Chapter 6
Preserve: Protecting Data for Long-Term Use

Robert B. Cook, Yaxing Wei, Leslie A. Hook, Suresh K.S. Vannan, and John J. McNelis

Abstract This chapter provides guidance on fundamental data management practices that investigators should perform during the course of data collection to improve both the preservation and usability of their data sets over the long term. Topics covered include fundamental best practices on how to choose the best format for your data, how to better structure data within files, how to define parameters and units, and how to develop data documentation so that others can find, understand, and use your data easily. We also showcase advanced best practices on how to properly specify spatial and temporal characteristics of your data in standard ways so your data are ready and easy to visualize in both 2-D and 3-D viewers. By following this guidance, data will be less prone to error, more efficiently structured for analysis, and more readily understandable for any future questions that the data products might help address.

6.1 Introduction

Preservation certainly encompasses the idea that there should be no loss of bits associated with a data product. In this chapter, we will expand this definition of preservation, to include all of the data management practices that will preserve the data at a high-enough level of quality so that it is usable well into the future. Well-curated and -preserved data will be easily discovered and accessed, understood by future users, and serve to enable others to reproduce the results of the original study. Preservation, in this broad sense, starts when the seed-ideas for a project are first pulled together, and continues until the data have been successfully finalized, curated, archived, and released for others to use (Whitlock 2011).

Proper preservation of the data files is an important part of a research project, as important as the sample design, collection, and analysis protocols in ensuring the overall success of a project. Often researchers do not spend enough effort ensuring that the data are properly managed, described, and preserved. Without well-

R.B. Cook • Y. Wei • L.A. Hook • S.K.S. Vannan (✉) • J.J. McNelis
Oak Ridge National Laboratory, Oak Ridge, TN, USA
e-mail: rbcook7@gmail.com; weiy@ornl.gov; hookla@ornl.gov; santhanavans@ornl.gov; mcnelisjj@ornl.gov

F. Recknagel, W.K. Michener (eds.), *Ecological Informatics*,
DOI 10.1007/978-3-319-59928-1_6

prepared data—no matter how carefully the sample design, collection, and analysis were done for a project—the research team may not be able to effectively use the data to test their hypotheses. And the data will not be useful for any potential future users.

Well-preserved ecological observations will continue to help us understand the functioning of the global ecosystem. More importantly, the data of ecological observations provide the foundation for advancing and sustaining economic, environmental, and social well being (Reid et al. 2010; IGBP 2012; USGEO 2015). Thus, well-preserved ecological data are critically needed to address global sustainability—what could certainly be considered the grand scientific challenge of the twenty-first century (Reid et al. 2010; IGBP 2012).

6.1.1 Preservation and Its Benefits

We will define preservation as preparing data packages—data, documentation, and metadata—for a user 20 years into the future (NRC 1991); some advocate even 100 years (Justice et al. 1995). The rationale is that those who generated the data initially or those who worked with the data when the data were first compiled will have forgotten the details of the data within a few years (Michener et al. 1997) (Fig. 5.1). Developing descriptive information for someone 20 or more years out who is unfamiliar with the project, methods, and observations will ensure that the information is preserved and, most importantly, usable (Fig. 6.1) (NRC 1991).

Well-managed and preserved data have many benefits. During the course of a project, investigators who make a habit of preparing organized and well-described data will spend less time doing data management and more time doing research.

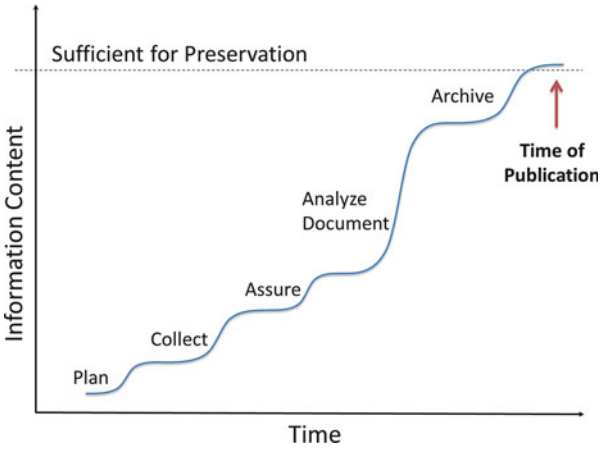

Fig. 6.1 With proper data management and preservation during the course of a project, information about the data is compiled during the data life cycle (plan, collect, assure, analyze, document, and archive; Strasser et al. 2012). Metadata and documentation are recorded so that future users will be able to find and use the data products

Researchers can pick up data files after being away from them for a period and immediately use the data without having to remember what the data means or how filters or analyses were done. Furthermore, researchers can hand off data and documentation to collaborators who can readily understand and use data files, without further explanation.

When the project has been completed and the data are finalized and properly curated, scientists outside your project can find, understand, and use your data to reproduce the findings of your research. Perhaps even more importantly, these data products can be used to address additional broader-scale research questions (Reid et al. 2010; Whitlock 2011; Michener 2017d; Schildhauer 2017). FLUXNET is an example of a project that started out studying the scientific mysteries of individual flux tower sites, but evolved to address larger scale questions across biomes and climate domains. Along with this scientific evolution, FLUXNET has experienced a data evolution in which the community has embraced standard methods for observations and processing, and has come to appreciate the advantages of placing data into common formats, with standard units and parameter names. This standardization facilitates combining data from 10s to 100s of flux towers to address broad questions that cannot be addressed by individual projects (Baldocchi et al. 2012; Papale et al. 2012). A common set of standards ultimately saves time, but requires buy-in, which takes time for investigators to realize the benefits.

Funding agencies protect their investment in Earth science research, through preservation of observations; many funding agencies require that data generated through their grants be shared over the long term (Whitlock 2011). The preserved observations provide the means to understand Earth processes, develop and test models, and provide information for decision makers. Not preserving data products so that they can effectively be used will decrease the return on research investment, and more importantly hinder our ability to advance Earth science.

Some journals (e.g., PNAS, Ecological Monographs), scientific societies (e.g., Ecological Society of America) now require that the data used in a paper be archived before the paper can be published, and others require that the data be shared (PLoS, Nature, Science; Michener 2015). In both cases, data citations with Digital Object Identifier (DOI) locators will allow readers to find the archived data (Cook et al. 2016). Following data management practices for long-term preservation will make it easier for authors to archive their data products associated with a submitted manuscript to meet this requirement.

Another benefit of data preservation is that others will use these well-curated data, resulting in the data producers getting credit. Data repositories have started to provide data product citations, each with a DOI (Parsons et al. 2010; Cook et al. 2016). A benefit of data preservation is that through data product citations (Cook et al. 2009, 2016), data authors get credit for archived data products and their use in other papers, in a manner analogous to article citations. In addition, readers of those articles can obtain the data used in an article (Cook et al. 2016) through the DOI locator.

6.2 Practices for Preserving Ecological Data

This chapter is written for a broad audience—for those who are creating data products, for those who may need to prepare the data products for archival, and for those who will access and use the archived data. Accordingly, we will present preservation activities that data contributors, data archives, and data users can perform to preserve data products and make them useful in the future. The focus will be on application of preservation principles, and less so with theoretical/academic aspects of preservation. We are orienting this chapter toward practical aspects, because ecologists may be willing to share their data, but they typically do not have knowledge and training of data management practices that they can use to facilitate sharing (Tenopir et al. 2011; Kervin et al. 2014).

