



Veg-X – an exchange standard for plot-based vegetation data

Susan K. Wisser, Nick Spencer, Miquel De Cáceres, Martin Kleikamp,
Brad Boyle & Robert K. Peet

Keywords

Data exchange; Data standard;
Ecoinformatics; Relevé; Taxonomic concept;
Vegetation plot

Abbreviations

dbh = diameter at breast height; FGDC = U.S.
Federal Geographic Data Committee;
sec = *secundum* according to (Berendsohn
1995); TDWG = Biodiversity Information
Standards organisation; XML = Extensible
Markup Language

Received 14 June 2010

Accepted 23 September 2010

Co-ordinating Editor: Jürgen Dengler

Wisser, S. (corresponding author, WisserS@landcare
research.co.nz) & **Spencer, N.** (SpencerN@
landcareresearch.co.nz): Landcare Research, PO
Box 40, Lincoln 7640, New Zealand

De Cáceres, M. (miquelcaceres@gmail.com):
Biodiversity and Landscape Ecology Laboratory,
Centre Tecnològic Forestal de Catalunya, Ctra
antiga St Llorenç km 2, ES–25280, Solsona,
Catalonia, Spain

Kleikamp, M. (Martin.Kleikamp@web.de):
Sieg Lindenweg 14, D–51469 Bergisch Gladbach,
Germany

Boyle, B. (bboyle@email.arizona.edu):
Department of Ecology and Evolutionary Biology,
The University of Arizona Herbarium, PO Box
210088, Tucson, Arizona 85721, USA

Peet, R.K. (peet@unc.edu): Department of Biology
CB#3280, University of North Carolina, Chapel Hill,
NC 27599-3280, USA

Introduction

The availability of large vegetation-plot databases provides powerful methods and capacity for addressing fundamental questions concerning plant community composition, diversity and distribution. In particular, the unprecedented geographic coverage of such data has

made possible global tests of allometric scaling theory (Enquist & Niklas 2001), deep-time palaeoclimate reconstruction (Boyle et al. 2008), and continental-scale analyses of plant range size (Weiser et al. 2007), wood density (Swenson & Enquist 2007; Russo et al. 2010),

Abstract

Question: Collaborative research efforts and synthetic vegetation analyses are often limited by difficulties in sharing or combining datasets. Can we facilitate these activities by means of an exchange standard for plot-based vegetation data?

Methods: In 2003, the Ecoinformatics Working Group and the Council of the International Association for Vegetation Science endorsed the development of a standard exchange schema for vegetation-plot data. In 2007, a first workshop was held to formulate a common set of goals, concepts, and terminology for plot-based vegetation data. At a second workshop in 2008, this ontology was developed into an XML (extensible markup language) schema representation designed to be maximally compatible with existing standards and databases.

Results: The exchange standard for plot-based vegetation data (Veg-X) allows for observations of vegetation at both individual plant and aggregated observation levels. It ensures that observations are fixed to physical sample plots at specific points in space and time, and makes a distinction between the entity of interest (e.g. an individual tree) and the observational act (i.e. a measurement). The standard supports repeated measurements of both individual organisms and plots, allows observations of entities to be grouped following predefined or user-defined criteria, and ensures that the connection between the entity observed and taxonomic concept associated with that observation are maintained.

Conclusions: Establishment of exchange standards followed by development of ecoinformatics tools built around those standards should allow scientists to efficiently combine plot data over extensive spatial and temporal gradients in order to perform analyses and make predictions of vegetation change and dynamics at local and global scales.

pollen spectra (Gosling et al. 2009), carbon storage (Hall et al. 2001), community assembly (Fridley et al. 2007; Manthey, Fridley & Peet, unpubl. data) and community phylogenetic structure (Swenson et al. 2007). Although there have been major efforts over the past two decades to ensure that vegetation data are well documented, archived and accessible (Wiser et al. 2001; Malhi et al. 2002; Jennings et al. 2009; Schaminée et al. 2009; Dengler et al. 2011), this information is stored in many different formats and data structures, with the consequence that it is not easily shared or aggregated. This limits potential for these data to be included in collaborative research initiatives and large-scale, synthetic vegetation analyses, and limits the general usability of tools developed for analysis of these data.

A primary technical impediment to large-scale sharing of vegetation data is the lack of a recognized international exchange standard for linking the panoply of tools and database implementations that exist among various organizations and individuals participating in vegetation research. In the absence of an exchange standard, the need for multiple, *ad hoc* mappings among databases and applications discourages merging of data and slows development of new analytical tools (Fig. 1a). By contrast, widespread use of a common exchange standard would avoid the need to repeatedly map data for synthetic projects by requiring only a single mapping between a given database or tool and the standard (Fig. 1b), thus facilitating data exchange and analysis. Application of an international exchange standard for vegetation data would form a critical part of the necessary infrastructure to allow these data to be combined for synthetic analysis at local and global scales.

Repeatability of research is a key tenet of the scientific method and standardization of methods and approaches facilitates repeatability and comparability among studies (Ellison 2010). In the absence of data exchange standards, data from past plot-based vegetation studies may be technically available, but the challenges and uncertainties associated with repeating large-scale data integration often present insurmountable impediments to data reuse and analytical repeatability. Moreover, even if plot data are well integrated in a single database system, long-term database maintenance is not guaranteed, whereas data marked up in a standard exchange format can be deposited in a digital archive with reasonable certainty of long-term preservation. Ultimately, exchange standards enable reuse of data, data management approaches, and analytical tools.

The need for standards to allow data interoperability is widely acknowledged (Jones et al. 2006). Organizations such as Biodiversity Information Standards (TDWG), Committee on Data for Science and Technology of the

International Council of Science (CODATA), the Open Geospatial Consortium (OGC) and the US Federal Geographic Data Committee (FGDC) consider the fostering of such standards a primary mission. Work on data exchange has been done by the biodiversity community and is represented in standards such as Darwin Core (<http://www.tdwg.org/standards/450/>) and Access to Biological Collection Data (ABCD; <http://www.tdwg.org/activities/abcd/>) for the exchange of spatial and temporal occurrence information for individual specimens from biological collection databanks, the Taxon Concept Schema (<http://www.tdwg.org/standards/117/>) for exchange of taxon concepts and their relationships, particularly as associated with any taxonomic determinations, and Structure of Descriptive Data (SDD) for the exchange of structured descriptive taxonomic data (<http://wiki.tdwg.org/wiki/bin/view/SDD/WebHome>). Exchange standards provide the ability to port data to analytical tools and other data users and systems. For example, the Global Biodiversity Information Facility (GBIF) consolidates species occurrence records sourced from myriad distributed data sources (Edwards et al. 2000), primarily through use of Darwin Core as an exchange schema.

