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

Eric M. Lind

Keywords: data management, database schema, distributed experiment, Nutrient Network, taxonomic resolution

1
2
3
4 14 *Abstract*

5
6 15 The rapidly growing number of grassroots ecological research networks demonstrates that
7
8
9 16 ecologists have embraced distributed data collection and experimentation as a new tool for
10
11 17 addressing global questions. A clear advantage of these networks is the ability to gather data at
12
13
14 18 larger spatial and temporal scales and at relatively lower cost than could be typically
15
16 19 accomplished by a single research team. However, a challenge arising from this structure is the
17
18
19 20 need to merge distributed datasets into a coherent whole. The Nutrient Network, a coordinated
20
21 21 distributed experiment entering its tenth year of data collection, has records from over 90 sites
22
23
24 22 worldwide to date. In this paper I present lessons learned about data management from this
25
26 23 project, focusing on such issues as standardization, storage, updates, and distribution of data
27
28
29 24 within the network. I provide a relational database schema and associated workflow that could be
30
31 25 generalized to many distributed ecological experiments or networked data observatories,
32
33 26 especially those with need for taxonomic reconciliation of species occurrences. The success of
34
35
36 27 distributed data collection efforts, especially long-term networks, will be proportional to the
37
38 28 ability to coordinate and effectively combine project datasets.
39
40

41 29

42
43 30
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

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 and Patterson, 2005). Likewise private efforts such as the Nature Conservancy’s Natural Heritage Network (now under the name NatureServe¹) has used data gathered in all US states and many other countries to create a standardized dataset of rare and threatened species (Stein et al., 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

¹ <http://www.natureserve.org/>

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 (Figure 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.

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, Figure 2). The design is analogous to the type of dimensional database used in compilations of plant trait data such as TRY (Kattge et al., 2010), wherein data are collected at multiple scales. The advantages of using a relational

1
2
3
4 145 database schema include being able to standardize fields (creating allowable values), adding
5
6 146 constraints on relationships, and storing, editing, or deleting each piece of information in a single
7
8
9 147 place (Codd, 1970). In addition to growing by adding rows of existing data types, the relational
10
11 148 structure also allows for new data types as they become available. Finally, the power of
12
13
14 149 relational databases is the ability to easily summarize, recombine, and reshape data to fit the
15
16 150 various analyses that might arise from a multidimensional project like NutNet.

17
18
19 151 Our schema (Figure 2) is arranged to match our hierarchical experimental structure, and
20
21 152 data associated with each level as appropriate. The spatial replication of the NutNet experiments
22
23
24 153 is based around sites, at which typically three blocks contain 8-10 replicate 5m x 5m plots. Core
25
26 154 data are collected within 2m x 2m subplots. The data structure in the schema likewise nests
27
28
29 155 subplots as many to one within plots, plots many to one within blocks, and blocks many to one
30
31 156 within sites. The advantage of this structure is the ability to easily link data at the appropriate
32
33
34 157 scale. For instance the geographic identifiers of the site (e.g., latitude and longitude) can be used
35
36 158 to query published datasets to retrieve ancillary data on long-term climate, monthly weather time
37
38 159 series, models of reactive nitrogen deposition, and so on. Blocks can have their own independent
39
40
41 160 geographic coordinates within sites. Plots have treatments applied and some plot-level covariates
42
43 161 like soil type. Data collected in each subplot are stored in their own tables with year identifiers.

44
45 162 Intersection tables serve to relate entities having the possibility of many-to-many
46
47
48 163 relationships, such as primary investigators to sites (PIs may have more than one site, and sites
49
50
51 164 may have more than one PI). A crucial advantage of storing data in a relational database instead
52
53 165 of assembling datasets on the fly is the use of intersection tables to handle complexities of plant
54
55 166 nomenclature and traits (discussed in more detail below). Intersection tables allow translation of
56
57
58
59
60
61
62
63
64
65

