**Data Management Plan**

Our team will collect a variety of data including biological specimens, ecological data, genetic sequences of bat hosts and viruses and metagenomic data. It will also develop analytical platforms that will be of value in bridging the ecology and virology scientific communities. Management of such diverse datasets requires a comprehensive data management plan that our research group has both the experience and capacity to implement. Pursuant to the Feb 2013 OSTP Memo “Increasing Access to the Results of Federally Funded Scientific Research”, we will store and distribute all data, analytical tools, and publications generated from this study using open-source and open-access resources and archives.

1. ***Expected Data***

This effort will generate a substantial body of data pertaining to **a)** bat specimens, **b)** batsamples (tissue, saliva, feces, urine, and blood) available for viral discovery and isolation, **c)** genetic sequence data from novel viruses, **d)** metagenomic data that will include host, commensal, and pathogen sequences, **e)** an analytical pipeline for genome analyses of viruses*,* and **f)** source code for ecological and evolutionary analytical models. All specimens and samples collected in this project will be curated with their associated metadata. This will include geographic locations of bat populations within the cave, records of co-occurring species, distribution maps of the bats in the cave, and habitat characterization of the cave and surrounding areas. Relevant guidelines and permits are in place from ongoing research efforts by our organizations (i.e. IACUC approvals, SEMARNAT (Secretariat of Environment and Natural Resources of Mexico) collection permits, CDC importation permits, etc.), but these will be updated and made specific to our proposed project upon funding.

1. ***Data Format***

**Bat voucher specimens** (collected as needed for identification) will be stored as whole bodies in formalin or ethanol, and/or as dried skins. Associated ***excreta samples*** (feces, saliva, urine, blood) will be maintained at -80°C in viral transport media and in liquid nitrogen in the field, ***wing*** ***tissue biopsies (DNA vouchers)*** and ***ectoparasite*** samples will be maintained in ethanol (96%). ***Gene sequences*** from all viruses and hosts will be saved as raw sequence data and in FASTA format (amino- and nucleic- acid), from which they can easily be converted into other common molecular analytical formats (.nex; .meg; etc.). Ecological and evolutionary models from our ***analytical pipeline*** will be coded in R, a statistical open-source framework. Raw datasets will be saved as text files (e.g. csv or txt), and phylogenetic trees saved as Nexus format compatible with open source software. Biodiversity and genomic data will be stored in a relational database implemented in the open source software SQLite. To manage our geospatial datasets, we will make use of GRASS GIS an open source software used for geospatial data management, analysis and spatial modelling.

1. ***Access to Data***

Each of the investigators contributing to this project agree to share data collected in this project freely and with other interested investigators as soon as practical, but no later than 2 months following project component completion.

Bat voucher specimens will be curated, catalogued and stored in the publicly accessible, physical collections at the ***American Museum of Natural History* (*AMNH*)** (see collaboration letter from Amato). These collections are curated with digital, online databases to ensure that they can easily be searched and shared with the international community. Bat wing tissue samples and nucleic acid will be stored in liquid nitrogen at the Ambrose Monell Cryo Collection at the American Museum of Natural History, and catalogued in the specimen collection. All bat samples (saliva, urine, feces, blood) for viral discovery will be stored, catalogued, and processed at the **Center for Immunity and Infection** at Columbia University’s BSL 3 facility. As we have done in previous projects, all genome and partial gene sequences will be submitted to publicly accessible molecular databases, such as ***Genbank*** in standard formats (e.g. FASTA). Phylogenetic trees and associated metadata will be submitted to **TreeBASE** (treebase.org), as our team has done with previous publications. Additional underlying data sets will be submitted to **Dryad** (http://datadryad.org/) at the time of publication. Analytical pipelines for genome analyses (i.e. processing HTS data for viral discovery), and our relational database will be deployed on ***Google App Engine***, a cloud-computing platform that will host our web applications where they will be freely accessible. All our analytical scripts will be handled using a distributed revision control (e.g. git) to keep track of all changes and inputs from all collaborators. Scripts will be freely available at our cloud-computing platform and will be linked to a separate **github** projectaccount (github.com) to assure redundancy. Github is a web-based hosting service for software development and data storage.

4. ***Sharing Practices and Policies, Re-use and Re-distribution***

EcoHealth Alliance was rated by Charity Navigator with 4 of 4 stars for transparency as an organization and is committed to open-access sharing of scientific data. This project will allow reuse, redistribution, and the production of derivatives of the data for the purpose of scientific research and education. Potential users need to inform the research team prior to use, and the data collectors reserve the right for their priority in publication and co-authorships of publications written by the data users. We have included budget to publish open access peer reviewed journal articles from this project (see Budget Justification).

5. ***Archiving of Data***

Specimens will be archived, curated and maintained with associated metadata following the guidelines and best practices of each respective institution (EHA, CU, UNAM). This will ensure long-term public access to the core primary data. We propose to use a 3-level distributed model for data archiving and preservation, as follows: 1) Each group will use its data center to collect, host, process, and distribute their datasets and metadata, which ensures flexibility, and timely documentation among various disciplines and groups. 2) The data products from individual groups will be assembled to support data synthesis, analysis and assimilation. We will ingest and archive all the data sets collected from this project into our database and we will develop a web-application to be deployed on the cloud. We will use this application to query, visualize and distribute data and to track and manage datasets and metadata. 3) All datasets collected in the project will be submitted into open-access digital archives including: Dryad, Genbank, TreeBASE, and all analytical code will available in our public repository in Github. We also recognize the value added to the broader research community by other data centers (for example, NCEAS Data Repository for ecological data). Following successful funding, we plan to work with NCEAS Data Repository to establish a mentoring and collaborative relationship that enables our team to develop long-term archival mechanisms.

6. ***Data sharing among labs and PIs during course of project***

Our format, access, sharing and archiving practices outlined above will ensure high-level coordination and collaboration among partners on this project. We will conduct all research, project development and analyses collaboratively within our multidisciplinary team. Coordination will occur via monthly conference calls with all project staff. In addition, an annual, in-person management meeting will be held at EHA’s offices in New York. Joint UNAM-EHA-CII teams will conduct field research, and travel between the US and Mexico will occur for both US and Mexican staff. These practices will ensure a mutual understanding of methods, ownership stakes in all data collected and analysed, and effective collaboration towards the goals of the project. Our team has a long history of outstanding collaboration implementing these principals.