Because plot-based vegetation data have certain consistent key components and relationships, standardized solutions can be applied to data storage and exchange, as well as to the development of tools for data entry, validation, quality assurance and control. The specimen-based standards cited above, however, are not adequate for community sampling because the information required goes beyond specimen and occurrence data, which record only the occurrence of an individual organism at a single place and time. The notion of ecological community is essentially grounded on species joint occurrence and its variation across space and time (Mueller-Dombois & Ellenberg 1974; Whittaker 1975). Moreover, community ecology commonly requires detailed biotic (e.g. vegetation structure, composition, growth, recruitment, mortality, plant traits) as well as abiotic information (e.g. climatic, soil, geomorphology data) in order to make inferences about pattern and process in vegetation systems.

Here we discuss some of the barriers to data exchange and integration and propose a technological solution to overcome them, an international exchange standard for vegetation-plot data that we call Veg-X. We then describe major components of plot-based vegetation data and their logical relationships, as well as their implementation within the Veg-X exchange standard. This analysis represents an extension of the logic and structure behind the XML (extensible markup language) schema developed for VegBank (<http://vegbank.org/vegdocs/xml/vegbank-xml-index.html>) and subsequently for VegetWeb (<http://www.floraweb.de/vegetation/aufnahmen.html>). Finally, we

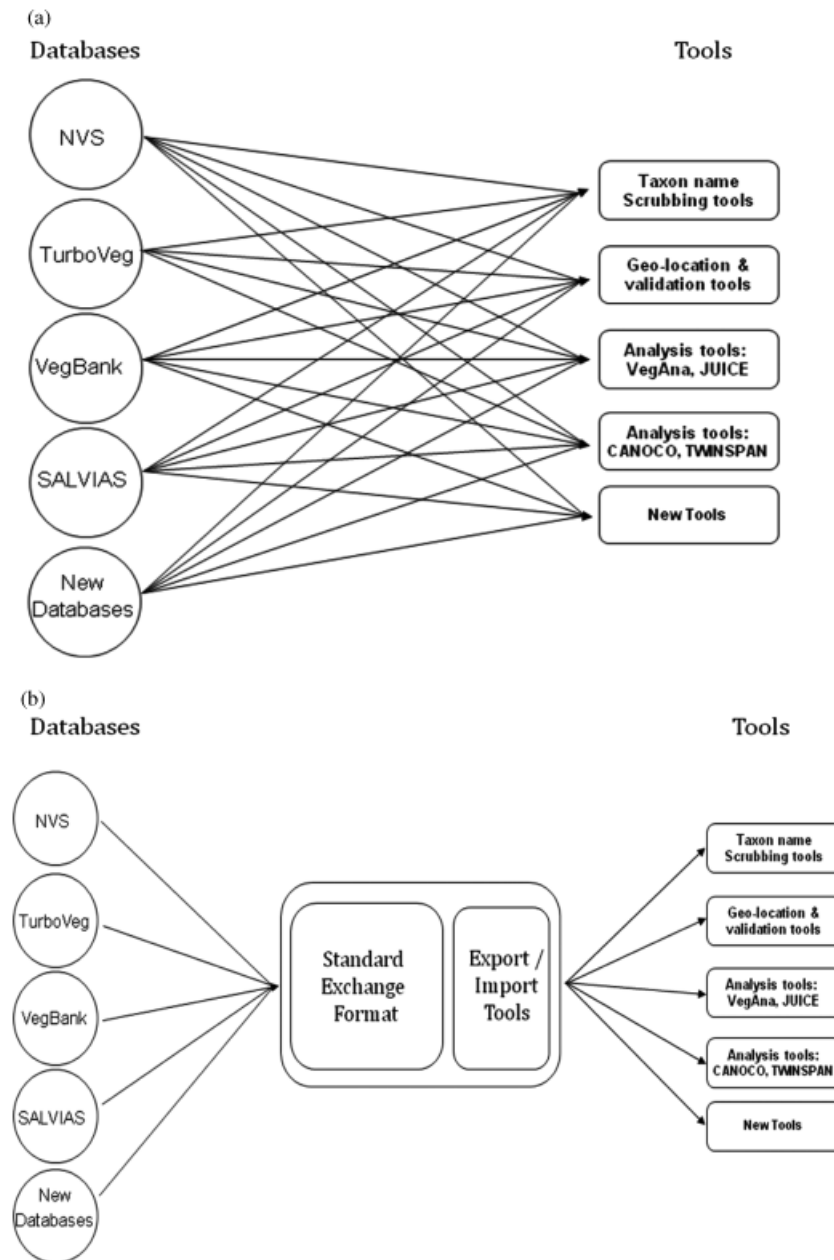


Fig. 1. Schematic diagram showing (a) problem: multiple tools and databases and (b) solution: sharing tools through a common standard.

discuss how vegetation exchange standards coupled with other developments may provide the immediate opportunity to begin organizing vegetation data and making them available to the wider ecological community for novel analyses.

Barriers to Exchange and Integration

Heterogeneous data collection methods

Standards exist for sampling and recording vegetation for specific purposes. For example, standards have been published for recording relevés and plots in field surveys

for vegetation classification (Mucina et al. 2000; Jennings et al. 2009), as have sampling guidelines for censusing stem-based tropical forest plots (Condit 1998). Further, many regions and countries have specific data collection protocols (e.g. Peet et al. 1998; Hurst & Allen 2007; Jennings et al. 2009). Despite the existence of such sampling standards, the use of different measurement scales (e.g. geographic projections), units, precision and techniques (e.g. instruments, protocols) is still a source of inconsistency that complicates data exchange (Hale et al. 2003). For example, several abundance-cover scales exist

for recording the abundance of plant taxa (van der Maarel 1979; Jennings et al. 2009). Sometimes the measurement methodologies used are well documented in databases, but in other cases methodologies are assumed and not explicitly captured. Exchange data formats must accommodate observations made using different measurement methodologies and document these differences, but this requires that the information on the methodology itself is explicit in source databases.

