Advertise your data using datacasting tools

To make your data available using standard and open software tools you should:
Use standard language and terms to clearly communicate to others that your data are available for reuse and that you expect ethical and appropriate use of your data
Use an open source datacasting (RSS or other type) service that enables you to advertise your data and the options for others to obtain access to it (RSS, GeoRSS, DatacastingRSS)

Description Rationale:
Additional Information:
Examples:

http://www.nsidc.org/libre/
http://datacasting.jpl.nasa.gov/
Tags: access, data services, discover

Assign descriptive file names

File names should reflect the contents of the file and include enough information to uniquely identify the data file. File names may contain information such as project acronym, study title, location, investigator, year(s) of study, data type, version number, and file type.

When choosing a file name, check for any database management limitations on file name length and use of special characters. Also, in general, lower-case names are less software and platform dependent. Avoid using spaces and special characters in file names, directory paths and field names. Automated processing, URLs and other systems often use spaces and special characters for parsing text string. Instead, consider using underscore (_) or dashes (-) to separate meaningful parts of file names. Avoid $ % ^ & # | : and similar.

If versioning is desired a date string within the file name is recommended to indicate the version.

Avoid using file names such as mydata.dat or 1998.dat.

Description Rationale:

Clear, descriptive, and unique file names may be important when your data file is combined in a directory or FTP site with your own data files or with the data files of other investigators. File names that reflect the contents of the file and uniquely identify the data file enable precise search and discovery of particular files.
Additional Information:

Hook, Les A., Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson. 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. Available online (http://daac.ornl.gov/PI/BestPractices-2010.pdf) from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/BestPractices-2010

Borer et al. 2009. Some Simple Guidelines for Effective Data Management. Bull. of ESA 90: 209-214.
Examples:

An example of a good data file name:

Sevilleta_LTER_NM_2001_NPP.csv
- Sevilleta_LTER is the project name
- NM is the state abbreviation
- 2001 is the calendar year
- NPP represents Net Primary Productivity data
- csv stands for the file type—ASCII comma separated variable

Instead of "data May2011" use "data_May2011" or "data-May2011"
Tags: access, describe, discover, format

Backup your data

To avoid accidental loss of data you should:
Backup your data at regular frequencies

When you complete your data collection activity
After you make edits to your data

Streaming data should be backed up at regularly scheduled points in the collection process

High-value data should be backed up daily or more often

Automation simplifies frequent backups

Backup strategies (e.g., full, incremental, differential, etc…) should be optimized for the data collection process
Create, at a minimum, 2 copies of your data
Place one copy at an “off-site” and “trusted” location

Commercial storage facility
Campus file-server
Cloud fire-server (e.g., Amazon S3, Carbonite)

Use a reliable device when making backups

External USB drive (avoid the use of “light-weight” devices e.g., floppy disks, USB stick-drive; avoid network drives that are intermittently accessible)
Managed network drive
Managed cloud file-server (e.g., Amazon S3, Carbonite)

Ensure backup copies are identical to the original copy

Perform differential checks
Perform “checksum” check

Document all procedures to ensure a successful recovery from a backup copy
Description Rationale:

To avoid data loss, and to retain information held in earlier versions, it is valuable to back-up your data at frequent intervals and in multiple locations.
Additional Information:

http://en.wikipedia.org/wiki/Backup
http://www.carbonite.com/
http://aws.amazon.com/s3/
http://en.wikipedia.org/wiki/Incremental_backup
Examples:
Tags: access, disaster recovery, preserve

Check data and other outputs for print and web accessibility

To maximize usability of your data or outputs, ensure that those with impairments or disabilities will still be able to access and understand them. The Web Accessibility Initiative, from the W3C, suggests that those producing content for others consider the following (text from their website):

Make your outputs perceivable

Provide text alternatives for non-text content.
Provide captions and other alternatives for multimedia.
Create content that can be presented in different ways, including by assistive technologies, without losing meaning.
Make it easier for users to see and hear content.

Make your outputs operable

Make all functionality available from a keyboard.
Give users enough time to read and use content.
Do not use content that causes seizures.
Help users navigate and find content.

Make your outputs understandable

Make text readable and understandable.
Make content appear and operate in predictable ways.
Help users avoid and correct mistakes.

Make your outputs robust

Maximize compatibility with current and future user tools.

Description Rationale:

By ensuring your data and other outputs are accessible by all, you are maximizing their potential for re-use and preventing misuse.
Additional Information:

Best Practices for Web Accessibility from W3C: http://www.w3.org/WAI/WCAG20/glance/
You can also check your images with Vischeck http://vischeck.com
Examples:
Tags: access, discover, standards

Choose and use standard terminology to enable discovery

Terms and phrases that are used to represent categorical data values or for creating content in metadata records should reflect appropriate and accepted vocabularies in your community or institution. Methods used to identify and select the proper terminology include:

Identify the relevant descriptive terms used as categorical values in your community prior to start of the project (ex: standard terms describing soil horizons, plant taxonomy, sampling methodology or equipment, etc.)
Identify locations in metadata where standardized terminology should be used and sources for the terms. Terminology should reflect both data type/content and access methods.
Review existing thesauri, ontologies, and keyword lists for your use before making up a new terms. Potential sources include: Semantic Web for Earth and Environmental Terminology (SWEET), Planetary Ontologies, and NASA Global Change Master Directory (GCMD)

Enforce use of standard terminology in your workflow, including:

Use of lookup tables in data-entry forms
Use of field-level constraints in databases (restrict data import to match accepted domain values)
Use XML validation
Do manual review

Publish metadata using Open Standards, for example:

z39.50
OGC Catalog Services for Web (CSW)
Web Accessible Directory (WAD)

If you must use an unconventional or unique vocabulary, it should be identified in the metadata and fully defined in the data documentation (attribute name, values, and definitions).
Description Rationale:

The consistent use of well-defined, referenced terminology in describing data products, their parameters, and access methods improves the ability to discover those products for specific uses and access via well-known methods. Determine if there are controlled vocabulary terms, scientific taxonomies, and ontologies used by your community, and use those when creating metadata for your dataset.
Additional Information:

Controlled vocabulary: http://www.nlm.nih.gov/mesh/meshhome.html
SWEET Web Site: http://sweet.jpl.nasa.gov/ontology/
Biological ontologies: http://www.obofoundry.org/
Taxonomy: http://www.biodiversitylibrary.org/

Ecological Informatics, Volume 2, Issue 3, October 2007, Pages 279-296 Meta-information systems and ontologies. A Special Feature from the 5th International Conference on Ecological Informatics ISEI5, Santa Barbara, CA, Dec. 4–7, 2006 - Novel Concepts of Ecological Data Management S.I.

Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. 2010. ANSI/NISO Z39.19. (http://www.niso.org/kst)

Olsen, L.M., G. Major, K. Shein, J. Scialdone, R. Vogel, S. Leicester, H. Weir, S. Ritz, T. Stevens, M. Meaux, C.Solomon, R. Bilodeau, M. Holland, T. Northcutt, R. A. Restrepo, 2007 . NASA/Global Change Master Directory

(GCMD) Earth Science Keywords. Version 6.0.0.0.0
Citing GCMD Keywords (http://gcmd.nasa.gov/Resources/valids/)
Examples:
Tags: controlled vocabulary, describe, documentation, metadata, ontologies, preserve, standards

---

Communicate data quality

Information about quality control and quality assurance are important components of the metadata:

Qualify (flag) data that have been identified as questionable by including a flagging_column next to the column of data values. The two columns should be properly associated through a naming convention such as Temperature, flag_Temperature.

Describe the qality control methods applied and their assumptions in the metadata. Describe any software used when performing the quality analysis, including code where practical. Include in the metadata who did the quality control analysis, when it was done, and what changes were

made to the dataset.

Describe standards or test data used for the quality analysis. For instance, include, when practical, the data used to make a calibration curve.

If data with qualifier flags are summarized to create a derived data set, include the percent flagged data and percent missing data in the metadata of the derived data file. High frequency observations are often downsampled, and it is critical to know how much of the data were rejected in the primary data.
Description Rationale:

Data quality and any methods used for quality control should be communicated so others can assess the data independently.
Additional Information:

Hook, L.A., Beaty, T.W., Santhana-Vannan, S., Baskaran, L. and Cook, R.B. 2007. Best practices for preparing environmental data sets to share and archive. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. (daac.ornl.gov/PI/bestprac.html)
Sheldon, W., Henshaw, D. and Ramsey, K. 2007. Final Report: Workshop to define quality management standards for data completeness in derived data products. Long Term Ecological Research Network Document Archive, University of New Mexico, Albuquerque, NM.
Examples:
Tags: assure, flag, quality

Confirm a match between data and their description in metadata

To assure that metadata correctly describes what is actually in a data file, visual inspection or analysis should be done by someone not otherwise familiar with the data and its format. This will assure that the metadata is sufficient to describe the data. For example, statistical software can be used to summarize data contents to make sure that data types, ranges and, for categorical data, values found, are as described in the documentation/metadata.

Description Rationale:

Sometimes mistakes in either data or metadata preparation cause discrepancies between the two. These can include missing (or extra) columns of data, mis-ordered columns of data, or discrepant values.
Additional Information:

Lin, C.C, Porter JH, Hsiao C.W, Lu S.S, Jeng M.R. Establishing an EML-based Data Management System for Automating Analysis of Field Sensor Data. Taiwan J For Sci. 23(3):279-285, 2008.
http://www.tfri.gov.tw/enu/pub_science_in.aspx?pid=895&catid0=43&catid1=...

Long, J.B. Validating Metadata at the VCR/LTER. LTER Databits, Spring 2011.
http://databits.lternet.edu/spring-2011/validating-metadata-vcrlter
Examples:

Metadata describes a dataset that has two columns, the first is defined to be StationID and should contain station codes "Station1" and "Station2." The second column contains temperature data with a range between -20 and 40 degrees Celsius. However, the data file contains three columns. The first contains the temperature, the second humidity and the third the StationID with stations labled "Stat1", "Stat2", and "Stat3". This sort of problem can occur if data is processed or added after initial metadata was created, or if there were simply mistakes made in the metadata preparation. Having a naive user use the metadata to ingest and analyze this data will make the problems clear and either the metadata or the data can be altered to make it so they correspond.
Tags: assure, data consistency, describe, documentation, metadata, quality

---

Consider the compatibility of the data you are integrating

The integration of multiple data sets from different sources requires that they be compatible. Methods used to create the data should be considered early in the process, to avoid problems later during attempts to integrate data sets. Note that just because data can be integrated does not necessarily mean that they should be, or that the final product can meet the needs of the study. Where possible, clearly state situations or conditions where it is and is not appropriate to use your data, and provide information (such as software used and good metadata) to make integration easier.
Description Rationale:

When using integrated data sets, it is crucial that the data are comparable and compatible to avoid mistakes in analyses and interpretation.
Additional Information:

Burley, T.E., and Peine, J.D., 2009, NBII-SAIN Data Management Toolkit, U.S. Geological Survey Open-File Report 2009-1170, 96 p. Available from: http://pubs.usgs.gov/of/2009/1170/
Examples:

Water-quality data collected by two separate agencies may be thematically similar but may have been sampled using completely different methods. Differences in such water-quality sample methods can include equipment, sampling method protocol, and lab analysis procedures. Analysis performed on integrated water-quality data that were collected using completely different methods would likely result in questionable results.
Tags: analyze, assure, database, integrate, quality, tabular

Create a data dictionary

A data dictionary provides a detailed description for each element or variable in your dataset and data model. Data dictionaries are used to document important and useful information such as a descriptive name, the data type, allowed values, units, and text description. A data dictionary provides a concise guide to understanding and using the data.
Description Rationale:

A data dictionary is an effective and concise way to describe the elements or variables that make up your dataset.
Additional Information:
Examples:
Tags: controlled vocabulary, describe, documentation, metadata, terminology, units

---

Create and document a data backup policy

A backup policy helps manage users' expectations and provides specific guidance on the "who, what, when, and how" of the data backup and restore process. There are several benefits to documenting your data backup policy:
Helps clarify the policies, procedures, and responsibilities
Allows you to dictate:
where backups are located
who can access backups and how they can be contacted
how often data should be backed up
what kind of backups are performed and
what hardware and software are recommended for performing backups
Identifies any other policies or procedures that may already exist (such as contingency plans) or which ones may supersede the policy
Has a well-defined schedule for performing backups
Identifies who is responsible for performing the backups and their contact information. This should include more than one person, in case the primary person responsible is unavailable
Identifies who is responsible for checking the backups have been performed successfully, how and when they will perform this
Ensures data can be completely restored
Has training for those responsible for performing the backups and for the users who may need to access the backups
Is partially, if not fully automated
Ensures that more than one copy of the backup exists and that it is not located in same location as the originating data
Ensures that a variety of media are used to backup data, as each media type has its own inherent reliability issues
Ensures the structure of the data being backed up mirrors the originating data
Notes whether or not the data will be archived

If this information is located in one place, it makes it easier for anyone needing the information to access it. In addition, if a backup policy is in place, anyone new to the project or office can be given the documentation which will help inform them and provide guidance.

