PI: Alan R. Franck, Ph.D.	**Department:** Cell Biology, Microbiology & Molecular Biology
Institution: University of South Florida	**Agency:** National Science Foundation

DATA MANAGEMENT PLAN

1. <u>Data Collected, Formats, and Standards</u>: Data from this project are derived directly from physical herbarium specimens collected from the 13-state SEUS region present at FTG and USF. Data comprise jpeg images and text metadata in a SQL database. With about 260,000 specimen images and their transcribed label data, we estimate about 375 GB of data. High-resolution, compressed jpegs (minimum 2912 x 4368 pixels, minimum 300 dpi, for 11.5 inch × 16.5 inch herbarium sheet) average 1.5 MB in size (375 GB total) and associated metadata comprises about 1 KB per specimen (0.25 GB total).

Metadata transcribed from labels are stored in two tables in a Microsoft SQL Server database management system, following Darwin Core metadata standards. Transcription data will be recorded through a login-based web portal, modified by the USF Water Institute from the Microsoft Access Virtual Herbarium Express software application originally developed by the New York Botanical Garden. One table contains all metadata except the scientific name (i.e. Country, State, County/Parish, Location, Habitat, Description, Collector(s), Date, Latitude and Longitude, Elevation, and Accession Number). Error warnings built-in to the Virtual Herbarium software occur for duplicate Accession Numbers, invalid Dates, and invalid Latitude and Longitude. These errors must be corrected immediately before allowing the user to save the record and advance to the next one. The other table contains all determinations/identifications which includes Family, Genus, Species, Authors, Rank, Infraspecies, Authors, Determiner, and Determination Date. This table allows for multiple scientific names (annotation history) found on some specimens to all be linked with the same specimen image. The presence of multiple scientific names on a specimen occasionally occurs when there are differing taxonomic concepts or misidentifications. Proper spelling of scientific names will be validated by cross-checking on TROPICOS, the International Plant Names Index, and The Plant List. One transcriber works on all specimens of one taxon to enhance familiarity and reduce spelling errors. Internal lists of recently transcribed specimens are also made available to allow curators to continuously check transcription quality and correct errors. A training module, developed by the USF Herbarium curator (Co-PI), ensures all data is entered using the same standards by all student and volunteer digitizers. With extensive experience in georeferencing, the USF Water Institute will perform the bulk of georeferencing to incorporate radii or polygons of uncertainty to allow for varying levels of accuracy to the actual historical (unknown) collection point. Locations that are rare or unique in the datasets and cannot be batch processed will be georeferenced by students and volunteers. Export of all metadata is allowed on the SERNEC Symbiota portal and USF herbarium database as comma separated value (CSV) file.

2. <u>Physical and Cyber Facilities for Storage and Preservation</u>: High-resolution images will be maintained in duplicate at 1) the CyVerse (formerly iPlant Collaborative) infrastructure and 2) the USF servers. A low-resolution triplicate image is also maintained by iDigBio. For the CyVerse infrastructure image storage, at least three copies of all data will be maintained and checksums will be generated and used to ensure the integrity of all data stored over the lifetime of the project. CyVerse provides replication through two active online copies (one at the Texas Advanced Computing Center at the University of Texas at Austin and one at the University of Arizona) and a tape copy housed in a separate data center at the Texas Advanced Computing Center.

The specimen metadata will be maintained in quadruplicate, with these four institutions: 1) iDigBio, 2) SERNEC portal MySQL database as hosted by Arizona State University, 3) USF servers in a Microsoft SQL Server database, and 4) GBIF. USF servers will store specimen metadata and images during and after the project. Data will be managed by the USF Water Institute, which has an agreement with the USF Information Technology Department to provide database, GIS, and web servers to support applications and technology as a component of the same virtual servers farm for all core USF applications and data. All USF servers are backed-up nightly to an off-site facility. A diesel generator ensures continuous power to the USF servers during potential power outages. USF servers have provided herbarium data online without any significant access problems since 2003 on the *Atlas of Florida Plants* (AFP).