Structural differences between vegetation databases

Vegetation database systems differ in their overall structure as a result of different emphases in different projects. For example, the choice between individual-based and aggregated organism observations is a key feature that allows vegetation databases to be coarsely grouped. A large number of vegetation databases store records of aggregated organism observations such as species cover (i.e. relevé databases). Examples of this kind are numerous, including all the vegetation databases based on the TURBOVEG program (Hennekens & Schaminée 2001), as well as many regional and national database systems, especially in Europe (Schaminée et al. 2009). Another major kind of vegetation database system focuses on storing individual-based observations generated by long-term studies and monitoring programmes, including international forest plot networks such as that of the Center for Tropical Forest Science (CTFS; <http://www.ctfs.si.edu/>), the Tropical Ecology Assessment and Monitoring Network (TEAM; <http://www.teamnetwork.org/en/>) and RAINFOR (<http://www.geog.leeds.ac.uk/projects/rainfor/>). The more complex database systems allow both individual-based and aggregated organism observations to be stored. Examples are the vegetation-plot databank of the Ecological Society of America's Panel on Vegetation Classification (VegBank; Jennings et al. 2009), New Zealand's National Vegetation Survey databank (Wiser et al. 2001), SALVIAS (<http://www.salvias.net/pages/index.html>), and BIOTABase (Finckh, M., Muche, M., Schmiedel, U. & Jürgens, N., unpublished data). Fortunately, the ability to group records of individuals of the same taxon into a single abundance value allows integrative analyses to be conducted where data is sourced from a range of database systems. Furthermore, studies requiring only measurements of taxon presence-absence or co-occurrence (e.g. Swenson et al. 2007; Boyle et al. 2008) can use data from both aggregate and individual-based samples.

Vocabularies

Just as inconsistent data structures and methods can prevent data sharing and integration, inconsistencies in

the use of terms or concepts can prevent integration of disparate or syntactically different data sources, even though they are semantically consistent. In the case of vegetation databases, different implementations vary in how they name core concepts such as observation, plot or method. For example, VegBank, New Zealand's National Vegetation Survey databank, and BIOTABase all have different terms for stratum: "Stratum", "Tier", and "Layer" respectively. Furthermore, different database implementations often vary considerably in how they name second-order concepts such as measured attributes of plots or plants. These inconsistencies are compounded by the use of different languages.

Such barriers can be difficult to overcome unless an attempt is made to either unify terminology or semantically map syntactically different terms to a common term or concept. Ultimately, attempting to unify terminology will fail; language and words are not static and change over time in both application and meaning. Design decisions for individual database instances inevitably lead to different structures and hence terms. Therefore, mapping disparate terms from multiple sources to a set of common concepts in an agreed data standard provides the only long-term solution to data integration barriers.

Taxonomy

Classifications of organisms and communities provide a complex set of problems that must be addressed in any information system containing references to biological taxa or ecological communities. The core problem is that taxonomic standards vary with time, place and investigator such that biological taxa and communities can have multiple names. Furthermore, the same name can represent multiple taxon concepts, or even entirely different entities (Berendsohn 1995; Kennedy et al. 2005; Franz et al. 2008; Franz & Peet 2009; Jansen & Dengler 2010). The traditional solution when integrating data of mixed provenance has been to develop a standard list of names and to map all the various data sources to that list. For example, within the USA there are several standard lists of plant taxa including Kartesz (1999), USDA PLANTS (<http://plants.usda.gov/>), and the Integrated Taxonomic Information System (ITIS; <http://www.itis.gov/>). However, this approach fails to allow effective dataset integration for several reasons: (1) online lists are periodically updated but usually are not simultaneously archived, with the consequence that the user cannot reconstruct the database for some arbitrary time in the past; (2) ambiguity arises from the fact that one name can be used for multiple taxonomic concepts and one concept can be labeled with multiple names; and (3) different parties have different perspectives on acceptable names and the meanings associated with them.

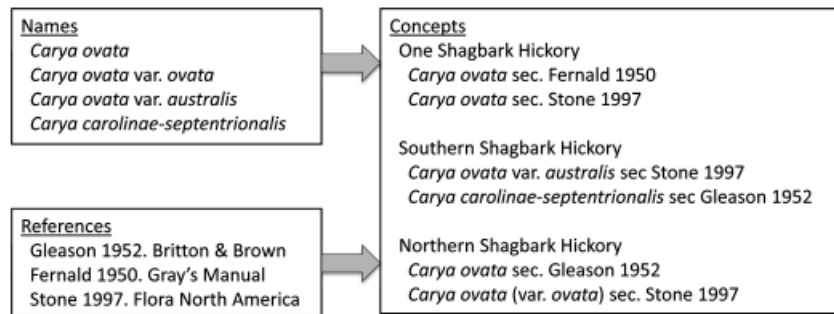


Fig. 2. Example of how taxon names and references combine to form taxon concepts. Shagbark hickory (*Carya ovata* s.l.) can be viewed as one or two taxa. Examining three references we find four names, three unique taxon concepts, and six name–reference couplets.

Part of the ambiguity arises from the requirement of biological nomenclature that when a taxon is split, the name continues to be applied to the taxon that includes the type specimen for the original name. Consider the case of shagbark hickory (*Carya ovata* (Miller) K. Koch s.l.), which some authors think of as a single entity and others think should be divided into two: northern shagbark and southern shagbark. If you encounter the name *Carya ovata* in a database, you cannot be sure whether it means all shagbarks (sec. Fernald 1950), or just northern shagbark (sec. Gleason 1952) (Fig. 2). Trees that Gleason recognizes as *Carya carolinae-septentrionalis* would be lumped within *Carya ovata* by a worker who follows Fernald. In addition, authors differ as to whether they believe the two types of shagbark should be recognized as distinct species or simply as varieties. Even if you know a worker follows Stone's treatment of *Carya* in Flora North America (1997), you cannot be sure whether plants identified as *Carya ovata* include just *C. ovata* var. *ovata*, (= *C. ovata* sec. Gleason 1952) or also include *C. ovata* var. *australis* (= *C. carolinae-septentrionalis*). Figure 2 shows the three names (technically four with inclusion of *C. ovata* var. *ovata*) associated with the shagbarks, the three references used above, and how names and references combine to form three concepts. Note that in this case there are two alternative and synonymous name–reference pairs for each of the three concepts.