Description Rationale:

Collecting information about backing data up before it is needed helps prevent problems and delays that may be encountered when a user needs data from a backup.
Additional Information:
Examples:

http://www.aub.edu.lb/info/data-backup.html
http://dept.wofford.edu/it/Data%20Backup%20Policy.pdf
Tags: backup, disaster recovery, plan, preserve, storage

Create, manage, and document your data storage system

Data files should be managed to avoid disorder. To facilitate access to files, all storage devices, locations and access accounts should be documented and accessible to team members. Use appropriate tools, such as version control tools, to keep track of the history of the data files. This will help with maintaining files in different locations, such as at multiple off-site backup locations or servers.

Data sets that result in many files structured in a file directory can be difficult to decipher. Organize files logically to represent the structure of the research/data. Include human readable "readme" files at critical levels of the directory tree. A "readme" file might include such things as explanations of naming conventions and how the structure of the directory relates to the structure of the data.
Description Rationale:

Keeping a managed file storage system will help prevent inconsistencies, e.g., duplicated, lost, or misplaced files.
Additional Information:
Examples:

A time series of image files from several remote cameras might be organized so that images from each camera are in different folders. These are, in turn, collected in a folder named "images". Each folder would be named with the identifier for the camera. The file names for images might reflect the time the image was taken.

A "readme" file would document the structure of this system, and document the name scheme to facilitate future curation and automated gathering of metadata.
Tags: access, documentation, file system, metadata, plan

Decide what data to preserve

The process of science generates a variety of products that are worthy of preservation. Researchers should consider all elements of the scientific process in deciding what to preserve:

Raw data
Tables and databases of raw or cleaned observation records and measurements
Intermediate products, such as partly summarized or coded data that are the input to the next step in an analysis
Documentation of the protocols used
Software or algorithms developed to prepare data (cleaning scripts) or perform analyses
Results of an analysis, which can themselves be starting points or ingredients in future analyses, e.g. distribution maps, population trends, mean measurements
Any data sets obtained from others that were used in data processing
Multimedia: documented procedures, or standalone data

When deciding on what data products to preserve, researchers should consider the costs of preserving data:

Raw data are usually worth preserving
Consider space requirements when deciding on whether to preserve data
If data can be easily or automatically re-created from raw data, consider not preserving. E.g. if data that have undergone quality control processes and were analyzed, consider preserving since reproduction might be costly
Algorithms and software source code cost very little to preserve
Results of analyses may be particularly valuable for future discovery and cost very little to preserve

Researchers should consider the following goals and benefits of preservation:

Enabling re-analysis of the same products to determine whether the same conclusions are reached
Enabling re-use of the products for new analysis and discovery
Enabling restoration of original products in the case that working datasets are lost

Description Rationale:

To meet multiple goals for preservation, researchers should think broadly about the digital products that their project generates, preserve as many as possible, and plan the appropriate preservation methods for each.
Additional Information:
Examples:
Tags: data archives, disaster recovery, image, preserve, storage

---

Define expected data outcomes and types

In the planning process, researchers should carefully consider what data will be produced in the course of their project.

Consider the following:

What types of data will be collected? E.g. Spatial, temporal, instrument-generated, models, simulations, images, video etc.
How many data files of each type are likely to be generated during the project? What size will they be?
For each type of data file, what are the variables that are expected to be included?
What software programs will be used to generate the data?
How will the files be organized in a directory structure on a file system or in some other system?
Will metadata information be stored separately from the data during the project?
What is the relationship between the different types of data?
Which of the data products are of primary importance and should be preserved for the long-term, and which are intermediate working versions not of long-term interest?

When preparing a data management plan, defining the types of data that will be generated helps in planning for short-term organization, the analyses to be conducted, and long-term data storage.

Description Rationale:

Considering data outcomes for the project helps anticipate budgetary, software, storage, and personnel needs, and for choosing an appropriate repository for long-term preservation.
Additional Information:

Graham, A., McNeill, K., Stout, A., & Sweeney, L. (2010). Data Management and Publishing. Last modified November 29, 2010.

http://libraries.mit.edu/guides/subjects/data-management/.

Van den Eynden, V., Corti, L., Woollard, M. & Bishop, L. (2011). Managing and Sharing Data: A Best Practice Guide for Researchers. Published May 2011. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf.
Examples:

"The project will result in spreadsheets of species abundance. One spreadsheet file (saved as .csv) will be generated for each site, and within each spreadsheet there will be data from multiple sampling dates. We will also generate text files that document observations by the researcher during data collection in the field. There will be a single text file for each site and each collection date."
Tags: data model, data source, format, plan

Define roles and assign responsibilities for data management

In addition to the primary researcher(s), there might be others involved in the research process that take part in aspects of data management. By clearly defining the roles and responsibilities of the parties involved, data are more likely to be available for use by the primary researchers and anyone re-using the data. Roles and responsibilities should be clearly defined, rather than assumed; this is especially important for collaborative projects that involve many researchers, institutions, and/or groups.

Examples of roles in data management:

data collector
metadata generator
data analyzer
project director
data model and/or database designer
computing staff responsible for backup and/or storage
staff responsible for running instruments
administrative support staff responsible for grant submission
specialized skills as defined in the plan (GIS, relational database design/implementation, computer programming of sensors/input forms, etc)
external data center or archive

Steps for assigning data management responsibilities:

For each task identified in your data management plan, identify the skills needed to perform the task
Match skills needed to available staff and identify gaps
Develop training/hiring plan
Develop staffing/training budget and incorporate into project budget
Assign responsible parties and monitor results

Description Rationale:

A successful data management plan requires that the appropriate staffing resources are available and trained. Identifying specific tasks and responsible parties will help with budgeting, implementation, and preservation of the data resources.
Additional Information:

Search on the web for information on "Project Planning: Roles and Responsibilities for developing specific staffing/training plans"

Van den Eynden, V., Corti, L., Woollard, M. & Bishop, L. (2011). Managing and Sharing Data: A Best Practice Guide for Researchers. Last updated May 2011. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf.
Examples:

Task: complete metadata record
Skills: knowledge of EML, and XML editing (Morpho), knowledge of the data to be documented
Responsible party: Jane Smith
Training needed: attend XML training on-line
Personnel time or %fte: 1/4 time for 1 month for Smith
Estimated cost: $1,000
Additional resources: $100 to travel to remote site to meet with researcher
Deadlines: one month after data is collected or end of fiscal year
Tags: data management plan, plan

Define the data model

A data model documents and organizes data, how it is stored and accessed, and the relationships among different types of data. The model may be abstract or concrete.

Use these guidelines to create a data model:

Identify the different data components- consider raw and processed data, as well as associated metadata (these are called entities)
Identify the relationships between the different data components (these are called associations)
Identify anticipated uses of the data (these are called requirements), with recognition that data may be most valuable in the future for unanticipated uses
Identify the strengths and constraints of the technology (hardware and software) that you plan to use during your project (this is called a technology assessment phase)
Build a draft model of the entities and their relations, attempting to keep the model independent from any specific uses or technology constraints.
Incorporate intended usage and technology constraints as needed to derive the simplest, most general model possible
Test the model with different scenarios, including best- and
worst-case (worst-case includes problems such as invalid raw data, user mistakes, failing algorithms, etc)
Repeat these steps to optimize the model

Description Rationale:

Considering and creating the data model helps with data planning and identifies potential problems that future data users might encounter.
Additional Information:

Wikipedia: http://en.wikipedia.org/wiki/Data_model
IBM: http://publib.boulder.ibm.com/infocenter/tivihelp/v8r1/index.jsp?topic=/...
Agile Data: http://www.agiledata.org/essays/dataModeling101.html
Examples:

Different types of data model examples can be found here: http://www.databaseanswers.org/data_models/index.htm
Tags: access, data model, describe, plan

Define the parameters

The parameters reported in the data set need to have names that clearly describe the contents. Ideally, the names should be standardized across files, data sets, and projects, in order that others can readily use the information.

The documentation should contain a full description of the parameter, including the parameter name, how it was measured, the units, and the abbreviation used in the data file.

A missing value code should also be defined. Use the same notation for each missing value in the data set. Use an extreme value (-9999) and do not use character codes in a numeric field. Supply a flag or a tag in a separate field to define briefly the reason for the missing data.

Within the data file use commonly accepted abbreviations for parameter names, for example, Temp for temperature, Precip for precipitation, Lat and Long for latitude and longitude. See the references in the Bibliography for additional examples. Some systems still have length limitations for column names (e.g.13 characters in ArcGIS); lower case column names are generally more transferrable between systems; Space and special characters should not be used in attribute names. Only numbers, letters and underscors (“_”) transfer easily between systems.

Also, be sure to use consistent capitalization (not temp, Temp, and TEMP in the same file).

Description Rationale:

In order for others to use your data, they must fully understand the parameters in the data set, including the parameter name, unit of measure, and format.
Additional Information:

Hook, Les A., Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson. 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. Available online (http://daac.ornl.gov/PI/BestPractices-2010.pdf) from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/BestPractices-2010.

Cook et al., 2001. Best Practices for Preparing Ecological and Ground-Based Data Sets to Share and Archive. Bulletin of ESA 82: 138-141.

Borer et al. 2009. Some Simple Guidelines for Effective Data Management. Bulletin of ESA 90: 209-214. doi:10.1890/0012-9623-90.2.205
Examples:
Tags: describe, documentation, metadata, parameter

Describe format for spatial location

Spatial coordinates should be reported in decimal degrees format to at least 4 (preferably 5 or 6) significant digits past the decimal point. An accuracy of 1.11 meters at the equator is represented by +/- 0.00001. This does not include uncertainty introduced by a GPS instrument.

Provide latitude and longitude with south latitude and west longitude recorded as negative values, e.g., 80 30' 00" W longitude is -80.5000.

