Collaborative Research Digitization TCN: Southwest Collections of Arthropods Network (SCAN): A Model for Collections Digitization to Promote Taxonomic and Ecological Research

Project Summary

Intellectual Merit. The Southwest Collections of Arthropods Network (SCAN) project will bring together resources from 10 small to large sized arthropod collections located in the megadiverse but taxonomically underexplored ecoregion of the southwestern United States and adjoining Mexico to create a virtual collection network. The region's high diversity of life zones, arthropod species, and susceptibility to climate change – combined with the varied historical trajectories of its regional collections – present unique challenges for specimen digitization and data integration. In particular, the lack of a single representative collection of regional arthropods and considerable distances between existing collections jointly hinder opportunities to make regional holdings available on-line. In addition, the strengths of each participating collection are highly varied in terms of regional and taxonomic coverage, level of identification, and digital documentation. As a result, hundreds of thousands of specimen records that are highly valuable for taxonomic and ecological research remain unavailable due to insufficient identification or lack of digitization and networking. To overcome these obstacles, SCAN will leverage new collaborations and institutional investments into collection resources to develop a dynamically structured, state-of-the-art digital platform designed to facilitate arthropod biodiversity and ecology projects in the southwestern US region. The project will focus on ground-dwelling arthropods (e.g., ants, selected beetle families, grasshoppers, spiders) because they are highly responsive to temporal and spatial environmental changes, taxonomically diverse, and among the most commonly monitored terrestrial arthropod taxa. We will use best museum stewardship practices and leading-edge informatics drawing on recent advances in collection cataloging (i.e., specimen-level data capture), imaging, networking, remote identification, and web delivery.

Specifically, we will **(1)** assess and develop mechanisms for integrating different database systems in operation by the participating institutions; **(2)** capture label data from over 750,000 specimens and image ~15,000 arthropod specimens in the collaborating SCAN collections; **(3)** develop and implement new cybertaxonomic practices, based on the Symbiota top-level software and the Filtered Push semantic model, to significantly increase the capacity of taxonomic experts to provide remote e-identifications; and **(4)** produce a coherent georeferenced dataset and virtual taxonomic identification library for southwestern ground-dwelling arthropod taxa, to be used for ecological monitoring and species distribution/climate change modeling. Thus, SCAN will help facilitate future taxonomic research on Southwest arthropods and usher in a new era of specimen-based biogeographic research in the Southwest by allowing researchers from multiple disciplines to quantify the ecological and evolutionary impacts of climate and land use on key arthropod groups.

Broader Impacts. SCAN will serve as a testbed to synergize systematic and applied ecological research through the integration of data from a regional group of arthropod collections of different sizes and strengths. The insights gained and products produced through this project – viz. multi-database integration, remote e-identifications, reciprocal distribution of value added to participating collections, and joint prioritization of datasets – will serve as a model for the development of future regional arthropod collection networks. Most immediately, a novel collection-driven approach will greatly promote the identification of existing specimens in smaller collections. SCAN will be interdisciplinary and will promote involve over 50 undergraduate students in cyberinfrastructure, systematics and ecology. SCAN datasets will be important for a number of ecological inventory and monitoring programs (e.g., LTER, NEON, NPS-Biodiversity Discovery), as well as climate impact studies that need historical and/or present-day occurrence data. Public outreach efforts will include **(1)** display- and presentation-based outreach to more than two million annual visitors to the participating collections and museums; **(2)** extension of a BugGuide-like website to serve as a regional image and identification resource library that will be made available to ongoing education-outreach programs; and **(3)** proactive engagement of working amateur entomologists and museums throughout the region to add their data to SCAN.

Section D. Project Description

1. Introduction

A. Project Challenge

The southwestern United States encompasses an exceptionally rich diversity of ecosystems that spans the four major North American deserts and the southern Rocky Mountains (DeBano *et al.*, 1995; Gottfried *et al.,* 2005; Marshall *et al.,* 2006). Overlaid on its general aridity, this ecoregion is characterized by strong elevation, temperature, and precipitation gradients that have resulted in the development of a patchwork of “sky islands,” which remain taxonomically and ecologically underexplored relative to many other areas of the United States (DeBano *et al.*, 1995; Molina Freaner & van Devender, 2011). The high plains, escarpments, bottomlands, and thousands of playas and rivers of the more eastern area add additional habitats of relatively rich biodiversity. These varied environmental gradients provide ideal outdoor laboratories for understanding climate change impacts on biodiversity (Mac *et al.,* 1998; Breshears *et al.,* 2008). However, the apparent disparity between (1) the region's high species richness and (2) the poor state of species-level exploration, poses severe challenges to large-scale ecological analyses (Jones *et al.,* 2006; Michener *et al.,* 2007). This is especially problematic for arthropods which constitute the world's most diverse lineage of multi-cellular organisms (Ødegaard, 2000), and includes >30,000 arthropod species in the Southwest. The relatively poor state of exploration of southwestern arthropods is mirrored in the presence of the small- to large-sized regional institutional collections whose respective specimen holdings and levels of curation are highly varied. Jointly, these conditions severely limit our ability to use collections to assess and manage the region's ecosystems.

The Southwest region covers approximately 1.6 million km^2 in area, comprising 20% of the contiguous United States. However, there are less than 15 institutionalized arthropod collections with active research and educational programs present in our region (**Figure 1**). Thus in relation to its large area, the density of collections is low as is reflected in a 764 km average distance between them. Historically, the wide geographic spacing has made cross-institutional integration of interests and projects difficult, and instead has led to most collections developing unique profiles that were typically narrowly tailored to their location and the research interests of the respective curator(s). As expected, the idiosyncratic trajectories of each collection have produced large differences in overall specimen numbers, regional and taxonomic concentration, activity and growth across longer time periods, species-level identification, and finally cataloging (i.e., specimen-level data capture), imaging, and the GBIF-compliant data presentation. For instance, more than 80-90% of specimens at Arizona State University (ASU) are identified to the level of species (see ledger at franz.lab.asu.edu/collection.html); however, this collection experienced very limited growth prior to 1960 and after 1995. In contrast, the similarly scoped collection at Northern Arizona

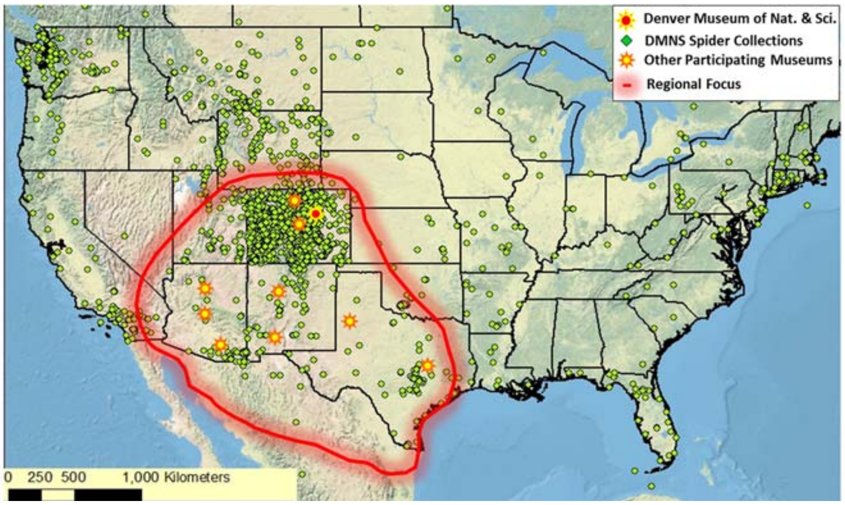
Denver Museum of Nat. & Sci.
DMNS Spider Collections
Other Participating Museums
Regional Focus
0 250 500 1,000 Kilometers

Figure 1. United States map showing regional focus (red line), location of participating museums (gold stars), and distribution of a subset of spider records from DMNS (green diamonds) illustrating the regional dispersion of SCAN arthropod collections relative to host member. A major goal will be to produce maps where the focus area is filled with green diamonds for all SCAN taxa.

University (NAU), has amassed tens of thousands of specimens in recent years through structured sampling efforts that offer valuable ecological insights, yet so far only 12% of these specimens have been expert-identified to the level of species. Other SCAN collections have consistently strong records of building collection capacity and specimen numbers; during the past 15 years the Texas A&M Insect Collection has added 50,000-100,000 research specimens per year (i.e., fully curated specimens).

Once the accumulated holdings of southwestern arthropod collections are networked through a unified e-portal, the aforementioned disparities between them will create a data environment where specific requests for expert identifications of thousands of specimens will become feasible, necessary, and likely very frequent. Such requests can be communicated through sets of provisional names, images, and other information. Fortunately, many members of this TCN proposal are either (1) taxonomic specialists in the targeted arthropod groups, and/or (2) have expertise with the necessary cyberinfrastructure to increase our capacity to identify existing specimens and discover new species. This means that we will have the capacity for *both* frequent occurrences of identification requests *and* people who can expertly address those. Finally, because of significant taxonomic overlap among the regional collections, there will be a need to *systematically and efficiently redistribute* the "value-added" species identification data among *all* collections; thus offering opportunities for implicit identifications of specimens not directly examined by an expert but implied by the original identification service.