Two key points emerge from this example. (1) A name–reference combination constitutes an assertion of a taxonomic concept. (2) Ideally, identifications or determinations (e.g. plot occurrences, specimen labels, treatments in authoritative works) should be by reference to a specific taxonomic concept by reference to a name as used by a specific author. The US National Vegetation Classification mandates use of concepts in identification of taxa (U.S. Federal Geographic Data Committee 2008) and the TDWG Taxon Concept Schema (<http://www.tdwg.org/standards/117/>) provides a framework for exchange of taxon concept and taxon concept relationship data. Franz et al. (2008) and

Franz & Peet (2009) provide guidance on how to integrate data where concepts are clearly reported. However, nearly all legacy data and most current data lack reference to concepts and contain only naked scientific names (Jansen & Dengler 2010). These must usually be interpreted as nominal concepts, meaning the broadest possible meaning of the name, which presents a major impediment to high-resolution-data integration. Note that the same problems apply to community names and concepts.

Process

At the first International Association for Vegetation Science (IAVS) Ecoinformatics Working Group meeting in 2003, vegetation scientists from 22 countries attended and unanimously endorsed creation by IAVS of an international exchange standard for vegetation-plot data. The Council of IAVS then invited this group to work with interested colleagues to develop a standard for approval and endorsement by IAVS.

A vegetation data exchange standard, Veg-X, was developed largely through two workshops that formulated requirements and allowed discussion and development of the draft schema. The first workshop (held in Apr 2007), which included 12 participants from six countries representing a range of interested groups and user perspectives, formulated a common set of goals, concepts, requirements and terminology for plot-based vegetation data that represented the majority of use-cases that could be identified. This included identification of the major components of plot-based vegetation data and a conceptual model of how these components relate to each other. Enumerating data components and how they are conceptually related to one another was a necessary first step towards defining the data structures that would accommodate data from different sources. A second workshop was held in Jun 2008 with a smaller group who developed an XML schema representation for consideration by the international vegetation science community. After Dec

2008 we conducted our discussion and developments under the umbrella of TDWG as a subgroup of the Observation Specimen Records Interest group called Vegetation Web (<http://wiki.tdwg.org/twiki/bin/view/Vegetation/WebHome>).

Major Components of Plot-Based Vegetation Data

The following set of component definitions, and their logical relationships, specifies the minimum level of data structuring required to map plot-based vegetation data from and to different database implementations.

Plots

A plot is a geographically delimited place or space, defined by one or more points, lines, polygons, or volumes that delimit the core vegetation observation (i.e. measurement). Its boundaries are assumed to be constant over time. Alternatively, plots may have no explicit bounds and refer to an area of inference, which is an indefinite but homogeneous area surrounding the centroid of the plot. A plot may be related to other plots in order to accommodate nestedness, contiguity and other types of spatial links. A plot record contains or references information about fundamental plot attributes that will not change over time (e.g. topography, spatial location, altitude). In some instances, a plot record may include a relative coordinate system to spatially locate objects (e.g. organisms) within its boundaries.

Plot observations

The essential unit of plot-based vegetation data is a plot observation – a set of measurements made or recorded on a given plot during a given period of time. A plot observation contains or references information about attributes that might differ between observation events (e.g. vegetation height, species presence or abundance, soil attributes), as opposed to fundamental plot attributes, which are considered static. Plot observations can include many component groups of records (e.g. floristic, climatic, edaphic). The basic unit of grouping plot observations themselves is typically temporal (i.e. observations made during a given period of observation) or spatial (i.e. observations falling within a coarser spatial unit).

Organism observations

A plot observation typically includes one or more observations about living or dead organisms occurring within the plot boundaries. In some cases organisms are recorded individually (optionally with their location within the plot) and measured attributes (e.g. phenology, morphological attributes such as diameter or height) comprise the information of interest. These measurements may be

repeated for the same individual on multiple occasions (censuses). We refer to such observations as individual organism observations. In other cases, measurement of individual organisms is considered either impractical (e.g. grasses, mosses and other small or clonal organisms) or too detailed given the aims of the research project, and the measurement act simply records an abundance value such as total number of individuals, biomass or cover (% or cover class; see Jennings et al. 2009, Table 3) for the set of all organisms considered to be of the same kind (usually a taxon, but may be a morphospecies or aggregate of taxa). We refer to the last kind of observations as aggregated organism observations.

Organism and community identifications

Organisms detected in plot observations can be interpreted as representing biological taxa that exist in various classifications. Similarly, the entire community can be the subject of classification and naming conventions (Weber et al. 2000; US Federal Geographic Data Committee 2008). Such identification events are a crucial scientific step because they allow organism (or community) observations to be linked to taxon (or community) attributes (biogeographical, ecological, evolutionary, physical).

Strata and stratum observations

When the aim is to describe the vertical structure of vegetation, vegetation strata (e.g. tree, shrub and herb layers) provide a simple way to group organism observations. For example, organism observations may be grouped into strata described by the lower and upper vertical bounds and an associated assessment of cover, or into diameter or height classes accompanied by counts or measurements that differ among classes (e.g. cover by species of plants ≤ 1 m high, counts by species of plants > 1 m high and ≤ 2.5 cm DBH, measurements of individual trees ≥ 2.5 cm DBH). Definitions of strata differ between observational protocols, such as one protocol defining a layer as all foliage between two heights (Hurst & Allen 2007) compared with another where the layer is based on all cover by those individuals with the majority of their foliage between two heights. Moreover, classes need not be strictly defined by non-overlapping height classes, but may also include structural classes that span multiple vertical strata, such as epiphytes and vines. Although stratum observations may be lacking in some plot-based vegetation data, often they are a major component. Defining strata acknowledges that the vegetation comprising different strata can be regarded as communities responding to different environmental factors (e.g. cryptogam layer versus canopy trees; see Gillet & Gallandat 1996; Dengler 2003) and that the relative abundance

of taxa in different strata can have important implications for community dynamics (Schnitzer & Bongers 2002).