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

1
2
3
4 189 The most challenging aspect of the data import process for organism-based survey data,
5
6
7 190 such as is generated in NutNet, can be the reconciliation of taxonomy of species names.
8
9 191 Taxonomic identification is crucial because the name given an organism serves as a link to all
10
11
12 192 other scientific knowledge about its evolution, ecology, and function (Patterson et al., 2010). But
13
14 193 these names also pose a significant obstacle, as synonymy (multiple names for a single
15
16 194 organism) and debates over lumping and splitting entities can make it difficult to assign an
17
18
19 195 organism a permanent and consistent name (Patterson et al., 2010). Within NutNet, many
20
21 196 questions revolve around the importance and functionality of plant biodiversity, both within sites
22
23
24 197 over time and among sites in the global network. Being able to determine even a simplistic
25
26 198 diversity metric like plant species richness in a plot, site, or group of sites relies entirely on the
27
28
29 199 ability to identify the number of unique plant taxa observed, and thus on resolving any
30
31 200 discrepancies or synonymy. These arise most frequently among sites, and among sampling years
32
33
34 201 within sites.

35
36 202 To resolve this issue we use a standard taxonomic checklist of worldwide vascular plants,
37
38 203 The Plant List² (Kalwij, 2012). The checklist assigns each name a ranking of “Accepted”,
39
40
41 204 “Synonym” or “Unresolved” (typically indicating the name has not yet been fully assessed)
42
43 205 according to taxonomic experts. Names submitted to NutNet are checked against this list using
44
45
46 206 the *Taxonstand* package (Cayuela et al., 2012) in R (R Core Team, 2015). Names which do not
47
48 207 match any listing in The Plant List are investigated case-by-case, usually for significant
49
50
51 208 misspellings. Names classified as unresolved are treated as accepted in our data. For names
52
53 209 classified as synonyms, the accepted name is treated as the NutNet standard name, and the
54
55
56 210 synonym retained as the name at the site in which the record was observed. Once resolved, the

60 ² <http://theplantlist.org>
61
62
63
64
65

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 (Figure 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

1
2
3
4 234 assembling from raw data each time also can be versioned by using programs such as git³ to
5
6 235 store versions of the script, instead of the data. In this case, the option exists to revert to previous
7
8
9 236 versions of the script (and thus the assembly process for the database).
10

11 237 12 13 14 238 *2.4 Extending the database: ancillary and add-on data* 15

16 239 With many ecological projects, data outside those directly collected by investigators can
17
18
19 240 be included in modeling and analysis. In the Nutrient Network a variety of data can help to
20
21 241 predict the differential response of plots within sites to the treatments, including long-term
22
23
24 242 climate, monthly weather, and landscape data such as distance to roads and human population
25
26 243 density. The relational data structure makes it straightforward to connect any additional ancillary
27
28
29 244 data by simply adding a new data table that is related to the site table by unique site identifier.
30
31 245 This also facilitates recombination and summarizing using SQL query views. A downside of
32
33
34 246 storing these data within the database structure is that as the ancillary data are created,
35
36 247 maintained, and updated by outside parties, regular updates or data checks are necessary to
37
38 248 ensure data integrity of the locally stored copies.
39

40
41 249 As the number of Nutrient Network collaborators and sites has grown, PIs have proposed
42
43 250 new project ideas not originally envisioned as part of the core investigations. The relational data
44
45
46 251 structure makes incorporating new data relatively straightforward, as long as the additional data
47
48 252 are collected at one of the existing spatial scales of the hierarchical organization of the network
49
50
51 253 (site, block, plot or subplot). New data tables and any associated lookup tables can then be
52
53 254 created with links through the appropriate unique identifier for the observation scale. One
54
55 255 example of this is a project in which arthropods were sampled from core subplots at several sites.
56
57
58
59

60 ³ <https://git-scm.com/>
61
62
63
64
65

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 observation 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).

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.

Figures

Figure 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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Figure 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.