B. Project Goals

Southwest **C**ollections of **A**rthropods **N**etwork project (**SCAN**) will jump-start emerging efforts to connect southwestern collections to form a virtual museum that will become a major component of the iDigBio infrastructure (idigbio.org/) and enable new kinds of biodiversity science. SCAN includes 10 small to large collections from the states of Arizona, Colorado, New Mexico, and Texas, with additional linkages to southern California and northern Mexico (Figure 1). We will focus on digitizing our entire specimen holdings of taxa that are typically collected in pitfall trap studies (Samways, 2005) – i.e., spiders, orthopterans, ground-dwelling beetles, ants, and other terrestrial arthropods totaling 162 families, many thousands of species, and over 1million specimens. Our digitization focus matches the complexity of the data on hand, and will allow us to produce high-quality information on southwestern arthropods that is readily usable in taxonomic, ecological, biogeographic, and climate change analyses (Sabu *et al.,* 2011). To achieve these high standards of data quality, we will implement new concepts and cyberinfrastructure from the Filtered Push project (Wang *et al.,* 2009; etaxonomy.org/mw/Filtered Push), and develop a distributed specimen identification request-and-redistribution platform that will increase the cross-institutional consistency of our data and serve as a testbed for similar functionalities for the iDigBio HUB and other collaborative projects. In particular, we will:

(1) Catalog 736,735 ground-dwelling arthropod specimens from 10 southwestern collections, thus reducing an enormous gap in taxonomic and geographic sampling of the most common invertebrate groups used for ecological monitoring;

(2) Produce high-resolution images of 15,125 arthropod specimens in order to extend our abilities for web delivery, remote identifications, and taxonomic research;

(3) Create a synthetic regional database using the Symbiota top-level software which also sustains the successful 1.5 million specimen SEINet herbarium network;

(4) Design a dedicated website to link SCAN members, primary digitization products, and other products that will enhance SCAN's visibility to biodiversity researchers and the public;

(5) Promote accessible, well-structured and taxonomically sound data for modeling climate change impacts on species distributions and ecological studies of arthropod communities; and

(6) Provide new remote specimen annotation and identification workflows through the SCAN network, based on the Filtered Push information model, with downstream benefits for the iDigBio HUB.

2. Background

A. Creation of the Southwest Collections of Arthropods Network

The Southwest Collections of Arthropods Network began in 2006 as an informal association of insect and arthropod collections in the southwestern United States. The network was initiated while planning for an All Taxa Biodiversity Inventory for 36 National Park Service lands of the Colorado Plateau region (www.mpcer.nau.edu/atbi/). Because of the collections' divergent strengths in size, taxon diversity and geographic representation, it quickly became clear that effective action in addressing issues associated with regional arthropod biodiversity would require the development of mechanisms that could support substantive collaborations among all participating regional institutions. Building on this ATBI background, representatives from many of the collections participated in a three-day organizational meeting in 2010 at which the strengths and limitations of Southwest arthropod collections were reviewed, individual collection goals and policies were assessed, and plans were developed to further strengthen collaborations among group members. Group discussions identified the current lack of a unified, regional, specimen-level database of arthropod taxa as the most critical impediment to future collaboration and the potential centerpiece of a future collaborative group initiative.

Sustained communications among group members, and new key personnel and infrastructure additions at several institutions since the 2010 meeting, have continued to synergize the network. In particular, several institutions (ASU, DMNS, NMSU, UNM, and UA) have recently made significant investments through the hiring of new collection-associated tenure-track faculty and/or new collection managers. Several institutions (ASU, NAU, UNM, UA) have also recently implemented new information technology infrastructure with capabilities for capturing and manipulating text and image data (see appended Facilities documents). These investments have already produced a culture change in inter-institutional relations among the SCAN collections. They are contributing to a revival in collection-based arthropod research at these institutions. Thus, our proposed digitization project is a logical, timely and important next step in our efforts to better network all Southwest arthropod collections for research and education.

B. Digitizing Efforts to Date

Progress towards specimen-level digitization is ongoing at all SCAN collections. Activities are partitioned as follows: (1) *cataloging* of specimen labels and associated data into a management database; (2) *imaging* of specimens and specimen lots using high-end imaging systems; and (3) *exploring* new procedures to increase the quality and speed of digitization, and incorporating data flows to increase the capacity to update and utilize data. To date 838,500 specimen records have already been entered into either museum databases or Excel sheets, of which 234,874 specimen records include our focus taxa (**Table 1**). Seven SCAN members have initiated imaging programs of both *in situ* referenced and vouchered specimens, totaling at least 10,195 specimen images. Four collections (ASU NAU, UA, and UNM) have acquired state-of-the-art imaging systems from Visionary Digital (VisDig). Several SCAN museums have significant holdings of arthropods from over 20 National Parks in the Southwest (note: NPS specimens not included in this proposal). All National Park Service (NPS) specimens are required to be cataloged, and we have so far digitized over 16,000 NPS arthropod records (CSU, NAU, and UNM). We have also initiated an image library for NPS specimens at cpbc.bio.nau.edu/CPMAB/NPS. We have conducted extensive reviews of external digitization efforts, database and web delivery systems (e.g. GBIF, 2010), so as to strategically incorporate best practices into the SCAN project.

The regional focus of SCAN complements other digitization efforts that are underway in adjacent parts of the U.S.; particularly the California-centered CalBug project and the Midwestern InvertNet initiative. Moreover, our plan to digitize ground-dwelling arthropods is: (1) designed to generate data that are ideal for ecological, geographic, and climate change analyses (e.g., Chown *et al.*, 2007; Kardol *et al.*, 2011; Pelini *et al.*, 2011); (2) calculated to provide a comprehensive regional coverage of the selected taxa through cumulative specimens records; and (3) set up to facilitate the design and implementation of remote specimen annotations. Jointly, these improvements will enhance our collections and serve as a globally applicable model to assist other digitization efforts.

Table 1. List of participating museums with descriptive statistics for both the entire collection and target taxa that will be the focus of SCAN. A total of 736,735 specimens will be cataloged (211,387) records to integrate, 525,349 new records) and 15,125 specimens representative of regional species will be imaged. The statistics emphasize the importance of the collections, prior commitment to digitization, and scope of the proposed project. [1] will adopt Specify; [2]based on total specimens in SCAN

	Entire Collection					Ground-Dwelling Arthropod Focus Taxa			
Institution	Collection Size (Specimens)	# of species	Management Database	# of specimens imaged	# of specimens cataloged	# of specimens cataloged for SCAN	# of specimens to be cataloged for SCAN	# of species to be imaged for SCAN	% of specimens identified to species
Arizona State University	700,000	12,000	Specify	275	0	0	56,705	1,658	77%
Colorado State University	2,650,000	40,000	Specify[1]	0	100,000	0	36,090	918	56%
Denver Museum of Nature & Science	1,346,000	25,000	KE EMu	120	30,000	27,000	34,123	3,235	47%
New Mexico State University	150,000	8,000	Specify	0	25,000	0	23,819	1,784	7%
Northern Arizona University	250,000	8,500	Specify	750	12,000	7,650	26,705	1,875	8%
Texas A&M University	2,600,000	43,700	TAMU	2,000	491,000	108,816	146,210	0	86%
University of Arizona	2,000,000	35,000	Specify	5,000	5,000	4,500	85,545	0	98%
University of Colorado at Boulder	700,000	14,500	Biota[1]	0	80,000	40,500	28,297	1,740	74%
University of New Mexico	120,000	10,000	Specify	850	30,500	5,580	30,544	1,740	50%
Texas Tech University	1,000,000	7,500	Specify[1]	1,200	65,000	17,340	57,312	2,175	54%
Total or Mean	11,516,000	20,420	10	10,195	838,500	211,387	525,349	15,125	65%

3. Rationale for Research Focus & Approach

A. Suitability and Delimitation of Target Taxa

We selected taxa that could be realistically digitized in three years and provide specimen data for use in the near-term by both ecologists and taxonomists. Specifically, we targeted arthropods that were (1) of taxonomic interest to regional systematists, (2) ecologically important and model taxa for climate change research; (3) sufficiently representative at each participating museum to make an overall impact on digitizing all collections; however (4) not so large as to decrease feasibility or compromise production of high-quality information that is readily applicable in ecological analyses. Since there is no universally accepted definition of "ground-dwelling arthropods", we included taxa that were commonly included in pitfall studies (Lightfoot *et al.*, 2008; Higgins, 2010; Holguin *et al.*, 2010). Using a "method of capture" criterion rather than using life-history traits, which may vary even at low taxonomic levels, provided us with an easy operational method for including the 162 arthropod families selected for this project.

Ground-dwelling arthropods are among the most targeted groups of arthropod community ecology studies, inventories, and monitoring efforts (Schowalter, 2011). They are ideal for monitoring biological diversity because they represent species-rich assemblages that reflect a wide range of ecosystem processes, provide a key food resource for many vertebrates, and respond to even small microhabitat changes (Beattie, 1985; Wilson, 1988), and they compose a large portion of the total arthropod species richness in most habitats (Agosti *et al.*, 2000; Leather, 2005).