Make sure that all location information in a file uses the same coordinate system, including coordinate type, datum, and spheroid. Document all three of these characteristics (e.g., Lat/Long decimal degrees, NAD83 (North American Datum of 1983), WGRS84 (World Geographic Reference

System of 1984)). Mixing coordinate systems [e.g., NAD83 and NAD27 (North American Datum of 1927)] will cause errors in any geographic analysis of the data.

If locating field sites is more convenient using the Universal Transverse Mercator (UTM) coordinate system, be sure to record the datum and UTM zone (e.g., NAD83 and Zone 15N), and the easting and northing coordinate pair in meters, to ensure that UTM coordinates can be converted to latitude and longitude.

To assure the quality of the geospatial data, plot the locations on a map and visually check the location.
Description Rationale:

Spatial information is necessary to describe where on Earth an observation was made.
Additional Information:
Examples:
Tags: access, describe, format, geospatial, location

Describe formats for date and time

For date, always include four digit year and use numbers for months. For example, the date format yyyy-mm-dd would appear as 2011-03-15 (March 15, 2011).

If Julian day is used, make sure the year field is also supplied. For example, mmm.yyyy would appear as 122.2011, where mmm is the Julian day.

If the date is not completely known (e.g. day not known) separate the columns into parts that do exist (e.g. separate column for year and month). Don't introduce a day because the database date format requires it.

For time, use 24-hour notation (13:30 hrs instead of 1:30 p.m. and 04:30 instead of 4:30 a.m.). Report in both local time and Coordinated Universal Time (UTC). Include local time zone in a separate field. As appropriate, both the begin time and end time should be reported in both local and UTC time. Because UTC and local time may be on different days, we suggest that dates be given for each time reported.

Be consistent in date and time formats within one data set.
Description Rationale:

The date and time are important pieces of contextual information for observations. A complete description of date and time allows the observation to be used and interpreted properly.
Additional Information:

ISO 8601 format date: YYYY-MM-DD time: HH:MM:SS datatime: YYYYMMDDTHHMMSS
Examples:
Tags: date, describe, format, standards, time

Describe measurement techniques

Data measurement descriptions should:

Describe data collection methods or protocols(can include diagrams, images, schematics, etc.)

How the data were collected
Measurement frequency and regularity

Describe instrumentation

Include manufacturer, model number, dates in use
Maintenance/repair history
Malfunction history
Calibration methods, scale, detection limits, and history

Document measurement uncertainty, including accuracy, precision, and reproducibility. Provide values in the context of the measurements, e.g., standard error, standard deviation, confidence limits.

Description Rationale:

Describe how data were collected in enough detail for others to reproduce the measurements, assess the appropriateness of the measurements, and/or assess comparability of the data with other datasets.
Additional Information:

R.H.G., J., Ter Braak, C. J., & Van Tongeren, O. F. (1995). Data Analysis in Community and Landscape Ecology. Cambridge University Press
Examples:
Tags: access, calibration

Describe method to create derived data products

When describing the process for creating derived data products, the following information should be included in the data documentation or the companion metadata file:

Description of primary input data and derived data
Why processing is required
Data processing steps and assumptions

Assumptions about primary input data
Additional input data requirements
Processing algorithm (e.g., volts to mol fraction, averaging)
Assumptions and limitations of algorithm
Describe how algorithm is applied (e.g., manually, using R, IDL)

How outcome of processing is evaluated

How problems are identified and rectified
Tools used to assess outcome
Conditions under which reprocessing is required

How uncertainty in processing is assessed

Provide a numeric estimate of uncertainty

How processing technique changes over time, if applicable

Description Rationale:
Additional Information:

Bourque, Linda B., Clark, Virginia A. Processing Data: The Survey Example (Quantitative Applications in the Social Sciences), Sage Publications, Inc. (December 14, 2008), ISBN 08056781901
Examples:
Tags: analyze, data processing, describe, provenance

Describe the contents of data files

A description of the contents of the data file should contain the following:

Define the parameters and the units on the parameter
Explain the formats for dates, time, geographic coordinates, and other parameters
Define any coded values
Describe quality flags or qualifying values
Define missing values

Description Rationale:

The contents of the data files should be described in the data file and in the documentation
Additional Information:
Examples:
Tags: describe, documentation, format, metadata, parameter, units

---

Describe the overall organization of your dataset

Data sets or collections are often composed of multiple files that are related. Files may have come from (or still be stored in) a relational database, and the relationships among the data tables or other entities are important if the data are to be reused. These relationships should be documented for a repository.

Describe the overall organization of your data set or collection. Often, a data set or collection contains a large number of files, perhaps organized into a number of directories or database tables. By describing and documenting this organization, files and data can be easily located and used.

At a minimum, the organization and relationships between the directories and files, or database tables and other supporting materials, need to be fully described. Use a description of the data set or collection (e.g, an abstract) to describe what tables contain, where the supporting material,

metadata, or other documentation are located, and/or descriptions of directory contents. Consider describing the logical relationships between data entities using an entity relationship diagram (ERD).

Associated specimens: if specimens (e.g., taxonomic vouchers, DNA samples) were collected with the data, include the name of the repository in which these specimens reside.
Description Rationale:

Relationships among data entities should be described documented to enable understanding by future users and repositories.
Additional Information:

Specimen repositories: http://www.biorepositories.org/
Describing data table constraints with Ecological Metadata Language (EML): http://knb.ecoinformatics.org/software/eml/eml-2.1.0/eml-constraint.html
Examples:
Tags: data model, database, describe, documentation, metadata

---

Describe the research project

The research project description should contain the following information:

Who: project personnel (principal investigator, researchers, technicians, others)
Where: location and description of study site or sites
When: range of dates for the project
Why: rational for the project (abstract)
How: description of project methods

Other useful information might include the project title, the overarching project (if any), institution(s) involved, and source of funding.

Description Rationale:

The project description provides essential background information and context for the dataset.
Additional Information:
Examples:
Tags: annotation, data creators, describe, geography, geospatial, measurement

---

Describe the sensor network

If your project uses a sensor network, you should describe and document that network and the instruments it uses. This information is essential to understanding and interpreting the data you use, and should be included as a part of the metadata generated for your project's data.

Describe the basic set-up of the sensor network installation, including such details as mount, power source, enclosures, wiring protection, etc.
Describe instrumentation, cameras and samplers (See "Describe measurement techniques" Best Practice in DataONEpedia)
Describe data loggers used by the network. Include the following:

Manufacturer, model, serial number, dates in use
Maintenance/repair history
Malfunction history
Deployment history
Replacement history

Ensure localization and time synchronization across data logger arrays
Archive copies of any custom scripts, software, or programs used. Scripts and programs should be accompanied by documentation that includes any information pertinent to their use (metadata).
As part of metadata, create a human-readable document that describes sampling frequency and data processing performed by the data logger

Description Rationale:

Data users need the ability to account for site conditions and equipment choices that may exert a strong influence on the character and/or quality of collected data; this includes significant data processing and selection that happens at the instrument and data logger levels, which is often forgotten.
Additional Information:

N.E.O.N. Site Prospectus: http://www.neoninc.org/sites/default/files/Prospectus_FINAL_med%20res.pdf
Examples:
Tags: calibration

Describe the spatial extent and resolution of your dataset

The spatial extent of your data set or collection as a whole should be described. The minimum acceptable description would be a bounding box describing the northern most, southern most, western most, and eastern most limits of the data.

If the entire collection is from a single location, use the same values for northerly/southerly limits and easterly/westerly values.
Be sure to specify in the metadata what units you choose to describe your spatial extent.
Use the following guidelines for quality control:

If the collection spans the north pole, the northerly limit should be 90.0 degrees
If the collection spans the south pole, the southerly limit should be -90.0 degrees
If the collection crosses the date line, the westerly limit should be greater than the easterly limit

If your data collection or dataset as a whole contains data acquired over a range of spatial locations during each collection period, it is important to document the spatial resolution of your dataset. Many metadata standards have standard terminology for describing data spacing or resolution (e.g. every half degree, 250 m resolution, etc.), but it may be necessary to describe complex data acquisition schemes textually.
Description Rationale:

Describing the spatial boundaries of a data collection as a whole allows users to assess the collection for applicability to their research needs without having to individually assess the relevance of each individual component of the collection.
Additional Information:

NASA: http://gcmd.nasa.gov/User/difguide/spatial_coverage.html
NASA: http://gcmd.nasa.gov/User/difguide/data_resolution.html
Examples:

A data collection consisting of a number of buoys floating in the Arctic ocean might have a spatial extent of...

- Northernmost latitude: 90.0
- Westernmost longitude: -180.0
- Easternmost longitude: 180.0
- Southernmost latitude: 66.0

250m gridded data:

- Data are measured every 250 m along a series of 1 km transects defined in a shape file

Tags: describe, documentation, geospatial, location, measurement, metadata

---

Describe the temporal extent and resolution of your dataset

The temporal extent over which the data within your dataset or collection was acquired or collected should be described. Normally this is done by providing

the earliest date of data acquisition
the date that the last data in the collection was acquired
Year, month, day, and time should be included in the description. If data collection is still ongoing, the end date can be omitted, though some statement about this should be placed in the dataset abstract. The status of the data set should indicate that data collection is still ongoing if the metadata standard being used supports this type of documentation.
Describe the temporal resolution of your dataset collection. The temporal resolution of your dataset is the frequency with which data is collected or acquired. While many metadata standards provide standard nomenclature for describing simple temporal resolutions (e.g., daily or monthly), more complex temporal collection patterns may need to be described textually.
Description Rationale:

Describing the temporal boundaries of a data collection as a whole allows users to assess the collection for applicability to their research needs without having to individually assess the relevance of each measurement within the collection. Describing the temporal resolution of your data set or collection allows users to assess the utility of your data set against their needs without examining each component of your data set individually.
Additional Information:

NASA: http://gcmd.nasa.gov/User/difguide/temporal_coverage.html
NASA: http://gcmd.nasa.gov/User/difguide/data_resolution.html
Examples:

Dates could be given in ISO 8601 date/time format (YYYY-MM-DDThh:mm:ss).

A data set collected daily since February 1, 1990 should be described as

- Start date: 1990-02-01

"Data measured every 5 minutes seasonally during the summer months of June - August"

If collection of data ended on March 19, 2002 the temporal extent would be:

- Start date: 1990-02-01
- Stop date: 2002-03-19

Tags: date, describe, documentation, measurement, metadata, time

---

Describe the units of measurement for each observation

The units of reported parameters need to be explicitly stated in the data file and in the documentation. We recommend SI units (The International System of Units) but recognize that each discipline has its own commonly used units of measure. The critical aspect here is that the units be defined so that others understand what is reported.

Do not use abbreviations when describing the units. For example the units for respiration are moles of carbon dioxide per meter squared per year.

Description Rationale:
Additional Information:

National Institute for Standards and Technology guide to SI units: http://physics.nist.gov/Pubs/SP330/sp330.pdf, BIPM Bureau International des Poids et Mesures (SI meaintenance agacy) http://www.bipm.org/en/si/
Examples:
Tags: describe, documentation, measurement, units

Develop a quality assurance and quality control plan

Just as data checking and review are important components of data management, so is the step of documenting how these tasks were accomplished. Creating a plan for how to review the data before it is collected or compiled allows a researcher to think systematically about the kinds of errors, conflicts, and other data problems they are likely to encounter in a given data set. When associated with the resulting data and metadata, these documented quality control procedures help provide a complete picture of the content of the dataset. A helpful approach to documenting data checking and review (often called Quality Assurance, Quality Control, or QA/QC) is to list the actions taken to evaluate the data, how decisions were made regarding problem resolution, and what actions were taken to resolve the problems at each step in the data life cycle. Quality control and assurance should include:

determining how to identify potentially erroneous data
how to deal with erroneous data
how problematic data will be marked (i.e. flagged)

For instance, a researcher may graph a list of particular observations and look for outliers, return to the original data source to confirm suspicions about certain values, and then make a change to the live dataset. In another dataset, researchers may wish to compare data streams from remote sensors, finding discrepant data and choosing or dropping data sources accordingly. Recording how these steps were done can be invaluable for later understanding of the dataset, even by the original investigator.

