Chapter 13
A Bigger Picture: Data Standards, Interoperability and Data Sharing

Sarah Cain Davidson

Abstract Data sharing is of growing interest in science and in ecology. Many research questions in ecology, particularly those addressing global change, require large, long-term data sets that cannot be collected by any one research group alone. Moreover, an increasing number of funding providers and publishers require that researchers make their data available in some form to other researchers or to the public. Benefits to sharing your data can include new collaborations and publications, increased citations of your research, expansion of successful wildlife management strategies to new areas or species, and fulfillment of journal and funding requirements for data sharing and management plans. As you develop your database, it is worth considering ways to share your data, either with specific collaborators or with the public, and to at minimum make a description of your data set publicly available. And, as we have emphasised throughout this book, the data organisation and documentation required for sharing data should be a standard part of data collection regardless of the end uses of your data. The goal of this chapter is to introduce you to existing ecological data standards and a variety of ways to make your database archivable and usable for additional analyses.

Keywords Data management · Data sharing · Data standards · Metadata

S. C. Davidson (✉)
Max Planck Institute for Ornithology, Am Obstberg 1, 78315 Radolfzell, Germany and
Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, 2070 Neil Ave, Columbus, OH 43210, USA
e-mail: sdavidson@orn.mpg.de

F. Urbano and F. Cagnacci (eds.), *Spatial Database for GPS Wildlife Tracking Data*,
DOI: 10.1007/978-3-319-03743-1_13,

Introduction

Data sharing is of growing interest in science and in ecology (Vision 2010; Reichman et al. 2011). Many research questions in ecology, particularly those addressing global change, require large, long-term data sets that cannot be collected by any one research group alone (Wolkovich et al. 2012). Therefore, scientists may wish to combine some or all of their data with those of collaborators, or with archived data collected in the past, to complete a new analysis. Moreover, an increasing number of funding providers and publishers require that researchers make their data available in some form to other researchers or to the public. Lastly, when designing and managing a database, it is important to remember that the specific study for which the database was created will one day be finished, and those who developed it will no longer be spending their days (and nights) thinking about all the details of the study. In order for a database to remain meaningful for possible future use, it is important to consider possibilities for long-term data preservation.

Benefits to sharing your data can include new collaborations and publications (Lacher et al. 2012), increased citations of your research (Piwowar et al. 2007), expansion of successful wildlife management strategies to new areas or species, and fulfillment of journal and funding requirements for data sharing and management plans (Whitlock 2011). In addition, it is our responsibility as scientists to promote new knowledge by making data available, as appropriate, to the rest of the scientific community and even to the public, who fund many wildlife tracking programmes and research studies.

As you develop your database, it is worth considering ways to share your data, either with specific collaborators or with the public, and to at minimum make a description of your data set publicly available. The goal of this chapter is to introduce you existing ecological data standards and ways to make your database archivable and usable for additional analyses.

Although shared data from many other fields, such as hydrology, meteorology and genetics, have been widely used for many years, many wildlife tracking researchers remain reluctant to share data. Common concerns are that data will be misunderstood or used without proper acknowledgment, that sensitive data will get into the wrong hands or that they do not have enough time or resources to properly share data. These concerns highlight the need for appropriate methods for sharing data, combined with good data management and thorough documentation. As is described in this chapter, existing methods for sharing data address each of these concerns, from enabling data citation, to limiting sharing to trusted users, to encouraging communication between data owners and users, to providing free tools and support. And, as we have emphasised throughout this book, data organisation and documentation is not only needed when data will be shared with others, but also should be a standard part of data collection regardless of the end uses of your data.

An essential component of data sharing is the use of standards. Given the heterogeneity of methods, data sets and software used in the field of wildlife tracking, combined with the potential benefits of collaboration, there is a need for internationally recognised standards for describing and sharing data. Several examples of existing standards are described later in this chapter. These standards can help to ensure compatibility between different software platforms, research groups and databases. In addition, data standards play an important role in improving data quality and can liberate data from the specific aim for which they were collected. Adhering to such standards ensures that data can be reused for a wide range of purposes, maximising the returns of research funding and facilitating multi-species, large-scale and long-term ecological studies.

Describing Data

In order for raw data files to be understandable to others, they need to be well described. The meaning and format of each term used in your database should be defined, including the following:

- terms describing the actual data set attributes, e.g. the reference coordinate system of locations, timestamp format, units and precision;
- terms describing entities like sensors and animals, such as sex, serial number or species name; and
- terms describing the entire database or discrete subsets of it, such as the title, authors, keywords, time and geographic range of the data set used in a particular analysis.