References

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

- Verma, T. Vesala, K. Wilson, and S. Wofsy., 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society* 82:2415–2434.
- Bechtold, W. A., and P. L. Patterson, editors., 2005. The enhanced forest inventory and analysis program - national sampling design and estimation procedures.
- Blackard, J. A., M. V. Finco, E. H. Helmer, G. R. Holden, M. L. Hoppus, D. M. Jacobs, A. J. Lister, G. G. Moisen, M. D. Nelson, R. Riemann, B. Ruefenacht, D. Salajanu, D. L. Weyermann, K. C. Winterberger, T. J. Brandeis, R. L. Czaplewski, R. E. McRoberts, P. L. Patterson, and R. P. Tymcio., 2008. Mapping US forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment* 112:1658–1677.
- Borer, E. T., W. S. Harpole, P. B. Adler, E. M. Lind, J. L. Orrock, E. W. Seabloom, and M. D. Smith., 2014. Finding generality in ecology: a model for globally distributed experiments. *Methods in Ecology and Evolution* 5:65–+.
- Borer, E. T., E. W. Seabloom, M. B. Jones, and M. Schildhauer., 2009. Some Simple Guidelines for Effective Data Management. *Bulletin of the Ecological Society of America* 90:205–214.
- Campbell, J. L., L. E. Rustad, J. H. Porter, J. R. Taylor, E. W. Dereszynski, J. B. Shanley, C. Gries, D. L. Henshaw, M. E. Martin, W. M. Sheldon, and E. R. Boose., 2013. Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data. *BioScience* 63:574–585.
- Carpenter, S. R., E. V. Armbrust, P. W. Arzberger, F. S. Chapin, J. J. Elser, E. J. Hackett, A. R. Ives, P. M. Kareiva, M. A. Leibold, P. Lundberg, M. Mangel, N. Merchant, W. W.

- 1
2
3
4 376 Murdoch, M. A. Palmer, D. P. C. Peters, S. T. A. Pickett, K. K. Smith, D. H. Wall, and A.
5
6 377 S. Zimmerman. 2009. Accelerate Synthesis in Ecology and Environmental Sciences.
7
8
9 378 BioScience 59:699–701.
10
11
12 379 Cayuela, L., I. Granzow-de la Cerda, F. S. Albuquerque, and D. J. Golicher., 2012.
13
14 380 TAXONSTAND: An R package for species names standardisation in vegetation
15
16 381 databases. Methods in Ecology and Evolution 3:1078–1083.
17
18
19 382 Cayuela, L., and J. Oksanen., 2014. Taxonstand: Taxonomic Standardization of Plant Species
20
21 383 Names.
22
23
24 384 Codd, E. F., 1970. A relational model of data for large shared data banks. Communications of the
25
26 385 ACM 13:377–387.
27
28
29 386 Falster, D. S., R. A. Duursma, M. I. Ishihara, D. R. Barneche, R. G. FitzJohn, A. Vårhammar, M.
30
31 387 Aiba, M. Ando, N. Anten, M. J. Aspinwall, J. L. Baltzer, C. Baraloto, M. Battaglia, J. J.
32
33 388 Battles, B. Bond-Lamberty, M. van Breugel, J. Camac, Y. Claveau, L. Coll, M. Dannoura,
34
35
36 389 S. Delagrangé, J.-C. Domec, F. Fatemi, W. Feng, V. Gargaglione, Y. Goto, A. Hagihara,
37
38 390 J. S. Hall, S. Hamilton, D. Harja, T. Hiura, R. Holdaway, L. S. Hutley, T. Ichie, E. J.
39
40
41 391 Jokela, A. Kantola, J. W. G. Kelly, T. Kenzo, D. King, B. D. Kloeppel, T. Kohyama, A.
42
43 392 Komiyama, J.-P. Laclau, C. H. Lusk, D. A. Maguire, G. le Maire, A. Mäkelä, L.
44
45
46 393 Markesteijn, J. Marshall, K. McCulloh, I. Miyata, K. Mokany, S. Mori, R. W. Myster, M.
47
48 394 Nagano, S. L. Naidu, Y. Nouvellon, A. P. O’Grady, K. L. O’Hara, T. Ohtsuka, N. Osada,
49
50
51 395 O. O. Osunkoya, P. L. Peri, A. M. Petritan, L. Poorter, A. Portsmouth, C. Potvin, J.
52
53 396 Ransijn, D. Reid, S. C. Ribeiro, S. D. Roberts, R. Rodríguez, A. Saldaña-Acosta, I. Santa-
54
55 397 Regina, K. Sasa, N. G. Selaya, S. C. Sillett, F. Sterck, K. Takagi, T. Tange, H. Tanouchi,
56
57
58 398 D. Tissue, T. Umehara, H. Utsugi, M. A. Vadeboncoeur, F. Valladares, P. Vanninen, J. R.
59
60
61
62
63
64
65