Datasets that contain similar and consistent data can be used as baselines against each other for comparison.
Obtain data using similar techniques, processes, environments to ensure similar outcome between datasets.
Provide mechanisms to compare data sets against each other that provide a measurable means to alert one of differences if they do indeed arise. These differences can indicate a possible error condition since one or more data sets are not exhibiting the expected outcome exemplified by similar data sets.

One efficient way to document data QA/QC as it is being performed is to use automation such as a script, macro, or stand alone program. In addition to providing a built-in documentation, automation creates error-checking and review that can be highly repeatable, which is helpful for researchers collecting similar data through time.
The plan should be reviewed by others to make sure the plan is comprehensive.

Description Rationale:

A plan for QA/QC is needed so that others can understand how to best use the data, and avoid potential mistakes that might occur due to use of poor quality data. Explicitly documenting how you check your data for errors and conflicts can also help resolve current and future questions about the data.
Additional Information:
Examples:

Water quality data for EPA is collected with a QA/QC plan often called QAPP Quality Assurance Project Plan.

See Quality Management Tools - QA Project Plans

A Quality Assurance Project Plan documents the planning, implementation, and assessment procedures for a particular project, as well as any specific quality assurance and quality control activities. It integrates all the technical and quality aspects of the project in order to provide a "blueprint" for obtaining the type and quality of environmental data and information needed for a specific decision or use. All work performed or funded by EPA that involves the acquisition of environmental data must have an approved Quality Assurance Project Plan.

http://www.epa.gov/QUALITY/qapps.html
Tags: assure, data consistency, flag, measurement, quality

Document and store data using stable file formats

File formats are important for understanding how data can be used and possibly integrated. The following issues need to be documented:

Does the file format of the data adhere to one or more standards?
Is that file standard an open (i.e. open source) or closed (i.e. proprietary) format?
Is a particular software package required to read and work with the data file? If so, the software package, version, and operating system platform should be cited in the metadata
Do multiple files comprise the data file structure? If so, that should be specified in the metadata

When choosing a file format, data collectors should select a consistent format that can be read well into the future and is independent of changes in applications.
Appropriate file types include:

Non-proprietary: Open, documented standard
Common usage by research community: Standard representation (ASCII, Unicode)
Unencrypted
Uncompressed

ASCII formatted files will be readable into the future

Use ASCII (comma-separated) for tabular data

For geospatial (raster) data the following provide a stable format:

GeoTIFF/TIFF
ASCII Grid
Binary image files
NetCDF
HDF or HDF-EOS
For image (Vector) data use the following file formats (these are mostly proprietary data formats; please be sure to document the Software Package, Version, Vendor, and native platform):

ARCVIEW software -- please store components of an ArcView shape file (*.shp, *.sbx, *.sbn, *.prj, and *.dbf files) ;
ENVI -- *.evf (ENVI vector file)
ESRI Arc/Info export file (.e00)

Description Rationale:

For long term preservation is it necessary to store data in file formats that will be readable in the future. It is also important to provide descriptive information on these data file types and formats. This will facilitate data retrieval and reuse.
Additional Information:

Data Management and Publishing (MIT Libraries) http://libraries.mit.edu/guides/subjects/data-management/

Burley, T.E., and Peine, J.D., 2009, NBII-SAIN Data Management Toolkit, U.S. Geological Survey Open-File Report 2009-1170, 96 p. Available from: http://pubs.usgs.gov/of/2009/1170/
Examples:

Certain file formats, for example a shapefile, can be made up of as many as 7 individual files. If one of those files is absent from the file assembly the shapefile data utility may be lost. Awareness of adherence to a particular file format standard can also be helpful for determining, for example, if a particular software package can read the data file. Awareness of whether that standard or format is open source or proprietary will also influence how and if the data file can be read.
Tags: documentation, format, metadata, preserve, storage, tabular

---

Document steps used in data processing

Different types of new data may be created in the course of a project, for instance visualizations, plots, statistical outputs, a new dataset created by integrating multiple datasets, etc. Whenever possible, document your workflow (the process used to clean, analyze and visualize data) noting what data products are created at each step. Depending on the nature of the project, this might be as a computer script, or it may be notes in a text file documenting the process you used (i.e. process metadata). If workflows are preserved along with data products, they can be executed and enable the data product to be reproduced.
Description Rationale:

To enable others to verify the quality of a given data product, and ideally, to reproduce it, it is critical that the steps followed to create that product

be properly documented.
Additional Information:

This best practice is also applicable to other categories including Analysis and Visualization and Data Documentation.

- Juliana Freire, Cláudio T. Silva, Steven P. Callahan, Emanuele Santos, Carlos Eduardo Scheidegger, Huy T. Vo: Managing Rapidly-Evolving Scientific Workflows. IPAW 2006: 10-18
- Juliana Freire, David Koop, Emanuele Santos, Cláudio T. Silva: Provenance for Computational Tasks: A Survey. Computing in Science and Engineering 10(3): 11-21 (2008)

Examples:
Tags: analyze, data processing, describe, integrate, provenance, replicable data

---

Document taxonomic information

Identification of any species represented in the data set should be as complete as possible.

Use a standard taxonomy whenever possible
Full taxonomic tree to most specific level available
Source of taxonomy should accompany taxonomic tree (if available)
References used for taxonomic identification should be provided, if appropriate (e.g. technical document, journal article, book, database, person, etc.)

Examples of standardized identification systems:

Integrated Taxonomic Information System (http://www.itis.gov/)
Species 2000 (http://www.sp2000.org/)
USDA Plants (http://plants.usda.gov/index.html)
Global Biodiversity Information Facility (http://www.gbif.org/informatics/name-services/using-names-data/)

Description Rationale:

Many ecological studies involve at least one species; identifying the species used in a study is critical for its evaluation. The most complete information possible should be provided since taxonomic names and classifications shift over time with new information.
Additional Information:
Examples:
Tags: describe, metadata, standards, taxonomy, terminology

Document the integration of multiple datasets

Document that steps used to integrate disparate datasets.

Ideally, one would adopt mechanisms to systematically capture the integration process, e.g. in an executable form such as a script or workflow, so that it can be reproduced
In lieu of a scientific workflow system, document the process, scripts, or queries used to perform the integration of data in documentation that will accompany the data (metadata)
Provide a conceptual model that describes the relationships among datasets from different sources
Use unique identifiers in the data records to maintain data integrity by reducing duplication
Identify foreign key fields in the data records which support the relationship between the data sources
When you use datasets and data elements from within those datasets as a source for new datasets, it is important to identify and document those data within the documentation of the new/derived dataset. This is known as dataset provenance; provenance describes the origin or source of something. Just as you would cite papers that are sources for your research paper, it is critical to identify the sources of the data used within your own datasets. This will allow for:

tracing the chain of use of datasets and data elements
credit and attribution to accrue to the creators of the original datasets
the possibility that if errors or new information about the original datasets or data elements comes to light, that any impact on your new datasets and interpretation of such could be traced

Description Rationale:

Provide enough information about the process used to integrate disparate datasets so that others can properly use your data and/or your process to integrate similar data sources.
Additional Information:
Examples:

Some proposed guidelines and methods of citing datasets and data elements can be found at:

- DataCite: Cite your data

- Dryad: Citing Data
- A Proposed Standard for the Scholarly Citation of Quantitative Data
- Dataverse Network Data Citation Standard

Tags: citation, data consistency, documentation, integrate, metadata, provenance

---

Document your data organization strategy

The following are strategies for effective data organization:

Sparse matrix: Optimal data models for storing data avoid sparse matrices, i.e. if many data points within a matrix are empty a data table with a column for parameters and a column for values may be more appropriate.

Repetitive information in a wide matrix: repeated categorical information is best handled in separate tables to reduce redundancy in the data table. In database design this is called normalization of data.

Column name is a value or repeating group: If the column name contains variable information, e.g. date or species name, the parameter/value organization of data is recommended as well for storage. Although the wide matrix is needed for statistical analysis and graphing it cannot be queried or subset in that format.

Description Rationale:

Data management requires an effective strategy for data organization.
Additional Information:

Borer, E. T., E. W. Seabloom, M. B. Jones, and M. Schildhauer. 2009. Some simple guidelines for effective data management. ESA Bulletin 90:205-214. http://www.esajournals.org/doi/abs/10.1890/0012-9623-90.2.205.
Examples:
Tags: data management plan, data model, data normalization, database, describe

---

Double-check the data you enter

Ensuring accuracy of your data is critical to any analysis that follows.

When transcribing data from paper records to digital representation, have at least two, but preferably more people transcribe the same data, and compare resulting digital files. At a minimum someone other than the person who originally entered the data should compare the paper records to the digital file. Disagreements can then be flagged and resolved.

In addition to transcription accuracy, data compiled from multiple sources may need review or evaluation. For instance, citizen science records such as bird photographs may have taxonomic identification that an expert may need to review and potentially revise.
Description Rationale:

Manually entered data can suffer from a high error rate. Double data entry and/or double checking of data can dramatically improve data quality.
Additional Information:
Examples:
Tags: assure, data consistency, quality

Ensure basic quality control

Quality control practices are specific to the type of data being collected, but some generalities exist:

Data collected by instruments:
Values recorded by instruments should be checked to ensure they are within the sensible range of the instrument and the property being measured. Example: Concentrations cannot be
Analytical results:
Values measured in the laboratory should be checked to ensure that they are within the detection limit of the analytical method and are valid for what is being measured. If values are below the detection limit, they should be properly coded and qualified. Any ancillary data used to assess data quality should be described and stored. Example: data used to compare instrument readings against known standards.

Observations (such as bird counts or plant cover):
Range checks and comparisons with historic maxima will help identify anomalous values that require further investigation.
Comparing current and past measurements help identify highly unlikely events. For example, it is unlikely that the girth of a tree will decrease from one year to the next.

Codes should be used to indicate quality of data.
Codes should be checked against the list of allowed values to validate code entries
When coded data are digitized, they should be re-checked against the original source. Double data entry, or having another person check and validate the data entered, is a good mechanism for identifying data entry errors.

Dates and times:
Ensure that dates and times are valid
Time zones should be clearly indicated (UTC or local)

Data Types:
Values should be consistent with the data type (integer, character, datetime) of the column in which they are entered. Example: 12-20-2000A should not be entered in a column of dates).
Use consistent data types in your data files. A database, for instance, will prevent entry of a string into a column identified as having integer data.

Geographic coordinates:
Map coordinates to detect errors

Description Rationale:

Quality control procedures identify potential problems with data that could affect its use.
Additional Information:

Chapman, Arthur D. 2005. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

US Environmental Protection Agency. 2002. Guidance on Environmental Data Verification and Data Validation. http://www.epa.gov/quality/qs-docs/g8-final.pdf.
Examples:
Tags: assure, coding, quality

Ensure datasets used are reproducible

When searching for data, whether locally on one's machine or in external repositories, one may use a variety of search terms. In addition, data are often housed in databases or clearinghouses where a query is required in order access data. In order to reproduce the search results and obtain similar, if not the same results, it is necessary to document which terms and queries were used.