Geospatial, or location, information is a fundamental component of ecological data. The preservation practices described here are primarily for geospatial data products, including tabular data as well as map and image data. Earlier best practices for data sharing focused almost exclusively on tabular data (Olson and McCord 2000; Cook et al. 2001), but the focus has expanded with improvements in sensors, standards, and processing software, and many ecologists are turning to geospatial data.

This chapter builds on the chapter on documentation and metadata (Michener 2017c). Because the metadata descriptors were thoroughly treated there, we will focus on human readable text documents that provide another view into the data. These text documents contain the contextual information about samples—under what conditions was the sample collected, what antecedent conditions influenced the sample, and what do others need to know about the sample context in order to understand the data.

The remainder of Sect 6.2 describes best data management practices that investigators can perform to improve the preservation and usability of their data for themselves and for future users.

6.2.1 Define the Contents of Your Data Files

The data compiled during a project is derived from the science plan (hypotheses/proposal) for that project. During the proposal writing stage, the investigator should identify the information needed to address the hypotheses and the best way to compile that information. Sometimes that compilation will be to collect samples and make measurements, other times it may be to run models to obtain output, or even fuse data from multiple sources to create a necessary product.

Also during the proposal writing stage, a Data Management Plan (DMP) (Michener 2017a) should be developed that lays out the content and organization of the data based on a comprehensive list of data required for the project. The environmental study will compile a suite of primary measurements along with

contextual and ancillary information that defines the study area (soil, landcover, plant functional types, weather, nutrient status, etc.).

Investigators should keep a set of similar measurements together in one *data file*. The similarity extends to the same investigator, site, methods, instrument, and time basis (all data from a given year, site, and instrument in one file). Data from a continental study of soil respiration at 200 plots could be one data file, but 30-min meteorological data from 30 sites over 5 years could be five data files (one per year) or 30 data files (one per site). We do not have any hard and fast rules about contents of each file, but we suggest that if the documentation/metadata for data are the same, then the data products should all be part of one data set.

6.2.2 Define the Parameters

Defining the name, units, and format used for each parameter within a project should be done with a clear view to the standards or guidelines of the broader community. Using widely accepted names, units, and formats will enable other researchers to understand and use the data. Ideally, the files, parameter names, and units should be based on standards established with interoperability in mind (Schildhauer 2017).

The SI (International System) should be used for units and ISO be used for formats. The ISO Standard 8601 for dates and time (ISO 2016) recommends the following format for dates:

yyyy-mm-dd or yyyymmdd, e.g., January 2, 2015 is 2015-01-02 or 20150102

which sorts conveniently in chronological order. ISO also recommends that time be reported in 24-h notation (15:30 hours instead of 3:30 p.m. and 04:30 instead of 4:30 a.m.).

In observational records, report in both local time and Coordinated Universal Time (UTC). Avoid the use of daylight savings time because in the spring the instrument record loses 1 h (has a gap of 1 h) and in the autumn, instrument records have a duplicate hour.

The components needed to define temporal information with sufficient accuracy for ecological data include the following: calendar used, overall start and end temporal representation of a data parameter, time point/period that each data value represents, and temporal frequency of a data parameter. As an important example, Daymet (Thornton et al. 2017), a 1-km spatially gridded daily weather data set for North America, uses the standard, or Gregorian, calendar and leap years are considered. But the years within Daymet always contain 365 days; Daymet does this by dropping December 31 from leap years. The documentation for Daymet defines this information (e.g., start and end times of each time step, which days are included and which days are not). Following the Climate and Forecast (CF) Metadata convention (Eaton et al. 2011) and the ISO 8601 Standard (ISO 2016), temporal information of Daymet is accurately defined.

CF Metadata, a convention for netCDF-formatted files, is becoming more common in ecological modeling and in some field studies. These conventions allow combination and ready analysis of data files, and importantly, facilitate the use of field data to parameterize and drive models with a minimum of conversions.

In addition to enabling integration of data files, standard units can be easily converted from one unit to another using a tool such as UDUNITS library (UCAR 2016).

For each data file, investigators should prepare a table that identifies the parameter, provides a detailed description of that parameter, and gives the units and formats (Table 6.1).

Table 6.1 Portion of a table describing contents and units (dos-Santos and Keller 2016)

Column heading	Units/format	Description
Site		Fazenda Cauaxi or Fazenda Nova Neonita. Both located in the Municipality of Paragominas
Area		Code names given to the site areas. The areas are PAR_A01 for the Fazenda Nova Neonita or CAU_A01 for the Fazenda Cauaxi
Transect		The transect ID number within an area. Transect = plot.
tree_number		Tree number assigned to each tree in each transect
date_measured	yyyy-mm-dd	Date of measurements
UTM_easting	m	X coordinate of tree individual location. Fazenda Cauaxi is in UTM Zone: 22S. Fazenda Nova Neonita is in UTM Zone: 23S
UTM_northing	m	Y coordinate of tree individual location. Fazenda Cauaxi is in UTM Zone: 22S. Fazenda Nova Neonita is in UTM Zone: 23S
common_name		Common name of tree. MORTA = dead tree
scientific_name		Scientific name of tree. NI = not identified. For common_name = MORTA (dead) or LIANA, scientific names are not provided.
DBH	cm	Diameter at breast height (DBH), 1.3 m above the ground. Measured on both live and standing dead trees.
height_total	m	Total Height (m), measured using a clinometer and tape as the height to the highest point of the tree crown. Measured on both alive and standing dead trees. Fazenda Cauaxi site 2012 only—not measured in 2014.

Table 6.2 Characteristics of sites from the Scholes (2005) study

Site name	Site code	Latitude	Longitude	Elevation	Date
Units		(deg)	(deg)	(m)	
Kataba (Mongu)	K	−15.43892	23.25298	1195	2000-02-21
Pandamatenga	P	−18.65651	25.49955	1138	2000-03-07
Skukuza Flux Tower	skukuza	−31.49688	25.01973	365	2000-06-15

Provide another table that describes each study site or area used in the data product [location, elevation, characteristics (climate or vegetation cover)], along with a formal site name (Table 6.2).

Once the metadata about a record is defined, be sure to use those definitions, abbreviations, units consistently throughout the data set and the project. For air temperature, pick one abbreviation and use it consistently. Do not use T, temp., MAT (mean annual temp), and MDT (mean daily temp) within a data set, if they all mean the same parameter; using one consistently will be much easier for users to understand, particularly as they write code to process the values.

When data values are not present in the data file, investigators should indicate this with a *missing value code*. We suggest that an extreme value never observed (e.g., −9999) be used consistently to indicate that the value is missing.

6.2.3 Use Consistent Data Organization

There are several different ways to organize data files. For *tabular data*, one way is similar to a spreadsheet table in which each row in a file represents a complete record, and the columns represent the parameters that make up the record. The table should have a minimum of two header rows, the first of which identifies the parameter names and the second header row identifies the parameter units and format (Table 6.3) (Cook et al. 2001). A suggestion for data files is that a column containing a unique id for each record be included for provenance tracking.

Another perfectly appropriate alternative for *tabular data* is to use the structure found in relational databases. In this arrangement, site, date, parameter name, value, and units are placed in individual rows; unique ids could also be placed in this row. This table is typically skinny (only 5 or 6 columns wide) and long, holding as many records (rows) as needed in the study (Table 6.4). This arrangement allows new parameters to be added to a project in the future without changing the tabular data columns.