- Wang, E. Wenk, R. Williams, F. de Aquino Ximenes, A. Yamaba, T. Yamada, T. Yamakura, R. D. Yanai, and R. A. York., 2015. BAAD: a Biomass And Allometry Database for woody plants. *Ecology* 96:1445–1445.
- Fraser, L. H., H. A. Henry, C. N. Carlyle, S. R. White, C. Beierkuhnlein, J. F. Cahill, B. B. Casper, E. Cleland, S. L. Collins, J. S. Dukes, A. K. Knapp, E. Lind, R. Long, Y. Luo, P. B. Reich, M. D. Smith, M. Sternberg, and R. Turkington., 2012. Coordinated distributed experiments: an emerging tool for testing global hypotheses in ecology and environmental science. *Frontiers in Ecology and the Environment*.
- Hedges, L. V., J. Gurevitch, and P. S. Curtis., 1999. The meta-analysis of response ratios in experimental ecology. *Ecology* 80:1150–1156.
- Jones, A. S., J. S. Horsburgh, S. L. Reeder, M. Ramírez, and J. Caraballo., 2015. A data management and publication workflow for a large-scale, heterogeneous sensor network. *Environmental Monitoring and Assessment* 187:1–19.
- Kalwij, J. M., 2012. Review of “The Plant List, a working list of all plant species.” *Journal of Vegetation Science* 23:998–1002.
- Kattge, J., K. Ogle, G. Boenisch, S. Diaz, S. Lavorel, J. Madin, K. Nadrowski, S. Noellert, K. Sartor, and C. Wirth., 2011. A generic structure for plant trait databases. *Methods in Ecology and Evolution* 2:202–213.
- Kratz, T. K., L. A. Deegan, M. E. Harmon, and W. K. Lauenroth., 2003. Ecological Variability in Space and Time: Insights Gained from the US LTER Program. *BioScience* 53:57–67.
- Limnology and Oceanography* Bulletin, August, 2013., 2013. . *Limnology and Oceanography: Bulletin* 22:69–100.

- 1
2
3
4 421 Madin, J., S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa., 2007. An
5
6 422 ontology for describing and synthesizing ecological observation data. Ecological
7
8
9 423 Informatics 2:279–296.
- 10
11 424 Michener, W., 2006. Meta-information concepts for ecological data management. Ecological
12
13
14 425 Informatics 1:3–7.
- 15
16 426 Michener, W., J. Brunt, J. Helly, T. Kirchner, and S. Stafford., 1997. Nongeospatial metadata for
17
18
19 427 the ecological sciences. Ecological Applications 7:342.
- 20
21 428 Ooms, J., D. James, S. DebRoy, H. Wickham, and J. Horner., 2016. RMySQL: Database
22
23
24 429 Interface and “MySQL” Driver for R.
- 25
26 430 Patterson, D. J., J. Cooper, P. M. Kirk, R. L. Pyle, and D. P. Remsen., 2010. Names are key to
27
28
29 431 the big new biology. Trends in Ecology & Evolution 25:686–691.
- 30
31 432 Peters DPC, Laney CM, Lugo AE, Collins SL, Driscoll CT, Groffman PM, Grove JM, Knapp
32
33
34 433 AK, Kratz TK, Ohman MD, Waide RB, and Yao J., 2013. Long-term trends in ecological
35
36 434 systems: a basis for understanding responses to global change. USDA Agricultural
37
38
39 435 Research Service Publication No., 1931. Washington, D.C.
- 40
41 436 R Core Team., 2015. R: A language and environment for statistical computing. R Foundation for
42
43 437 Statistical Computing, Vienna, Austria.
- 44
45 438 Reichman, O. J., M. B. Jones, and M. P. Schildhauer., 2011. Challenges and Opportunities of
46
47
48 439 Open Data in Ecology. Science 331:703–705.
- 49
50
51 440 Running, S. W., 2012. A Measurable Planetary Boundary for the Biosphere. Science 337:1458–
52
53 441 1459.
- 54
55 442 Steffen, W., K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R.
56
57
58 443 Carpenter, W. de Vries, C. A. de Wit, C. Folke, D. Gerten, J. Heinke, G. M. Mace, L. M.