Projects

Data from plots are typically collected as part of a larger project to sample vegetation data for a particular area and purpose. Projects are typically undertaken over a specific period, such as a fieldwork season. Plots within a project may be remeasured at a later date in a subsequent project or even as part of the same project. A project may consist of many plots, and use more than one data collection method across the plots sampled (e.g. contain both relevés and diameter measurements of individual trees). However, plots within a project are typically collected using the same methodology, by the same investigator or group of investigators, and represent both a unit of intellectual property as well as a cohesive data set referenced by one or more publications. For these reasons, the ability to group by project is essential for capturing homogeneous attributes of sets of plots referenced by the same metadata.

Metadata

The ability of potential users to understand ecological data has been identified as the most critical factor in determining whether they are reused (Jones et al. 2006; Zimmerman 2008). Metadata provide descriptive information about the data that allow this critical understanding. Comprehensive metadata should describe what data are stored, why and how they were collected and their quality (Michener et al. 1997; Michener 1998). Metadata are essential for two primary reasons (Conley & Brunt 1991; Stafford 1993; Hale 2000): (1) metadata provide the information required for long-term use and reuse of a dataset (Colwell 1995), and (2) metadata allow users to ensure that their use of the data is not beyond the bounds of the questions that the data can address (Chrisman 1994). This is especially important when a user is attempting to combine data from different projects to address questions at large spatial or temporal scales.

Overview of Veg-X

Our goal was to produce a standard specifically for vegetation-plot data that is relatively simple to read and use. To achieve this we avoided highly nested structures and included major vegetation data components (e.g. plot attributes, plot observations, organisms) as top-level elements that are referenced by each other through unique identifiers (e.g. a unique numerical ID) that allow the integrity of the original linkage to be captured. Although the main logical structure of vegetation data (i.e. the logical relationships between major data components) is fixed, we allowed for alternative, user-defined ways of

grouping observations. As such, the standard can accommodate projects that are linked across time as well as longitudinal measures of plots or individuals to the extent that these are referenced in the original dataset through appropriate unique identifiers in those original sources.

The standard accommodates different data collection protocols by allowing specific aspects of data collection methods to be captured, such as whether plots were located subjectively or randomly, plot dimensions, definitions of cover-abundance scales, references to published methods, etc. The standard also allows for the original units of measurement to be retained. All elements in the standard are clearly defined. This allows synonymous terms in source datasets to be mapped to a common set of concepts, thus overcoming the problems caused by inconsistent terminologies.

The *plotObservation* is the central Veg-X element, resulting from sampling a physical plot at a specific point in time, and can be related to one or more research projects (Fig. 3). The information about a sampled plot that is fixed over time (e.g. altitude, plot identifier or name, dimensions, aspect, slope, geology) and references to related plots (e.g. a parent plot) are stored in the separate element *plot*. By structuring the plot data in this way, repeat measures and nested plots can be accommodated in the standard. For the top-level element *project* we adopted the Ecological Metadata Language (EML; <http://knb.ecoinformatics.org/software/eml/>) project module. EML provides structured, well-defined categories for documenting the key aspects of ecological datasets and is supported by a number of institutions (Jones et al. 2006).

Specific observations, either biotic or abiotic, are components of the plot observation element. The standard allows for observations of vegetation at both the individual specimen (*individual OrganismObservation*) and aggregated organism (*aggregateOrganismObservation*) levels. The standard maintains a clear distinction between the entity of interest (e.g. an individual organism, plot or stratum) and the observation act (e.g. a measuring event applied to it). Together with unique identifiers that maintain the integrity of references between individual records within each component (e.g. between a plot and all the measuring events applied to it), the separation of components allows the standard to store multiple observations of the same entity (e.g. a plot or a tree). Analogously, a single observation event (e.g. a plot observation) may apply to multiple entities, thereby providing explicit grouping of entity observations. Each entity of interest (e.g. a tree) may have multiple observed properties (e.g. height, DBH) whose values are determined through measurement using a specific procedure or a method belonging to a particular protocol. Unlike individual organism observations, aggregated organism observations