For whichever organization chosen, be consistent in file organization and formatting throughout the entire file (Porter 2017). The file should have a separate set of header rows that describes the content of the file. For example, the first row of the file should contain file name, data set title, author, date, and any related companion file names (Table 6.5) (Hook et al. 2010). Within the body of the file,

Table 6.3 An arrangement of content in which all of the information for a particular site and date (e.g., site, date, parameter name, value and unit) is placed into one row

Station	Date	Temp.	Precip.
Units	YYYYMMDD	C	mm
HOGI	20121001	12	0
HOGI	20121002	14	3
HOGI	20121003	19	−9999

Table 6.4 An arrangement of information in which each row in a file represents a complete record, and the columns represent the parameters that make up the record

Station	Date	Parameter	Value	Unit
HOGI	20121001	Temp.	12	C
HOGI	20121002	Temp.	14	C
HOGI	20121001	Precip.	0	mm
HOGI	20121002	Precip.	3	mm

do not change or re-arrange the columns or add any notes in marginal cells. Additional features provided by specific software, such as colored highlighting or special fonts (bold, italicized, etc.) that indicate characteristics to humans are not useful for computers, and any information contained in the colors or fonts will not be preserved.

Spatial data files containing vector data, such as ESRI's Shapefile format, treat each point, line, or polygon as a unique record described by a set of common attributes. Records within a shapefile are organized in tabular format where each row corresponds to a feature representing the location or area to which the row's attributes pertain. Tabular data stored inside ESRI shapefiles are limited by character count and cannot contain special characters so it is good practice to maintain a complementary data dictionary file that defines the parameters, abbreviations, and units.

6.2.4 Use Stable File Formats

A key aspect of preservation is to ensure that computers can read the data file well into the future. Experience has shown that proprietary and non-standard formats often become obsolete and difficult or even impossible to read. Operating systems, the proprietary software, and the file formats will no longer be supported and researchers are left with useless bits.

Over the short term, usually during the course of the research, it is fine to use familiar proprietary data formats. But be sure that those formats can be exported into an appropriate format (without loss of information) suitable for long-term preservation.

Standardized, self-describing, and open data formats are recommended for long-term preservation of ecological data (Table 6.6). Standardized formats increase interoperability of data and lower the barrier of integrating heterogeneous data (Schildhauer 2017). Self-describing formats make data easier to use by a wide range of users. More importantly, open formats ensure consistent support and improvement from user communities and increase longevity of ecological data. Standardized and open formats also serve as a solid basis for developing data access, subsetting, and visualization tools.

Table 6.5 An example of a well-organized portion of a file with a set of header rows that describe the file (file name, contributor, citation, date, and any relevant notes)

File name	NGEE_Arctic_Barrow_Soil_Incubations_2012							
Date modified:	2015-10-27							
Contact:	Colleen Iversen (iversencm@ornl.gov)							
Data set DOI	doi:10.5440/1185213							
Notes	For more information, see data set DOI							
Region	Locale	Latitude	Longitude	Date_sampled	Thaw_Depth	Soil_Horizon	Carbon_concentration_of_soil_layer	Nitrogen_concentration_of_soil_layer
		Decimal_degrees	Decimal_degrees	yyyy-mm-dd	cm		Percent	Percent
North Slope	Barrow	71.280867	−156.611606	2012-08-01	36.8	O	42.94	2.66
North Slope	Barrow	71.280867	−156.611606	2012-08-01	36.8	O	42.94	2.66
North Slope	Barrow	71.280867	−156.611606	2012-08-01	36.8	O	42.94	2.66
North Slope	Barrow	71.280867	−156.611606	2012-08-01	36.8	Mi	39.36	2.24
North Slope	Barrow	71.280867	−156.611606	2012-08-01	36.8	Mi	39.36	2.24
North Slope	Barrow	71.280867	−156.611606	2012-08-01	36.8	Mi	39.36	2.24
North Slope	Barrow	71.280867	−156.611606	2012-08-01	36.8	DO	31.6	1.72
North Slope	Barrow	71.280867	−156.611606	2012-08-01	36.8	DO	31.6	1.72
North Slope	Barrow	71.280867	−156.611606	2012-08-01	36.8	DO	31.6	1.72
North Slope	Barrow	71.280891	−156.61164	2012-08-01	39.4	O	34.42	1.76
North Slope	Barrow	71.280891	−156.61164	2012-08-01	39.4	O	34.42	1.76
North Slope	Barrow	71.280891	−156.61164	2012-08-01	39.4	O	34.42	1.76

The body of the file is consistently organized and completed (Iversen et al. 2015)

Table 6.6 Recommended formats for ecological data preservation (ESO 2016; Edinburgh Data Share 2015)

Format	Description
Text/CSV	Suitable for representing tabular data such as field observations and site characteristics.
Shapefile	Most widely used open format for representing vector data, such as points, lines, and polygons.
GeoTIFF	Open and popular format for storing geospatial raster imageries.
HDF/ HDF-EOS	A feature-rich format suitable for storing complex multi-dimensional and multi-parameter scientific data. The HDF format and its EOS extension (HDF-EOS) have been widely used for NASA earth observation mission data for many years.
netCDF	Similar to HDF but simpler; ideal for storing multi-dimensional and multi-parameter data. Combined with Climate & Forecast (CF) convention, netCDF data files can be standardized and self-describing, which can greatly advance data interoperability. netCDF is gaining popularity in many research communities.

6.2.5 *Specify Spatial Information*

Almost all ecological data are location-relevant and many also have an associated time component. For example, photos taken of field sites should be associated with the accurate location, elevation, direction, and time information; otherwise they will not be suitable for research. There are many other spatial and temporal data types, for example, soil respiration observations across the world, MODIS Leaf Area Index (LAI) maps, and global 0.5-degree monthly Net Ecosystem Exchange (NEE) simulations generated from terrestrial biosphere models. When preparing ecological data for use or long-term preservation, their spatial ("where") and temporal ("when") information need to be accurately defined.

Two critical components of spatial information include the Spatial Reference System (SRS) used and the spatial extent, boundary, resolution, and scale under the given SRS. For example, Daymet v3 (Thornton et al. 2017) provides daily weather parameters at 1-km spatial resolution for North America from 1980 to 2016. It uses a special SRS called Lambert Conformal Conic and its definition using the Open Geospatial Consortium (OGC) Well-Known Text (WKT) standard is shown in Table 6.7.

Under this SRS, X and Y coordinates of each of the 1-km grid cells are accurately defined following the CF convention in the netCDF files where Daymet data are stored.

6.2.6 *Assign Descriptive File Names*

Even desktop personal computers can have large hard drives, and it can be very easy to lose files and information on such large drives. To prevent time spent

Table 6.7 Example Spatial Reference System, showing the projection, spatial extent, boundary, resolution and scale

```
PROJCS["North_America_Lambert_Conformal_Conic",
    GEOGCS["GCS_North_American_1983",
        DATUM["North_American_Datum_1983",
            SPHEROID["GRS_1980",6378137,298.257222101]],
        PRIMEM["Greenwich",0],
        UNIT["Degree",0.017453292519943295]],
    PROJECTION["Lambert_Conformal_Conic_2SP"],
    PARAMETER["False_Easting",0],
    PARAMETER["False_Northing",0],
    PARAMETER["Central_Meridian",-96],
    PARAMETER["Standard_Parallel_1",20],
    PARAMETER["Standard_Parallel_2",60],
    PARAMETER["Latitude_Of_Origin",40],
    UNIT["Meter",1],
    AUTHORITY["EPSG","102009"]]
```

searching for files, organize the information in a directory or folder structure based on project or activity. The directory structure and file names need to be both human- and machine-readable and so the names should contain text characters only and contain no blank spaces (Cook et al. 2001; Hook et al. 2010). Carefully check for any operating or database system limitations on characters (upper or lowercase, special characters, and file name lengths).