B. Expected Impacts towards Advancing Community Ecology

More than1,600 published studies since 1963 have used pitfall traps to assess the abundance and diversity of ground-dwelling arthropods (ISI Web of Knowledge). Pitfall traps are (1) used by the National Ecological Observatory Network (NEON) to assess the abundance and diversity of ground beetles (Carabidae); (2) typically employed in arthropod surveys at most Long-Term Ecological Research (LTER) sites (Lightfoot *et al.,* 2008; Parmenter *et al.,* 2011;); and (3) were chosen for a pilot monitoring program by the National Park Service (Cobb & Higgins, 2011). Despite the large number of studies, most pitfall studies conducted in North America have exceedingly poor taxonomic resolution. This limits our ability to understand the ecology of any single species, since most studies either only identify taxa to supra-specific levels or designate "morphospecies" codes (Krell, 2004). Confirmed species-level identification allows for various forms of meta-analyses designed to elucidate spatial and temporal patterns in occurrence and identify traits which are critical for assessing species' responses to climate change (Jones *et al.,* 2006). The deficiency in the taxonomic resolution of North American arthropod community studies is illustrated in **Figure 2,** which compares pitfall community studies that did *not* engage taxonomists versus those studies that *did* utilize taxonomists. Eleven percent of the "species" that were identified in "non-taxonomist" studies were identified to the lowest level possible, while 89% were coded as "morphospecies". However, studies using taxonomists properly identified 97% of the species in the study, comparable to levels of identification achieved in many European studies. Furthermore, "non-taxonomist" studies only recorded 60-70% as many total species compared to taxonomist-assisted studies. We propose that the consistently finer taxonomic resolution in European studies is due to their strong historical commitment to taxonomy. It is unlikely that the number North American arthropod taxonomists will increase in the near future. However, by precisely mapping species occurrences, coupled with image catalogs and novel eTaxonomy resources, we can greatly increase our capacity to provide higher resolution data for inventories, community-level studies and monitoring programs.

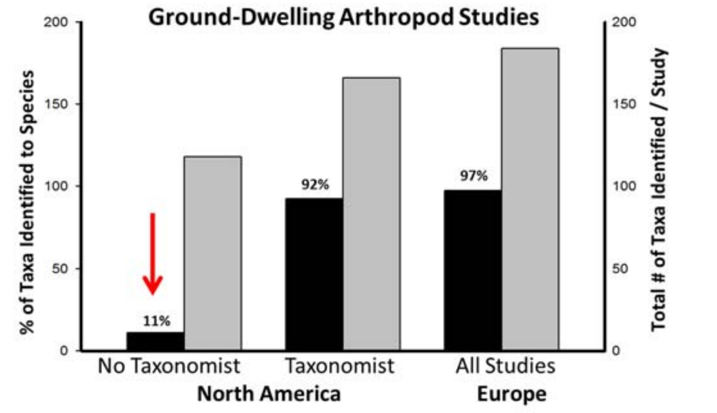

Figure 2. Comparison of taxonomic resolution among three types of arthropod community studies using pitfall traps. Studies from North America that either did not use taxonomists (n=16), or used taxonomists (n=8). The third category included six randomly selected pitfall studies from Europe (n=6).

C. Where are the Data? Climate Change & Arthropod Species Distributions

Arthropods are poorly represented in climate change impact-species distribution modeling studies because only a few taxa have adequate occurrence data required to populate distribution models; and no North American studies are comparable in data richness to their European analogs (Settele *et al.,* 2008). Even well-studied taxa such as ants, where global projections have been performed (Jenkins *et al.,* 2011), lack adequate data for the Southwest to model individual species responses (Gary Alpert, pers. comm.). Only 15% of 278 studies that modeled species distribution responses under different climate change scenarios used arthropods and yet they comprise ~70% of all described species. Arthropod occurrence data are fundamentally different from vascular plants and vertebrates in that they almost exclusively reside in *collections,* and are only very rarely based on field observations. We will address these shortcomings by digitizing and georeferencing existing specimen data in our collections.

D. Expected Impacts towards Advancing eTaxonomy

SCAN's digitization and networking goals. Southwestern ground-dwelling arthropods are highly diverse and are of great interest to SCAN curators and collection associates. SCAN's digitization and networking goals will not only (1) reduce a very significant and longstanding gap in our collective knowledge of their basic species-level diversity and distribution, but (2) facilitate profound *procedural* changes (viewed as a "game changer") in *how* smaller regional collections tackle the taxonomic challenges related to identifying and revising these groups through *pioneering* involvement of global expert resources.

Scope of Taxonomic Challenges. We still only have an approximate diversity estimate for even the most well studied Southwest arthropod taxa. For instance, more than 330 ant species have been documented to occur in Arizona alone, however, (1) at least 50 of these remain unidentified and many are likely new to science (Johnson, 1996; Cover & Johnson, 2011). New species records are regularly being discovered as part of the annual Ant Course in Portal, AZ, and the Navajo Ant Project (http://navajonature.org/). Nearly 1,200 species of darkling beetles have been recorded from the United States, and ~ 85% of them occur in the Southwest. Arizona alone is estimated to have 600 darkling beetle species, including several Mexican species that have recently been collected in the southern half of the state (A.D. Smith, unpublished data). More than 1,000 weevil species are thought to occur in the Sonoran desert (C.W. O'Brien, pers. data), of which only ~550 species have been officially recorded (O'Brien & Wibmer, 1982). Virtually all larger weevil genera have new species in the southwestern region and require new taxonomic revisions. In this context, SCAN will leverage information on tens of thousands of taxonomically under-determined arthropod specimens, providing both georeferenced locality data and images that will be accessible to expert taxonomists world-wide.

"Game-Changing" Approach to Engaging Taxon Experts. We clearly recognize that digitizing our collection holdings within a regional management database, while essential to modern biodiversity research, is by itself not sufficient to achieve the highest desirable resolution of SCAN's taxonomic data (Moore, in press). Indeed, at the global level we face a daunting need to identify and describe several millions of new arthropod species in a span of few decades, as a prerequisite for sustained management of our natural resources (Wheeler *et al.*, 2011). In this context, the SCAN member collections embody an all-to-common set of conditions (see also Section 1A) that have historically limited their relative contributions to large-scale taxonomic revisions. Too often, many of their most valuable specimens remain *less than optimally identified,* thus not offering the greatest returns on investments in physical storage and in terms of usefulness for biodiversity research. Under the traditional approach, such specimens tend to remain haplessly underused *unless* specimen requests/loans are made proactively, or a specialist happens to visit the collection to carry out on-site curatorial work. The difficulty of bringing in national and international experts to perform such tasks affects all research collections; but smaller, geographically remote collections are disproportionally affected.

We view the SCAN model as a critical step towards paradigmatic change in how such small collections interact with taxonomic experts, by making the requests for identifications *collection-driven* as opposed to *expert-driven.* In particular, individual SCAN curators, collection associates, and even students will have the capability to readily *assemble* sets of specimen records and high-quality images, and *package* these sets as requests to one or more taxonomic experts using the Filtered Push/Symbiota cyberinfrastructure (see details below). Other SCAN members can learn about these requests and their outcomes. The threshold for contacting experts will decrease significantly, since no shipments and paperwork among collections and experts are involved. The visibility of the taxonomic composition of such requests, and of the experts providing resolution, will increase, thereby highlighting the critical needs and who is credited with addressing them.

By making the e-identifications public, permanently retrievable, and suitable for redistribution through the Filtered Push system, a mutually enhancing revisionary taxonomy network can emerge. In other words, individual collections can potentially benefit from *any* identification service made to *any* of the networked member collections. From the perspective of the working taxonomic expert, this outlook should provide additional incentive to provide identifications, since their impact may be pushed far beyond the original source of specimens.

4. SCAN Organization & Expected Outcomes

A. Organization & Responsibilities

Neil Cobb (NAU) will be responsible for overall project performance, as well as collaborations with other digitization efforts, especially the existing TCNs and the national iDigBio HUB. He will provide coordination among participating SCAN museums with regard to *cataloging* and *imaging* activities. Cobb will also chair an oversight committee comprised of PIs from SCAN institutions. Paul Heinrich (NAU) will be responsible for coordinating database entry procedures and integration of member data into a regional database and the SCAN Image Repository. Nico Franz (ASU) will assume the primary responsibility for coordinating Filtered Push-

Symbiota work, and will coordinate with Paul Heinrich on the Specify-Symbiota integration tasks. Additional SCAN tasks will be tailored to match the combination of expertise, experience, capacity, needs, and affordability at each member institutions; so that each can create a team to carry out the day-to-day digitization workflow tasks. Each respective PI, co-PIs, and senior personnel will accept leadership responsibilities depending on their levels of expertise in taxonomy, digitization and IT infrastructure, ecology, and outreach to the public. All relevant best practices for specimen digitization will be summarized and shared at SCAN's all-hands meetings.

B. Taxonomic and Geographic Scope

We have identified 100 focal families in 22 arthropod orders to be digitized (Grimaldi & Engel, 2005). We included taxa that spend at least one stage of their life history in or on the ground. Although our regional emphasis is on the Southwest, we included specimens located outside this region *unless* they represented large special collections from other regions, especially those located outside of North America. This strategy ensures that a very high percentage of our digitized holdings will pertain to the focal ecoregion. Likewise, we were fairly inclusive taxonomically and deliberately chose to digitize all specimens within a selected family; *except* for in very specific cases were large numbers of specimens are from subfamilies that are not known to be sampled by pitfalls (e.g., Curculionidae holdings at ASU and TAMU). Based on previous digitizing experiences, we want to avoid any “checkerboard effect” in cataloging our collections that would impair long-term curation needs. This strategy follows the ADBC goal of digitizing all specimens in biological collections. We will significantly impact all collections by digitizing 5.4% of our cumulative number of *specimens* and an estimated 7% of the number of *species*. The data pipelines we build for this project can be readily exported to other arthropod groups as resources permit.