There are several general rules to follow when describing terms in a data set:

- Use controlled vocabularies (a set of predefined words or terms) where possible. For example, if you are classifying migration stage for each record in your data set, allow only a discrete list of terms, such as ‘stopover’, ‘northward migration’ and ‘breeding grounds’. This supports consistent classification, prevents spelling errors and allows for easier analysis. Database tools such as lookup tables and constraints can be helpful in implementing these vocabularies.
- Never use a term twice if the definition is not exactly the same. If you are using two types of sensors, label them ‘GPS sensor’ and ‘activity sensor’, for example, rather than calling them both ‘sensor’ and risking confusion or errors, even if it is clarified by contextual information (e.g. the name of the table or of the schema where the information is stored).
- Where possible, data values should follow common standards—for example, providing timestamps in Coordinated Universal Time (UTC) and using species names from a published taxonomy.

- If you use codes or ambiguous shortened names (e.g. ‘CC’ for ‘*Capreolus capreolus*’), be sure to include tables that provide a full translation of codes used and that these tables are always included with any data transfer.
- To the extent possible, rely on standard database design approaches. Consider using the data and metadata standards described below, and follow generic table formats like those we present in this guidebook. If data structure and definition are unnecessarily complex or specific to the original context for which a database was developed, merging data sets and linking your data to external analysis tools becomes more complicated.
- Most importantly, make sure to maintain a written definition of all terms in your database that is available to all users. The definition should have a text description of the term along with any units, valid ranges, example values or controlled lists. The written definitions should explain where the values come from, such as the source of altitude estimates that may come from a DEM or from the GPS unit, or the method for determining habitat or behaviour. You could create this as a separate table in the database or as a plain text file.

Data and Metadata Standards

Several standards or schemas have been designed to deal specifically with describing ecological and geospatial data. These standards support description, discovery and integration of biological and geospatial data and are used by a wide range of research institutes, universities, museums, government agencies and other organisations. Standards provide relevant terms and definitions, have policies governing how to maintain and use the terms and document the history of changes to the standard. Where possible, it may be helpful to use terms from one or more of these standards in your database. This allows you to use and reference existing definitions, rather than writing your own, and would make it easier to share your data or metadata with databases such as the Global Biodiversity Information Facility (GBIF[1]) or the earth observation database DataONE[2].

Note that the difference between ‘data’ and ‘metadata’ is not clearly defined and will vary depending on the context. In general, metadata refers to ‘data about data’ or information that describes a data set. For example, descriptions of study animals might be considered ‘metadata’ describing your tracking data in one context, while in another, this information might be a part of your data set, with ‘metadata’ referring to a description of the entire study (such as title, authors and the time period of data collection). For our purposes, it may be helpful to think of metadata standards as useful for *finding* data and data standards as useful for *integrating or combining* data.

---

1 http://www.gbif.org/.

2 http://www.dataone.org/.

Three metadata standards specific to biology are Darwin Core, Access to Biological Collections Data (ABCD) and Ecological Metadata Language (EML). These standards are currently in use around the world and are freely available. In addition, they have support for geographic and temporal information.

The **Darwin Core** and **ABCD** standards are developed by Biodiversity Information Standards or TDWG (formerly the Taxonomic Database Working Group). Initially developed for use with natural history collections, **Darwin Core**[3] is widely used and includes terms for describing species occurrence data, including physical specimens, observations and digital records. It is focused primarily on terms that are generically applicable to natural history collections[4].

The **ABCD** standard[5] supports species occurrence data and includes around 1,200 terms (they refer to these as 'concepts'). It includes a larger number of terms and a more complex structure than Darwin Core, making it able to describe data and relationships between them more thoroughly, but requiring more technical expertise to fully implement (Wieczorek et al. 2012)[6].

EML[7] is a metadata standard developed by the Knowledge Network for Biocomplexity for describing ecological data (Higgins et al. 2002; Fegraus et al. 2005). It is open source and implemented by voluntary project members. It was designed primarily to describe data sets and other digital resources. EML consists of several modules that can be adopted by users as needed, including modules to support detailed descriptions of methods, attributes, tables in relational databases, and raster and vector geographic information[8].

In addition, the following other data and metadata standards may be useful:

The **Content Standard for Digital Geospatial Metadata** (**CSDGM**[9]), developed by the US Federal Geographic Data Committee (FGDC), is the current Federal metadata standard in the USA for geospatial data. This standard has a Biological Data Profile to provide additional support for biological data. Although it has been widely used throughout the USA, the FGDC now supports the transition to the ISO 19115 standard (see below).