Use descriptive file names that are unique and reflect the contents of the files. In the metadata and documentation define the terms and acronyms in the file names. Examples of good file names include "daymet_v3_tmax_annavg_1988_na.nc4", a Daymet version 3 file containing daily maximum and annual average maximum temperature in 1988 for North America (na) in netCDF-4 format (Thornton et al. 2017).

Names should also be clear both to the user and to those with whom the files will be shared. File names like "Mydata.xls," "2001_data.csv," and "best version.txt" do not adequately describe the file and would not be useful to understand the contents.

While the name should be descriptive and unique, the file name is not the location for all of the metadata associated with a file. A standard metadata record in XML format is a much more useful location for detailed information about a data file, and will be accessible by APIs. See Michener (2017c) on metadata.

6.2.7 Document Processing Information

To preserve your data and its integrity, save your raw data in a "read-only" form (Strasser et al. 2012). By doing so, the raw data will not be affected by any changes, either purposeful or inadvertent. Some spreadsheet type software allows cells to be deleted inadvertently with the slip of a finger on a keyboard. Read only files will prevent those sorts of changes.

Use a scripted language such as "R", "SAS" or "MATLAB" to process data in a separate file, located in a separate directory (Hook et al. 2010; Strasser et al. 2012). The scripts you have written are an excellent record of data processing, can also easily and quickly be revised and rerun in the event of data loss or requests for edits, and have the added benefit of allowing a future worker to follow-up or reproduce your processing. The processing scripts serve as the basis for a provenance record. An example R script and some figures generated from the script are captured in Appendix of this chapter.

Scripts can be modified to improve or correct analyses, and then rerun against the raw data file. This approach can be especially beneficial when preparing manuscripts. Two or three months after the analyses have been run and written up, reviewers may want to have changes made (new filtering or statistical analysis, additional data, etc.). Scripts saved along with data files serve as a record of the analysis and can quickly be modified to meet the reviewer's need. If they were not saved, authors may have difficulty resurrecting the exact formula and perhaps even the data used in the analysis.

6.2.8 Perform Quality Assurance

Quality assurance pertains not only to the data values themselves, but also to the entire data package. All aspects of the data package need to be checked including parameter names, units, documentation, file integrity, and organization, as well as the validity and completeness of data values. One can think of quality assurance of a data set like the careful steps authors go through to finalize an accepted paper for publication.

There are a number of specific checks that researchers can perform to ensure the quality of data products (Cook et al. 2001; Hook et al. 2010; Michener 2017b). The organization within data files has to be consistent. Data should be delimited, lining up in the proper column (Cook et al. 2001). Key descriptors, like sample identifier, station, time, date, and geographic location, should not be missing. Parameter names should follow their definition, and the spelling and punctuation should not vary. Perform an alphabetical sort of the parameter names to identify discrepancies. Check the content of data values through statistical summaries or graphical

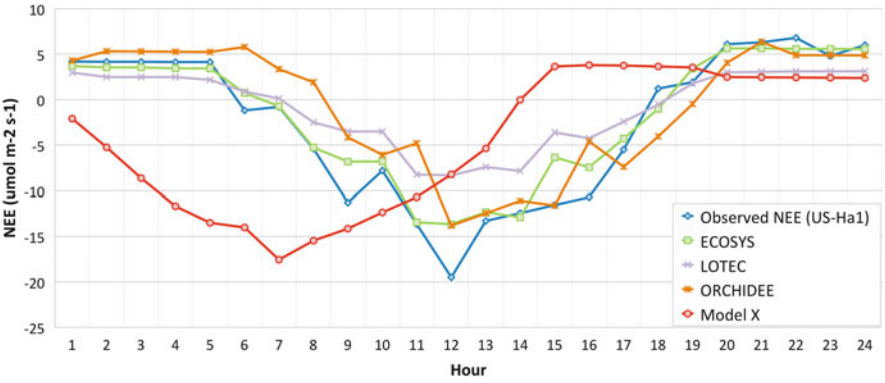

Fig. 6.2 Comparison of diurnal Net Ecosystem Exchange (NEE) for the Harvard Forest Flux Tower with terrestrial biosphere model output of NEE used to quickly identify quality issues. While most of the models are consistent with the timing and magnitude of noontime NEE, the onset and conclusion of the phytoperiod shows some variation among models, especially Model X, which was an outlier because of improper documentation. It was run with UTC time instead of Eastern US time but was not labeled carefully and was mistakenly plotted with a peak NEE 5 h earlier than the tower or other models (Ricciuto et al. 2013)

approaches to look for anomalous or out of range values. A number of different graphical approaches (leaf diagram, box and whisker diagram, histograms, scatterplots, etc.) are described in Michener (2017b). Another approach is to generate plots of time-series data to check for the physical reasonableness of the values and to ensure that the time zone is correct (Fig. 6.2). Plot the data on a map to make sure that the site locations are as expected (Cook et al. 2001). Common errors in spatial data are placing sites in the wrong hemisphere by not including the correct sign of latitude or longitude or providing the spatial accuracy required to place the site correctly on a shoreline, rather than mistakenly in a lake or coastal ocean (e.g., Fig. 6.3).

There is no better quality assurance than to use the data files in an analysis. Issues with the files, units, parameters, and other aspects of the data products will become evident and draw the attention of the analysts.

6.2.9 Provide Documentation

The documentation accompanying a data set should describe the data in sufficient detail to enable users to understand and reuse the data. The documentation should describe the goals of the project, why the data were collected, and the methods used for sample collection and analysis, and data reduction. The description should be detailed enough to allow future researchers to combine that data with other similar data across space, time, and other disciplines (Rüegg et al. 2014).

