Ecological Informatics 36 (2016) 231–236

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Unified data management for distributed experiments: A model for collaborative grassroots scientific networks

Eric M. Lind

Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN 55108, United States

ARTICLE INFO

Article history:
Received 27 May 2016
Received in revised form 10 August 2016
Accepted 14 August 2016
Available online 18 August 2016

Keywords:
Data management
Database schema
Distributed experiment
Nutrient network
Taxonomic resolution

ABSTRACT

The rapidly growing number of grassroots ecological research networks demonstrates that ecologists have embraced distributed data collection and experimentation as a new tool for addressing global questions. A clear advantage of these networks is the ability to gather data at larger spatial and temporal scales and at relatively lower cost than could be typically accomplished by a single research team. However, a challenge arising from this structure is the need to merge distributed datasets into a coherent whole. The Nutrient Network, a coordinated distributed experiment entering its tenth year of data collection, has records from over 90 sites worldwide to date. In this paper I present lessons learned about data management from this project, focusing on such issues as standardization, storage, updates, and distribution of data within the network. I provide a relational database schema and associated workflow that could be generalized to many distributed ecological experiments or networked data observatories, especially those with need for taxonomic reconciliation of species occurrences. The success of distributed data collection efforts, especially long-term networks, will be proportional to the ability to coordinate and effectively combine project datasets.

1. Introduction

The discipline of ecology is challenged to predict consequences of global change at scales relevant to the biosphere and society (Steffen et al., 2015). To fulfill this in times of increasingly limited funding, ecologists have been turning to a variety of emerging techniques, each of which represents a variety of tradeoffs. For example, remote sensing can provide data with truly global coverage at high frequency (Running, 2012), though without a complete understanding of the local dynamics. Long-term research programs exemplified by the US Long-term Ecological Research sites (Kratz et al., 2003; Peters et al., 2013) provide unprecedented understanding of both long-term trends, as well as how those trends might be changing, in different ecosystems. However it is not always clear how or when the insights gleaned from long-term research in one location are predictive of responses elsewhere, even in similar habitats.

Another approach is to use meta-analysis, in which evidence from multiple locations and studies is used to infer effects at larger spatial and temporal scales (Hedges et al., 1999). This is a specific form of quantitative synthesis, or bringing together disparate data sources to test or reveal generalities in ecological systems (Carpenter et al., 2009). Typically meta-analysis standardizes the effects observed across multiple sites rather than standardizing the underlying data, which may have significant methodological deviations or differing experimental treatments. These differences, in turn, can limit confidence in the resulting inference.

One way to strengthen analysis and inference across multiple sites is to create a single coherent dataset, so that "apples-to-apples" comparisons and single statistical models can be used. Multiple approaches can be used to construct a coherent dataset. One approach is to connect and partially standardize data being collected similarly but independently by researchers in a given system. This approach is exemplified by international efforts like FluxNet (Baldocchi et al., 2001), CTFS-ForestGEO (Anderson-Teixeira et al., 2015), and GLEON (Weathers et al., 2013). In these examples standard equipment such as flux towers measuring micrometeorological gas exchange or buoys recording water temperature and chemistry are deployed for local research at many sites, but these data can be assembled across sites into coherent datasets due to the similar methodologies and dimensionality of the data.

A more restrictive cross-site data aggregation framework relies on identical methodology of data collection and experimentation to generate a single dataset. In this way researchers from many sites contribute to an expanding dataset but under a common structure. This can help provide insight into local dynamics, at many places simultaneously (Fraser et al., 2012). Such an approach is not totally new — the US Forest Service Forest Area Inventory (FIA), for example, has been using standardized sampling to record data on tree communities for 85 years (Bechtold, 2005). Likewise private efforts such as the Nature Conservancy's Natural Heritage Network (now under the name NatureServe[1]) has used data gathered

[1] http://www.natureserve.org/.

http://dx.doi.org/10.1016/j.ecoinf.2016.08.002
1574-9541/

in all US states and many other countries to create a standardized dataset of rare and threatened species (Stein, 2000). Nonetheless, ecologists have recently increasingly embraced the model of distributed data collection, because it allows inference on a regional and global scale, while remaining relatively cost-efficient. Data collection efforts and costs can be distributed among many researchers. While there are constraints on the type of information that can be gathered, distributed data collection and especially distributed experimentation have significant advantages over other approaches in terms of ecological inference (Fraser et al., 2012; Borer et al., 2014).

