Data Management and Access Plan

Below we describe how we will assemble, manage, analyze and disseminate information generated through the network. An asterisk* after an individual's name indicates that there is a corresponding letter of collaboration.

1. Computer network and technical support
   The cooperative agreement between the USGS and Polistes Foundation will provide the computer infrastructure and technical support that we will need. This infrastructure includes a network of powerful Sun Solaris and X86 servers at the University of Georgia and an off-site location (Simpson*).
2. Long-term data preservation
   Should Polistes not be able to maintain Discover Life in the future, it is legally obligated by this cooperative agreement to transfer its tools and data to either another NGO or to a governmental agency. In addition, we will make all specimen data widely available through regular uploads to GBIF and other interested organizations.
3. Nightly backups
   To ensure data integrity, Discover Life maintains at least three copies of all files and keeps time-stamped copies of file updates and web transactions. Each night a 'mirror' utility automatically copies new and edited files from 'original' servers to a 'production' server and an off-site archive.
4. Database structure
   Discover Life does not use a database application, per se. Instead it uses custom software to manage millions of files that are stored and accessed via the operating system through Apache, mod-perl and other means. All original files are archived, converted to formats that can be shared rapidly and securely via HTTP, HTTPS and SCP, depending on users' requirements, and then indexed for rapid retrieval and analysis. Each night automated programs re-index datasets that have changed and build a master index across all datasets. While unconventional, this approach has major advantages over conventional databases: it is faster, requires less storage space, and scales well in integrating very large datasets. The mapper, for example, plots 22,000 points per second and can display maps with over 3 million points.
5. Data import
   Over the past two decades, the PI has developed tools to support the import of many file formats, data schema, and *un*controlled vocabularies. Rather than requiring contributors to submit their information in a standard format such as Darwin Core, we encourage them to upload the data in their existing formats. We use a variety of tools to convert contributions and integrate them across datasets. In addition to supporting Darwin Core, we will integrate data from non-standard formats from spreadsheets and databases. In addition to supporting decimal latitude and longitude degrees required by Darwin Core, for example, Discover Life supports over 10 additional ways of documenting latitude-longitude and UTM coordinates. The software assigns permanent, globally unique identifiers to images and records as they are uploaded and processed. These unique identifiers reflect contributors' specimen identifiers wherever possible. They facilitate user feedback and error correction.
6. Data export
   Discover Life's existing tools will enable the project to export data in different formats, including tab delimited text files, XML, and HTML. We will serve images as JPEGs at 5 resolutions, sound recordings as MP3 files, and videos at 3 resolutions in both Flash and QuickTime formats.

7. Software tools
   Discover Life's integrated tools enable users to upload, manage, analyze and download information. The tools include online albums of images and associated data (20p), a global mapper (20m), and identification guides and checklists (20q). These three tools work together so that records in individual albums contribute to maps and checklists, which customize identification guides by location, which in turn simplify the determination of species in the albums.
8. Web services
   The 20p, 20m and 20q software can be called as services by other websites and software programs. When customized these services return deep links to other websites. For example, Brown* and Hogue* call 20m with HTML iframe tags to map points that their Hover Fly Survey of LA County project has uploaded to Discover Life.
9. Documentation
   We document web services and other features of Discover Life at www.discoverlife.org/help. We include information on metadata, ownership, terms of use, and credits with pages, images and individual records as appropriate or required by contributors.
10. Automated data processing
    We will run automated programs using 'cron' to process information, update indices, output reports, and ready files for download by other websites and individual users.
11. Error checking and correction
    We will use automated programs, gazetteers, taxonomic authority lists, natural language processors and feedback from users to detect and correct errors.
12. Security
    All datasets on Discover Life require a password or specific IP address to change them. As required by contributors, we restrict access to certain information or to the means by which it is presented. While we will make the data collected by our high school and undergraduate teams globally available, we will continue to abide by the terms of use of other contributors' information. General web users will be able to contribute images and associated data. However, their contributions will be restricted to 'purgatory' and not visible until reviewed by project staff to make sure that they are appropriate for the site.
13. Restricted access to sensitive data
    In certain cases, such as with the exact location of rare and endangered species, we will not make all data publicly available but rather restrict them to researchers and land managers on a need-to-know basis.
14. Terms of use
    We will make the data collected by the students, in both raw and summarized forms, available on the condition that users credit contributors and project funders, in a manner similar to used by GBIF. We will expect authors of published papers to cite this NSF project and the USGS for their support.
15. DATA COLLECTION (see Project Management Plan[15])
    The undergraduate and high school teams will upload images and other data to their personal albums. They will document information such as where, when and how data were collected.
16. IDENTIFICATION (see Project Management Plan[16])
    As described in Species Identification (4.1.2.2), we will identify the species associate with records, or when impossible, the genus or higher taxa. For fungi that are not in our target 100 species, we will make images and associated data available to the Mushroom Observer so that their members can identify additional taxa (Hollinger*, N. Wilson*).
17. DATA INTEGRATION (see Project Management Plan[17])
    In addition to the photographs and other data that our teams will collect, we propose to

make NOAA, NASA, NEON (Gram*), GBIF, ITIS and other data available through Discover Life as permitted by the terms of use of such contributors. We will integrate such data with our field data to facilitate answering our research questions and also make them available in various forms for investigators not directly associated with the project.

18. ANALYSIS (see Project Management Plan[18])
    Our Science and Analysis Sub-committee includes the PI, Co-PI LeBuhn, Hargrove*, Hubbell*, Kjar* and others. These individuals, working with colleagues and students, will analyze our contemporary biological data in conjunction with abiotic data and historical biological information from collections, field sites and the literature. Hargrove and the PI propose to fit day-degree models to the flight periods of moths and flowering periods of plants to evaluate how the phenology of different species respond to temperature patterns. Their modeling will complement what they plan with MODIS satellite seasonal data around NOAA weather stations. They will extend Hochberg, Pickering and Getz's (1986) methods to evaluate phenology models using field data. They will use Monte Carlo methods to conduct a sensitivity analysis of their models' predictions. Hubbell, Hargrove, the PI and their colleagues will analyze alpha and beta diversity at our field sites both within and across the context of the NEON domains. LeBuhn and her student will analyze how temperature affects plant-pollinator interactions and seed set. Co-PI Stephenson will work with our modelers to analyze the impact of abiotic factors on slime molds, fungi, and lichen growth rates. Our analysis will include a hierarchical model framework that will allow us to separate the data acquisition model from the ecological process model. Our initial approach will be to define an observation process model (relating to data acquisition) and a process model (relating to the question being addressed). Our simplest models will examine the occurrence of species and their phenology. They will incorporate recent statistical methods to account for imperfect detection.
19. DISSEMINATION (see Project Management Plan[19])
    Through Discover Life we will make raw data and summary reports readily available to web users as we process and update them nightly. We will work with partner websites and other projects to make the information available through other outlets (NPN, Crimmins*; NEON, Gram*; USDA Forest Service, Hargrove*; Mushroom Observer, Hollinger*; EOL, Parr* and N. Wilson*; Floral Report Card, Schwarz*).
20. Physical samples and collections
    Although the vast majority of the data that we propose to collect will be digital, we envision collecting some physical samples to supplement this information for species and groups that cannot be identified by digital means alone. In these cases, specimens will be deposited in public collections.
21. Integrating new technology and software tools
    We will enable other researchers and web developers to extend our existing web services and integrate new statistical, graphical and other functions to process and disseminate the data.