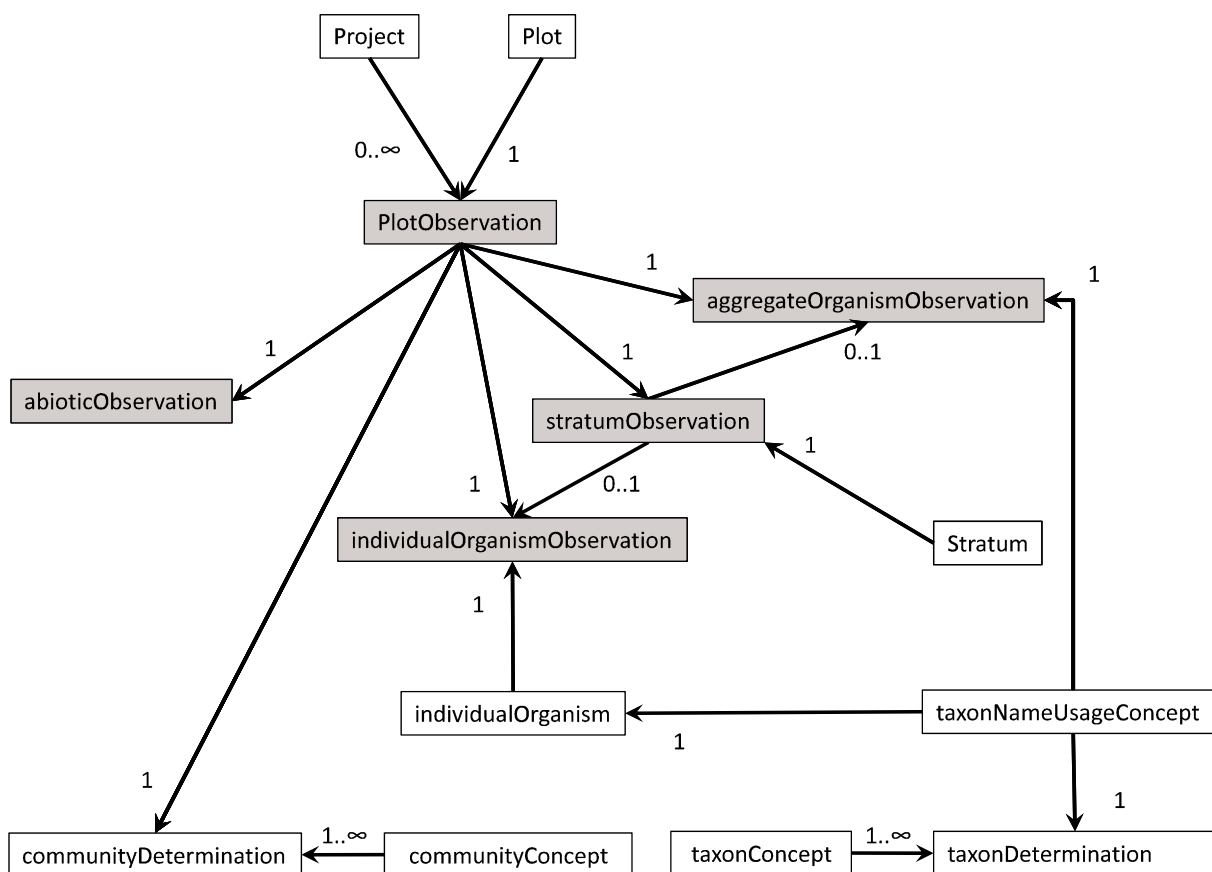


Fig. 3. Main Veg-X elements and their logical relationships. Arrows indicate that an identifier of the origin element is referenced in the destination element. Accompanying numbers indicate the number of instances of the origin element that are allowed to be referenced in the destination element. Observations are in tinted boxes.

do not relate to a specific physical entity but provide estimates of the importance of a (abstract) taxonomic entity within the plot, such as through a cover estimate. Strata can be the subject of stratum observations (e.g. percentage of tree cover, tree height) and can be linked to both individual and aggregated organism observations (Fig. 3).

The standard also maintains a distinction between taxonomic identity (the taxon or taxon concept) and how these identities are applied to particular observations of organisms. This is done through three top-level elements: (1) *taxonConcept* is a specific taxonomic concept (i.e. a name-reference combination); (2) *taxonNameUsageConcept* is a particular type of organism referencing one or more individual or aggregate occurrences in one or more plots and typically associated with a taxon name string used by the author of the dataset; (3) *taxonDetermination* is an assertion, made by a party, linking one or more *taxonConcepts* to a *taxonNameUsageConcept*. All the organism observations referencing a given *taxonNameUsageConcept* are affected by determination events applied to it. This allows different determinations and taxonomic concepts

to be associated with a vegetation entity so temporal changes in opinion regarding identification (i.e. “determination history”) can be recorded and both formal (i.e. taxon names) and informal (e.g. “field names”, “morphospecies”) names applied simultaneously to a particular organism observation can be preserved. The fact that the *taxonNameUsageConcept* is not nested within observations permits the same name string to be reused within the scope of the dataset. Community determinations are handled in a similar way: *communityDetermination* elements allow a given plot observation to be related to one or multiple community concepts, although the lack of informal field names removes the need for a *communityNameUsageConcept* entity.

Although the standard supports fully specified taxonomic concepts, it does not require them. This is important as the full concept is unspecified and, furthermore, unrecoverable for most legacy data. Conversely, because the schema can accommodate determination information (who did the identification, when, and with what reference), in theory it could be possible to recover concepts

for many legacy datasets – in particular, tropical forest plots where such information is commonly preserved in the form of herbarium voucher specimens.

Veg-X is written as an XML schema, which is a definition of user-defined tags to structure textual information in order to create self-describing datasets. XML is an open standard, and XML files are both machine and human-readable (they are stored in plain-text ASCII format). These characteristics help to ensure that data in this format will be accessible in the future. We made use of existing XML schema definitions, which we incorporated as modules of our schema. Specifically, we used parts of the EML to define entities such as projects or protocols. Similarly, we used TDWG standards such as the Darwin Core geospatial extension (<http://wiki.tdwg.org/twiki/bin/view/DarwinCore/GeospatialExtension>) to deal with georeferences, and the Taxon Concept Schema (<http://www.tdwg.org/standards/117/>) for taxon names and concepts. Updated versions of the Veg-X XML schema, as well as a prototype import/export utility, can be found at the TDWG Vegetation Web group page (<http://wiki.tdwg.org/twiki/bin/view/Vegetation/WebHome>).

Implementation

Some authors contend that “logistical barriers to data sharing are illusory” and imply that the main reason ecologists do not share data is because they are withholding data for their own future gain (Parr & Cummings 2005). We contend that at least for vegetation-plot data, the major impediment to data sharing and aggregation has been the lack of an accepted exchange standard for these data. Although some database systems are already capable of exporting vegetation data in XML formats (VegBank XML, TurboVeg XML, ESVeg), the corresponding XML schemas are structurally a copy of the database structure they came from. As such, they are useful to exchange data among people sharing the same database system but do not allow data exchange across different systems. In contrast, Veg-X has been conceived as a standard to facilitate exchange of plot data between vegetation databases, and from databases to analytical tools (Fig. 1b). As a result, it should outperform the former schema definitions by being flexible enough to exchange and combine data that were originally sampled with different needs and stored in different data structures. Furthermore, its flexibility and readability makes it a good candidate for archiving vegetation data underpinning scientific papers and for guiding the design of future vegetation-plot databases. Like a data warehouse, Veg-X allows vertical integration of data and thus facilitates data exchange and sharing. It is more general than specific vegetation-plot databases, but cannot accommo-

date all content and data management requirements of all vegetation-plot databases (i.e. it does not maintain referential integrity between entities and is not a suitable platform for live edits of data). With appropriate metadata, however, Veg-X can provide a suitable platform for archiving raw vegetation-plot data.

As data standards such as Veg-X are developed and adopted, resolving issues surrounding data access, ownership and intellectual property rights (Frankel 1999; Dresser 2000; Jones et al. 2006) will become increasingly urgent. Whereas preparing a dataset for archiving requires considerable effort to ensure that it is properly documented, hard copies or ancillary information are available, and the data are properly organized (Wiser et al. 2001; Jones et al. 2006; Piwowar et al. 2007), so using these data may simply require requesting or downloading them. There are clear advantages to being a data user; less clear are the benefits to data providers. Despite these issues, a recent survey showed that 95% of ecologists and evolutionary biologists believe that data should be publicly archived (Whitlock et al. 2010). A number of journals currently require or are poised to require that data supporting published results be stored in an appropriate public archive (Nelson 2009; Whitlock et al. 2010). There are further steps that would encourage archiving and sharing of data, the practice of which is currently not well rewarded in the science system (Olson & McCord 2000; Jones et al. 2006). These steps include databanks having clear data use and citation policies (Olson & McCord 2000) as well as mechanisms to restrict data from public use for a fixed period (Porter & Callahan 1994; Hale et al. 2003; Toronto Data Release Workshop 2009). Further, in a study of papers concerning cancer clinical trials, those that archive their data had a substantially higher citation rate than those that do not (Piwowar et al. 2007).