I focus the rest of this paper on the details of administering and managing a database derived from collaborative, distributed data collection under a single methodology. I use experience developed as coordinator of the Nutrient Network ("NutNet"; Borer et al., 2014), a coordinated, distributed grassland experiment being replicated at over 90 sites across six continents, to review current and future ecoinformatics challenges facing NutNet and other groups with existing or planned distributed experiments.

2. Challenges and solutions: NutNet as a case study

In attempting to effectively compile and manage data gathered from a distributed ecological network, there are several general challenges. Multiple solutions exist for each of these challenges. I illustrate the challenge and potential solution sets for the following areas:

1. assembling a coherent dataset;
2. standardization especially with respect to taxonomy;
3. versioning data;
4. incorporating new data types;
5. providing data access to internal partners.

2.1. Assembling a coherent dataset

The fundamental advance of distributed experimental networks with respect to building a coherent dataset, is that the data are collected with identical methodology. In the NutNet experiments, the treatments and the data collection at each site are conducted using identical, commonly used field methods in grassland ecology (Borer et al., 2014). The primary investigators at each site are responsible for ensuring adherence to the protocols, and transcribing data into a standardized data sheet (Appendix 1). The data sheet has separate tabs for each core dataset, each formatted in "long-form" (Wickham, 2014). The core datasets include: (a) site geographic location and descriptors (elevation, slope, aspect, etc); (b) a site "plan" describing the block and plot layout, and treatments applied to each plot; (c) a table of percent aerial cover by species observed in each plot; (d) a table of plant taxa observed including higher taxonomy, provenance (native/introduced), lifespan, and lifeform (if known); (e) a table of aboveground biomass by functional type collected from each plot; and (f) a table of photosynthetically active radiation (PAR) intercepted above and below the canopy in each plot (see Borer et al., 2014 for more complete methodology). These data are collected annually, and submitted to the NutNet data coordinator for incorporation into the larger database.

The basic principle used in handling the data is that original data submissions should not be meaningfully altered. Thus any errors or confusion in derived datasets can always be ultimately traced back to the original submission (Borer et al., 2009; Campbell et al., 2013). Other than saving the tables in the submitted datasheets as individual comma-delimited text files, the data from each NutNet site is stored as it was submitted. We use scripting languages (e.g. R, SQL) to act upon the raw data, to make any needed alterations to the data to ensure the data conform to our desired standardized format. Scripting is a key component of building and maintaining high-quality data (Borer et al., 2009). Any alterations between the submitted data and any final data product are documented in the script, both as the executed lines of the script, as well as in meta-code comments in the script.

Two main approaches can be used to take the data from each site-year set of observations and combine them with all other sites and years. The first dataset building approach is to use processing scripts to build a dataset each time an analysis is to be conducted, and the second approach is to process and store data in a separate database format. The essential difference between the two approaches is whether the processed and assembled datasets are assembled on-the-fly, or assembled and stored for later use. We have used both approaches in NutNet and discuss advantages and disadvantages of each.

One main advantage of using scripts to build larger datasets from raw data each time it is to be used is that any changes to the data content and organization, and the reasoning underlying those decisions, can be revisited and altered if necessary. This works well when the desired compiled dataset depends on choices in how the data will be summarized or combined. The power of this approach was demonstrated recently by Falster et al.'s (2015) BAAD dataset of woody plant allometry, especially by sharing the complete scripts used to create the dataset from its heterogeneous sources. However, the need for decision-making to achieve standardization when combining heterogeneous data is greater than that needed when combining data derived from standard methodology. Speed of assembly of the dataset is also a consideration, given that, as the network expands with additional sites, the number of submitted data files needed to construct the dataset grows even faster, since previously existing sites are also contributing new data each year (Fig. 1). Additionally, a drawback related to this approach is that the data cannot be combined across data types, queried, or summarized without creating the full dataset as a first step.