The **ISO 19115 standard** is the International Organization for Standardization's (ISO[10]) metadata standard for describing geographic data. Unlike the other resources listed here, ISO standards are not freely available (i.e. must be paid for).

---

3 http://rs.tdwg.org/dwc.

4 Terms are currently described at http://rs.tdwg.org/dwc/terms/index.htm.

5 http://www.tdwg.org/activities/abcd.

6 Terms are currently described at http://wiki.tdwg.org/twiki/bin/view/ABCD/AbcdConcepts.

7 http://knb.ecoinformatics.org/software/eml.

8 Terms and modules are described at http://knb.ecoinformatics.org/software/eml/eml-2.1.1/index.html.

9 http://www.fgdc.gov/metadata/geospatial-metadata-standards#csdgm, http://www.fgdc.gov/metadata/csdgm/.

10 http://www.iso.org.

The **Open Geospatial Consortium** (OGC[11]) provides publicly available interoperability standards (see below). While these are not specifically data or metadata standards, some of their standards do include relevant data/metadata terms and schema that are used to implement interoperability.

In addition to these resources, manufacturers of tagging equipment typically provide data using fairly standardised formats and data attributes. Simply following the format delivered by popular manufacturers or other data providers, such as Argos, may be a simple way to make large amounts of data easy to share and compile. However, keep in mind that formats, attributes and units differ between manufacturers and sensors and that these companies are in the business of selling equipment, not maintaining databases. Combining data from different manufacturers and sensor types, each with their own specialised terms and data structure, can require significant effort. In addition, you will find that manufacturer-provided data formats often change over time and can be ambiguous, resulting in data files that misleadingly appear to be in the same format—for example, in some cases, users are allowed to choose a time zone in which data are delivered, although this choice is not indicated anywhere in the data file.

Lastly, you may want to look at data formats used by existing online animal tracking databases, such as those listed at the end of this chapter, in particular if you intend to use one of these databases for sharing or analysis. For example, see the Movebank Attribute Dictionary[12]. While these do not constitute official standards, the managers of these databases have developed data formats that are shared by large groups of data owners and could be extended or modified to meet your specific requirements.

Interoperability

Data that use shared metadata and data standards may be stored in diverse data formats and in online and offline databases. To allow others to locate and access metadata in a shared format, there must be ways to search metadata for species name, location, time period of data collection, etc., using Web-based databases or search engines. To actually combine multiple data sets for analysis requires additional work and may be time-consuming or impossible to do manually. In order to properly and efficiently search for or integrate data sets, we require interoperability.

In this context, interoperability can be generally defined as the ability for multiple databases, analysis software or other relevant systems to work together.

[11] http://www.opengeospatial.org/.

[12] https://www.movebank.org/node/2381.

For example, consider the interoperability of your database with an external client software for analysis:

- Not interoperable: Your data are stored in a proprietary format that cannot be read by the analysis software, and there is no easy way to export the data for use with the software.
- Somewhat interoperable: You can easily export all or part of your database as a .csv file, which can be read by the software.
- Very interoperable: You can query your database directly from the software, run an analysis and automatically store the results in your database.

There are several general ways in which you can make your database more interoperable with other systems. For example, you may use database software like PostgreSQL and PostGIS, which are open source, widely support international standards, and are likely to be maintained in the future. Using common non-proprietary file formats such as .csv rather than .xlsx when needed will minimise the chances of files becoming unreadable by contemporary software. Lastly, using established data and metadata standards such as those described above will make it more likely that your data can be understood and integrated with other data sets and software in the future.

Full interoperability with software tools, other databases and search engines requires implementation of more specific technical standards, which is beyond the scope of this guide. These include specifications for exchanging information using data/metadata exchange file formats, such as Extensible Markup Language (XML) or Resource Description Framework (RDF), and transfer protocols, such as TAPIR[13]. These standards are necessary, for example, to allow computers to automatically read data or metadata, retrieve search results, and present them in a way understandable to the user.

The **OGC**[14] is an international consortium that develops voluntary standards for interoperability of GIS data. PostGIS, used in this guide, follows the OGC's 'Simple Feature Access—Part 2: SQL Option' specification[15] and has been certified compliant with the 'Simple Features—SQL—Types and Functions 1.1' specification.

Publish Your Metadata

An alternative to publishing a data set in full is to make a description, in the form of metadata about your data set, available to a wider community. This makes it easier for other researchers to find out about your research and contact you about

13 http://www.tdwg.org/activities/tapir/specification.

14 http://www.opengeospatial.org/.