3. <u>Media and Data Dissemination</u>: Both USF and SERNEC seek to make data publicly available immediately. This will be accomplished by an IPT to continually transfer data from FTG and USF directly

to iDigBio, the SERNEC Symbiota portal, and GBIF after data has been reviewed by herbarium curators to ensure standards are followed. Current protocols regularly ingest data from USF every ~2–3 weeks to iDigBio and GBIF. By the completion of this project, all digitized collections (including those outside the SEUS) from FTG and USF will be publicly available on the SERNEC Symbiota portal, iDigBio, and GBIF and will also be available on the specimen search database hosted on the AFP. To increase access to data, taxa-specific links will be provided on each species page at the AFP (with >100,000 users annually) to the SERNEC portal taxa page. The scientific name present on the AFP containing the genus, species and, potentially, infraspecies will be used to produce a link in this format: http://sernecportal.org/portal/taxa/index.php?taxon=”Genus”%20”species”%20”rank”.%20”infraspecies”.

4. Data Sharing Policies and Public Access: No personal data will be shared as part of this project. As a publicly-funded project, no claim to copyright or database ownership will be made of aggregated data in the SERNEC TCN. Licensing follows CC BY-NC rights (Creative Commons Attribution Non-Commerical) requiring attribution and free non-commercial use including sharing and adapting the information for reuse. Location information of sensitive taxa (based on the SERNEC rare, threatened and sensitive species list) will be hidden from the public on iDigBio, SERENC, AFP, and GBIF but made available to herbarium curators and PIs through an administrator login system in SERNEC and the AFP.

5. Roles and Responsibilities of Personnel: The USF Water Institute will be responsible for the transcription application, the bulk of georeferencing, maintaining data, transferring data to iDigBio, the SERNEC Symbiota portal, and GBIF, and ensuring continual online accessibility of metadata and images hosted on USF servers. The Co-PIs (Franck & Jestrow) will be responsible for ensuring metadata follow specified standards and images are of high-quality (in-focus, properly orientated), and will be responsible for confirming or correcting the identification of the specimens.

6. Local Sustainability Plan: Both FTG and USF will continue digitizing all incoming SEUS specimens and continue sharing them with iDigBio, SERNEC, and GBIF beyond the life of this 3-year project. The USF Water Institute will maintain data transferring protocols beyond the scope of the project in order to continue enhancing the SERNEC dataset with new accessions added to the FTG and USF herbaria. Digitization is now integral to herbarium management and specimen organization at both institutions. To maintain an organized herbarium, the protocol of FTG and USF is to digitize systematically, by geography and taxonomy. Once the SEUS specimens of FTG and USF are digitized, any new specimens from the SEUS will be immediately digitized to maintain organization in the herbarium beyond the life of this 3-year project. USF organizes its herbarium into Florida specimen folders and SEUS specimen folders. For example, once all of USF’s Asteraceae from the SEUS are digitized, any newly accessioned SEUS Asteraceae are automatically digitized before filing. Similarly, FTG is organized by Florida folders and USA folders (very few specimens outside of Florida). Once a taxon-geographic folder unit is digitized, the protocol is to keep that folder unit digitized indefinitely.

Both FTG and USF will continually train multiple new volunteers to digitize in the herbarium, as has been done since digitization efforts first began in these institutions. This training will continue, regardless of available funding, to maintain digitization initiatives to seek digitization of the the complete herbarium specimen holdings, inherent goals of both herbaria.

Data access, sharing, and usage are central to the mission of each herbarium. FTG and USF continually seek to expand their purpose and broaden their utility to increase their valuation in the global community. FTG and USF will maintain a close partnership with the digitization community to sustain its data sharing with iDigBio, SERNEC, and GBIF to continue to be impactful beyond the life of this project. Networks with the Association of Southeastern Biologists, Society of Herbarium Curators, SERNEC, and iDigBio will enable us to develop collaborative long-term solutions for maintenance, access, and continued digitization of our herbarium data.

FTG and USF are tightly integrated with local community environmental organizations, such as native plant societies, water management districts, and county, state, and federal park districts. The majority of the feedback from these communities refer to the utility of digitized specimens. Digitization will remain as core objectives to both institutions, as we both service to the local community. These organizations will be sought for continued funding to maintain online transcription tools, imaging, and data management, in order to maintain public access to our increasingly digitized collections.