Note the location of the originating data set
Document which search terms were used
Document any additional parameters that were used, such as any controls that were used (pull-down boxes, radio buttons, text entry forms)
Document the query term that was used, where possible
Note the database version and/or date, so you can any limit newly-added data sets since the query was last performed
Note the name of the website and URL, if applicable

Description Rationale:

In order to reproduce a data set or result set, it is necessary to document which terms were originally used to capture that data. By documenting this information while the search is being conducted, one greatly enhances the chance of being able to reproduce the results at a later date.
Additional Information:
Examples:

A query example can either be formatted as a URL string (http://www.google.com/#sclient=psy&hl=en&site=&source=hp&q=data+backup+p...) or a database string (select * from database_name where collection_data IS NOT NULL).
Tags: analyze, assure, data archives, data processing, discover, provenance, replicable data

---

Ensure flexible data services for virtual datasets

In order for a large dataset to be effectively used by a variety of end users, the following procedures for preparing a virtual dataset are recommended:
Identify data service users
Define data access capabilities needed by community(s) of users. For example:

Spatial subsetting
Temporal subsetting
Parameter subsetting
Coordinate transformation
Statistical characterization

Define service interfaces based upon Open Standards. For example:

Open Geospatial Consortium (OGC WMS, WFS, WCS)
W3C (SOAP)
IETF (REST – derived from Hypertext Transfer Protocol [HTTP])

Publish service metadata for published services based upon Open Standards. For example:

Web Services Definition Language (WSDL)
RSS/Atom (see Service Casting reference below for an example of a model for publishing service metadata for a variety of service types)

Description Rationale:

Some datasets are too large to efficiently deliver in their entirety, or are not directly usable by some users. To enable their effective use by a variety of end users, data collections may be published as "virtual" datasets that are extracted and/or processed based upon source data and pre-defined functions that deliver products derived from the source data.
Additional Information:

Web Service Definition Language: http://www.w3.org/TR/wsdl
Service Casting via RSS/Atom: http://wiki.esipfed.org/index.php/Atom_ServiceCasting_to_advertise_Web_S...
Examples:
Tags: data archives, data services, describe, preserve

Ensure integrity and accessibility when making backups of data

For successful data replication and backup:

Users should ensure that backup copies have the same content as the original data file.

Calculate a checksum for both the original and the backup copies and compare; if different back up the file again MD5: algorithm to determine check sum http://en.wikipedia.org/wiki/MD5

Compare files to ensure that there are no differences

Document all procedures (e.g., compression / decompression process) to ensure a successful recovery from a backup copy
To check the integrity of the backup file, periodically retrieve your backup file, open it on a separate system, and compare to the original file.
A data backup is only valuable if it is accessible. When access to a data backup is required, the owner of the backup may not be available. It is important that others know how to access the backup, otherwise the data may not be accessible for recovery. It is important to know the "who, what, when, where, and how" of the backups:

Have contact information available for the person responsible for the data
Ensure that those who need access to backups have proper access
Communicate what data is being backed up
Note how often the data is backed up and where that particular backup is located including

physical location (machine, office, company)
file system location

Be aware that there may be different backup procedures for different data sets:

Not all backups may be located in the same location
Depending upon the backup schedule, each iteration of the backup may be located in different locations (for example, more recent backups may be located on-site and older backups may be located off-site)

Have instructions and training available so that others know how to pull the backup and access the necessary data in case you are unavailable

Description Rationale:

For successful preservation a backup data file should contain the same information as the original.
Additional Information:

Data Management and Publishing (MIT Libraries) http://libraries.mit.edu/guides/subjects/data-management/
Examples:
Tags: access, backup, data archives, preserve, quality, restore

---

Ensure the reliability of your storage media

All storage media, whether hard drives, discs or data tapes, will wear out over time, rendering your data files inaccessible. To ensure ongoing access to both your active data files and your data archives, it is important to continually monitor the condition of your storage media and track its age. Older storage media and media that show signs of wear should be replaced immediately. Use the following guidelines to ensure the ongoing integrity and accessibility of your data:

Test Your Storage Media Regularly: As noted in the “Backup Your Data” best practice, it is important to routinely perform test retrievals or

restorations of data you are storing for extended periods on hard drives, discs or tapes. It is recommended that storage media that is used infrequently be tested at least once a year to ensure the data is accessible.

Beware of Early Hardware Failures: A certain percentage of storage media will fail early due to manufacturing defects. In particular, hard drives, thumb drives and data tapes that have electronic or moving parts can be susceptible to early failure. When putting a new drive or tape into service, it is advisable to maintain a redundant copy of your data for 30 days until the new device "settles in."

Determine the Life of Your Hard Drives: When purchasing a new drive unit, note the Mean Time Between Failure (MTBF) of the device, which should be listed on its specifications sheet (device specifications are usually packaged with the unit, or available online). The MTBF is expressed in the number of hours on average that a device can be used before it is expected to fail. Use the MTBF to calculate how long the device can be used before it needs to be replaced, and note that date on your calendar (For example, if the MTBF of a new hard drive is 2,500 hours and you anticipate having the unit powered on for 8 hours a day during the work week, the device should last about 2 years before it needs to be replaced).

Routinely Inspect and Replace Data Discs: Contemporary CD and DVD discs are generally robust storage media that will fail more often from mishandling and improper storage than from deterioration. However lower quality discs can suffer from delamination (separation of the disc layers) or oxidation. It is advisable to inspect discs every year to detect early signs of wear. Immediately copy the data off of discs that appear to be warping or discolored. Data tapes are susceptible both to physical wear and poor environmental storage conditions. In general, it is advisable to move data stored on discs and tapes to new media every 2-5 years (specific estimates on media longevity are available on the web).
Handle and Store Your Media With Care: All storage media types are susceptible to damage from dust and dirt exposure, temperature extremes, exposure to intense light, water penetration (more so for tapes and drives than discs), and physical shock. To help prolong its operational life, store your media in a dry environment with a comfortable and stable room temperature. Encapsulate all media in plastic during transportation. Provide cases or plastic sheaths for discs, and avoid handling them excessively.

Description Rationale:

Successful preservation depends in great part on storage media that are in good physical and operational condition.
Additional Information:

"Longevity: How Long Do CDs/DVDs/Tapes Last?" DigitalFAQ.com. Last updated April, 2008.
http://www.digitalfaq.com/guides/media/longevity.htm

"Reliability and Availability Basics." EventHelix.com. Accessed 2011-05-11.
http://www.eventhelix.com/RealtimeMantra/FaultHandling/reliability_avail...
Examples:
Tags: access, backup, data archives, disaster recovery, preserve, restore, storage

---

Identify and use relevant metadata standards

Many times significant overlap exists among metadata content standards. You should identify those standards that include the fields needed to describe your data. In order to describe your data, you need to decide what information is required for data users to discover, use, and understand your data. The who, what, when, where, how, why, and a description of quality should be considered. The description should provide enough information so that users know what can and cannot be done with your data.

Who: The person and/or organization responsible for collecting and processing the data. Who should be contacted if there are questions about your data?
What: What parameters were measured or observed? What are the units of your measurements or results?
When: A description of the temporal characteristics of your data (e.g., time support, spacing, and extent).
Where: A description of the spatial characteristics of your data (e.g., spatial support, spacing, and extent). What is the geographic location at which the data were collected? What are the details of your field sensor deployment.
How: What methods were used (e.g., sensors, analytical instruments, etc.). Did you collect physical samples or specimens? What analytical methods did you use to measure the properties of your samples/specimens? Is your result a field or laboratory result? Is your result an observation or a model simulation?
Why: What is the purpose of the study or the data collection? This can help others determine whether your data is fit for their particular purpose or not.
Quality: Describe the quality of the data, which will help others determine whether your data is fit for their purpose or not.

Considering a number of metadata content standards may help you fine-tune your metadata content needs. There may be content details or elements from multiple standards that can be added to your requirements to help users understand your data or methods. You wouldn't know this unless you consider multiple content standards.

If the project or grant requirements define a particular metadata standard, incorporate it into the data management plan
If the community has a recommended or has a most commonly used metadata standard, use it
Consider using a metadata standard that is interoperable with many systems, repositories, and harvesters
If the community's preferred metadata standard is not widely interoperable, consider creating metadata using a simple but interoperable standard, e.g. Dublin Core, in addition to the main standard.

Useful Definitions:

Metadata Content Standard: A Standard that defines elements users can expect to find in metadata and the names and meaning of those elements.
Metadata Format Standard: A Standard that defines the structures and formats used to represent or encode elements from a content standard.

Description Rationale:

The metadata standards that you use determine the communities that can easily use your data.
Additional Information:

Beall, Jeffrey. 2007. "Metadata for Digitization Projects: Discrete Criteria for Selecting and Comparing Metadata Schemes - Reflecting further on schema selection, Jeffrey enumerates twelve points of comparison to help one decide which of the many schemas available best suits one digital project". Against the Grain. 19 (1): 28.

Eichenlaub, N. 2010. "Metadata for Digital Resources: Implementation, Systems Design and Interoperability, by Muriel Foulonneau and Jenn Riley". CATALOGING AND CLASSIFICATION QUARTERLY. 48 (4): 348-351.

UK Digital Curation Centre (DCC) Disciplinary Metadata Catalog
http://www.dcc.ac.uk/resources/metadata-standards

Seeing Standards: A Visualization of the Metadata Universe
http://www.dlib.indiana.edu/~jenlrile/metadatamap/

Xu W, and M Okada. 2007. "EBM metadata based on Dublin Core better presenting validity of clinical trials". Journal of Medical Systems. 31 (5): 337-43.
Examples:
Tags: controlled vocabulary, describe, documentation, format, metadata, preserve

---

Identify data sensitivity

Steps for the identification of the sensitivity of data and the determination of the appropriate security or privacy level are:

Determine if the data has any confidentiality concerns
Can an unauthorized individual use the information to do limited, serious, or severe harm to individuals, assets or an organization's operations as a result of data disclosure?
Would unauthorized disclosure or dissemination of elements of the data violate laws, executive orders, or agency regulations (i.e., HIPPA or Privacy laws)?
Does the data have any integrity concerns?
What would be the impact of unauthorized modification or destruction of the data?
Would it reduce public confidence in the originating organization?
Would it create confusion or controversy in the user community?
Could a potentially life-threatening decision be made based on the data or analysis of the data?
Are there any availability concerns about the data?
Is the information time-critical? Will another individual or system be relying on the data to make a time-sensitive decision (i.e. sensing data for earthquakes, floods, etc.)?

Document data concerns identified and determine overall sensitivity (Low, Moderate, High)
Low criticality would result in a limited adverse effect to an organization as a result of the loss of confidentiality, integrity, or availability of the data. It might mean degradation in mission capability or result in minor harm to individuals.
Moderate criticality would result in a serious adverse effect to an organization as a result of the loss of confidentiality, integrity, or availability of the data. It might mean a severe degradation or loss of mission capability or result in significant harm to individuals that does not involve loss of life or serious life threatening injuries.
High criticality would result in a severe or catastrophic adverse effect as a result of the loss of confidentiality, integrity, or availability of the data. It might cause a severe degradation in or loss of mission capability or result in severe or catastrophic harm to individuals involving loss of life or serious life threatening injuries.

Develop data access and dissemination policies and procedures based on sensitivity of the data and need-to-know.
Develop data protection policies, procedures and mechanisms based on sensitivity of the data.