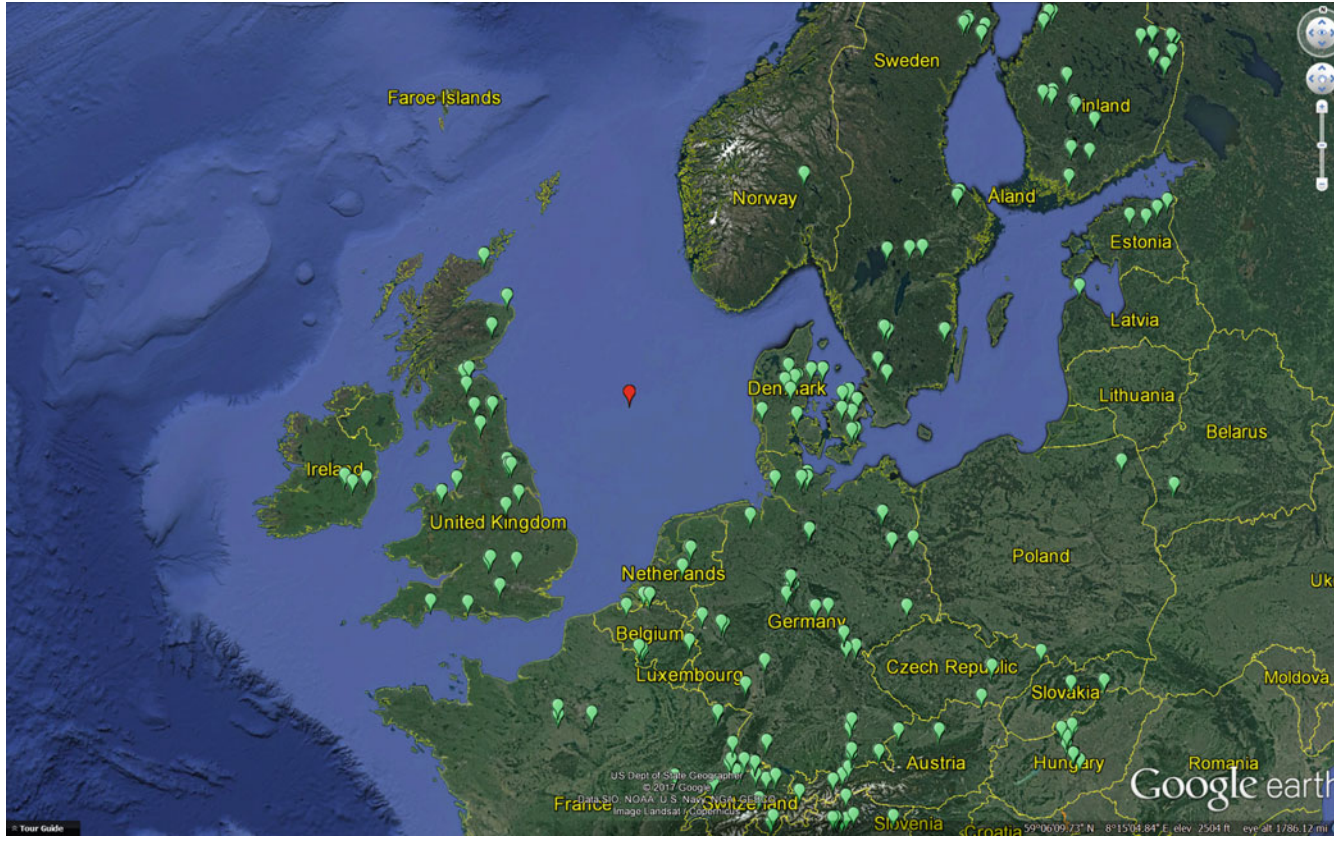

Fig. 6.3 Map showing locations of terrestrial sites where soil respiration was measured. The coordinates for the site in the middle of the North Sea need to be checked (Bond-Lamberty and Thomson 2014)

A data set document should contain the following information:

- **What** does the data set describe?
- **Why** was the data set created?
- **Who** produced the data set?
- **When** and how frequently were the data collected?
- **Where** were the data collected and with what spatial resolution?
- **How** was each parameter measured?
- **How** reliable are the data (e.g., what is the uncertainty and measurement precision and accuracy? what problems remain in the data set?)?
- **What** assumptions were used to create the data set (e.g., spatial and temporal representativeness)?
- **What** is the use and distribution policy of the data set?
- **How** can someone get a copy of the data set?
- **Provide** any references to use of data in publication(s)

Often a data set is a collection of multiple files. Each file should be described, including file names, temporal and spatial extent of the data, and parameters and units. If all of the files are the same, this information can be contained in the data set metadata and data set documentation. If each file is unique, in terms of contents, then each should be described separately with file-level metadata record and a file description document. The purpose for such a description is so that an investigator can use an automated method to search for an individual file or even part of the file that is required (e.g., XML metadata record or even self-describing file, like netCDF or HDF; Michener 2017c). If each file is not named in a descriptive manner or described in a document, then a user would have to manually view each file to obtain the required data, something that no one would want to do, especially for big data collections.

6.2.10 Protect Your Data

Everyone knows the sickening feeling when files are lost, due to hard drive crashes or from other problems. They have either experienced the feeling themselves or know someone who has lost their drives or files. A desktop, laptop, or server is fine one day and the next a problem has come up with the hard drive, and the files have disappeared. Backups are the key to surviving such losses. If you do not have backups, then you cannot retrieve the information and your files are not preserved.

Researchers—really anyone using computers—should create back-up copies often, to ensure that information is not lost. Ideally researchers should create three copies, the original, one on-site, and one off-site (Brunt 2010). The off-site storage prevents against hazards that may affect an institution such as fire, floods, earthquakes, and electrical surges. Data are valuable and need to be treated

accordingly, with appropriate risk mitigation. Cloud-based storage is becoming a valid option for storing and protecting data, especially as an off-site backup solution.

Frequency of the backups is based on need and risk. If you are compiling data from a high frequency sensor, then frequent (e.g., 30-min or hourly) backups are warranted to ensure that the information is not lost due to a disk crash. One can develop a backup strategy that relies on a combination of sub-daily, daily, weekly, and monthly backups, to cut back on the number of individual backups saved but still maintain sufficient backup so that no work is lost.

A critical aspect of any backup is that it be tested periodically, so that you know that you can recover from a data loss. After the initial shock of losing data, there is nothing worse than having the false hope of a backup that is not intact and is corrupted.

Another aspect of protecting your data deals with data transfers (e.g., over the Internet, such as large files, large numbers of files, or both). Ensure that file transfers are done without error by reconciling what was sent and received, using checksums and lists of files.

6.3 Prepare Your Data for Archival

The practices in Sect. 6.2 should have provided the background needed to prepare consistently structured, thoroughly defined, and well-documented data products. During the course of the project, the data should have been easy-to-share with team members, and readily analyzed to address the project's science questions.

At the end of the project, the data products need to be turned over to a data archive for curation and long-term storage (Fig. 6.4). Transitioning the data to an archive should have been part of the initial project planning conducted during the proposal writing stage, when a Data Management Plan (DMP) (Michener 2017a) was developed. The Plan should have identified the data center responsible for curating and archiving the data, and the investigators should have made initial contact with the archive before the proposal was submitted. The DMP should have included some information about the archive, their requirements, and a statement of collaboration by the archive. Because of space restrictions, the two-page DMP would not have included much detailed information. During the research project, the team should have interacted with data center personnel to inform them of the types and formats of data products being produced. Key characteristics that the data center needs to know are the volume and number of files, the delivery dates, and any special needs for the data (viewers or other tools, restricted access, etc.). Suggest a title that is concise in its description of the data set's scientific content and that indicates its spatial and temporal coverage. The data center will have requirements, and the project should identify what those are early in the project to ensure those requirements are incorporated before the data are submitted to the archive.

