Basic Recommendations for Preparing Data for Archive

* Data should be in a csv text file
  + If starting with an Excel spreadsheet, please make sure it does not contain any formulas and comments on cells. If you need comments put them in their own column.
  + If data were used in a database and major table linking is necessary to analyze, please de-normalize into a [flat file](https://en.wikipedia.org/wiki/Flat-file_database), not just database table exports.
  + When naming your csv file, use descriptive names and avoid spaces (e.g. bad naming convention “file 1.csv”; good naming convention “PepperwoodBreedingBirdSurvey\_2007to2019.csv”). Pepperwood would like the following naming convention format for csv files; “Pepperwood\_ProjectName\_CollectionYear(s).csv”
* Consider your data table format. Should your table be wide or long?
  + In a “wide”, or matrix-style data model each variable has a separate column.
  + In a “long”, or attribute-value data model there is one column containing the names of the variables and another column for the variables’ values. More information can be found [here](https://environmentaldatainitiative.org/five-phases-of-data-publishing/phase-1/long-or-wide/#Recommendations)
* Clean your data. EDI has resources describing best practices [here](https://environmentaldatainitiative.org/five-phases-of-data-publishing/phase-2/) and a presentation with examples [here](https://github.com/EDIorg/tutorials/blob/master/data_cleaning/How_to_clean_and_format_data.pdf)
  + Pepperwood would like your data to use the following date, time or date-time formats
    - YYYY-MM-DD
    - hh:mm:ss (24-hour time)
    - YYYY-MM-DD hh:mm:ss
  + Use consistent naming conventions. For example, LikeThis, like.this, or like\_this. Choose one and stick with it throughout the dataset
  + Any blank cells in your data should be filled with “NA” (also acceptable “NaN”, or “-999”). Specify a definition for each missing value code in document below
* Organize your data into appropriate packages for archiving. Considerations for organization can be found [here](https://environmentaldatainitiative.org/five-phases-of-data-publishing/phase-1/#Scientific%20domain)
  + Note: a data package can include raw data (no formulas, etc.), Spatial data (often includes additional metadata for sites, plots, etc.), Photographs (pdf file of images), Maps (pdf), rcode, etc.
  + [Best Practices for Data Package Design](https://ediorg.github.io/data-package-best-practices/datapackage-design/index.html#entities-to-include-in-a-data-package) - Helpful resource made by EDI