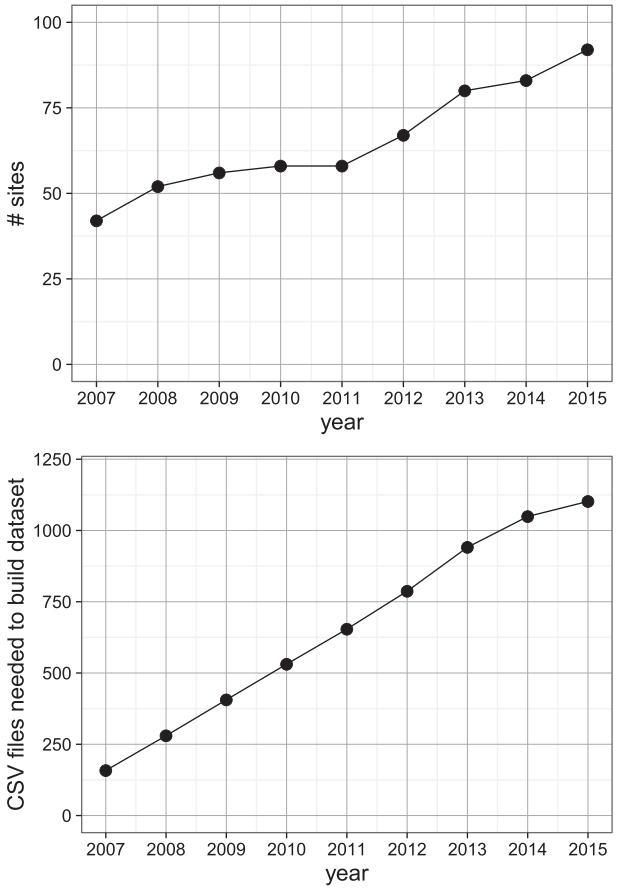

Fig. 1. Growth of the Nutrient Network through time. Top: number of sites in the network. Bottom: number of files (tables) needed to construct the full NutNet dataset.

In contrast, the second approach is to process the data making decisions about standardization and quality control using scripts as above, but storing the resulting coherent dataset as a data product that can be updated and queried. In our case we use a relational database framework, which maps onto the hierarchical structure of our data, and the association of attributes at multiple levels of the hierarchy (Codd, 1970, Madin et al., 2007, Fig. 2). The design is analogous to the type of dimensional database used in compilations of plant trait data such as TRY (Kattge et al., 2011), wherein data are collected at multiple scales. The advantages of using a relational database schema include being able to standardize fields (creating allowable values), adding constraints on relationships, and storing, editing, or deleting each piece of information in a single place (Codd, 1970). In addition to growing by adding rows of existing data types, the relational structure also allows for new data types as they become available. Finally, the power of relational databases is the ability to easily summarize, recombine, and reshape data to fit the various analyses that might arise from a multidimensional project like NutNet.

Our schema (Fig. 2) is arranged to match our hierarchical experimental structure, and data associated with each level as appropriate. The spatial replication of the NutNet experiments is based around sites, at which typically three blocks contain 8–10 replicate 5 m × 5 m plots. Core data are collected within 2 m × 2 m subplots. The data structure in the schema likewise nests subplots as many to one within plots, plots many to one within blocks, and blocks many to one within sites. The advantage of this structure is the ability to easily link data at the appropriate scale. For instance the geographic identifiers of the site (e.g., latitude and longitude) can be used to query published datasets to retrieve ancillary data on long-term climate, monthly weather time series, models of reactive nitrogen deposition, and so on. Blocks can have their own independent geographic coordinates within sites. Plots have treatments applied and some plot-level covariates like soil type. Data collected in each subplot are stored in their own tables with year identifiers.

Intersection tables serve to relate entities having the possibility of many-to-many relationships, such as primary investigators to sites (PIs may have more than one site, and sites may have more than one PI). A crucial advantage of storing data in a relational database instead of assembling datasets on the fly is the use of intersection tables to handle complexities of plant nomenclature and traits (discussed in more detail below). Intersection tables allow translation of names into common entities, and allow for the reality of differences in key traits such as provenance (native/exotic) for a single species across sites.

2.2. Standardizing submitted data to match schema and taxonomy

Creating standard submission forms and a common data structure does not necessarily ensure a simple process when combining newly submitted data with already-processed datasets. Newly submitted data must be checked for concordance with both dataset structure and existing data content. NutNet follows a scripted quality control (QC) procedure with each newly submitted dataset. In order to import submitted data into the relational database, progressive QC benchmarks must be met, especially with respect to new data from existing sites. These benchmarks are as follows:

(1) submitted data use standard tables and fields
(2) incoming location (site, plot, subplot) and treatment data match existing records
(3) incoming plant species observations match an accepted name
(4) units of data (e.g. g m^{-2}) are verified and match the data type.

