**Habitat Selection by Grassland Birds, Their Habitat Requirements and Informing Grassland Bird Conservation: A Research Data Management Case**

Compiled by Regina Raboin, Tufts University

Edited by Andrew Creamer and Nancy LaPelle, University of Massachusetts Medical School

**Summary of Teaching Points**

**Module 2: Types, Formats & Stages of Data**

* Has raw, analyzed, and modeling data stored in proprietary software using file naming conventions that were not clear
* Software stored data in its own application database that needed to be exported to a common format for analysis, storage, etc.
* The need to update proprietary software as new releases become available to maintain support from vendor and keep data formats current
* Data in digital and paper format

**Module 3: Contextual Details**

* No use of validity checks on data entry
* No data dictionary
* No metadata standards

**Module 4: Data Storage, Backup and Security**

* No plan for syncing data stored on university servers and external drive

**Module 5: Legal and Ethical Issues**

* Use of data sets from the Internet

**Module 6: Data Sharing and Re-Use**

* Researcher wants to be informed of reuse
* Researcher wants to share with fellow graduate students

**Module 7: Plan for Archiving and Preservation of Data**

* Use of an online archive
* Need for preservation-friendly formats
* Need for documentation to contextualize data
* Need to investigate minimal core metadata set for uploading to archive

A PhD student in biology has received NSF funding to conduct research on grassland bird habitat quality. He would like to create a tool that could potentially be used to assess the suitability of habitat management sites, and inform decisions on which sites to attempt to improve, and what parts of the sites would benefit most from management. It could also potentially be used to assess when the benefits of management for grassland birds would be relatively minor (since managing habitat for grassland birds can be the opposite of what other species of concern may need).

The student is working with his thesis advisor (in this advisor’s lab), and the lab has four field assistants. The lab is collaborating with two additional researchers and their teams at universities in another state. For his research project, the student uses a rectangular sampling tool called a quadrat, which is a grid that separates an area into 20 symmetrical plots. He records data based on the measurements and observations of organisms and habitat within these plots: the types of vegetation in the field, the vegetation height, leaf litter depth, among others. He also records data on the flight initiation distance, which is the distance at which an approaching stimulus (predators, humans, etc.) will cause grassland birds to move away.

The student also collects data related to “openness” (measures of a landscape’s boundaries, which he contextualizes with observations of the habitat’s relationships within this larger landscape, and how energy flows in and out, etc.). He uses a portable GPS to record coordinates for location, angles to the horizon, distances to the habitat’s edges, and compass bearings, etc.). He also creates simulations using these data to create data models, which generate random data points that he can compare with the probability of obtaining the observed data relative to random chance.

The student has already collected several hundred GBs of data. Early in his research, he recorded data using hard copy data sheets that he then had to enter into a computer. One of the field assistants would check over the data sheets to ensure the accuracy of his data entry. Now the student enters his data directly using a portable computer (Trimble Juno GPS). He enters observational and GIS data into MS Excel™ spreadsheets and a proprietary mapping software program, ArcGIS™. Data were examined in MS Excel™ or ArcGIS™, and analyzed using SAS™ software. In addition to the data he collects from the field, he has downloaded a large amount of existing GIS datasets from the Internet, and processed these files using ArcGIS™ and Python™ software, and analyzed them using Python™ and SAS™ software. He uses .docs files generally for data descriptions and write-ups; his .xls and .xlsx files contain data tables and spreadsheets; his .shp files contain geographic information with associated data tables; his .csv files contain the same data as the .xls, .xlsx, and .shp files, but he finds that it is a format easier to read with R™ or Python™. His .py, .sas, and .r files contain scripts that processed and analyzed other data or computed statistics.

The student analyzes his data in different ways (e.g., bird surveys are entered into a distance software program, corrected for sampling bias, and then used as a dependent variable in the analyses of patch area, patch openness, and edges). He stores his hard copy data sheets in binders, with copies of these binders in his PI’s lab. Electronic files are stored on his lab’s space on the university’s servers, which the university secures and backs up. He also has a backup copy in an external hard drive that he regularly backs up.

The student does not have a formal protocol for metadata, but he has created “guide” files (with file names leading with 0s so they are at the top of the directory). They have some documentation and description of the files in each of his folders. Unfortunately he has not made a guide file for all of his files and folders, and he does not always keep them up to date.

The student recognizes that he is benefitting from the availability of public data, which he is repurposing for his own research, and he would also like to share his data. There are no formal restrictions placed on his data for reuse, but he feels that he would like to be informed if someone else wanted to reuse his data, which he feels is the unspoken rule in his discipline. He wants to be contacted because he feels he would provide necessary context and input on any future uses of his data. He has already been contacted by another graduate student who feels that some of his datasets recorded from habitats in Hawaii would aid her research. He also feels some of his data related to climate change may be useful to other students interested in global climate change, such as his PET data (Potential Evapotranspiration), which is a measurement of how much water is lost through evaporation and also from transpiration (given off by plants as part of photosynthesis).

After he finishes his project, the student intends to enter some of his relevant field data into ebird.org, an online archive of ornithological sightings maintained by Cornell University.

1. Types of data
   1. What types of data are being collected for this study?
   2. What would be needed in a data management plan to describe use of proprietary equipment or software?
   3. What analytical methods and mechanisms will be applied to the researcher’s data either prior to or post integration?
   4. What type of outcome data will be generated?
2. Contextual details
   1. What file formats and naming conventions will be used for the separate data sources and for the integrated file used for analysis?
   2. What impact would the naming conventions have on later data access?
   3. What other contextual details would the researcher specifically need to document to make his data meaningful to others?
   4. In what form will the researcher capture these details?
3. Data Storage, Backup, Security

a. Where and on what media will the data from each data source be stored?

b. How, how often and where will the data from each source be backed up?

c. How long following the completion of his study will he need to store the data?

d. What impact would the proprietary software, and software updates have on storage and access?

1. Data protection/privacy

a. Who will own any copyright or intellectual property rights to the data from each source?

b. How will the datasets be licensed if rights exist?

1. Policies for reuse of data
   1. Will the data be restricted to be re-used only for certain purposes or by specific researchers?
   2. Are there any reasons not to share or re-use data?
2. Policies for access and sharing
   1. Will some kind of contribution or fee be charged for subsequent access to this data?
   2. What process should be followed to gain future access to the researcher’s study data?
3. Archiving and preservation

a. What is the long-term strategy for maintaining, curating and archiving the data?

b. What data will be included in an archive?

c. Where and how will it be archived?

d. What other contextual data or other related data will be included in the archive?

e. How long will the data be kept beyond the life of the project?