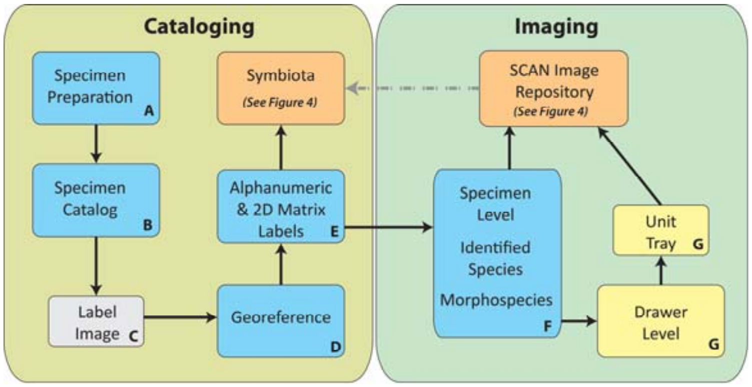

Figure 3. Digitization workflow steps for *cataloging* and *imaging* specimens (A) Organize collection taxonomically; (B) enter label data & field notes; (C) third party future collaboration, Paul Tinerella (UMN) for OCR & speech to text solutions; (D) process data through GEOLocate; (E) apply institutional catalog labels and 2D matrix labels; (F) image representative specimens; and (G) image specimen lots. Data is uploaded to the SCAN Image Repository with linkages to Symbiota and local databases. Yellow boxes indicate actions that will occur at a subset of collections.

C. Cataloging New Specimen Data

The two entry processes of the digitization effort, i.e. *cataloging* and *imaging,* are diagrammed in **Figure 3**. We will develop a priority list for digitizing, focusing first on key taxa represented in all museums (e.g., Carabidae, Tenebrionidae and Formicidae). Specific institutions will develop digitization standards for taxa with special curation and imaging needs (e.g., UNM for Araneae). The cataloging of 525,349 new specimen records will be operationally divided into three processes: (1) organizing specimens, including unidentified material for ease in processing; (2) recording label data and accessory data, including georeferencing through GEOLocate; and (3) applying catalog and barcode or 2D matrix labels. Some curatorial work is anticipated to maximize efficiency during the digitizing process, and there will be extensive preparation of the database prior to entering data.

Nomenclature. We have developed a shared taxonomic nomenclature stemming primarily from the *Nomina Nearctica* checklist, with select updates (e.g. Bouchard *et al.,* 2011). For non-insect groups we have adopted specific up-to-date authority files for each group; for example, we will use Chilobase for centipedes, the Hoffman catalog for millipedes, and Platnick's World Spider Catalog. Our taxonomic authority hierarchies will be shared and homogenized using Symbiota (see also Section 4G). Our current nomenclature files include all North American arthropod species. We have mapped 95 common data entry fields to a shared a Specify schema for three collections

(NAU, UA, and UNM). This schema will be shared with other SCAN collections that will use Specify and will be aligned with DMNS and TAMU who will continue to use their respective software platforms.

Georeferencing & Habitat Data. All museums will georeference records through GEOLocate, either within Specify or separately through a stand-alone version of GEOLocate. We have averaged 80% correct matching of latitude-longitude coordinates with descriptive label localities using GEOlocate. Manual determination of coordinates for the remaining 10-20% of records will depend on the data and resources available at the respective institutions. For example, a large series from a single collector with extensive field notes will be manually georeferenced. Additional habitat or collection data not on specimen labels will also be recorded if they are readily available from publications or field notes.

Labeling. All digitized specimens will receive a unique catalog number that will consist of an alphanumeric catalog label that includes the institution code and unique identifier numeric codes (example ASUC-054321). The alphanumeric code will typically be included on a matrix code label. Most museums will follow the 2D matrix code procedure developed by UNM (see www.barcode-labels.com/[barcoding-bugs]). Our intention is to minimize the number and size of labels and maximize the efficiency in being able to scan specimens for loans or inventory.

D. Incorporating Previously Cataloged Data

We will incorporate an additional 211,387 specimen records that exist in various databases or Excel sheets. Several of these legacy data sets are relatively complete (i.e.., georeferenced and Darwin Core compliant), while others will require additional work. For example, very few records from TTU are georeferenced, and all still need to be reviewed for accuracy in transcription. UCB plans to convert data from Biota to Specify. We will also inform all other North American museums about our project and inquire about possibilities of sharing data they have on ground-dwelling arthropods from the Southwest and the status of their cataloging efforts for our target groups. We have already acquired a large ant database courtesy of Michael Weiser (University of North Carolina); and are confident that we can add to our records through the inclusion of future SCAN collaborators.

E. Imaging Representative Specimens of Species

The primary objectives of imaging activities will be to increase the rate of *species identifications* for specimens that are not identified to species (cf. Krell, 2004) as well as providing *representative images* for known species to aid in future ecological studies and eTaxonomy. To this end we will create a complete image set (e.g., dorsal, lateral) for 26,077 specimens, including adult male/female sets for species that display significant sexual dimorphism. We will coordinate among institutions to avoid unnecessary duplication of images for the same species. Our target is to provide 1-2 examples of images for each of the identified species found at all museums. Our first all-hands meeting will set SCAN standards for these activities, following best practices that have already been developed and received consensus either within SCAN or at a global scale (e.g. Häuser *et al.,* 2005). Eight SCAN institutions will be engaged in imaging, while TAMU and UA will only conduct specimen cataloging due to the large number of specimens targeted for cataloging at their collections.

We will use portable or desktop imaging systems from Visionary Digital (VisDig) for all imaging. Three methods will be employed to produce and present high-quality images. First, the Visionary Digital Imaging System employs Zerene Stacker to montage multiple images taken at different focal points into one image with sharp focus over the entire specimen. This is extremely difficult with conventional macrophotography equipment and greatly enhances the value of web-presented images. Second, we will use Zoomify to display these large images on the web. Zoomify allows web presentation of extremely large images which can be zoomed, panned and annotated (cpbc.bio.nau.edu/cpmab/nps/). Finally, we will use GigaPan's photomosaic imaging tools to create very high resolution images of insect drawers and present them on the web. We have a GigaPan robot for drawer-level imaging (NAU), using cameras and lenses from the Visionary Digital systems. We have experimented with both VisDig and GigaPan imaging options and currently prefer VisDig. Regardless of the system used, we will obtain levels of image quality comparable to other GigaPan drawer imagery (http://blog.insectmuseum.org/?p=2467).

For institutions running Specify we will link images to cataloged specimens and automate the delivery of new images to the SCAN Image Repository. In addition, all images will be posted to Morphbank and other relevant portals operated by taxon-specific user groups (e.g., BugGuide). There are multiple routes to posting images to

MorphBank – i.e., directly from Specify, or through web-based image up-loading and we will explore each to determine which works best with our institution-specific workflows.

F. Training

Each institution will be responsible for training and day-to-day student mentoring. We will develop standard digitization protocols through consensus within SCAN. In order to facilitate the training of students and standardize techniques, we will create digital pamphlets and short "how-to" videos which can be viewed on the SCAN website. A major goal of the ADBC program is to promote increased efficiency in the digitization process. To that end, we will continuously assess procedures used either within SCAN or by other digitization projects that could significantly increase our efficiency.

G. SCAN Cyberinfrastructure

SCAN will be supported by an innovative and carefully coordinated cyberinfrastructure (**CS**) that will facilitate high levels of inter-institutional data communication and reciprocal enhancement. The SCAN CS (**Figure 4**) is comprised of four main interacting domains: (A) Symbiota, a centralized top-level data portal that allows harvest and search of all member catalogs: (B) the various institutional specimen-level databases (in most cases Specify 6); (C) the Filtered Push annotation technology which will allow us to push identification requests and services provided by taxonomic experts through the network; and (D) a centralized SCAN Image Repository linked to both the local databases and the main portal. In order to accomplish this, the Symbiota portal software (see Figure 4A) will automatically harvest the member institutions' local databases and import their data into a centralized portal. Symbiota is specifically designed for providing top-level database integration and value added (such as semi-automated error correction, annotations, user and public interfaces, etc.). Symbiota presently sustains the successful SEINet herbarium network with 30 member collections, and is also the data portal for the recently funded lichen TCN. Therefore, in working with Symbiota, SCAN is setting up preconditions for the integration of plant/arthropod taxonomic and distributional information throughout the Southwest. Symbiota lead developers are on board to support the integration of their software into SCAN through the development of data harvesting tools compatible with each SCAN member database.

Local Collection Database Management Systems (Figure. 4B). Most SCAN members will adopt or are already using Specify 6 as their local collection software. Specify is a NSF-supported, free, open-source software that embodies modern back-end standards and accessible front-end graphic interfaces. Two SCAN members will continue to use their current database systems (DMNS: KE Emu; TAMU: OZ, a Specify precursor), whereas UCB plans to migrate their data from Biota to Specify. Data will be entered and managed locally at each member institution. Most SCAN member institutions have adopted a shared database schema and shared taxonomy for ground-dwelling arthropods; this will be a focal point of discussion at the first all-hands meeting. Once the Symbiota portal is set up, we will periodically harvest data from all member institutions into a centralized Symbiota portal. Symbiota will provide the data and functionalities for portal-wide searches, downloads and visualizations (maps) of taxon distributions. Symbiota can be configured to automatically harvest the data from each local database using the TDWG Darwin Core standard. We will also institute a data security policy to protect sensitive species records (see Data Management Plan).