These benchmarks can be met using scripts by importing the incoming data into temporary tables in the database that can be evaluated against the existing data. An example of such a script in pseudocode is presented as Appendix 2. The script can be run to flag problems with

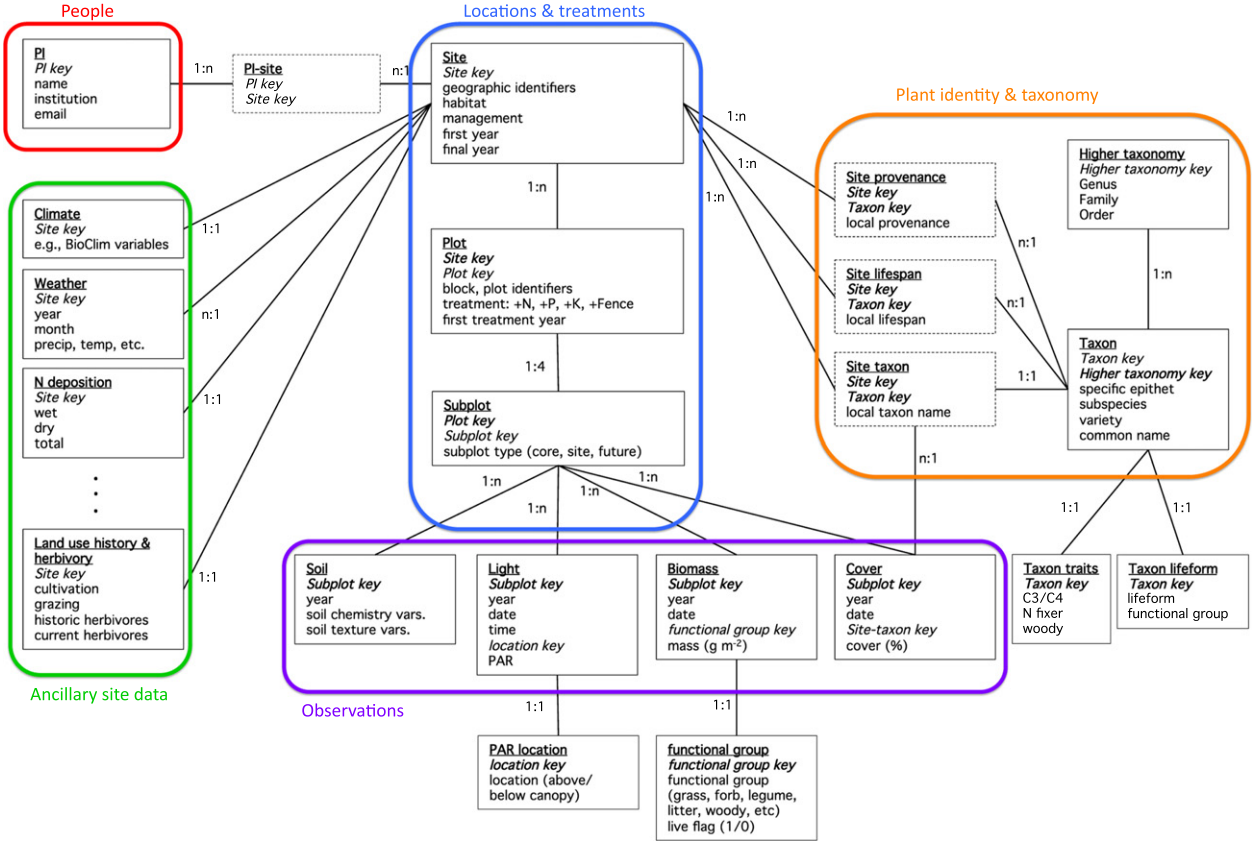

Fig. 2. Conceptual overview of NutNet relational database schema. Boxes indicate data tables, with intersection tables in dotted outline. Table names are underlined. Lines connect tables linked by relational keys, which are in italics. Relationships as numbers of rows in linked identities are 1:1, many to one (n:1), or one to many (1:n) as indicated.

incoming data, or mismatches between incoming and existing records (for instance, treatment assignments to plots).

The most challenging aspect of the data import process for organism-based survey data, such as is generated in NutNet, can be the reconciliation of taxonomy of species names. Taxonomic identification is crucial because the name given an organism serves as a link to all other scientific knowledge about its evolution, ecology, and function (Patterson et al., 2010). But these names also pose a significant obstacle, as synonymy (multiple names for a single organism) and debates over lumping and splitting entities can make it difficult to assign an organism a permanent and consistent name (Patterson et al., 2010). Within NutNet, many questions revolve around the importance and functionality of plant biodiversity, both within sites over time and among sites in the global network. Being able to determine even a simplistic diversity metric like plant species richness in a plot, site, or group of sites relies entirely on the ability to identify the number of unique plant taxa observed, and thus on resolving any discrepancies or synonymy. These arise most frequently among sites, and among sampling years within sites.