15 http://www.opengeospatial.org/standards/sfs.

possible collaborations. Databases with searchable study metadata often exist at the level of the research group, university, region or country. To be broadly useful, these typically require only a minimal number of descriptive terms that are applicable to all studies in the database.

To reach the widest possible audience, here are two global online databases where metadata about your wildlife tracking database could be shared:

- **GBIF** (**Global Biodiversity Information Facility**) publishes metadata about primary biodiversity occurrence data.[16] A list of contacts is available at http://www.gbif.org/communications/directory-of-contacts/regional-nodes
- **DataONE** (**Data Observation Network for Earth**) publishes metadata about earth observational data[17]. A list of member nodes is available at www.dataone.org/current-member-nodes

To contribute, you must be associated with a member node. Member nodes may include research institutes, government agencies and other organisations. In the case of GBIF, most countries also have a national node, and so if you are not affiliated with an existing member organisation, you could contact your national node to find out how to get involved.

It is relatively easy to store metadata within your PostgreSQL database. In addition to storing definitions for each term in the database (see **Describing data** above), you can create descriptive metadata for the database itself and for subsets within it, such as the set of records used for a specific publication or analysis. These metadata can include required terms for external databases and can be stored within your database in an XML format that complies with XML schema for the metadata standards described above using the XML data type in PostgreSQL.

In addition, many software programs exist to help researchers write and publish metadata in interoperable formats. One example is **Morpho**[18], a free, user-friendly software tool developed to help researchers write and publish metadata without special knowledge of technical interoperability requirements. Morpho allows you to write detailed metadata about ecological data sets and individual data tables that comply with the EML standard in XML format. After creating metadata, you can store files locally or upload the metadata, and even data tables, to the Knowledge Network for Biocomplexity, where they are searchable and available to other registered members (it is possible to restrict access to specific collaborators). This program is available from the Knowledge Network for Biocomplexity[19] (Higgins et al. 2002).

16 http://www.gbif.org.

17 http://www.dataone.org.

18 https://knb.ecoinformatics.org/morphoportal.jsp.

19 https://knb.ecoinformatics.org.

Publish and Share Your Data

Publishing your data allows you to share them in a more formal and structured way than exchanging files individually with collaborators and can make it accessible to a much wider research community. It also makes it easier for others to properly cite your data and allows you to list these citations in your CV as valuable research products in their own right. Depending on how you publish your data, you can make them available to the public or to specific user groups, define explicit terms of use, and ensure that some or all of a data set is permanently archived and remains accessible. Publishing data commonly involves the following:

- a review process to ensure the quality of data and related documentation;
- assignment of a persistent identifier such as a Digital Object Identifier (DOI) or Life Science Identifier (LSID) to ensure that the item will remain permanently available; and
- data licences that provide explicit conditions for reuse, such as those offered by Creative Commons[20] and Open Data Commons[21, 22].

If you have a completed data set that you would like to make available to the public and scientific community, you can submit it for review and publication. Several journals publish ‘data papers’, which include a biological data set along with a written description of the data, for research in biology and ecology. These include **Biodiversity Data Journal**, **Dataset Papers in Science**, **Ecological Archives**, **and Scientific Data**. In addition, there are databases that publish ‘data packages’ or sets of non-proprietary files associated with a written publication (see Penev et al. 2011). These include **Data Dryad**[23], which publishes data sets in the life sciences, and the **Movebank Data Repository**[24], which publishes animal tracking data in a standardised format.

Share Your Data Without Publishing

In some cases, the formality and permanence of publishing your data as described will not be the best option. For example, publicly revealing precise breeding or foraging locations of endangered populations may put them at risk. More commonly,

---

[20] http://www.creativecommons.org.

[21] http://opendatacommons.org.

[22] Note that making data files available as supplementary material along with a written article in general does not fit this definition of publication. In most cases the files are not part of the review process, and there is no guarantee by the publisher that the files will remain available (Anderson et al. 2006).

[23] http://www.datadryad.org.

[24] http://datarepository.movebank.org.

