LTER Project Update

Andrew Bibian

March 15, 2016

```
#-------------------------------------------------------------------------------
# The following block of code was used to generate the statistics reported in the summary
# below.
#-------------------------------------------------------------------------------
# This is part of the working directory for the project
setwd("C:/Users/MillerLab/Box Sync/LTER/Site Specific Data List")

# Webscrape is a dataframe generated from the LTER Data Portal website.
# It contains information about the total potential data entities available
# through the Data Portal as of the listed date
# (source:https://portal.lternet.edu/nis/home.jsp)
webscrape <- read.csv(
  paste0("../R Scripts for Automated Synthesis of LTER Data",
         " Portal/List_Of_Studies_On_Data_Portal_Mon_Aug_31_104450_2015.csv"))

data_available <- sum(webscrape$data_obj_count)

# Manual is a dataframe containing informatoin that
# was gathered from individual LTER web portals. Using their available
# metadata, detailed information about data entities was used to collect and
# categorize and identity relevant data.
manual <- read.csv(
  "All_Expanded.csv")

#Number of sites reviewed
reviewed <- length(unique(manual$lter))

# Quality is a qualitative factor assigned to identified datasets. A proctcol
# exist that defines the categorical cutoffs.
quality <- c('low', 'high')

# Data that were identified to have suffient resolution are contain
# population dynacmis data
popdata <- sum(na.omit(manual$pop_data[manual$quality %in% quality]))

# Number of data entities that have been summarized, categorized
# and grouped
data_summarized<- nrow(manual)

# Number of potential data entities that are remain to be
# characterized
data_remaining<- sum(webscrape$data_obj_count)-nrow(manual)

percent_complete<- round(data_summarized/data_available,2)*100

# Estimate of potential datasets we will encounter through the end of the project
potentialpop<- round((popdata/data_summarized)*data_available)
```

Summary

The first year of the data synthesis phase associated with the LTER EARGER grant (P.I. Tom Miller) is currently nearly a third of the way finished. Using the information that is available through the LTER data portal, we have identified 9621 potential datasets available to scientist (as of August 31, 2015). These data entities are reported from across all 26 LTER locations.

Through a systematic reivew of the data catalogs at 7 LTERs (SBC, SEV, SGS, VCR, AND, NWT, BNZ), we have cataloged 2923 (30% of available) datasets. Of these catalogued data, 191, have been identified as containing information with high enough spatial and temporal resolution for infering population dyanimcs of various organisms. There currently remains about 6698 data entities that must be identified and catalogued.

We have developed methodoly to standardize and format data by separating the components of a study into units of temporal and spatial replication. We have begun to populate a PostgreSQL database that will serve as a central repository for the scientific community. Currently we estimate that the database will contain 629 independent datasets relevent to population ecologist.

Although we have developed the protocols, methods, and tools to fulfill the data synethsis phase of the grant, more time will be needed to extend the useability and accessability of repository we are creating. These additional steps for which funding is required will involve the development of R packages and web applications to help others in access and using the resource we begun to create.