To resolve this issue we use a standard taxonomic checklist of worldwide vascular plants, The Plant List[2] (Kalwij, 2012). The checklist assigns each name a ranking of "Accepted", "Synonym" or "Unresolved" (typically indicating the name has not yet been fully assessed) according to taxonomic experts. Names submitted to NutNet are checked against this list using the *Taxonstand* package (Cayuela et al., 2012) in R (R Core Team, 2015). Names which do not match any listing in The Plant List are investigated case-by-case, usually for significant misspellings. Names classified as unresolved are treated as accepted in our data. For names classified as synonyms, the accepted name is treated as the NutNet standard name, and the synonym retained as the name at the site in which the record was observed. Once resolved, the taxon names at a site may also change from year to year, as different or better identification becomes available (for instance if a perennial plant is observed flowering in one year, it may be more precisely or differently identified, and thus require altering prior year's identifications in that same permanent quadrat). The database schema (Fig. 2) is designed to hold both pieces of information by employing an intersection table called "Site-Taxon" in which a name used at a site is connected to a standard taxon record. In most cases the names in the standard taxon and local site taxon records are the same, but where synonymy occurs, the local name is retained while still allowing a single standard taxon list to be used across the network. When summarizing data to be compared across sites, the standard taxon names are used.

2.3. Altering and versioning

As with building datasets, a challenge for updating a database is recording the process of decision-making behind changes, as well as the execution of any changes to the data (Jones et al., 2015). With script-based approaches and diligent commenting, a transparent record of these decisions and actions is available. With some relational database software (e.g. MySQL), there is also the ability to implement transactional records of changes so that all alterations are stored. We use SQL scripts to act on the NutNet database in order to alter or update data, as required. These scripts are themselves stored in a commonly accessible folder that is regularly archived.

As data are added and altered, the entire database changes state, making regular backups essential. Automated, regularly scheduled snapshots of the entire NutNet database are made and stored by date. This practice is essential for the ability to recreate analyses performed on earlier versions of the database, as the data can be restored and queried from a snapshot even after the underlying data have been amended or altered. The alternative data-building approach of assembling from raw data each time also can be versioned by using programs such as git[3] to store versions of the script, instead of the data. In this case, the option exists to revert to previous versions of the script (and thus the assembly process for the database).

2.4. Extending the database: Ancillary and add-on data

With many ecological projects, data outside those directly collected by investigators can be included in modeling and analysis. In the Nutrient Network a variety of data can help to predict the differential response of plots within sites to the treatments, including long-term climate, monthly weather, and landscape data such as distance to roads and human population density. The relational data structure makes it straightforward to connect any additional ancillary data by simply adding a new data table that is related to the site table by unique site identifier. This also facilitates recombination and summarizing using SQL query views. A downside of storing these data within the database structure is that as the ancillary data are created, maintained, and updated by outside parties, regular updates or data checks are necessary to ensure data integrity of the locally stored copies.

As the number of Nutrient Network collaborators and sites has grown, PIs have proposed new project ideas not originally envisioned as part of the core investigations. The relational data structure makes incorporating new data relatively straightforward, as long as the additional data are collected at one of the existing spatial scales of the hierarchical organization of the network (site, block, plot or subplot). New data tables and any associated lookup tables can then be created with links through the appropriate unique identifier for the observation scale. One example of this is a project in which arthropods were sampled from core subplots at several sites. This effort resulted in an Arthropod sample table, which holds information on the taxonomic identity (linked to a standardized look-up table), number of individuals, and total mass of each sample.

2.5. Rearranging and recombining data; data access & sharing

The single strongest benefit of pursuing the effort needed to construct and maintain a relational database structure for large-scale collaborative scientific networks is the ability to combine, rearrange, and summarize data into the most desirable forms for a wide variety of analyses. This is the original purpose of the SQL family of languages, and most implementations are efficient at operating across hierarchies and data types as well as handling large table sizes via indexing. Queries link data through the relational structure, and hierarchical data can either be returned in full or summarized at certain levels of the hierarchy. For instance, a query can return each observation of plant biomass by functional type within each plot (multiple rows of data per plot), or a query could return the sum of these observations to give a single production value per plot (one row per plot), or the mean of all plots within a site can be returned (one row per site).