Table 13.1 A summary of several online databases that host animal tracking data

Database name	Description	Data access	Host
Australian Animal Tagging and Monitoring System (http://imos.aodn.org.au/webportal)	Part of the Integrated Marine Observing System, which includes other types of observational oceanic data; animal tracking data include marine animals from Australia to Antarctica	All data are available to the public	The Australian Government
Eurodeer (European Roe DEER Information System) (www.eurodeer.org)	A database used and maintained by researchers studying European roe deer (*Capreolus capreolus*) to support collaborative analysis	Data are shared between participating research groups; access can be requested, subject to agreement with 'terms of use'	Fondazione Edmund Mach
Global Procellariiform Tracking Database (www.seabirdtracking.org)	A compilation of existing seabird tracking data, obtained through personal communication with data owners to support the development of distribution maps and conservation recommendations	Controlled by data owners, with some summary information visible to the public	BirdLife International
Global Tagging of Pelagic Predators (www.gtopp.org)	A compilation of marine tracking and related bio-logged data	Select data are available to the public	An international collaboration of government and research institutions and equipment manufacturers
Movebank (www.movebank.org)	A data storage, sharing, analysis, and archiving platform for animal movement and related sensor data (the Movebank Data Repository)	Controlled by data owners; can be shared with select users or the public	The Max Planck Institute for Ornithology and the University of Konstanz
OBIS-SEAMAP (seamap.env.duke.edu)	A component of the Ocean Biogeographic Information System that allows the public to search for and aggregate information about marine species distributions, including marine mammals, seabirds, and sea turtles	All data are available to the public	Duke University

(continued)

Table 13.1 (continued)

Database name	Description	Data access	Host
Seaturtle and WildlifeTracking (www.seaturtle.org, www.wildlifetracking.org)	A data storage, sharing, and analysis platform for data collected using the Argos satellite system	Controlled by data owners	A private 501(c)3 tax-exempt organisation in the US
Wireless Remote Animal Monitoring (WRAM) (www.slu.se/WRAM)	A data storage, networking, sharing, and analysis platform for movement and other sensor data from fish and wildlife	Controlled by data owners; can be shared between participating research groups	The Swedish University of Agricultural Sciences

when research is ongoing and results of analysis are unpublished, you will likely want to share data only with specific people, such as collaborators or funding providers. Data sharing for the purpose of ongoing collaborative analysis requires different tools. These tools should allow you to define access rights to specific people, allowing them to view and/or add and edit data, and provide an infrastructure that helps multiple researchers to put their data into the same format.

In Chap. 2 and elsewhere in this guide, we have briefly described how you can use PostgreSQL to define multiple users and distinct access rights for each within your database[25] and allow remote connections. These features can support data sharing with colleagues, along with all the analysis and database design options PostgreSQL provides.

In addition, there are many existing online databases for sharing animal tracking data. One or more of these databases could provide a useful resource to complement your personal research database. These databases have varying Data-sharing options and are available to different sets of users and study types. They provide a way for a wide range of researchers, educators and conservation groups to find out about your research. Some of these databases also provide data-sharing and data-collection services, as well as a visualisation and analysis tools, that could be particularly useful for those with limited technical facilities. Brief descriptions of several of these are given in Table 13.1.

References

Anderson NR, Tarczy-Hornoch P, Bumgarner RE (2006) On the persistence of supplementary resources in biomedical publications. BMC Bioinf 7:260. doi:10.1186/1471-2105-7-260

Fegraus EH, Andelman S, Jones MB, Schildhauer M (2005) Maximizing the value of ecological data with structured metadata—an introduction to Ecological Metadata Language (EML) and principles for metadata creation. Bull Ecol Soc Am 86(3):158–168. doi:10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2

Higgins D, Berkley C, Jones MB (2002) Managing heterogeneous ecological data using Morpho. In: Proceedings of the 14th international conference on scientific and statistical database management, pp 69–76. doi:10.1109/SSDM.2002.1029707

Lacher TE, Boitani L, da Fonseca GAB (2012) The IUCN global assessments—partnerships, collaboration and data sharing for biodiversity science and policy. Conserv Lett 5:327–333. doi:10.1111/j.1755-263X.2012.00249.x

Penev L, Mietchen D, Chavan V, Hagedorn G, Remsen D, Smith V, Shotton D (2011) Pensoft data publishing policies and guidelines for biodiversity data. Pensoft Publishers. www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

Piwowar HA, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. PLoS ONE 2(3):e308. doi:10.1371/journal.pone.0000308

Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and opportunities of open data in ecology. Science 331(703):703–705. doi:10.1126/science.1197962

Vision TJ (2010) Open data and the social contract of scientific publishing. Bioscience 60(5):330–331. doi:10.1525/bio.2010.60.5.2

[25] http://www.postgresql.org/docs/9.2/static/user-manag.html.

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

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Roberson T, Vieglais D (2012) Darwin Core—An evolving community-developed biodiversity data standard. PLoS ONE 7(1):e29715. doi:10.1371/journal.pone.0029715

Wolkovich EM, Regetz J, O'Connor MI (2012) Advances in global change research require open science by individual researchers. Glob Change Biol 18:2102–2110. doi:10.1111/j.1365-2486.2012.02693.x