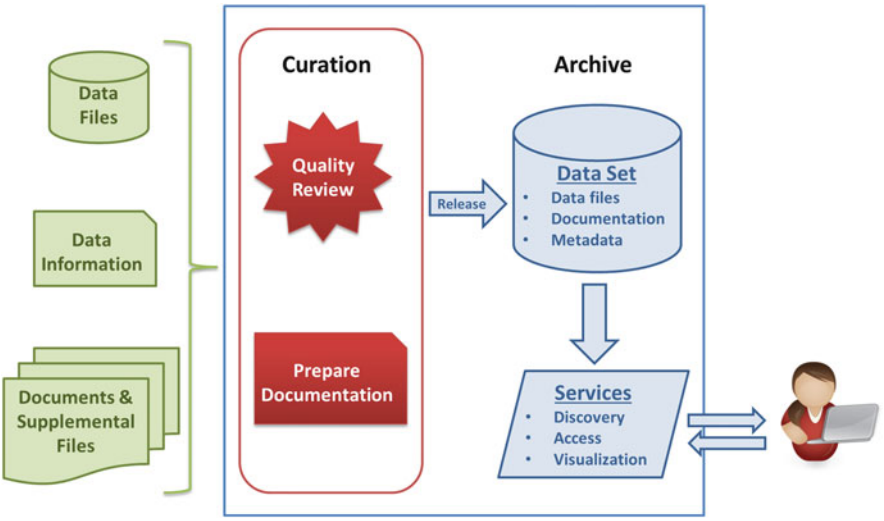

Fig. 6.4 Flow of data and documentation from the investigator team to the archive, where quality checks are performed and documentation is compiled. After the data are released, users can search for, download, and use data of interest

Data archives will need data files, documentation that describes the files and the content (Sect. 6.2.9), and, if possible, standardized metadata records (Michener 2017c). As part of the data package, some data centers require supplemental information such as sample design, sample collection and analysis methods, algorithms, code, and data analysis methods, description of field sites, photographs, articles using the data, etc. All of this information will provide context for those who are trying to understand and use the data, especially many years into the future.

6.4 What the Archive Does

After the data sets have been finalized, and the project team has transmitted them to the archive, the archive staff begins the process of curation leading to long-term preservation (Lavoie 2000). This section will briefly describe what typically happens to a data set during curation and the services that investigators receive when they archive a data set.

The archive is selected based on a number of factors. The agency that funded the research may have a designated archive, perhaps based on science area. In recent years, a principal investigator's institution, often the library, will provide long-term stewardship. Some journals and scientific societies have preferences for where data associated with published articles should be archived.

6.4.1 Quality Assurance

A data center goes through the following general steps during curation, summarized in the following list:

1. *Files received as sent*
   After the data have been received, the archivist will check the numbers of files and the "checksum" to ensure that the files were received as sent (see Sect. 6.2.10). At this time, staff will also make sure that the file type is appropriate for long-term storage and use (see Sect. 6.2.4).
2. *Documentation describes files*
   The archivist will read the documentation and any manuscript associated with the data product to get an understanding of why the data were produced and what the workflow is. If there are a number of unique files, a table will be generated that identifies the contents of each file or group of files. The archivist will check the filenames to ensure they are descriptive and appropriate based on the file content, date, spatial extent, etc. (see Sect. 6.2.6).
3. *Parameters and units defined*
   The documentation and the data files should provide the parameter definitions and the units. For tabular data, the data provider should have created a table that defines the column names and units; if not, the archivist could generate this useful table. Often the original investigator will be contacted to identify "mystery" parameters that are not identified, defined, or are unitless (see Sect. 6.2.2).
4. *File content is consistent*
   For *spatial data* files, the analyst may view the file or a sample of the files in a GIS tool for consistency. The datum, projection, resolution, and spatial extent will be exported from all files and checked for consistency (see Sect. 6.2.5).
   For *tabular data*, the archivist will ensure that the parameter definitions and units are consistent across all files (see Sect. 6.2.2).
5. *Parameter values are physically reasonable*
   The maximum and minimum value will be exported and the range checked for reasonableness (see Sect. 6.2.2).
   Geospatial tabular data will be loaded onto a basemap for visual inspection of proper overlay (see Sect. 6.2.8).
   Staff will check that missing values and other flags are reasonable and consistent (see Sect. 6.2.2). If a scale factor is applied, the archivist will make sure that it is defined.
6. *Reformat and reorganize data files if needed*
   The archivist will judge if the formats and organization of the received data files are the most appropriate based on their data stewardship expertise in the relevant research fields and the interactions with data providers. If needed, received data files will be reformatted and reorganized to ease the usage and maximize the future interoperability of data.

6.4.2 Documentation and Metadata

The archive will often generate two types of documentation. One is a metadata record in standardized format to describe the data and also to find data within a large archive (Michener 2017c). The second is a data set document (a readme type document) that provides a description of the data (what, where, when, why, who) and references to manuscripts using data (see Sect. 6.2.9).

Sometimes the investigator drafts a metadata record (Michener 2017c) using metadata-editing tools, but more often the data center will compile the metadata record.

Often the archivist will generate a data set document that defines all of the parameters and units, based on information or manuscripts provided by the investigators. Each file or type of file in the data set will be described and the spatial and temporal domain and resolution will be provided. The document should also describe the methods and limitations and estimates of quality or uncertainty. The data set document will also include browse images or figures that effectively illustrate the data set contents.

The investigator provides documents with contextual information (see Sect. 6.3) that is archived along with the data files and data set documentation.

A key part of data curation is to generate a data citation that gives credit to the data contributors and the archive, as well as provide a DOI that allows others to find and use the data. Data product citations have structures similar to manuscript citations and include authors, date released, data set title, data center, and DOI (ESIP 2014; Starr et al. 2015; Cook et al. 2016).

6.4.3 Release of a Data Set

After curating and archiving data, data centers can perform a number of services that benefit both the data users, the data providers, and the funders of the archive as well as funders of the research project. The following list contains a summary of archive activities after the data have been released:

1. Advertise data through email, social media, and website
2. Provide tools to explore, access, visualize, and extract data
3. Provide long-term, secure archiving (back-up and recovery)
4. Address user questions, and serve as a buffer between users and data contributors
5. Provide usage statistics and data citation statistics
6. Notify users when newer versions/updates of data products are available, particularly users who have downloaded the out-of-date data.

Data derived from research is advertised and made available through discovery and access tools. The data can be used to address other hypotheses and when those results are reported in a paper, the original data are cited, which can be used as a measure of the impact of that work and the data center on science.

6.5 Data Users

The key responsibility for the users of archived data is to give proper credit to the data contributors. Using other's ideas and research products, including data, requires proper attribution. Data used should be cited in a manner similar to articles, with callouts in the text, tables or figures, and a complete citation with DOI locator in the list of references. Compilation of all of the citations of a data set into a data citation index will ensure that the data authors are given credit for all of the effort associated with making the measurements and compiling a well-preserved data product.

A secondary responsibility of data users is to identify any issues with the data files or documentation or discovery or access tools. Feedback to the data center and to the data contributor on these issues will improve the quality of the data and services at the archive.

6.6 Conclusions

Data management is important in today's science, especially with all of the advances in information technology. Sensors and other data sources can generate voluminous data products, storage devices can safely store data files for rapid access, and compute capabilities are sufficient to analyse and mine the big data. Internet transfer speeds are catching up, but in the short-term, cloud computing and storage has enabled access and analysis to occur within the same cluster.

Well-managed and organized data will enable the research team to work more efficiently during the course of the project, including sharing data files with collaborators so that they can pick up the files and begin using them with minimal training. Data that is thoroughly described and documented can potentially be re-used in ways not imagined when originally collected. For example, well-preserved Earth observations are important for understanding the operation of the Earth system and provide a solid foundation for sustaining and advancing economic, environmental, and social well-being.

Because of the importance of data management, it should be included in the research workflow as a habit, and done frequently enough that good data products are generated. The steps outlined in this and related chapters in this book will ensure that the data are preserved for future use.

Appendix: Example R-Script for Processing Data

This R script (Table 6.8) analyzes a CSV data file of the ORNL DAAC-archived data set: "LBA-ECO CD-02 Forest Canopy Structure, Tapajos National Forest, Brazil: 1999–2003" (Ehleringer et al. 2011).