Each of these combinations of the data can be scripted and called as needed so that analyses always use the most updated data version, but can also be stored as database objects called "views". Views are virtual tables that are constructed from the actual tables typically based on frequently used queries. In the Nutrient Network, we use views to output standard data tables that are commonly desired for many different research questions. These include a summary of all main variables by plots within sites; a full table of biomass records; a full table of percent cover by plant records; and a table of soil nutrient information. One major advantage of creating views for standard data products is the ability to associate standardized metadata with these tables, to indicate the sources of the fields used in analysis across the network (Michener et al., 1997; Michener, 2006).

[2] http://theplantlist.org.

[3] https://git-scm.com/.

Ultimately the scientific output of Nutrient Network relies on modeling and analysis of the data in the database, which means accessing data within other software. In practice within the Nutrient Network, automated scripts in a command line shell call views from the database and pipe them into comma-separated files, which are stored in a commonly accessible online shared folder. These include the system date in the title of the file, and older versions are archived and stored so that there is only a single, "live" shared dataset at any one time within the network, defined as the most recently created. Direct access to the database itself through interfaces like the *RMySQL* package (Ooms et al., 2016) in R allows creation of data objects from an SQL query. While this approach to obtaining data is limited by account validation to connect to the database and the need to know the schema and SQL, it offers a direct way to move from optimal data storage to optimal analysis software without having to use a text file or other intermediate data format.

3. Concluding remarks

As ecologists continue to develop the power of distributed science and long-term research to answer the most important and pressing research questions, data management practices will determine whether the expense and effort of these collaborations result in usable data, publishable results, and solid inference. In the Nutrient Network, we have adopted a relational database schema and associated processes that result in a standardized, coherent, versioned, tracked dataset that can flexibly change to incorporate each new site and year of data collection. I recognize that there are many methodologies for assembling and managing data from distributed networks, and the processes detailed here may not be appropriate for all cases. However, for efficient hierarchical data storage, the ability to reconcile and store alternative taxonomies, and the ease of recombining and summarizing data, the relational data storage approach has performed well for the Nutrient Network. I strongly advocate for the use of this approach and management of data via a scripted language and believe that this approach can serve as a model for storage and access of high quality data for emerging distributed ecological efforts.

Acknowledgments

I thank Elizabeth Borer and Eric Seabloom for ongoing discussion and feedback on these processes and this manuscript. Habacuc Flores-Moreno provided insightful comments on an earlier draft. This work was generated using data from the Nutrient Network (http://www.nutnet.org) experiment, funded at the site-scale by individual researchers. Coordination and data management have been supported by funding to E. Borer and E. Seabloom from the National Science Foundation Research Coordination Network (NSF-DEB-1042132) and Long Term Ecological Research (NSF-DEB-1234162 to Cedar Creek LTER) programs, and the Institute on the Environment (DG-0001-13). I also thank the Minnesota Supercomputer Institute for hosting project data and the Institute on the Environment for hosting Network meetings.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ecoinf.2016.08.002.

References

Anderson-Teixeira, K.J., Davies, S.J., Bennett, A.C., Gonzalez-Akre, E.B., Muller-Landau, H.C., Joseph Wright, S., Abu Salim, K., Almeyda Zambrano, A.M., Alonso, A., Baltzer, J.L., Basset, Y., Bourg, N.A., Broadbent, E.N., Brockelman, W.Y., Bunyavejchewin, S., Burslem, D.F.R.P., Butt, N., Cao, M., Cardenas, D., Chuyong, G.B., Clay, K., Cordell, S., Dattaraja, H.S., Deng, X., Detto, M., Du, X., Duque, A., Erikson, D.L., Ewango, C.E.N., Fischer, G.A., Fletcher, C., Foster, R.B., Giardina, C.P., Gilbert, G.S., Gunatilleke, N., Gunatilleke, S., Hao, Z., Hargrove, W.W., Hart, T.B., Hau, B.C.H., He, F., Hoffman, F.M., Howe, R.W., Hubbell, S.P., Inman-Narahari, F.M., Jansen, P.A., Jiang, M., Johnson, D.J., Kanzaki, M., Kassim, A.R., Kenfack, D., Kibet, S., Kinnaird, M.F., Korte, L., Kral, K., Kumar, J., Larson, A.J., Li, Y., Li, X., Liu, S., Lum, S.K.Y., Lutz, J.A., Ma, K., Maddalena, D.M., Makana, J.-R., Malhi, Y., Marthews, T., Mat Serudin, R., McMahon, S.M., McShea, W.J., Memiaghe, H.R., Mi, X., Mizuno, T., Morecroft, M., Myers, J.A., Novotny, V., de Oliveira, A.A., Ong, P.S., Orwig, D.A., Ostertag, R., den Ouden, J., Parker, G.G., Phillips, R.P., Sack, L., Sainge, M.N., Sang, W., Sri-ngernyuang, K., Sukumar, R., Sun, I.-F., Sungpalee, W., Suresh, H.S., Tan, S., Thomas, S.C., Thomas, D.W., Thompson, J., Turner, B.L., Uriarte, M., Valencia, R., Vallejo, M.I., Vicentini, A., Vrška, T., Wang, X., Wang, X., Weiblen, G., Wolf, A., Xu, H., Yap, S., Zimmerman, J., 2015. CTFS-ForestGEO: a worldwide network monitoring forests in an era of global change. Glob. Chang. Biol. 21, 528–549.

