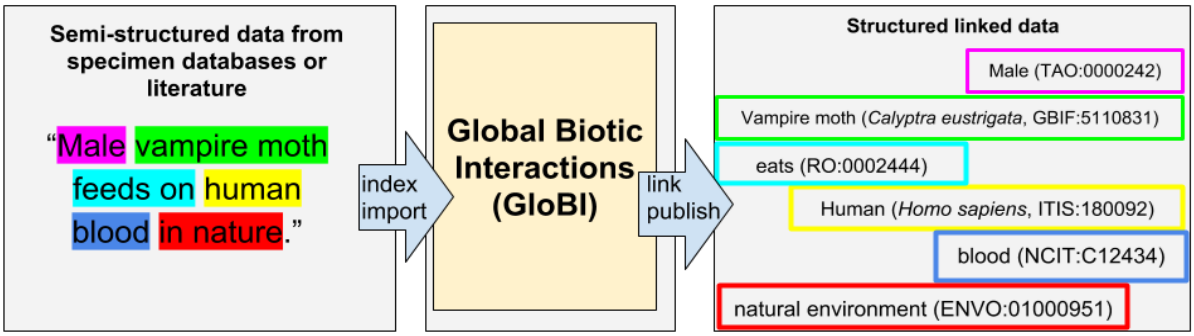

Figure 8. GloBI takes the semi-structured data from specimens and literature, maps them to ontologies and name resolution services, and exposes them as linked text. Biotic interactions are merged, even though they may come into the TPT portal in different formats. Color of text boxes on the left correspond with colors of boxes on the right.

EnvO; Mungall et al. 2012; Buttigieg et al. 2013; Seltmann et al. 2013; Pafilis et al. 2015; OBO Foundry 2016; Simons & Poelen 2017), but data sharing is not dependent on resolution of a specific data sharing model or consensus over terminology. Biotic interactions will come from a variety of sources, including the TPT Portal and other data portals (i.e., Walter Reed Biosystematics' VectorMap). The TPT Project Coordinator will work closely with network participants to identify and process literature data for input into GloBI.

VII. Training and Workshops

We will host a workshop for TPT participants during all three years of the project period. **Year 1: Project Kick Off.** We will host a three-day workshop focused on **i)** project management, **ii)** best practices and standardization of digital data capture, and **iii)** workflow development. The developers of our primary imaging system (i.e., Macroscopic Solutions) will provide hands-on image training demonstrations using Macropod Systems. During this workshop, PI Seltmann and consultant Jorrit Poelen (GloBI) will provide specific guidance for capturing biotic-association data in Symbiota (TPT) and in other data entry platforms. **Year 2: Interaction Data Mobilization**. This workshop will be a dedicated meeting for up to 10 project personnel to work on evaluating and improving the interaction data sharing using GloBI. The goal of this workshop will be to evaluate existing pipelines from TPT as use cases and publish concrete examples of interaction data sharing between repositories. Prior to the meeting, PI Seltmann and consultant Poelen will provide several examples to data providers using the Symbiota and Arctos databases, spreadsheets, literature, and other project databases. The group will discuss how the

information is shared, the use of interaction terms found in relevant ontologies (Relations Ontology, UBERON), work through examples of data sharing as a group, and provide data to GloBI at the meeting. The last portion of the meeting will include VertNet participants John Wieczorek and David Bloom, who will focus on integration with vertebrate data, and capturing a discussion on the future challenges for use and sharing of species interaction data. **Year 3: Train the Trainers.** A final workshop will be held towards the end of the project period and will provide learning and discussion sessions on **i)** functionality of TPT web portal, **ii)** final data processing, crowdsourcing and georeferencing, **iii)** interaction data research use in GloBI, and **iv)** sustainability. Trainers will consult with colleagues at their home institutions and give presentations to their respective research communities in the TPT Network. Workflows, guides, and videos will be made available via the TPT Portal for other users throughout the project period and beyond, setting the stage for consistent digitization efforts of parasite collections worldwide.

VIII. Broader Impacts

Public Engagement. TPT will partner with the MPM Education and Programs department to educate the public about arthropod parasites and the importance of natural history collections. We will use TPT as a platform for enhancing the impact of visitor engagement activities such as: science-focused lectures (e.g., "*Lunch and Lecture*" program), MPM's "*Exploring Life on Earth*" exhibit, youth summer camp series, and informal presentations in the "*Bugs Alive*" outreach area. The MPM serves over 500,000 visitors annually, including youth from low-income families (up to 36% poverty rate; Smeeding & Thornton 2016) and other underrepresented groups of all ages in Wisconsin. We will also work with PI Seltmann at UCSB to develop new online educational resources for teachers at intercity schools (i.e., Milwaukee Public School District, a Community Eligibility Program *CEP* participant; 83% of MPS students eligible) and for the award-winning Kids in Nature (KIN) program at UCSB (i.e., a federally recognized Hispanic-serving institution that provides resources for connecting research opportunities to underrepresented students). KIN has been working with K-12 teachers and students for over 16 years to develop environmental science lesson plans that follow the Next Generation Science Standards.

In partnership with the Learning Center at the FMNH, we will develop a series of *Meet the Scientist* activities that communicate the central themes of the project while fostering an understanding of the importance of museum collections and potentially assisting with data entry. *Meet the Scientist* is an established, scientist-driven, weekly program that past digitization projects have successfully used to promote crowdsourcing and that FMNH staff use to share their research with museum goers in the main hall of the museum. These tools will also be made available to other TPT natural history museums (UNL, UNM), centers (NAU), and outreach coordinators (PU). Results from the TPT project will also be disseminated broadly through a YouTube video series hosted at the FMNH, *The Brain Scoop*. This award-winning YouTube program reaches over 430,000 subscribers and to date has had over 20 million views. One episode will be produced in Year 2 and will feature the ways in which scientists use collections data and introduce the viewers to the unique and intriguing life histories of arthropod parasites.

Increasing Undergraduate Literacy in Education. The PIs, in collaboration with the NSF-funded Biodiversity Literacy in Undergraduate Education (BLUE) RCN-UBE team, will develop, assess, implement, and disseminate a module addressing core concepts and competencies as per the AAAS Vision and Change Document in data and biodiversity literacy. The BLUE Data Network is a diverse and inclusive network of biodiversity researchers, data scientists, and biology educators focused on undergraduate data-centric biodiversity education. PIs from TPT will work with BLUE to develop a module that aligns with current content in introductory biology courses and is designed for broad scale adoption. The module will include generating research questions using digitized collections-based data, exposing students to the data pipeline of digitizing a set of museum specimens of vector taxa, transcribing the data from the labels, and georeferencing. The students will use the data, in combination with archived digitized data, to develop preliminary niche models, with the goal of answering hypothesis-driven research questions, such as '*What were the historical distributions of the vectors? Do the past data accurately predict the present day distributions?'* Using distribution points, students can ask questions about the ecological nature of disease transmission: '*Where have diseases been reported? Does that correspond with the distribution of the vector?'* This module will be assessed to determine the degree of retention of biological knowledge and understanding. Finally, this module will be disseminated to all institutions associated with this TCN, made available online, and discussed in the education section at the Evolution meetings. We have already identified PIs at TPT institutions (e.g., UNR, UWSP, UU, UCSB) who will use this module in their classes or modify it for museum internship programs (e.g., CAS, FMNH, MPM).