- Persson, V. Ramanathan, B. Reyers, and S. Sörlin., 2015. Planetary boundaries: Guiding human development on a changing planet. *Science*:1259855.
- Stein, B. A., L. S. Kutner, and J. Adams, editors., 2000. *Precious Heritage*.
- Weathers, K.C., P.C. Hanson, P. Arzberger, J. Brentrup, J. Brookes, **C.C. Carey**, E. Gaiser, D.P. Hamilton, G.S. Hong, B. Ibelings, E. Jennings, B. Kim, T. Kratz, F.-P. Lin, K. Muraoka, C. O'Reilly, C. Piccolo, K.C. Rose, G. Zhu. The Global Lake Ecological Observatory Network (GLEON): the evolution of grassroots network science. , 2013. *Limnology and Oceanography Bulletin*. 22:71-73.
- Wickham, H., 2014. Tidy Data. *Journal of Statistical Software* 59:1–23.

Figure 1

Figure 2

Graphical Abstract (for review)
distributed data collection...

... leading to coherent unified data products.

	A	B	C	D	E
1	site_code	country	plot	live biomass g m ⁻²	species richness m ⁻²
2	burren.ie	IE	1	216.65	27
3	burren.ie	IE	2	247.6	22
4	burren.ie	IE	3	254.4	26
5	chilcas.ar	AR	1	595.5	14
6	chilcas.ar	AR	2	435.5	13
7	chilcas.ar	AR	3	427	14
8	hopl.us	US	1	156.649995	15
9	hopl.us	US	2	87.6	22
10	hopl.us	US	3	123.999995	20
11	ping.au	AU	1	263.464	8
12	ping.au	AU	2	172.494	8
13	ping.au	AU	3	187.396	9
14	sereng.tz	TZ	1	223	16
15	sereng.tz	TZ	2	195.5	13
16	sereng.tz	TZ	3	225.5	13

A screenshot of a spreadsheet application (likely Excel) showing a data table. The table has columns labeled A through I. The data rows show site codes, country codes, plot numbers, subplot numbers, taxa names, mass values, dates, and note_biomass values. The table is part of a larger document titled 'NutNet_CoreMeasurements.xls'.

	A	B	C	D	E	F	G	H	I
		block	plot	subplot	taxa	mass	date	note_biomass	
9	7	5	3	1	D	bryophytes	0.00	2015-07-03	
10	8	6	4	2	D	dead	168.50	2015-07-03	
11	9	7	5	3	D	forbs	342.16	2015-07-03	
12	10	8	6	4	D	graminoids	1.47	2015-07-03	
13	11	9	7	5	D	legumes	0.00	2015-07-03	
14	12	10	8	6	D	woody	0.00	2015-07-03	
15	13	11	9	7	2A	bryophytes	0.00	2015-07-03	
16	14	12	10	8	2A	dead	46.88	2015-07-03	
17	15	13	11	9	2A	forbs	198.17	2015-07-03	
18	16	14	12	10	2A	graminoids	32.55	2015-07-03	
19	17	15	13	11	2A	legumes	0.00	2015-07-03	
20	18	16	14	12	2A	woody	0.00	2015-07-03	
21	19	17	15	13	3B	bryophytes	0.00	2015-07-03	
22	20	18	16	14	3B	dead	272.70	2015-07-03	
23	21	19	17	15	3B	forbs	732.92	2015-07-03	
24	22	20	18	16	3B	graminoids	27.26	2015-07-03	
25	23	21	19	17	3B	legumes	0.00	2015-07-03	
26	24	22	20	18	3B	woody	0.00	2015-07-03	
27	25	23	21	19	4D	bryophytes	0.00	2015-07-03	
28	26	24	22	20	4D	dead	69.58	2015-07-03	
29	27	25	23	21	4D	forbs	214.53	2015-07-03	
30	28	26	24	22	4D	graminoids	210.57	2015-07-03	
31	29	27	25	23	4D	legumes	0.00	2015-07-03	
32	30	28	26	24	4D	woody	0.00	2015-07-03	
33	31	29	27	25	5C	bryophytes	0.00	2015-07-03	
34	32	30	28	26	5C	dead	35.43	2015-07-03	
35	33	31	29	27	5C	forbs	253.70	2015-07-03	
36	34	32	30	28	5C	graminoids	11.76	2015-07-03	
37	35	33	31	29	5C	legumes	0.00	2015-07-03	
38	36	34	32	30	5C	woody	0.00	2015-07-03

... facilitates scripted database integration...

... through standardized data input...