Baldocchi, D., Falge, E., Gu, L.H., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X.H., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Pilegaard, K.U.K., Schmid, H.P., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2001. FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bull. Am. Meteorol. Soc. 82, 2415–2434.

Bechtold, W.A., Patterson, P.L (Eds.), 2005. The Enhanced Forest Inventory and Analysis Program — National Sampling Design and Estimation Procedures.

Borer, E.T., Seabloom, E.W., Jones, M.B., Schildhauer, M., 2009. Some simple guidelines for effective data management. Bull. Ecol. Soc. Am. 90, 205–214.

Borer, E.T., Harpole, W.S., Adler, P.B., Lind, E.M., Orrock, J.L., Seabloom, E.W., Smith, M.D., 2014. Finding generality in ecology: a model for globally distributed experiments. Methods Ecol. Evol. 5 65–+.

Campbell, J.L., Rustad, L.E., Porter, J.H., Taylor, J.R., Dereszynski, E.W., Shanley, J.B., Gries, C., Henshaw, D.L., Martin, M.E., Sheldon, W.M., Boose, E.R., 2013. Quantity is nothing without quality: automated QA/QC for streaming environmental sensor data. Bioscience 63, 574–585.

Carpenter, S.R., Armbrust, E.V., Arzberger, P.W., Chapin, F.S., Elser, J.J., Hackett, E.J., Ives, A.R., Kareiva, P.M., Leibold, M.A., Lundberg, P., Mangel, M., Merchant, N., Murdoch, W.W., Palmer, M.A., Peters, D.P.C., Pickett, S.T.A., Smith, K.K., Wall, D.H., Zimmerman, A.S., 2009. Accelerate synthesis in ecology and environmental sciences. Bioscience 59, 699–701.

Cayuela, L., Granzow-de la Cerda, I., Albuquerque, F.S., Golicher, D.J., 2012. TAXONSTAND: an R package for species names standardisation in vegetation databases. Methods Ecol. Evol. 3, 1078–1083.

Codd, E.F., 1970. A relational model of data for large shared data banks. Commun. ACM 13, 377–387.

Falster, D.S., Duursma, R.A., Ishihara, M.I., Barneche, D.R., FitzJohn, R.G., Vårhammar, A., Aiba, M., Ando, M., Anten, N., Aspinwall, M.J., Baltzer, J.L., Baraloto, C., Battaglia, M., Battles, J.J., Bond-Lamberty, B., van Breugel, M., Camac, J., Claveau, Y., Coll, L., Dannoura, M., Delagrange, S., Domec, J.-C., Fatemi, F., Feng, W., Gargaglione, V., Goto, Y., Hagihara, A., Hall, J.S., Hamilton, S., Harja, D., Hiura, T., Holdaway, R., Hutley, L.S., Ichie, T., Jokela, E.J., Kantola, A., Kelly, J.W.G., Kenzo, T., King, D., Kloeppel, B.D., Kohyama, T., Komiyama, A., Laclau, J.-P., Lusk, C.H., Maguire, D.A., le Maire, G., Mäkelä, A., Markesteijn, L., Marshall, J., McCulloh, K., Miyata, I., Mokany, K., Mori, S., Myster, R.W., Nagano, M., Naidu, S.L., Nouvellon, Y., O'Grady, A.P., O'Hara, K.L., Ohtsuka, T., Osada, N., Osunkoya, O.O., Peri, P.L., Petritan, A.M., Poorter, L., Portsmuth, A., Potvin, C., Ransijn, J., Reid, D., Ribeiro, S.C., Roberts, S.D., Rodríguez, R., Saldaña-Acosta, A., Santa-Regina, I., Sasa, K., Selaya, N.G., Sillett, S.C., Sterck, F., Takagi, K., Tange, T., Tanouchi, H., Tissue, D., Umehara, T., Utsugi, H., Vadeboncoeur, M.A., Valladares, F., Vanninen, P., Wang, J.R., Wenk, E., Williams, R., de Aquino Ximenes, F., Yamaba, A., Yamada, T., Yamakura, T., Yanai, R.D., York, R.A., 2015. BAAD: a Biomass And Allometry Database for woody plants. Ecology 96, 1445.