Description Rationale:

The identification of the sensitivity and importance of data or information processed on an information system is essential to the determination of the appropriate security and privacy considerations to ensure the confidentiality, integrity, and availability of the data as well as data sharing decisions.
Additional Information:

FIPS Pub 199, Standards for Security Categorization of Federal Information and Information Systems:
http://csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199-final.pdf

Guide for Mapping Types of Information and Information Systems to Security Categories: (2 Volumes) - Volume 1: Guide Volume 2: Appendices
http://csrc.nist.gov/publications/nistpubs/800-60-rev1/SP800-60_Vol1-Rev...

http://csrc.nist.gov/publications/nistpubs/800-60-rev1/SP800-60_Vol2-Rev...
Examples:
Tags: access, data archives, plan, preserve

Identify data with long-term value

As part of the data life cycle, research data will be contributed to a repository to support preservation and discovery. A research project may generate many different iterations of the same dataset - for example, the raw data from the instruments, as well as datasets which already include computational transformations of the data.

In order to focus resources and attention on these core datasets, the project team should define these core data assets as early in the process as possible, preferably at the conceptual stage and in the data management plan. It may be helpful to speak with your local data archivist or librarian in order to determine which datasets (or iterations of datasets) should be considered core, and which datasets should be discarded. These core datasets will be the basis for publications, and require thorough documentation and description.

Only the datasets which have significant long-term value should be contributed to a repository, requiring decisions about which datasets need to be kept.
If data cannot be recreated or it is costly to reproduce, it should be saved.
Four different categories of potential data to save are observational, experimental, simulation, and derived (or compiled).
Your funder or institution may have requirements and policies governing contribution to repositories.

Given the amount of data produced by scientific research, keeping everything is neither practical nor economically feasible.
Description Rationale:

Decisions about what data to keep will help to focus project resources on those data that should be stored for long-term preservation.
Additional Information:

Whyte, Angus. Appraise and Select Research Data for Curation. Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides/appraise-select-research-data
Examples:
Tags: data archives, preserve, storage

Identify missing values and define missing value codes

Missing values should be handled carefully to avoid their affecting analyses. The content and structure of data tables are best maintained when consistent codes are used to indicate that a value is missing in a data field. Commonly used approaches for coding missing values include:

Use a missing value code that matches the reporting format for the specific parameter. For example, use ""-999.99"", when the reporting format is a FORTRAN-like F7.2.
For character fields, it may be appropriate to use ""Not applicable"" or ""None"" depending upon the organization of the data file.
It might be useful to use a placeholder value such as ""Pending assignment"" when compiling draft information to facilitate returning to incomplete fields.
Do not use character codes in an otherwise numeric field.

Whatever missing value is chosen, it should be used consistently throughout all data associated files and identified in the metadata and/or data description files.
Description Rationale:

Missing values are common in environmental data and affect the interpretation, analysis and calculations. Therefore, they need to be carefully defined and properly described. In addition many instruments will automatically add missing value codes in their datastream which will have to be dealt with for storage and analysis.
Additional Information:

Monitoring programs like EPA, USGS, etc. have online documentation on how do handle missing values and can be consulted.

L. A. Hook, L.A., T.W. Beaty, S. SanthanaVannan, L. Baskaran, and R. B. Cook. 2007. Best Practices for Preparing Environmental Data Sets to Share and Archive.
Cook et al., 2001 ""Best Practices for Preparing Ecological and Ground-Based Data Sets to Share and Archive“ Bulletin of ESA 82: 138-141.
Borer et al. 2009. Some Simple Guidelines for Effective Data Management. Bull. of ESA 90: 209-214.
Examples:
Tags: assure, coding, missing values

Identify most appropriate software

Follow the steps below to choose the most appropriate software to meet your needs.

Identify what you want to achieve (discover data, analyze data, write a paper, etc.)
Identify the necessary software features for your project (i.e. functional requirements)
Identify logistics features of the software that are required, such as licensing, cost, time constraints, user expertise, etc. (i.e. non-functional requirements)
Determine what software has been used by others with similar requirements

Ask around (yes, really); find out what people like
Find out what software your institution has licensed
Search the web (e.g. directory services, open source sites, forums)
Follow-up with independent assessment

Generate a list of software candidates
Evaluate the list; iterate back to Step 1 as needed
As feasible, try a few software candidates that seem promising

Description Rationale:

By carefully considering software choice before a project begins, costs, requirements, and limitations can be addressed early in the data life cycle.
Additional Information:

Open source sites: sourceforge.net, http://git-scm.com, etc.
Forums: http://stackoverflow.com

Examples:
Tags: analyze, data processing, data services

Identify outliers

Outliers may not be the result of actual observations, but rather the result of errors in data collection, data recording, or other parts of the data life cycle. The following can be used to identify outliers for closer examination:

Statistical determination:
Outliers may be detected by using Dixon's test, Grubbs test or the Tietjen-Moore test.

Visual determination:
Box plots are useful for indicating outliers
Scatter plots help identify outliers when there is an expected pattern, such as a daily cycle

Comparison to related observations:
Difference plots for co-located data streams can show unreasonable variation between data sources. Example: Difference plots from weather stations in close proximity or from redundant sensors can be constructed. Comparisons of two parameters that should covary can indicate data contamination. Example: Declining soil moisture and increasing temperature are likely to result in decreasing evapotranspiration.

No outliers should be removed without careful consideration and verification that they are not representing true phenomena.
Description Rationale:

Outliers may represent data contamination, a violation of the assumptions of the study, or failure of the instrumentation. Although outliers may be valid observations it is important to identify and examine their validity.
Additional Information:

V. Barnett and T. Lewis, Outliers in Statistical Data (John Wiley & Sons, 2d ed., New York, NY, 1985).

Edwards, D. 2000. Data Quality Assurance. Pages 70-91 in: Ecological Data: design, management, and processing. Michener, W. and Brunt, J., eds. Blackwell Science Ltd. (ISBN: 0-682-05231-7)
Examples:
Tags: analyze, annotation, assure, quality

Identify suitable repositories for the data

Shaping the data management plan towards a specific desired repository will increase the likelihood that the data will be accepted into that repository and increase the discoverability of the data within the desired repository. When beginning a data management plan:

Look to the data management guidelines of the project/grant for a required repository
Ask colleagues what repositories are used in the community
Determine if your local institution has a repository that would be appropriate (and might be required) for depositing your data
Check the DataONE website for a list of potential repositories.

Description Rationale:

Shaping the data management plan towards a specific desired repository will increase the likelihood that the data will be accepted into that repository and increase the discoverability of the data within the desired repository.
Additional Information:

Digging into Data List of Repositories: http://www.diggingintodata.org/Repositories/tabid/167/Default.aspx
Registry of Research Data Repositories: http://www.r3data.org
Simmons list of repositories: http://oad.simmons.edu/oadwiki/data_repositories
Examples:

Scenario: An ornithologist is creating a data management plan for migratory bird data. She would like to ultimately put her data in eBird. She looks at the repository to find its rules for data management to incorporate in her plan.
Tags: access, data archives, plan, preserve, storage

Identify values that are estimated

Data tables should ideally include values that were acquired in a consistent fashion. However, sometimes instruments fail and gaps appear in the records. For example, a data table representing a series of temperature measurements collected over time from a single sensor may include gaps due to power loss, sensor drift, or other factors. In such cases, it is important to document that a particular record was missing and replaced with an estimated or gap-filled value.

Specifically, whenever an original value is not available or is incorrect and is substituted with an estimated value, the method for arriving at the estimate needs to be documented at the record level. This is best done in a qualifier flag field. An example data table including a header row follows:

Day, Avg Temperature, Flag
1, 31.2, actual
2, 32.3, actual
3, 33.4, estimated
4, 35.8, actual

Description Rationale:

Correct interpretation of the content of a data table typically depends on knowing which data values were actually recorded in the field or estimate using other approaches.
Additional Information:

Hook, Les A., Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson. 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. Available online (http://daac.ornl.gov/PI/BestPractices-2010.pdf) from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/BestPractices-2010
Examples:
Tags: analyze, assure, flag, quality

Maintain consistent data typing

Choose the right data type and precision for data in each column. As examples: (1) use date fields for dates; and (2) use numerical fields with decimal places precision. Comments and explanations should not be included in a column that is meant to include numeric values only. Comments should be included in a separate column that is designed for text. This allows users to take advantage of specialized search and computing functionality and improves data quality. If a particular spreadsheet or software system does not support data typing, it is still recommended that one keep the data type consistent within a column and not mix numbers, dates and text.
Description Rationale:

Strict data typing provides quality control and enables extended analytical procedures such as date calculations and quality assurance procedures.
Additional Information:
Examples:
Tags: database, describe, documentation, format, metadata

Mark data with quality control flags

As part of any review or quality assurance of data, potential problems can be categorized systematically. For example data can be labeled as 0 for unexamined, -1 for potential problems and 1 for "good data." Some research communities have developed standard protocols; check with others in your discipline to determine if standards for data flagging already exist.

The marine community has many examples of quality control flags that can be found on the web. There does not yet seem to be standards across the marine or terrestrial communities.
Description Rationale:

Data quality should be able to be assessed by potential data users so that any problems are identified early in a project.

Additional Information:
Examples:

Some marine examples - Several standards and cross walk between standards at
solar radiation data http://solardat.uoregon.edu/QualityControlFlags.html
www.oceandatastandards.org
Tags: assure, coding, data quality, flag

Plan data management early in your project

A Data Management Plan should include the following information:

Types of data to be produced and their volume

Who will produce the data

Standards that will be applied

File formats and organization, parameter names and units, spatial and temporal resolution, metadata content, etc.

Methods for preserving the data and maintaining data integrity

What hardware / software resources are required to store the data
How will the data be stored and backed up
Describe the method for periodically checking the integrity of the data

Access and security policies;

What access requirements does your sponsor have
Are there any privacy / confidentiality / intellectual property requirements
Who can access the data:

During active data collection
When data are being analyzed and incorporated into publications
When data have been published
After the project ends
How should the data be cited and the data collectors acknowledged

Plans for eventual transition of the data to an archive after the project ends

Identify a suitable data center within your discipline
Establish an agreement for archival
Understand the data center's requirements for submission and incorporate into data management plan

Description Rationale:

When you develop hypotheses and the design of sample collection for your new project, you should also plan for data management. Careful planning for data management before you begin your research and throughout the data's life cycle is essential to improve the data's usability, and ensure data's preservation and access both during the project and well into the future.
Additional Information:

Australian National University Data Management Planning: http://ilp.anu.edu.au/dm/
MIT Data Management Planning: http://libraries.mit.edu/data-management/
UK Data Archive Data Management and Sharing Plan: http://www.data-archive.ac.uk/news/publications/managingsharing.pdf
Examples:
Tags: backup, plan, preserve

Plan for effective multimedia management

Multimedia data present unique challenges for data discovery, accessibility, and metadata formatting and should be thoughtfully managed. Researchers should establish their own requirements for management of multimedia during and after a research project using the following guidelines. Multimedia data includes still images, moving images, and sound. The Library of Congress has a set of web pages discussing many of the issues to be considered when creating and working with multimedia data. Researchers should consider quality, functionality and formats for multimedia data. Transcriptions and captioning are particularly important for improving discovery and accessibility.

Storage of images solely on local hard drives or servers is not recommended. Unaltered images should be preserved at the highest resolution possible. Store original images in separate locations to limit the chance of overwriting and losing the original image.

Ensure that the policies of the multimedia repository are consistent with your general data management plan.

There are a number of options for metadata for multimedia data, with many MPEG standards (http://mpeg.chiariglione.org/), and other standards such as PBCore (http://pbcore.org/).