The full utility of a vegetation data exchange standard will be realized by the development of related ecoinformatic tools (Fig. 1). Tools will be required for translating data from multiple sources into the exchange standard and for performing tasks such as data validation, standardization of units, taxon name correction, etc. Some of this work is already under way: both the NVS databank and VegBank have developed import–export routines using Veg-X, and the NVS databank has begun using Veg-X to import a range of legacy data formats. Other tools will be needed to export data from the exchange standard to analytical software such as statistical packages or geographic information systems. Such applications would alleviate the expensive task of developing multiple import–export routines for system-specific data formats. An additional advantage of this architecture is that data providers will be able to share and archive their data in a single format. Ultimately, the

development of data mining and discovery tools based on schemas such as Veg-X will allow millions of observations from multiple sources to be queried and merged across an unprecedented range of spatial and temporal scales, enabling novel analyses and predictions of vegetation function, distribution and change.

Acknowledgements

This work was financially supported by the New Zealand Foundation for Research, Science and Technology (NVS IO - C09X0502 Ecosystem resilience OBI, <http://www.frst.govt.nz/>), the New Zealand Terrestrial and Freshwater Biodiversity Information Systems program (<http://www.biodiversity.govt.nz/land/nzbs/tfbis/tfbis/>), the ARC-NZ Research Network for Vegetation Function (Australian Research Council and Landcare Research, NZ; <http://www.vegfunction.net/>), the US National Evolutionary Synthesis Center (<http://www.nescent.org/>), the US National Center for Ecological Analysis and Synthesis (<http://nceas.org>) and the US National Science Foundation (grants DBI-9905838 and DBI-0213794 to R.K.P.). Matt Bolton, Stephan Hennekens, Matt Jones, Michael Lee, Kristin Snow and Jerry Cooper contributed to the formulation of the ideas underlying Veg-X at the first Vegetation Exchange Standard workshop in 2007. We thank the Botanical Information and Ecology Network (<http://www.nceas.ucsb.edu/featured/enquist>), hosted by the National Center for Ecological Synthesis and Analysis (<http://www.nceas.ucsb.edu/>) and Matt Wheeler, in particular, for testing and improving upon Veg-X, and Christine Bezar for final editing and formatting.

References

- Berendsohn, W.G. 1995. The concept of "potential taxa" in databases. *Taxon* 44: 207–212.
- Boyle, B., Meyer, H.W., Enquist, B.J. & Salas, S. 2008. Higher taxa as paleoecological and paleoclimatic indicators: a search for the modern analog of the Florissant fossil flora. *Geological Society of America Special Paper* 435: 33–51.
- Chrisman, N.R. 1994. Metadata required to determine the fitness of spatial data for use in environmental analysis. In: Michener, W.K., Brunt, J.W. & Stafford, S.G. (eds.) *Environmental information management and analysis: ecosystems to global scales*. pp. 177–190. Taylor & Francis, London, UK.
- Colwell, R.K. 1995. Ecological Society of America special committee on ESA communications in the electronic age. *Bulletin of the Ecological Society of America* 76: 120–131.
- Condit, R. 1998. *Tropical forest census plots: methods and results from Barro Colorado Island, Panama and a comparison with other plots*. Springer-Verlag, Berlin, DE.
- Conley, W. & Brunt, J.W. 1991. An institute for theoretical ecology? Part V: practical data management for cross-site analysis and synthesis of ecological information. *Coenoses* 6: 173–180.
- Dengler, J. 2003. Entwicklung und Bewertung neuer Ansätze in der Pflanzensoziologie unter besonderer Berücksichtigung der Vegetationsklassifikation [Archiv naturwissenschaftlicher Dissertationen no. 14]. Galunder, Nümbrecht, DE.
- Dengler, J., Chytrý, M., Ewald, J., Glöckler, F., Jansen, F., De Cáceres, M., Finckh, M., Hennekens, S.M., Mucina, L., Oldeland, J., Peet, R.K., Rodwell, J.S., Schaminée, J.H.J. & Spencer, N. 2011. The Global Index of Vegetation-Plot Databases (GIVD): a new resource for vegetation science. *Journal of Vegetation Science* 21, doi: 10.1111/j.1654-1103.2011.01265.x.
- Dresser, R. 2000. Accountability in science and government: is access the answer? *Hasting Center Report* 2000(May–June): 29–30.
- Edwards, J.L., Lane, M.A. & Nielsen, E.S. 2000. Interoperability of biodiversity databases: biodiversity information on every desktop. *Science* 289: 2312–2314.
- Ellison, A.M. 2010. Repeatability and transparency in ecological research. *Ecology* 91: 2536–2539.
- Enquist, B.J. & Niklas, K.J. 2001. Invariant scaling relations across tree-dominated communities. *Nature* 410: 655–660.
- Fernald, M.L. 1950. *Gray's Manual of Botany*. 8 ed. American Book Co., New York, NY, US.
- Frankel, M.S. 1999. Public access to data. *Science* 283: 1114.
- Franz, N.M. & Peet, R.K. 2009. Towards a language for mapping relationships among taxonomic concepts. *Systematics and Biodiversity* 7: 5–20.
- Franz, N.M., Peet, R.K. & Weakley, A.S. 2008. On the use of taxonomic concepts in support of biodiversity research and taxonomy. Symposium proceedings. In: Wheeler, Q.D. (ed.) *The new taxonomy*. pp. 63–86. [Systematics Association Special Volume 74.] Taylor & Francis, Boca Raton, FL, US.
- Fridley, J.D., Vandermaast, D.B., Kuppinger, D.M., Manthey, M. & Peet, R.K. 2007. Co-occurrence-based assessment of habitat generalists and specialists: a new approach for the measurement of niche width. *Journal of Ecology* 95: 707–722.
- Gillet, F. & Gallandat, J.D. 1996. Integrated synusial phytosociology: some notes on a new, multiscale approach to vegetation analysis. *Journal of Vegetation Science* 7: 13–18.
- Gleason, H.J. 1952. *The new Britton and Brown illustrated flora of the Northeastern United States and adjacent Canada*. Hafner Press, New York, NY, US.
- Gosling, W., Mayle, F., Tate, N. & Killeen, T. 2009. Differentiation between Neotropical rainforest, dry forest, and savannah ecosystems by their modern pollen spectra and implications for the fossil pollen record. *Review of Palaeobotany and Palynology* 153: 70–85.
- Hale, S.S. 2000. How to manage data badly (part 2). *Bulletin of the Ecological Society of America* 81: 101–103.
- Hale, S.S., Miglarese, A.H., Bradley, M.P., Belton, T.J., Cooper, L.D., Frame, M.T., Friel, C.A., Harwell, L.M., King, R.E., Michener, W.K., Nicolson, D.T. & Peterjohn, B.G. 2003.