Specify will be the local database management program for most collections. Specify 6 contains several advanced features, including the integration of GEOLocate for georeferencing of specimens from locality text and LifeMapper for geospatial data modeling, visualization and analysis. Specify users will therefore be able to render maps of their own collection data and also see how localities of their specimens compare to the known distributions and projected models of those species based on data in GBIF (James H. Beach, pers. comm., see letter of collaboration).

Symbiota – Specify - Filtered Push Integration. Through the Symbiota-based SCAN portal we will also deliver data to iDibBio and GBIF using the Darwin Core-based upload features in Symbiota. Both Symbiota, which already provides a range of cross-collection data quality assessment and correction functions and the Specify Software Group (**SSG**) are working with the Filtered Push (**FP**) development team to include FP technology (**Figure 5**). Specifically, Symbiota and FP are collaborating to include new functionalities to detect and annotate problematic records, and to provide messaging capabilities that would enable their correction. SSG is working with FP to update Specify database content semi-automatically using FP. In addition, Symbiota supports the integration of the Google Mapping API, which facilitates both map displays and spatial search of occurrence records.

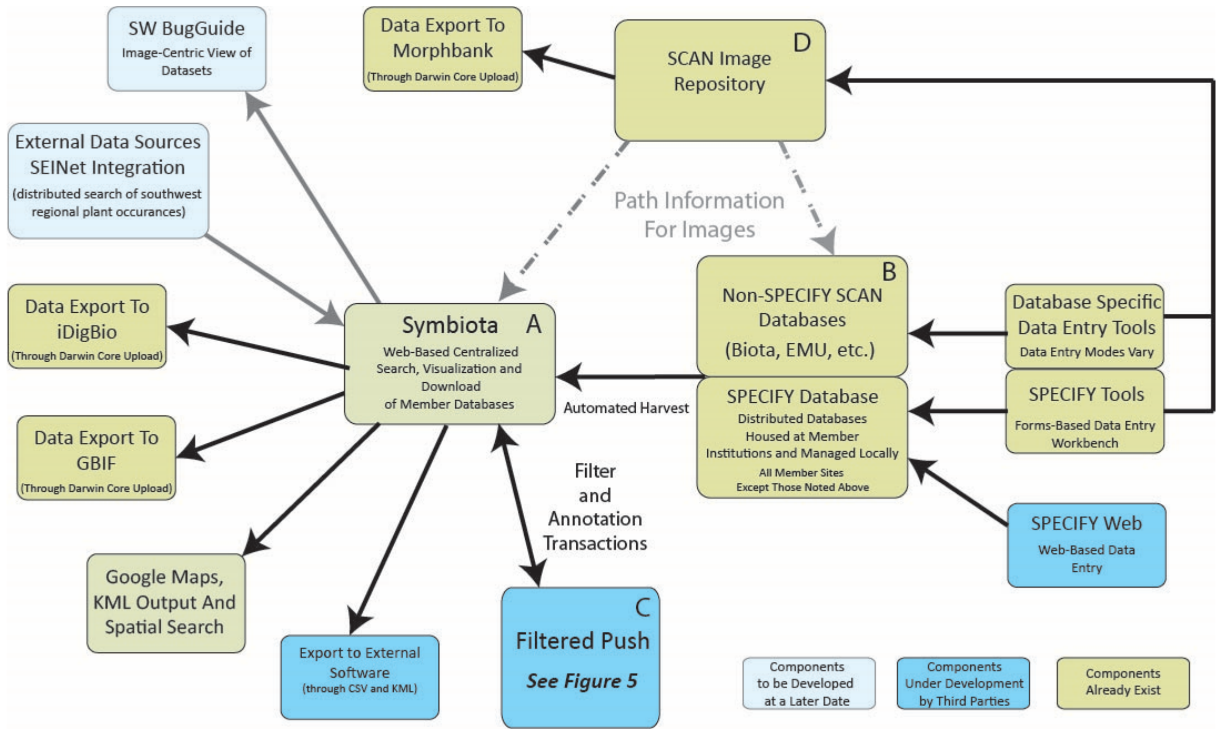
SW BugGuide
Image-Centric View of Datasets
Data Export To Morphbank
(Through Darwin Core Upload)
SCAN Image Repository
D
External Data Sources
SEINet Integration
(distributed search of southwest regional plant occurances)
Path Information For Images
Non-SPECIFY SCAN Databases
(Biota, EMU, etc.)
B
Database Specific Data Entry Tools
Data Entry Modes Vary
Data Export To iDigBio
(Through Darwin Core Upload)
Symbiota
A
Web-Based Centralized Search, Visualization and Download of Member Databases
Automated Harvest
SPECIFY Database
Distributed Databases Housed at Member Institutions and Managed Locally
All Member Sites Except Those Noted Above
SPECIFY Tools
Forms-Based Data Entry Workbench
Data Export To GBIF
(Through Darwin Core Upload)
Filter and Annotation Transactions
SPECIFY Web
Web-Based Data Entry
Google Maps, KML Output And Spatial Search
Export to External Software
(through CSV and KML)
C
Filtered Push
See Figure 5
Components to be Developed at a Later Date
Components Under Development by Third Parties
Components Already Exist

Figure 4. SCAN Cyberinfrastructure comprised of Servers, Software and Processes. (A) Search, aggregated download and limited mapping are provided by the centralized portal system. (B) Data are housed and managed locally. (C) FilteredPush technology at the portal allows efficient and automated discovery and correction of a wide range of data problems at the record level. (D) Images are uploaded to the repository directly and path information is recorded in the local and central databases. Images are contributed to external repositories from the SCAN repository while catalog data is contributed through the portal.

Filtered Push - Network Supported Specimen Identification and Redistribution. We will implement and refine new ontology-based specimen data annotation tools being developed by the Filtered Push project (Morris *et al.*, 2009; Wang *et al.*, 2009; Dou *et al.*, 2011). We view Filtered Push as a pioneering technology that promises to revolutionize remote data quality improvement in distributed specimen collection networks (see Figure 5). In this regard, SCAN will benefit the greater collections community by becoming one of the first TCNs to apply this emerging technology, using Symbiota (http:\\symbiota.org/tiki/tiki-index.php) as the top-level environment for annotation requests, delivery, and redistribution to all network members. We have engaged both the lead developers of Filtered Push and Symbiota to achieve these goals (see ASU budget and justification). We will concentrate specifically on *annotations of identifications* within SCAN, thereby addressing a central need to add taxonomic value to tens of thousands of insufficiently identified specimens (Table 1). The envisioned workflow pipeline has been identified by iDigBio HUB leaders as a necessary tool that nevertheless remains outside their present development activities and therefore complements them (Lawrence M. Page, pers. comm.). Our internal evaluations strongly indicate that Filtered Push and Symbiota are the most adequate platforms to achieve this functionality, in part because these projects are now approaching each other with the same purpose. In other words, Filtered Push, so far primarily a computer science focused project, will find a relevant, motivated user community in SCAN. Symbiota already offers a range of functionalities to identify and correct data errors or inconsistencies that are not strictly taxonomic in nature.

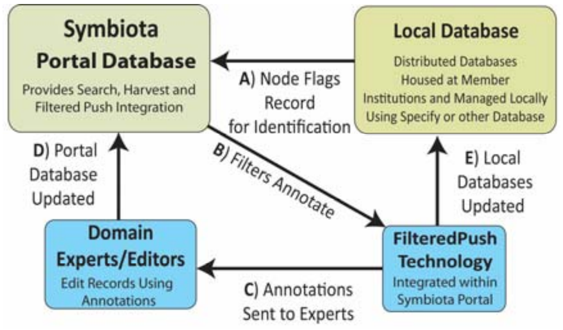
Symbiota
Portal Database
Provides Search, Harvest and Filtered Push Integration
Local Database
Distributed Databases Housed at Member Institutions and Managed Locally Using Specify or other Database
A) Node Flags Record for Identification
B) Filters Annotate
D) Portal Database Updated
E) Local Databases Updated
Domain Experts/Editors
Edit Records Using Annotations
FilteredPush Technology
Integrated within Symbiota Portal
C) Annotations Sent to Experts

Figure 5. Filtered Push deployed in the SCAN framework will specifically; (A) detect records for identification or correction; (B) create annotations describing the problem; (C) route annotated data to domain experts for editing; (D) once edited, update the Symbiota; and (E) update the local database. Local collections managers can then accept or reject these changes or request clarification.

Our primary FP use case (see Figure 5) is that of a taxon specialist performing a remote identification, either deliberately or following a collection-driven request. This use case is directly tied to the overall justification and goals of SCAN (see Section 1A), and is well documented by Filtered Push (etaxonomy.org/mw/). In this scenario, an individual collection will *flag* sets of database records and imaged specimens as insufficiently identified in their database (Fig. 5A). These records-to-receive-identification-annotations will be *harvested* by Symbiota which also stores links to the corresponding images in Morphbank. Filtered Push *recognizes* these requests (Fig. 5B) and sends them in a tabulated form to the pre-identified taxon expert (Fig. 5C). Taxon can come into contact with SCAN through a variety of opportunities, and can set up specific interest profiles to receive such requests. Once the taxon expert makes their identification the portal database is updated (Fig. 5D), and subsequently FP updates the local databases (Fig. 5e). The identification request-provision-redistribution loop is then closed through the regular harvests of updated database holdings to the Symbiota portal. Presently, both Morphbank and Symbiota are developing graphic user interfaces to also facilitate such expert identification annotations. These options will provide two alternative entry pathways to engage experts and increase the taxonomic identification value of SCAN holdings.