Table 6.8 R-script that processes data from a read-only file and generates two figures

```
################################################################################################
# Example R script to process data file of an ORNL DAAC-archived data set: "LBA-ECO CD-02  #
# Forest Canopy Structure, Tapajos National Forest, Brazil: 1999-2003"                     #
# Input data file is stored in directory called "original", which is assigned with         #
# read-only permission.                                                                    #
# All output files, including processed data and plots, will be stored in another          #
# directory called "analysis".                                                             #
#                                                                                          #
# version: 0.1                                                                             #
################################################################################################

# Set working directories
# Input CSV data file is in sub-dir "original", on which this script has read permission
# Outputs from this script are stored in sub-dir "analysis", on which this script has both
# read and write permission
setwd('/Users/ywi/Workspace/Temp/R_Scripts/ds1009/')
input_dir  <- 'original'
output_dir <- 'analysis'

# Read in original CSV data file and save data into variable "lai"
# Notes: skip first 16 comment lines
#        parse headers on the 17th line
#        column separator is ","
input_file <- 'CD02_LAI_measurements_TNF.csv'
lai <- read.csv(paste(input_dir, '/', input_file, sep=''), header=TRUE, sep=',', quote='\"',
 dec = '.', skip = 16)

# Show summary information about variable lai
summary(lai)

# Select data records associated with site "km 67 - Primary Forest Tower" and save them
# into variable lai_site1
site_id <- 'km 67 - Primary Forest Tower'
lai_site1 <- subset(lai, lai$Site_ID == site_id)

# Filter out records with non-positive tree height and non-positive LAI value
lai_site1 <- subset(lai_site1, lai_site1$Height > 0 & lai_site1$LAI > 0)

# Save variable lai_site1 into CSV file CD02_LAI_measurements_TNF_Primary_Forest_Tower.csv,
# which is located in the analysis sub-directory.
output_file <- 'CD02_LAI_measurements_TNF_Primary_Forest_Tower.csv'
write.csv(lai_site1, file=paste(output_dir, '/', output_file, sep=''), quote=TRUE);

# Create a histogram plot of the LAI values of all trees near the Primary Forest Tower
# Also, overlay a density distribution line on top of the histogram plot
# The plot will be save into file Primary_Forest_Tower_Histogram_LAI.png in the
# "analysis" sub-directory
output_plot1 <- 'Primary_Forest_Tower_Histogram_LAI.png'
par(mar=c(3, 3, 0, 0), mgp=c(2, 1, 0), cex=5)
png(paste(output_dir, '/', output_plot1, sep=''), width=800, height=800, units = 'px')
hist(lai_site1$LAI, main='', xlab='LAI', col='green', breaks=16, prob=TRUE)
lai_density <- density(lai_site1$LAI)
lines(lai_density, col='red', lwd=10)
dev.off()

# Create a scatter plot showing the relationship between tree heights and LAI values
# Also, overlay a regression line of tree heights and LAI values on top of the scatter plot
# The plot will be save into file Primary_Forest_Tower_LAI_Height_Plot.png in the
# "analysis" sub-directory
output_plot2 <- 'Primary_Forest_Tower_LAI_Height_Plot.png'
png(paste(output_dir, '/', output_plot2, sep=''), width=800, height=800, units = 'px')
par(mar=c(3, 3, 1, 1), mgp=c(2, 1, 0), cex=5)
plot(lai_site1$Height, lai_site1$LAI, main='', xlab='Tree Height (m)', ylab='LAI',
 col='blue', type="p")
lai_height_reg <- lm(lai_site1$LAI~lai_site1$Height)
abline(lai_height_reg, col='red', lwd=10)
dev.off()
```

The script retrieves data records with positive height and LAI values for trees near the site designated “Primary Forest Tower.” After determining a frequency histogram (Fig. 6.5), it then analyzes the relationship between tree height and LAI

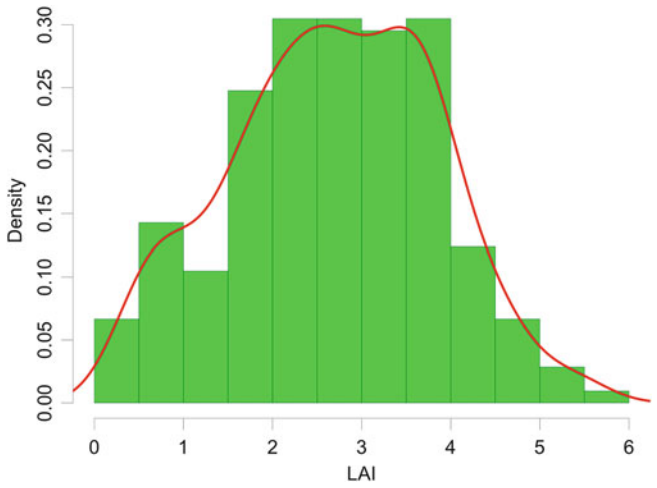

Fig. 6.5 Histogram of LAI for trees in primary forest near flux tower site

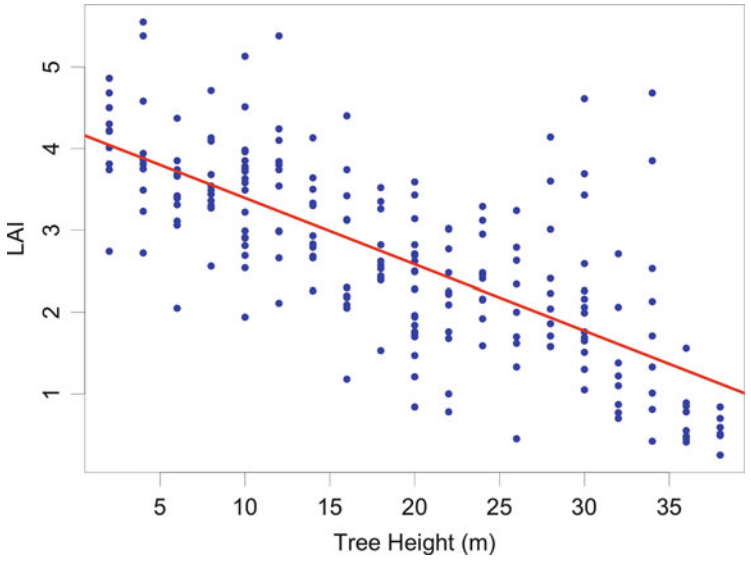

Fig. 6.6 Plot of LAI versus Tree Height for primary forest tress near the flux tower site

values. As revealed by the output plot (Fig. 6.6), height and LAI values have negative correlation for trees near site "Primary Forest Tower".

Input CSV data file of this R script is stored in directory "original", on which the script has only read-only permission. All outputs of this script are saved in directory "analysis", for which the script has both read and write permission.