- Managing troubled data: coastal data partnerships smooth data integration. *Environmental Monitoring and Assessment* 81: 133–148.
- Hall, G.M.J., Wiser, S.K., Allen, R.B., Beets, P.N. & Goulding, C.J. 2001. Strategies to estimate national carbon biomass from forest inventory data: the 1990 New Zealand baseline. *Global Change Biology* 7: 389–403.
- Hennekens, S.M. & Schaminée, J.H.J. 2001. Turboveg, a comprehensive database management system for vegetation data. *Journal of Vegetation Science* 12: 589–591.
- Hurst, J.M. & Allen, R.B. 2007. *A permanent plot method for monitoring indigenous forests – field protocols*. Manaaki Whenua – Landcare Research, Lincoln, NZ.
- Jansen, F. & Dengler, J. 2010. Plant names in vegetation databases – a neglected source of bias. *Journal of Vegetation Science* 21: 1179–1186.
- Jennings, M.D., Faber-Langendoen, D., Loucks, O.L., Peet, R.K. & Roberts, D. 2009. Standards for associations and alliances of the U.S. national vegetation classification. *Ecological Monographs* 79: 173–199.
- Jones, M.B., Schildhauer, M.P., Reichman, O.J. & Bowers, S. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution and Systematics* 37: 519–544.
- Kartesz, J. 1999. *Synthesis of the North American flora*. North Carolina Botanical Garden, Chapel Hill, NC, US.
- Kennedy, J., Kukla, R. & Paterson, T. 2005. Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In: Ludäscher, B. & Raschid, L. (eds.) *Data integration in the life sciences: Proceedings of the Second International Workshop*. San Diego, CA, July 20–22, 2005. DILS LNBI 3615, 2005, 80–95.
- Malhi, Y., Phillips, O.L., Lloyd, J., Baker, T.R., Wright, J., Almeida, S., Arroyo, L., Frederiksen, T., Grace, J., Higuchi, N., Killeen, T., Laurance, W.F., Leão, C., Lewis, S.L., Meir, P., Monteagudo, A., Neill, D., Núñez Vargas, P., Panfil, S.N., Patiño, S., Pitman, N., Quesada, C.A., Rudas-Ll, A., Salomão, R., Saleska, S., Silva, N., Silveira, M., Sombroek, W.G., Valencia, R., Vásquez Martínez, R., Vieira, I.C.G. & Vinceti, B. 2002. An international network to monitor the structure, composition and dynamics of Amazonian forests (RAINFOR). *Journal of Vegetation Science* 13: 439–450.
- Michener, W.K. 1998. Ecological metadata. In: Michener, W.K., Porter, J.H. & Stafford, S.G. (eds.) *Data and information management in the ecological sciences: a resource guide*. pp. 47–51. Long Term Ecological Research Network Office, University of New Mexico, Albuquerque, NM, US.
- Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B. & Stafford, S.G. 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications* 7: 330–342.
- Mucina, L., Schaminée, J.H.J. & Rodwell, J.S. 2000. Common data standards for recording relevés in field survey for vegetation classification. *Journal of Vegetation Science* 11: 769–772.
- Mueller-Dombois, D. & Ellenberg, H. 1974. *Aims and methods of vegetation ecology*. John Wiley and Sons, New York, NY, US.
- Olson, R.J. & McCord, R.A. 2000. Archiving ecological data and information. In: Michener, W.K. & Brunt, J.W. (eds.) *Ecological data: design, management and processing*. Blackwell Science, Oxford, UK.
- Nelson, B. 2009. Empty archives. *Nature* 461: 160–163.
- Parr, C.S. & Cummings, M.P. 2005. Data sharing in ecology and evolution. *Trends in Ecology and Evolution* 20: 362–363.
- Peet, R.K., Wentworth, T.R. & White, P.S. 1998. A flexible, multipurpose method for recording vegetation composition and structure. *Castanea* 63: 262–274.
- Piwowar, H.A., Day, R.S. & Fridsma, B. 2007. Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2: e308, doi: 10.1371/journal.pone.0000308.
- Porter, J.H. & Callahan, J.T. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. In: Michener, W.K., Brunt, J.W. & Stafford, S.G. (eds.) *Environmental information management and analysis: ecosystems to global scales*. pp. 193–202. Taylor & Francis, London, UK.
- Russo, S.E., Jenkins, K.L., Wiser, S.K., Uriarte, M., Duncan, R.P. & Coomes, D.A. 2010. Interspecific relationships among growth, mortality, and xylem traits of 59 woody species from New Zealand. *Functional Ecology* 24: 253–262.
- Schaminée, J.H.J., Hennekens, S.M., Chytry, M. & Rodwell, J.S. 2009. Vegetation-plot data and databases in Europe: an overview. *Preslia* 81: 173–185.
- Schnitzer, S.A. & Bongers, F. 2002. The ecology of lianas and their role in forests. *Trends in Ecology & Evolution* 17: 223–230.
- Stafford, S.G. 1993. Data, data everywhere but not a byte to read: managing monitoring information. *Environmental Monitoring and Assessment* 26: 1125–1141.
- Stone, D. 1997. *Carya*. In: Flora of North America Editorial Committee. (ed.) *Flora of North America North of Mexico*. 3. pp. 416–425. