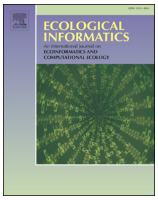
Contents lists available at ScienceDirect

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 xxxx

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/© 2016 Elsevier B.V. All rights reserved.

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.

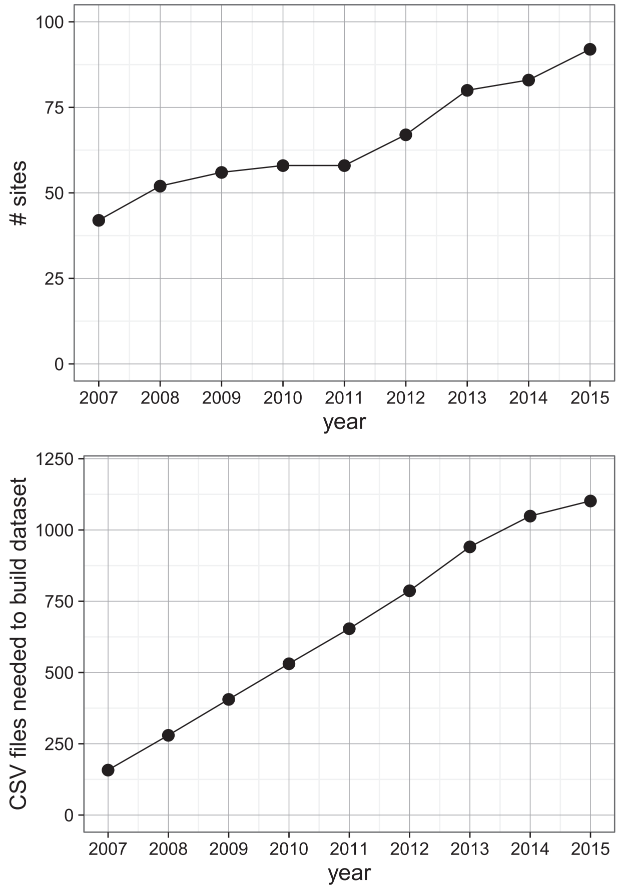


Fig. 1. Growth of the Nutrient Network through time. Top: number of sites in the network. Bottom: number of files (tables) needed to construct the full NutNet dataset.