The following web pages have sections describing considerations for quality and functionality and formats for each of still images, sound (audio) and moving images (video).

Sustainability of Digital Formats Planning for Library of Congress Collections:

Still Images
Sound
Moving Images

Online, generic multimedia repositories and tools (e.g. YouTube, Vimeo, LIFE)

provide domain-specific metadata fields and controlled vocabularies customized for expert users
are highly discoverable for those in the same domain
can provide assistance in curating metadata
optimize scientific use cases such as vouchering, image analysis
rely on research or institutional/federal funding
may require high-quality multimedia, completeness of metadata, or restrict manipulation
may not be open to all
may provide APIs for sharing or re-use for other projects
are recognized as high-quality, scientific repositories
may migrate multimedia to new formats (e.g. analog to digital)
may have restrictions on bandwidth usage

Some institutions or projects maintain digital asset management systems, content management systems, or other collections management software (e.g. Specify, KE Emu) which can manage multimedia along with other kinds of data

projects or institutions should provide assistance
may be mandated by institution
may be more convenient, e.g. when multiple data types result from a project
may not be optimized for discovery, access, or re-use
usually not domain-specific
may or may not be suitable for long-term preservation

Description Rationale:

Multimedia metadata is particularly important for discovery as the objects do not contain text that can be indexed.
Additional Information:

"Multimedia Semantics - The Role of Metadata": http://www.springer.com/engineering/computational+intelligence+and+compl...
"Multimedia Semantics: Metadata, Analysis and Interaction": http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470747005.html
Examples:

http://www.metadataworkinggroup.org/pdf/mwg_guidance.pdf
http://images.nbii.gov/life/pdf/Acceptable.pdf
http://macaulaylibrary.org/building-the-archive
Tags: documentation, format, image, metadata, plan, preserve, storage

---

Preserve information: keep your raw data raw

In order to preserve the raw data for future use:

Do not make any changes / corrections to the original raw data file

Use a scripted language (e.g., R) or a software language that can be documented (eg., C, Java, Python, etc.) to perform analysis or make corrections and save that information in a separate file

The code, along with appropriate documentation will be a record of the changes
The code can be modified and rerun, using the raw data file as input, if needed

Consider making your original data file read-only, so it cannot be inadvertently altered.
Avoid spreadsheet software and other Graphical User Interface-based software. They may seem convenient, but changes are made without a clear record of what was done or why. Spreadsheets provide incredible freedom and power for manipulating data, but if used inappropriately can create tremendous problems. For this reason special attention needs to be paid to adhering to best practices in organizing data in spreadsheets. Particularly important best practices that are also highlighted elsewhere are:

Data should be organized in columns with each column representing only a single type of data (number, date, character string. An exception to this is that sometimes a header line containing column names (sometimes called variable or field names) may be placed at the top of a column.
Each data line should be complete, that is, each line of the data should contain data for each column. Sometimes in spreadsheets, to promote human readability, values will be provided only when they change. However, if the data is sorted, the relationships would become scrambled. An exception to this rule is if a data item is really missing (and not just omitted for human readability)a missing value code might be used.

Additional best practices regarding consistent use of codes for categorical variables, and informative field names also apply, but keeping the data in consistent and complete columns are the most important.

A key test is whether the data from a spreadsheet can be exported as a delimited text file, such as a comma-separated-value (.csv) file that can be read by other software. If columns are not consistent the resulting data may cause software such as relational databases (e.g., MySQL, Oracle, Access) or statistical software (e.g., R, SAS, SPSS) to record errors or even crash.

As a general rule, spreadsheets make a poor archival data format. Standards for spreadsheet file formats change frequently or go out of fashion. Even within a single software package (e.g., Excel) there is no guarantee that future versions of the software will read older file versions. For this reason, and as specified in other best practices, generic (e.g., text) formats such as comma-separated-value files are preferred.

Sometimes it is the formulae embedded in a spreadsheet, rather than the data values themselves that are important. In this case, the spreadsheet itself may need to be archived. The danger of the spreadsheet being rendered obsolete or uninterpretable may be reduced by exporting the spreadsheet in a variety of forms (e.g., both as .xls and as .xlsx formats). However the long-term utility of the spreadsheet may still depend on periodic opening of the archived spreadsheet and saving it into new forms.

Upgrades and new versions of software applications often perform conversions or modifications to data files produced in older versions, in many cases without notifying the user of the internal change(s).

Many contemporary software applications that advertise forward compatibility for older files actually perform significant modifications to both visible and internal file contents. While this is often not a problem, there are cases where important elements like numerical formulas in a spreadsheet, are changed significantly when they are converted to become compatible with a current software package. The following practices will help ensure that your data files maintain their original fidelity in the face of application updates and new releases:

Where practical, continue using the version of the software that was originally used to create the data file to view and manipulate the file contents (For example, if Excel 97 was used to create a spreadsheet that contains formulas and formatting, continue using Excel 97 to access those data files as long as possible).
When forced to use a newer version of a software package to open files created with an older version of the application, first save a copy of the original file as a safeguard against irretrievable modification or corruption.
Carefully inspect older files that have been opened/converted to be compatible with newer versions of an application to ensure data fidelity has been carried forward. Where possible, compare the converted files to copies of the original files to ensure there have been no data modifications during conversion.

Description Rationale:

Preserve the information content of your original raw data.
Additional Information:

Borer et al. 2009. Some Simple Guidelines for Effective Data Management. Bull. of ESA 90: 209-214
Examples:
Tags: collect, data consistency, format, preserve

Provide a citation and document provenance for your dataset

For appropriate attribution and provenance of a dataset, the following information should be included in the data documentation or the companion metadata file:

Name the people responsible for the dataset throughout the lifetime of the dataset, including for each person:

Name
Contact information
Role (e.g., principal investigator, technician, data manager)

According to the International Polar Year Data and Information Service, an author is the individual(s) whose intellectual work, such as a particular field experiment or algorithm, led to the creation of the dataset. People responsible for the data can include: individuals, groups, compilers or editors.

Description of the context of the dataset with respect to a larger project or study (include links and related documentation), if applicable.
Revision history, including additions of new data and error corrections.
Links to source data, if the data in one dataset were derived from data in another dataset.
List of project support (e.g., funding agencies, collaborators, material support).
Describe how to properly cite the dataset. The data citation should include:

All contributors
date of dataset publication
Title of dataset
media or URL
Data publisher
Identifier (Digital Object Identifier)

Description Rationale:

Documenting the dataset origin, history, and contact information allows for proper citation of datasets. By encouraging the proper citation of datasets, data providers and publishers receive appropriate credit for their efforts.
Additional Information:

The Oak Ridge National Laboratory Distributed Active Archive Center has guidance and rational for citing data sets:
Editorial: Citations to Published Data Sets.

Buneman P, Khanna S, Tan W. 2001. Why and Where: A Characterization of Data Provenance. Pp. 316-330 in Lecture Notes in Computer Science. Springer Berlin/Heidelberg.
Osterweil LJ, Clarke LA, Ellison AM, Boose E, Podorozhny R, Wise A. 2010. Clear and precise specification of ecological data management processes and dataset provenance. IEEE Transations on Automation Science and Engineering 7(1):189-195.
Simmhan YL, Plale B, Gannon D. 2005. A survey of data provenance in e-science. ACM SIGMOD 34(3):31-36.
Examples:

Turner, D.P., W.D.Ritts, and M. Gregory. 2006. BigFoot NPP Surfaces for North and South American Sites, 2002-2004. Data set. Available on-line (http://daac.ornl.gov) from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/750.
Tags: citation, data creators, data source, describe, preserve, provenance

---

Provide budget information for your data management plan

As a best practice, one must first acknowledge that the process of managing data will incur costs. Researchers should plan to address these costs and the allocation of resources in the early planning phases of the project. This best practice focuses on data management costs during the life cycle of the project, and does not aim to address costs of data beyond the end of the project.

Budgeting and costing for your project is dependent upon institutional resources, services, and policies. We recommended that you verify with your sponsored project office, your office of research, tech transfer resources, and other appropriate entities at your institution to understand resources available to you.

There are a variety of approaches to budgeting for data management costs. All approaches should address the following costs in each phase:

short-term costs
long-term costs
internal/external costs
equipment/services (ie. compute cycles, storage, software, and hardware) costs
overhead costs
time costs
human resource costs

Methods for Managing Costs

In-sourced costs: items that are managed directly within the research group.
Out-sourced costs: items that are contracted or managed outside of the research group.

Phases of the Data Life Cycle (see Primer on Data Management on the DataONE website for a description of the life cycle)

Collect - Likely both in-sourced and out-sourced costs.
Coordinate with central IT services or community storage resources to ensure appropriate data storage environment and associated costs during this phase or throughout the life of the project.
Assure - Likely in-sourced costs. This phase is primarily focused on quality assurance/control, and costs will primarily be incurred around time and personnel.

Describe - Likely in-sourced costs. This phase includes initial and ongoing documentation as well as continuous development of metadata. Documentation captures the entire structure of the project, all configurations/parameters, as well as all processes during the course of the entire project. See the Documentation and Metadata best practices for more detail on what should be addressed.
Deposit - Likely both in-sourced and out-sourced costs.
Preserve - Likely both in-sourced and out-sourced costs. Coordinate with central IT services or community repository environments that are equipped to provide preservation services. This phase will be tied closely to the costs of the collection phase.
Discover - Likely in-sourced costs. Coordinate with librarians, IT service providers, or repository providers to identify and access data sources.
Integrate - Likely in-sourced costs. Coordinate with IT service providers or other service groups to merge and prepare data sources for analysis phase.
Analyze - Likely in-sourced costs. Coordinate with central IT services or other workspace providers to connect data sources with appropriate analysis and visualization software.

Description Rationale:

This best practice activity is intended to articulate the necessary steps involved in budgeting and resourcing a research data management plan.
Additional Information:

- Charles Beagrie Ltd and JISC. “Keeping Research Data Safe Factsheet: Cost issues in digital preservation of research data.” Neil Beagrie’s Blog. Last modified Sepetember 18, 2010. http://www.beagrie.com/KRDS_Factsheet_0910.pdf
- “Activity-based data management costing tool for researchers.” UK Data Archive. Last modified http://www.data-archive.ac.uk/media/257647/ukda_jiscdmcosting.pdf

Examples:
Tags: data management plan, integrate, plan

Provide capabilities for tagging and annotation of your data by the community

People have different perspectives on what data means to them, and how it can be used and interpreted in different contexts. Data users ranging from community participants to researchers in different domains can provide unique and valuable insights into data through the use of annotation and tagging. The community-generated notes and tags should be discoverable through the data search engine to enhance discovery and use.