References

Baldocchi D, Reichstein M, Papale D et al (2012) The role of trace gas flux networks in the biogeosciences. Eos Trans 93:217–218. doi:10.1029/2012EO230001

Bond-Lamberty BP, Thomson AM (2014) A global database of soil respiration data, version 3.0. ORNL DAAC, Oak Ridge, TN. doi:10.3334/ORNLDAAC/1235

Brunt JW (2010) Protecting your digital research data and documents: LTER cybersecurity briefing #1. http://intranet2.lternet.edu/content/protecting-your-digital-research-data-and-documents. Accessed 25 Jan 2015

Cook RB, Olson RJ, Kanciruk P et al (2001) Best practices for preparing ecological and ground-based data sets to share and archive. Bull Ecol Soc Am 82:138–141. http://www.jstor.org/stable/20168543

Cook RB, Post WM, Hook LA et al (2009) A conceptual framework for management of carbon sequestration data and models. In: McPherson BJ, Sundquist ET (eds) Carbon sequestration and its role in the global carbon cycle, AGU Monograph Series 183. American Geophysical Union, Washington, DC, pp 325–334. doi:10.1029/2008GM000713

Cook RB, Vannan SKS, McMurry BF et al (2016) Implementation of data citations and persistent identifiers at the ORNL DAAC. Ecol Inf 33:10–16. doi:10.1016/j.ecoinf.2016.03.003

dos-Santos MN, Keller MM (2016) CMS: forest inventory and biophysical measurements, Para, Brazil, 2012-2014. ORNL DAAC, Oak Ridge, TN. doi:10.3334/ORNLDAAC/1301

Eaton B, Gregory J, Drach R et al (2011) NetCDF climate and forecast (CF) metadata conventions (Vers. 1.6). CF conventions and metadata. http://cfconventions.org/cf-conventions/v1.6.0/cf-conventions.pdf. Accessed 10 May 2016

Edinburgh Data Share (2015) Recommended file formats. http://www.ed.ac.uk/files/atoms/files/recommended_file_formats-apr2015.pdf Accessed 10 May 2016

Ehleringer J, Martinelli LA, Ometto JP (2011) LBA-ECO CD-02 forest canopy structure, Tapajos National Forest, Brazil: 1999–2003. ORNL DAAC, Oak Ridge, TN. doi:10.3334/ORNLDAAC/1009

ESIP (Earth Science Information Partners) (2014) Data citation guidelines for data providers and archives. doi:10.7269/P34F1NNJ

ESO (ESDIS Directory Standards Office) (2016) Standards, requirements and references. https://earthdata.nasa.gov/user-resources/standards-and-references. Accessed 20 Apr 2016

Hook LA, Vannan SKS, Beaty TW et al (2010) Best practices for preparing environmental data sets to share and archive. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, TN. doi:10.3334/ORNLDAAC/BestPractices-2010

IGBP (International Geosphere Biosphere Program) (2012) The Merton Initiative: towards a global observing system for the human environment. http://www.igbp.net/publications/themertoninitiative.4.7815fd3f14373a7f24c256.html. Accessed 7 Mar 2016

ISO (2016) Date and time format - ISO 8601. http://www.iso.org/iso/home/standards/iso8601.htm. Accessed 18 Apr 2016

Iversen CM, Vander Stel HM, Norby RJ et al (2015) Active layer soil carbon and nutrient mineralization, Barrow, Alaska, 2012. Next generation ecosystem experiments arctic data collection, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, TN. doi:10.5440/1185213

Justice CO, Bailey GB, Maiden ME et al (1995) Recent data and information system initiatives for remotely sensed measurements of the land surface. Remote Sens Environ 51:235–244. doi:10.1016/0034-4257(94)00077-Z

Kervin K, Cook RB, Michener WK (2014) The backstage work of data sharing. In: Proceedings of the 18th international conference on supporting group work (GROUP), Sanibel Island, FL, ACM, New York. doi:10.1145/2660398.2660406

Lavoie B (2000) Meeting the challenges of digital preservation: the OAIS reference model. OCLC. http://www.oclc.org/research/publications/library/2000/lavoie-oais.html. Accessed 21 Aug 2015

Michener WK (2015) Ecological data sharing. Ecol Inform 29:33–44. doi:10.1016/j.ecoinf.2015.06.010

Michener WK (2017a) Project data management planning, Chapter 2. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

Michener WK (2017b) Quality assurance and quality control (QA/QC), Chapter 4. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

Michener WK (2017c) Creating and managing metadata, Chapter 5. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

Michener WK (2017d) Data discovery, Chapter 7. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

Michener WK, Brunt JW, Helly J et al (1997) Non-geospatial metadata for ecology. Ecol Appl 7:330–342. doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

NRC (National Research Council) (1991) Solving the global change puzzle: A U.S. strategy for managing data and information, Report by the Committee on Geophysical Data, Geosciences, Environment and Resources, National Research Council. National Academy Press, Washington, DC. http://dx.doi.org/10.17226/18584

Olson RJ, McCord RA (2000) Archiving ecological data and information. In: Michener WK, Brunt JW (eds) Ecological data: design, management and processing. Blackwell Science, Oxford, pp 117–130

Papale D, Agarwal DA, Baldocchi D et al (2012) Database maintenance, data sharing policy, collaboration. In: Aubinet M, Vesala T, Papale D (eds) Eddy covariance: a practical guide to measurement and data analysis. Springer, Dordrecht, pp 411–436. doi:10.1007/978-94-007-2351-1

Parsons MA, Duerr R, Minster J-B (2010) Data citation and peer-review. Eos Trans 91 (34):297–298. doi:10.1029/2010EO340001

Porter JH (2017) Scientific databases for environmental research, Chapter 3. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

Reid WV, Chen D, Goldfarb L et al (2010) Earth system science for global sustainability: grand challenges. Science 330:916–917. doi:10.1126/science.1196263

Ricciuto DM, Schaefer K, Thornton PE et al (2013) NACP site: terrestrial biosphere model and aggregated flux data in standard format. ORNL DAAC, Oak Ridge, TN. doi:10.3334/ORNLDAAC/1183

Rüegg J, Gries C, Bond-Lamberty B et al (2014) Completing the data life cycle: using information management in macrosystems ecology research. Front Ecol Environ 12:24–30. doi:10.1890/120375

Schildhauer M (2017) Data integration: principles and practice, Chapter 8. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

Scholes RJ (2005) SAFARI 2000 woody vegetation characteristics of Kalahari and Skukuza sites. ORNL DAAC, Oak Ridge, TN. doi:10.3334/ORNLDAAC/777

Starr J, Castro E, Crosas M et al (2015) Achieving human and machine accessibility of cited data in scholarly publications. Peer J Comp Sci 1:e1. doi:10.7717/peerj-cs.1

Strasser C, Cook RB, Michener WK et al (2012) Primer on data management: what you always wanted to know about data management, but were afraid to ask. California Digital Library. http://dx.doi.org/doi:10.5060/D2251G48

Tenopir C, Allard S, Douglass K et al (2011) Data sharing by scientists: practices and perceptions. PLoS One 6:e21101. doi:10.1371/journal.pone.0021101

Thornton PE, Thornton MM, Mayer BW et al (2017) Daymet: daily surface weather data on a 1-km grid for North America, Version 3. ORNL DAAC, Oak Ridge, TN. doi:10.3334/ORNLDAAC/1328

UCAR (University Corporation for Atmospheric Research) (2016) UDUNITS. http://www.unidata.ucar.edu/software/udunits/. Accessed 18 Apr 2016

USGEO (US Group on Earth Observation) (2015) Common framework for earth – observation data. US Group on Earth Observation, Data Management Working Group, Office of Science and Technology Policy. https://www.whitehouse.gov/sites/default/files/microsites/ostp/common_framework_for_earth_observation_data_draft_120215pdf. Accessed 25 Jan 2015

Whitlock MC (2011) Data archiving in ecology and evolution: best practices. Trends Ecol Evol 26 (2):61–65. doi:10.1016/j.tree.2010.11.006