Expert identification annotations provided to an individual SCAN member will be accessible to all other collections in the network (using the Filtered Push Mapper for non-Specify institutions). The new identifications are furthermore *pushed back* to the Specify database requesting them. Filtered Push will generate institution-specific aggregate statistics ("metrics of data quality improvement") of the numbers of requests, annotations, specimens, taxa (including taxonomic ranks), and experts involved, thus allowing us to continuously track ongoing identification activities throughout the network and identify deficiencies in key groups. We believe that this functionality can serve as a model for engaging and recognizing taxonomic expert contributions to insufficiently identified arthropod collections at a national and even global scale.

SCAN Project Website. We will create a dedicated website to post all policies, procedures, and updates as well as other supporting content, including news from other digitization efforts. The ASU School of Life Sciences Visualization Laboratory (see ASU budget justification) will design the front-end for the SCAN website. Content placed on the SCAN web site will be coordinated by NAU, who will also maintain the server, hard drives, and back-up storage system (see NAU Facilities and Data Management Plan).

H. SCAN Complementing other Collection Digitization Efforts

We have established communications with as many groups as possible that would ensure that we are compatible with, and complementary to, other collection digitization efforts in terms of geographic (Calbug) or taxonomic emphasis (SEINet, iDigBio TCNs). Likewise these collaborations will be important to ensure that we do not "reinvent the wheel" and can incorporate lessons learned from other efforts. SCAN will maintain compatibility with other TCNs and the iDigBio HUB by utilizing the same data standards (Darwin Core) and compatible software environments. We have corresponded with each of the three existing TCNs in order to understand their processes so that we can maximize compatibility wherever possible, and will conversely invite other TCNs to collaborate with us at all levels. We will take advantage of prior experiences made by these TCNs, to refine SCAN's own training to researchers and undergraduate students. In relation to each of these, SCAN has a distinct taxonomic, regional, ecological, and collection-service oriented emphasis, which flows readily from the unique challenges of SCAN (see Section 1) and complements existing efforts to document other taxa (e.g. bee species [Cornell & AMNH])

5. SCAN Timeline and Sustainability

Timeline. The three-year project timeline will allow for thorough coordination among SCAN members and collaborators, while still maintaining a strong focus in finalizing digitization of collections and serving data. Major issues to solidify and reinforce at project initiation include standardization of digitization and IT practices where necessary, and implementation of relevant new approaches from individual museums and other projects. These tasks will be prioritized for the project's initial all-hands meeting at NAU and reviewed at annual meetings. Throughout the project, we will ensure that each museum is keeping pace. Development of the new Filtered Push functions will begin in the second half of the first project year, and become fully implemented in year 2 when the SCAN Symbiota portal is fully functional and populated with specimens requiring identifications.

TIMELINE	Summer Year 1	Fall-Spring Year 1	Summer Year 2	Fall - Spring Year 2	Summer Year 2	Fall - Spring Year 3	Summer Year 3
SCAN (All/Many Hands) Meetings	▓			▓			▓
Catalog Specimen Data	▓	▓	▓	▓	▓	▓	░
Image Species		░	▓	▓	▓	▓	░
Symbiota Integration & Filtered Push		░	▓	░	▓	▓	░
SCAN Image Repository (Web)			░	▓	▓	░	░
Data Sharing (GBIF, MorphBank)				▓	░	░	▓
Reports, Publications & Metadata				▓	░	░	▓
Assessment & Evaluation	░	▓	░	▓	░	▓	▓

The timeline chart reflects programing of key activities, dark grey indicates times of strong focus and light grey indicates a less intense focus.

Sustainability of SCAN. Our commitment to a regional and virtual collections network is underscored by our collective progress in cataloging, imaging, database development, and workflow implementation. The long-term prospects of SCAN are promising in light of the significant number of recent hires of tenure-track curators and full-time collection managers at various member institutions. Sustainability is also ensured by NAU's long-term committed support of the Geospatial Research and Information Lab (see NAU Facilities). SCAN will put in place concepts, infrastructure, and practices (Figs. 4 & 5) that can readily be expanded to include other taxa, as well as new institutional and personal collections. In particular, our adherence to Symbiota as a Darwin Core compliant top-level software, and use of open source software such as Specify, Morphbank, and Filtered Push, will allow easy entry points for new data and information providers. (see also Data Management Plan)

6 Broader Impacts: Research, Mentoring & Outreach

SCAN will provide a wide range of broader impacts, ranging from genuine conceptual and technological advances in eTaxonomy to museum outreach. Several key examples are described below.

A. Enhancing Virtual Capabilities

Species Exploration. Over the past year we have concentrated our efforts on developing the capacity for virtual research and education, spanning across the full range of our interests in taxonomy, ecology and cyberinfrastructure. Increasing capacity for virtual research and education in smaller regional museums is critical in the Southwest where the physical distance among museums is often prohibitive. For instance, we will take advantage of the global virtual popularity of the International Institute of Species Exploration lead by Quentin Wheeler (IISE; http://species.asu.edu/), and use SCAN data as a testbed for new IISE products. In particular, Wheeler and Larimer (VisDig) are developing new remotely controlled imaging technologies to increase our capacity for off-site involvement of taxonomic experts. Such technologies are critical for smaller collections with low levels of species identifications to establish meaningful e-collaborations with specialists worldwide (Wheeler, 2009). We are extending this work by offering a first digital imaging course in collaboration with VisDig in 2012. We expect to expand this course in 2013 to include all SCAN participants and solicit participation from other imaging projects.

Trans-Border Impacts. Unfortunately, at present there are no suitable and actively maintained arthropod collections located in the states of Sonora and Chihuahua, Mexico (T.R. van Devender, Sky Island Alliance MABA Project, pers. comm.). However, several SCAN collections (ASU, UA, UNM, TAMU) already include tens of thousands of specimens from northern Mexico whose arthropod communities remain vastly undersampled and poorly known (Bailowitz & Palting, 2011; Castrezana, 2011). By providing the necessary cyberinfrastructure to integrate specimen-level data among collections, our project will also assist Mexican scientists in identifying ground-dwelling arthropods and therefore enable international collaborations related to systematics, biogeography, and changing distributions of these taxa.

SCAN-SEINet Interactions SCAN data will be well positioned to interface with the SEINet data portal. SEINet is a distributed network of ~1.5 million southwestern plant collection records at 30 institutions – all searchable through the SEINet Portal. In the long-term, we envision full integration of plant and arthropod specimen data and related functionalities through SEINet and Symbiota. Working through Symbiota will position SEINet and SCAN as nation-wide leaders in terms of providing plant/arthropod specimen data for biodiversity research and outreach.

eTaxonomy. We have outlined the potentially game-changing impact of SCAN for eTaxonomy in Sections 3D and 4G. The novel, ontology-based, remote Filtered Push tools for specimen identifications will aid both the collections and systematics communities. We expect that, as more biodiversity data on diverse and difficult taxa are brought on-line, the need for requesting, providing, storing, and redistributing *authoritative data annotations* will match, or even exceed, that of just adding more data. Taxonomists will benefit immensely from receiving alerts about potentially new species material tailored to their specific interest. The SCAN model is a nucleus of a system for connecting experts with annotated specimen data records and images derived directly from an individual collection's needs. It is therefore a novel and broadly impacting solution to the "insufficient identification" challenge, using *both* an expert- *and* collection-driven approach. SCAN, in this sense, may act as a precursor to how the developed nations will soon interact with museums located in places like the Neotropics. Lastly, the taxonomic identification value added to the SCAN data will have positive impacts on derivative studies focusing on biodiversity, ecology, and climate change.

B. Community Ecology Studies and Climate Change Impacts

A major SCAN outcome will be the availability of data for ecologists and climate change impact modelers interested in both species distributions and biodiversity models (see Jones *et al.,* 2006; Saupe *et al.,* 2011). For instance, there are at least 13 recently finished or ongoing pitfall studies conducted by six SCAN institutions that will be able to leverage the full scale of SCAN data for their respective research themes and products. Given the proposed scope, hundreds or thousands of individual data points for many thousand arthropod species will be readied for such analyses. Moreover, we have been developing data for several years on hundreds of plant species in the Southwest, and are thus well-positioned to simultaneously serve plant and arthropod data to biodiversity and climate change impact modelers (Garfin *et al.,* 2011; Jenkins *et al.,* 2011).

C. Student Mentoring & Outreach to General Public

SCAN will require extensive training of contributing graduate and undergraduate in curation, databases, digital imaging, GIS, web development, and virtual collaboration within a museum network. At least **50 undergraduates** will be funded through SCAN, and each will receive direct training through the project leaders, and/or through courses in specimen imaging, GIS, informatics, and museum stewardship. Starting in year 2 we will offer a one-credit virtual class for both graduate and undergraduate students, which will cover all aspects of digitization process, including strategies for effective virtual collaborations and developing outreach material.

Outreach to the general public is a strong component of the mission of every SCAN member collection, as reflected (in part) in their web presence. We will incorporate SCAN outreach activities that are already in place at all of our museums, such as regular tours and presentations. In addition, we will provide content for complementary programs at our institutions (e.g., Insect Discovery at UA, and CAP-LTER-ASU). We will prepare and publish digitization best-practice manuals, posters, displays, brochures, and other training and outreach documents to enhance the project's visibility.