When providing capabilities for community tagging and annotations, you should consider the following:

Differentiate between the metadata developed by the creator and additional tags or annotations to the data or metadata
Allow for community tags and annotations to be indexed as part of the terms or text that is indexed in a search
Provide easy-to-understand examples of the kinds of tagging or annotation that will promote the discovery of your data
Consider whether or not a review process for community tagging is needed
Consider whether controlled vocabularies will be used for tags
Provide clear guidelines for the addition of tags and construction of annotations
Make tags accessible via an application programming interface (API)

Description Rationale:

Others' perceptions and views of your dataset might differ; by providing the ability for the community to annotate your data, the data are more likely to be discovered by a broad range of potential users.
Additional Information:
Examples:

http://www.galaxyzoo.org/
http://www.myexperiment.org/
http://www.flickr.com/
http://www.plosbiology.org/static/ratingGuidelines.action
http://www.youtube.com/
Tags: annotation, controlled vocabulary, describe, documentation, metadata

---

Provide identifier for dataset used

In order to ensure replicable data access:

Choose a broadly utilized Data Identification Standard based on specific user community practices or preferences

DOI
OIDs
ARKs
LSIDs
XRIs
URNs/URIs/URLs
UUIDs)

Consistently apply the standard
Maintain the linkage
Participate in implementing infrastructure for consistent access to the resources referenced by the Identifier

Description Rationale:

Digital objects should utilize a standard, stable identifier to access a specific data product to ensure data consistency among analyses that use the "same" data product.
Additional Information:

ESIP Federation Preservation and Stewardship Cluster: http://wiki.esipfed.org/index.php/Preservation_and_Stewardship#Data_Iden...
Examples:
Tags: access, data consistency, describe, preserve, provenance, replicable data

---

Provide version information for use and discovery

Provide versions of data products with defined identifiers to enable discovery and use

Items to consider when versioning data products:

Develop definition of what constitutes a new version of the data, for example:

New processing algorithms
Additions or removal of data points
Time or date range
Included parameters
Data format
Immutability of versions

Develop standard naming convention for versions with associated descriptive information
Associate metadata with each version including the description of what differentiates this version from another version

Description Rationale:

Data products potentially change through time as they are developed using new or improved algorithms or based upon different source data. Providing versions with defined identifiers will enable users to determine the appropriate version of the data to use for their particular application and allow them to properly cite the data used in an analysis allowing others to replicate their work.
Additional Information:

NASA Modis versioning using processing levels and collections:

- https://lpdaac.usgs.gov/lpdaac/products/modis_products_table
- https://lpdaac.usgs.gov/lpdaac/products/modis_products_table/vegetation_...
- http://landweb.nascom.nasa.gov/QA_WWW/forPage/MOD13_VI_C5_Changes_Docume...

Examples:
Tags: assure, documentation, metadata, provenance, quality

Recognize stakeholders in data ownership

When creating the data management plan, review all who may have a stake in the data so future users of the data can easily track who may need to give permission. Possible stakeholders include but are not limited to:

Funding body
Host institution for the project
Home institution of contributing researchers
Repository where the data are deposited

It is considered a matter of professional ethics to acknowledge the work of other scientists and provide appropriate citation and acknowledgment for subsequent distribution or publication of any work derived from stakeholder datasets. Data users are encouraged to consider consultation, collaboration, or co-authorship with original investigators.
Description Rationale:

Researchers are more willing to share their data if they are receive appropriate recognition for their contributions, and project sponsors, host institutions, data repositories, and researchers' home institution may all have a stake in data ownership and require acknowledgment.
Additional Information:

Eisenberg, Rebecca S. R. A. K. "Harnessing and Sharing the Benefits of State-Sponsored Research: Intellectual Property Rights and Data Sharing in California's Stem Cell Initiative." Berkeley Technology Law Journal. 21.3 (2006): 1187.

Evans, James. "Industry Collaboration, Scientific Sharing, and the Dissemination of Knowledge." Social Studies of Science. 40.5 (2010): 757-791.
Examples:

Example of citation for a database: Gregory, S. 2010. Aquatic Vertebrate Population Study, Mack Creek, Andrews Experimental Forest. Long-Term Ecological Research. Forest Science Data Bank, Corvallis, OR. [Database]. Available: http://andrewsforest.oregonstate.edu/data/abstract.cfm?dbcode=AS006 (11 May 2011).
Tags: access, preserve, provenance

Revisit data management plan throughout the project life cycle

The plan will be created at the conceptual stage of the project. It should be considered a living document and a road map for the project, and should be closely followed. Any changes to the data management plan should be made deliberately, and the plan should be updated throughout the data life cycle.

Data management planning provides crucial guidance to all stages of the data life cycle. It provides continuity for operations within the research group. The data management plan will define roles for all project participants and workflows for data collection, quality assurance, description, and deposit for preservation and access. The data management plan is a tool to communicate requirements and restrictions to all members of the project team, including researchers, archivists, librarians, IT staff and repository managers. The plan governs the active research phase of the project life cycle and makes provisions for the hand-off to a repository for preservation and data delivery.

Funding agencies and institutions require data management plans for project funding and approval.
Description Rationale:

The data generated by research is seen as an increasingly valuable research output, and careful planning is required to optimize future use and reuse.
Additional Information:

DataONE. Data Management Plans. http://www.dataone.org/plans

Inter-university Consortium for Political and Social Research. http://www.icpsr.umich.edu/icpsrweb/ICPSR/dmp/resources.jsp#a06

Rice University. Office of Sponsored Research. http://osr.rice.edu/dataManagementPlans.cfm

University of California, San Diego. Research Cyberinfrastructure. http://rci.ucsd.edu/dmp/examples.html
Examples:
Tags: data management plan, documentation, plan

Separate data values from annotations

A separate column should be used for data qualifiers, descriptions, and flags, otherwise there is the potential for problems to develop during analyses. Potential entries in the descriptor column:
Potential sources of error
Missing value justification (e.g. sensor off line, human error, data rejected outside of range, data not recorded
Flags for values outside of expected range, questionable etc.

Description Rationale:

Mixing numeric and textual data in one column will cause problems with analysis. Sometimes it is necessary to have both, the value and a qualifier.
Additional Information:
Examples:
Tags: annotation, describe, documentation, flag, format, metadata

---

Sharing data: legal and policy considerations

All research requires the sharing of information and data. The general philosophy is that data are freely and openly shared. However, funding organizations and institutions may require that their investigators cite the impact of their work, including shared data. By creating a usage rights statement and including it in data documentation, users of your data will be clear what the conditions of use are, and how to acknowledge the data source.

Include a statement describing the "usage rights" management, or reference a service that provides the information. Rights information encompasses Intellectual Property Rights (IPR), copyright, cost, or various Property Rights. For data, rights might include requirements for use, requirements for attribution, or other requirements the owner would like to impose. If there are no requirements for re-use, this should be stated.

Usage rights statements should include what are appropriate data uses, how to contact the data creators, and acknowledge the data source. Researchers should be aware of legal and policy considerations that affect the use and reuse of their data. It is important to provide the most

comprehensive access possible with the fewest barriers or restrictions.

There are three primary areas that need to be addressed when producing sharable data:

Privacy and confidentiality: Adhere to your institution's policy
Copyright and intellectual property (IP): Data is not copyrightable. Ensure that you have the appropriate permissions when using data that has multiple owners or copyright layers. Keep in mind that information documenting the context of data collection may be under copyright.
Licensing: Data can be licensed. The manner in which you license your data can determine its ability to be consumed by other scholars. For example the Creative Commons Zero License provides for very broad access.

If your data falls under any of the categories below there are additional considerations regarding sharing:

Rare, threatened or endangered species
Cultural items returned to their country of origin
Native American and Native Hawaiian human remains and objects
Any research involving human subjects

If you use data from other sources, you should review your rights to use the data and be sure you have the appropriate licenses and permissions.
Description Rationale:

When sharing data, or using data shared by others, researchers should be aware of any policies that might affect the use of the data. Including a usage rights statement makes clear to data repository users what the conditions of use are, and how to acknowledge the data source.
Additional Information:

ICPSR Confidentiality Review
17 U.S.C. Sec. 102
Feist Publications, Inc v. Rural Telephone Service Co., Inc., 499 U.S. 340, 245 (1991)
NIH Data Sharing Policy
University of Virginia Library's Data Rights and Responsibilities Guidance
http://www.lternet.edu/data/netpolicy.html
Examples:

Copyright 2001 Regents of the University of California Santa Barbara. Free for use by all individuals provided that the owners are acknowledged in any use or publication.
Tags: access, citation, data creators, data source

Store data with appropriate precision

Data should not be entered with higher precision than they were collected in (e.g if a device collects data to 2dp, an Excel file should not present it to 5 dp). If the system stores data in higher precision, care needs to be taken when exporting to ASCII. E.g. calculation in excel will be done to the highest possible precision of the system, which is not related to the precision of the original data.
Description Rationale:
Additional Information:
Examples:
Tags: analyze, measurement, preserve, storage

Understand the geospatial parameters of multiple data sources

Understand the input geospatial data parameters, including scale, map projection, geographic datum, and resolution, when integrating data from multiple sources. Care should be taken to ensure that the geospatial parameters of the source datasets can be legitimately combined. If working with raster data, consider the data type of the raster cell values as well as if the raster data represent discrete or continuous values. If working with vector data, consider feature representation (e.g., points, polygons, lines). It may be necessary to re-project your source data into one common projection appropriate to your intended analysis. Data product quality degradation or loss of data product utility can result when combining geospatial data that contain incompatible geospatial parameters. Spatial analysis of a dataset created from combining data having considerably different scales or map projections may result in erroneous results.

Document the geospatial parameters of any output dataset derived from combining multiple data products. Include this information in the final data product's metadata as part of the product's provenance or origin.
Description Rationale:

Awareness and proper use of geospatial data parameters can affect the ability to interpret or produce sound analysis results.
Additional Information:

Burley, T.E., and Peine, J.D., 2009' NBII-SAIN Data Management Toolikit, U.S. Geological Survey Open-File Report 2009-1170, 96p. Available from: http://pubs.usgs.gov/of/2009/1170/
Examples:

Combining geospatial data created at a 1:24,000 scale with data that were created at a 1:100,000 scale results in data that are only as accurate as

the least accurate input (essentially resolution is lost). Combining geospatial data that have considerably differing map projections can also result in spatial errors and potentially erroneous results. Combining raster data with a 10m resolution with raster data that have a 100m resolution will result in data that are only as accurate as the least accurate input (essentially resolution is lost).
Tags: analyze, documentation, geography, geospatial, integrate, metadata, provenance

---

Use appropriate field delimiters

Delimit the columns within a data table using commas or tabs; these are listed in order of preference. Semicolons are used in many systems as line end delimiters and may cause problems if data are imported into those systems (e.g. SAS, PHP scripts). Avoid delimiters that also occur in the data fields. If this cannot be avoided, enclose data fields that also contain a delimiter in single or double quotes.

An example of a consistently delimited data file with a header row:

Date, Avg Temperature, Precipitation
01Jan2010, 32.3, 0.0
02Jan2010, 34.1, 0.5
03Jan2010, 31.4, 2.5
04Jan2010, 33.2, 0.0
Description Rationale:

Consistent use of preferred field delimiters (e.g., comma separated variables) enables data tables to be easily incorporated into analytical and other software programs and ensures that the data content and structure are preserved.
Additional Information:

Best Practices for Preparing Environmental Data Sets to Share and Archive, (formerly Cook et al., 2001) Updated by L. A. Hook, T. W. Beaty, S. Santhana-Vannan, L. Baskaran, and R. B. Cook. June 2007. http://daac.ornl.gov/PI/bestprac.html
Examples:
Tags: access, collect, describe, format

Use consistent codes

Be consistent in the use of codes to indicate categorical variables, for example species names, sites, or land cover types. Codes should always be the same within one data set Pay particular attention to spelling and case; most frequent problems are with abbreviations for species names and sites.

Consistent codes can be achieved most easily by defining standard categorical variables (codes) and using drop down lists (excel, database). Frequently a code is needed for ‘none of the above’ or ‘unknown’ or ‘other’ to avoid imprecise code assignment.
Description Rationale:

Inconsistencies in coding schemes can create confusion on the part of data users. For example, a typical user might not understand why "T", "temp", and "t" are used interchangeable within a single data file.
Additional Information:
Examples:
Tags: coding, collect, controlled vocabulary, describe, ontologies