Fraser, L.H., Henry, H.A., Carlyle, C.N., White, S.R., Beierkuhnlein, C., Cahill, J.F., Casper, B.B., Cleland, E., Collins, S.L., Dukes, J.S., Knapp, A.K., Lind, E., Long, R., Luo, Y., Reich, P.B., Smith, M.D., Sternberg, M., Turkington, R., 2012. Coordinated distributed experiments: an emerging tool for testing global hypotheses in ecology and environmental science. Front. Ecol. Environ.

Hedges, L.V., Gurevitch, J., Curtis, P.S., 1999. The meta-analysis of response ratios in experimental ecology. Ecology 80, 1150–1156.

Jones, A.S., Horsburgh, J.S., Reeder, S.L., Ramírez, M., Caraballo, J., 2015. A data management and publication workflow for a large-scale, heterogeneous sensor network. Environ. Monit. Assess. 187, 1–19.

Kalwij, J.M., 2012. Review of "The plant list, a working list of all plant species.". J. Veg. Sci. 23, 998–1002.

Kattge, J., Ogle, K., Boenisch, G., Diaz, S., Lavorel, S., Madin, J., Nadrowski, K., Noellert, S., Sartor, K., Wirth, C., 2011. A generic structure for plant trait databases. Methods Ecol. Evol. 2, 202–213.

Kratz, T.K., Deegan, L.A., Harmon, M.E., Lauenroth, W.K., 2003. Ecological variability in space and time: insights gained from the US LTER program. Bioscience 53, 57–67.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F., 2007. An ontology for describing and synthesizing ecological observation data. Ecol. Inf. 2, 279–296.

Michener, W., 2006. Meta-information concepts for ecological data management. Ecol. Inf. 1, 3–7.

Michener, W., Brunt, J., Helly, J., Kirchner, T., Stafford, S., 1997. Nongeospatial metadata for the ecological sciences. Ecol. Appl. 7, 342.

Ooms, J., James, D., DebRoy, S., Wickham, H., Horner, J., 2016. RMySQL: Database Interface and "MySQL" Driver for R.

Patterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L., Remsen, D.P., 2010. Names are key to the big new biology. Trends Ecol. Evol. 25, 686–691.

Peters, D.P.C., Laney, C.M., Lugo, A.E., Collins, S.L., Driscoll, C.T., Groffman, P.M., Grove, J.M., Knapp, A.K., Kratz, T.K., Ohman, M.D., Waide, R.B., Yao, J., 2013. Long-term trends in

ecological systems: a basis for understanding responses to global change. USDA Agricultural Research Service Publication No., 1931. Washington, D.C.
R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Running, S.W., 2012. A measurable planetary boundary for the biosphere. Science 337, 1458–1459.
Steffen, W., Richardson, K., Rockström, J., Cornell, S.E., Fetzer, I., Bennett, E.M., Biggs, R., Carpenter, S.R., de Vries, W., de Wit, C.A., Folke, C., Gerten, D., Heinke, J., Mace, G.M., Persson, L.M., Ramanathan, V., Reyers, B., Sörlin, S., 2015. Planetary boundaries: guiding human development on a changing planet. Science 1259855.
Stein, B.A., Kutner, L.S., Adams, J. (Eds.), 2000. Precious Heritage.
Weathers, K.C., Hanson, P.C., Arzberger, P., Brentrup, J., Brookes, J., Carey, C.C., Gaiser, E., Hamilton, D.P., Hong, G.S., Ibelings, B., Jennings, E., Kim, B., Kratz, T., Lin, F.-P., Muraoka, K., O'Reilly, C., Piccolo, C., Rose, K.C., Zhu, G., 2013. The Global Lake Ecological Observatory Network (GLEON): the evolution of grassroots network science. Limnol. Oceanogr. Bull. 22, 71–73.
Wickham, H., 2014. Tidy data. J. Stat. Softw. 59, 1–23.