D. Involvement of Underrepresented Groups

Our project includes four females that are PIs or Co-PIs (DMNS, UA, UCB, and UNM). All SCAN universities have significant numbers of minority students, especially Hispanic and/or Native American students, and all have specific programs for engaging minority students.

SCAN Data Management Plan

Cyberinfrastructure

Collection records will be stored locally for each SCAN member collection, eight nodes will use the open source Specify 6 and MySQL database engine. This software combination is mature and cross-platform; it can run on Linux, Windows or Mac workstations or servers, granting maximum versatility for installation configuration to participating institutions. Two of the collections (Denver Museum of Nature and Science (KE EMu) and Texas A&M (in house but may adopt KE EMu), will continue to use their existing database solutions. Specify and Symbiota both store images of specimens and of labels as links to files in the filesystem. The Specify Project has a collaboration underway with the Morphbank Project to integrate Specify image archiving with Morphbank's planned distribute image repositories. Access to the collective holdings of the member institutions will be provided through a centralized data portal housed by the Geospatial Research and Information Laboratory at Northern Arizona University. This portal will be built using Symbiota, which is already in use with SEINet (a large network of herbarium collections). Using Symbiota will allow the automated harvest of records from member node database and storage of those records in the central portal database. This system is intended to allow easy integration into iDigBio's activities through the utilization of appropriate standards and processes.

User management, access control: SCAN will make use of the already existing user management and access control built into Specify and Symbiota. This will ensure that only approved users have the ability to modify record data, although Specify also has the added ability of keeping track of editorial changes. Monitoring of imaging and data processing activities at each collaborating collection will be the responsibility of that institution under the supervision of the project data manager. We envision that all editing of specimen records will be done through the local database system tools. Centralized searching of all catalogs through the Symbiota web portal will be read-only and will not allow alteration of records. Web Portal-users will have the option of viewing search results or downloading records in XML, Excel, or CSV files.

Data Management

Data Standards: Several of the proposed members of SCAN have already adopted a standardized Specify data schema for ground-dwelling arthropods. These collections have also devised a shared taxonomy for the most common insect taxa in our region. These schema are compliant with the Darwin Core Standard through schema mapping within Specify. TAPIR and DiGIR connectivity are available through Specify and Symbiota directly. Individual sites will have the responsibility of implementing TAPIR and DiGIR access to their collections. Taxonomic authorities will conform to the International Code of Zoological Nomenclature. TDWG-ratified georeferencing protocols and standards (http://wiki.tdwg.org/Geospatial/) will be followed whenever possible. In general, position data should be recorded in web- and GPS-friendly WGS 84 and decimal degrees. Finally, we will document each member collection in our network using the Natural Collections Descriptions standard version 0.9 or later (http://wiki.tdwg.org/NCD/). This standard allows standardized representation of collection, institution and contact information details.

Accessibility and electronic dissemination: All data will be served to the public using the Symbiota web portal. Most curators and their assistants will access the databases using a combination of Specify Web or the standard Specify Client. Locality data for species of special concern will be protected from the general user by specific settings in the database and portal software. Approved users will be able to view and map the full data for these records. This functionality is already built into the Specify software package. Information (including images) for these records will not be accessible to the general public until the locality details have been altered or removed to make site identification impossible. By adding some level of randomization (perhaps 10km) to locations, it should be possible to protect rare species locations while still allowing some reasonable level of species distribution mapping. Decisions on what information to suppress will be made based on local knowledge of species occurrence and national red lists. Both species records and site locations will be protected.

Data use tracking: Data-use tracking will be provided by the Specify database application and the Symbiota web portal analytics. Statistics for each Specify collection will show the number of searches against their records as well as number of downloads. General web site access will be tracked using Google analytics.

Data Quality Control and Assurance

Quality Assurance (QA) and Quality Control (QC) will be implemented at all nodes in as standard a way as possible, with an understanding that there can be more than one method for achieving results. A major aspect of this project will be the development of data quality assurance policies and methods (see documentation section). Specify's use of controlled taxonomy tables and automatic population of data entry forms for people, entities and locations from the database will be used to reduce errors in these parameters. Once an entity, person or location is added to the database and approved by a collection's manager, these data can be offered to data editors as pull-down selections as can taxon information and localities.

In addition to these automations we will use Filtered Push from within Symbiota to inform interested parties of both data quality and taxonomic questions. The Filtered Push approach allows messages regarding data quality or taxonomic identification to be pushed to collection managers or domain experts who have subscribed to alerts about specific taxa, geographic regions or data quality issues—i.e., alerts may be routed only to individuals who have expressed specific interests. Once the domain expert makes changes to the record, Symbiota and FP will be able to update the original record in the node's Specify database (with the local collection manager's approval). The Filtered Push developers are currently working with both the Symbiota and Specify developers to add these capabilities and plan to demonstrate Symbiota/Specify/FP capabilities in Spring 2012. . A major element of our QA/QC workflow will be the utilization of the QA\QC tools available in Specify, Symbiota and Filtered Push (see details in Cyberinfrastructure Section). This integration of software QA/QC at three levels will greatly improve the quality of occurrence data.

Data Sustainability

MPCER/GRAIL operates a data center comprised of multiple database and webservers with over 50TB of data storage including (see the facilities section for details). We recently received funding from the NAU Office of the Vice-President for Research to purchase a digital imaging system and a database/web server with 13TB of attached storage specifically for the archiving and dissemination of collections data. This server will host our Symbiota Portal and Image Repository. NAU and MPCER are committed to the goal of making research data more available to the research community and to the general public. SCAN is one means we are pursuing to attain this goal.

Project Documentation / Intellectual Property Rights / Data Security

We will develop the following formal documents during the first six months of the project. First a data ownership and sharing policy will be developed using input from the member institutions. This document will detail the rights and responsibilities of member institutions and the centralized portal. Our data ownership and sharing policies will follow the most recent recommendations of the Organization of Biological Filed Stations (OBFS) and NSF. Second we will develop a set data QA/QC policies, workflows and tools which will be used by all members. Finally, we will develop a data security schema which will formalize the process of access control for individual records. This security schema will be designed to allow curators to follow a simple set of rules to determine whether access to a specimen's record should be restricted or public. All project documentation will be publically available for download through our portal system. Any software or media developed with project funding will be made publically available through use of the appropriate open source license.

References Cited

Agosti, D., J.D. Majer, L.E. Alonso, and E. Schultz (Eds.). 2000. Ants: Standard Methods for Measuring and Monitoring Biodiversity. Smithsonian Institution Press, Washington, D.C.

Bang, C., and S.H. Faeth. 2011. Variation in arthropod communities in response to urbanization: seven years of arthropod monitoring in a desert city. Landscape and Urban Planning 103: 383–399. doi:10.1016/j.landurbplan.2011.08.013

Beattie, A.J. 1985. The Evolutionary Ecology of Ant-Plant Mutualisms. Cambridge University Press, Cambridge, UK.

Bailowitz, R.A., and J. Palting. 2011. Biodiversidad de los insectos con especial énfasis en Lepidoptera y Odonata. In: F.E. Molina Freaner, and T.R. van Devender (Eds.); Diversidad Biológica de Sonora, Universidad Nacional Autónoma de México, Hermosillo; pp. 315–338.

Bouchard, P., Y. Bousquet, A.E. Davies, M.A. Alonso-Zarazaga, J.F. Lawrence, C.H.C. Lyal, A.F. Newton; C.A.M. Reid, M. Schmitt, S.A. Ślipiński, and A.B.T. Smith. 2011. Family-group names in Coleoptera (Insecta). ZooKeys 88: 1–972.

Breshears, D.D., T.E. Huxman, H.D. Adams, C.B. Zou, and J.E. Davison. 2008. Vegetation synchronously leans upslope as climate warms. Proceedings of the National Academy of Sciences of the United States of America 105: 11591–11592.

Castrezana, S.J. 2011. Artrópodos terrestres no-hexápodos. In: F.E. Molina Freaner, and T.R. van Devender (Eds.); Diversidad Biológica de Sonora, Universidad Nacional Autónoma de México, Hermosillo; pp. 293–314.

Chown, S.L., S. Slabber, M.A. McGeoch, C. Janion, and H.P. Leinaas. 2007. Phenotypic plasticity mediates climate change responses among invasive and indigenous arthropods. Proceedings of the Royal Society of London, Series B 274: 2531–2537. doi: 10.1098/rspb.2007.0772

Cobb, N., and J. Higgins J. 2011. Monitoring ground-dwelling arthropods on National Park Service lands: pilot project at Mesa Verde National Park. Final Report submitted to the Southern Colorado Plateau Inventory and Monitoring Program, National Park Service.

Cover, S.P., and R.A. Johnson. 2011. Checklist of Arizona ants; revised 8-VIII-2011. Accessible on-line at http://www.asu.edu/clas/sirgtools/AZants-2011%20updatev2.pdf

DeBano, L.F., P.H. Folliott, A. Ortega-Rubio, G.J. Gottfried, R.H. Hamre, and C.B. Edminster (Eds.). 1995. Biodiversity and Management of the Madrean Archipelago: the Sky Islands of Southwestern United States and Northwestern Mexico; September 19–23, 1994; Tucson, Arizona. United States Department of Agriculture, Forest Service, Rocky Mountain Research Station. General and Technical Report RM-GTR-264.

Dou, L., J. Hanken, B. Ludaescher, J.A. Macklin, T.M. McPhillips, P.J. Morris, R.A. Morris, and Z. Wang. 2011. Building specimen-data curation pipelines using Kepler workflow technology in a Filtered Push network. Society for the Preservation of Natural History Collections Annual Meeting, Program and Abstract, 46. (see http://www.youtube.com/watch?v=DEkPbvLsud0)

Garfin, G.M., J.K. Eischeid, M.T. Lenart, K.L. Cole, K. Ironside, and N. Cobb. 2010. Downscaling climate projections in topographically diverse landscapes of the Colorado Plateau in the arid southwestern United States. In: van Riper III, C., B.F. Wakeling, and T.D. Sisk (Eds.); The Colorado Plateau IV: Shaping Conservation through Science and Management. University of Arizona Press, Tucson, pp. 22–43.

GBIF – Global Biodiversity Information Facility. 2010. GBIF Best Practice Guide for "Data Discovery and Publishing Strategy and Action Plans", Version 1.0. V.S. Chavan, R.K. Sood, and A.R. Arino. Global Biodiversity Information Facility, Copenhagen, 29 pp. ISBN: 87-92020-12-7. Accessible on-line at http://www.gbif.org

Gottfried, G.J., B.S. Gebow, L.G. Eskew, and C.B. Edminster (Eds.). 2005. Connecting Mountain Islands and Desert Seas: Biodiversity and Management of the Madrean Archipelago II; May 11–15, 2004; Tucson, Arizona. United States Department of Agriculture, Forest Service, Rocky Mountain Research Station. Proceedings RMRS-P-36. Fort Collins, CO. 631 pp.

Grimaldi, D., and M.S. Engel, 2005. Evolution of the Insects. Cambridge University Press, New York. 755 pp.

Häuser, C.L., A. Steiner, J. Holstein, and M.J. Scoble (Eds.). 2005. Digital Imaging of Biological Type Specimens: a Manual of Best Practice. European Network for Biodiversity Information, Stuttgart. 309 pp. Accessible on-line at http://www.gbif.org/orc/?doc_id=2429

Higgins, J. 2011. Ground dwelling arthropod responses to successional endpoints: burned versus old growth pinyon-juniper. M.Sc. thesis. Northern Arizona University, Flagstaff, AZ.

Holguin, C.M., F.P.F. Reay-Jones, J.R. Frederick, P.H. Adler, J-H. Chong, and A. Savereno. 2010. Insect diversity in switchgrass grown for biofuel in South Carolina. Journal of Agricultural and Urban Entomology 27: 1–19.

Jenkins, C.N., N. J. Sanders, A.N. Andersen, X. Arnan, C.A. Brühl, X. Cerda, A.M. Ellison, B.L. Fisher, M.C. Fitzpatrick, N.J. Gotelli, A.D. Gove, B. Guénard, J.E. Lattke, J.-P. Lessard, T.P. McGlynn, S.B. Menke, C.L. Parr, S.M. Philpott, H.L. Vasconcelos, M.D. Weiser, and R.R. Dunn. 2011. Global diversity in light of climate change: the case of ants. Diversity and Distributions 17: 652–662. doi:10.1111/j.1472-4642.2011.00770.x

Johnson, R.A. 1996. Arizona ants. Arizona Wildlife Views, June: 2–5. Accessible on-line at: http://www.asu.edu/clas/sirgtools/arizona_wildlife_views_1995.pdf

Jones, M.B., M.P. Schildhauer, O.J. Reichman, and S. Bowers. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics 37: 519–544.

Kardol, P., W.N. Reynolds, R.J. Norby, and A.T. Classen. 2011. Climate change effects on soil microarthropod abundance and community structure. Applied Soil Ecology 47: 37–44.

Krell, F.T. 2004. Parataxonomy versus taxonomy in biodiversity studies – pitfalls and applicability of "morphospecies" sorting. Biodiversity and Conservation 13: 795–812.

Leather, S.R. (Ed.). 2005. Insect Sampling in Forest Ecosystems. Blackwell Publishing, Oxford. 303 pp.

Lightfoot, D.C., S.L. Brantley, and C.D. Allen. 2008. Geographic patterns of ground-dwelling arthropods across an ecoregion transition in the North American Southwest. Western North American Naturalist 68: 83–102.

Lightfoot, D.C., A.D. Davidson, C.M. McGlone, and D.G.Parker. 2010. Rabbit abundance relative to rainfall and plant production in northern Chihuahuan Desert grassland and shrubland habitats. Western North American Naturalist 70: 490–499.

Mac, M.J., P.A. Opler, C.E. Puckett-Haecker, and P.D. Doran. 1998. Status and Trends of the Nation's Biological Resources, Volumes 1 & 2. United States Department of the Interior, United States Geological Survey, Reston, VA. 964 pp.

Marshall, R., M. List, and C. Enquist. 2006. Ecoregion-Based Conservation Assessments of the Southwestern United States and Northwestern Mexico: a Geodatabase for Six Ecoregions, Including the Apache Highlands, Arizona-New Mexico Mountains, Colorado Plateau, Mojave Desert, Sonoran Desert, and Southern Rocky Mountains. The Nature Conservancy, Tucson, AZ, 37 pp. Accessible on-line at www.azconservation.org

Michener, W.K., J.H. Beach, M.B. Jones, B. Ludaescher, D.D. Pennington, R.S. Pereira, A. Rajasekar, and M.A. Schildhauer. 2007. A knowledge environment for the biodiversity and ecological sciences. Journal of Intelligent Transformation Systems 29: 111–126.

Molina Freaner, F.E., and T.R. van Devender. 2011. Diversidad Biológica de Sonora. Universidad Nacional Autónoma de México, Hermosillo.

Moore, W. In press. Biology needs cyberinfrastructure to facilitate specimen-level data acquisition for insects and other hyperdiverse groups. Zookeys. doi: 10.3897/zookeys.@@.1944

Morris, P.J., M. Kelly, D.B. Lowery, J.A. Macklin, R. Morris, D. Tremonte, and Z. Wang. 2009. Filtered Push: annotating distributed data for quality control and fitness for use analysis. Eos Transactions AGU, 90(52), Fall Meeting Supplement, Abstract IN34B-08. Available on-line at http://etaxonomy.org/mw/File:Morris_AGU_2009.pdf

O'Brien , C.W., and G.J. Wibmer. 1982. Annotated checklist of the weevils (Curculionidae *sensu lato*) of North America, Central America, and the West Indies (Coleoptera: Curculionoidea) . Memoirs of the American Entomological Institute 34: 1–382 .

Ødegaard, F. 2000. How many species of arthropods? Erwin's estimate revisited. Biological Journal of the Linnean Society 71: 583–597.

Parmenter, R.R., M. Kreutzian, D.I. Moore, and D.C. Lightfoot. 2011. Short-term effects of a summer wildfire on a desert grassland arthropod community in New Mexico. Environmental Entomology 4:1051–1066.

Pelini, S.L., F.P. Bowles, A.M., Ellison, N.J. Gotelli, N.J. Sanders, and R.R. Dunn. 2011. Heating up the forest: open-top chamber warming manipulation of arthropod communities at Harvard and Duke Forests. Methods in Ecology and Evolution 2: 534–540.

Sabu, T.K., R.T. Shiju, K.V. Vinod, and S. Nithya. 2011. A comparison of the pitfall trap, Winkler extractor and Berlese funnel for sampling ground-dwelling arthropods in tropical montane cloud forests. Journal of Insect Science 11: 1–28.

Samways, M.J. 2005. Insect Diversity Conservation. Cambridge University Press, New York. 342 pp.

Saupe, E.E., M. Papes, P.A. Selden, and R.S. Vetter. 2011. Tracking a medically important spider: climate change, ecological niche modeling, and the Brown Recluse (*Loxosceles reclusa*). PLoS ONE 6(3): e17731. doi:10.1371/journal.pone.0017731

Schowalter, T.D. 2011. Insect Ecology, Third Edition: an Ecosystem Approach. Academic Press, Burlington, MA. 650 pp.

Settele J., O. Kudrna, A. Harpke, I. Kühn, C. van Swaay, R. Verovnik, M. Warren, M. Wiemers, J. Hanspach, T. Hickler, E. Kühn, I. van Halder, K. Veling, A. Vliegenthart, I. Wynhoff , and O. Schweiger. 2008. Climatic risk atlas of European butterflies. BioRisk 1: 1–710. doi:10.3897/biorisk.1

Wang Z., H. Dong, M. Kelly, J.A. Macklin, P.J. Morris, and R.A. Morris. 2009. Filtered-Push: a map-reduce platform for collaborative taxonomic data management. 2009 WRI World Congress on Computer Science and Information Engineering 3: 731–735.

Wheeler, Q.D. 2009. The science of insect taxonomy: prospects and needs. In: Foottit, R., and P. Adler (Eds.); Insect Biodiversity: Science and Society. Blackwell Publishing, New York; pp. 359–380.

Wheeler, Q.D. *et al.* 2011. Mapping the biosphere: origin, organization, and sustainability. Trends in Ecology and Ecology. (in review, acceptance pending)

Wilson, E.O. (Ed.) 1988. Biodiversity. National Academy of Sciences, Smithsonian Institution. D.C. 521 pp.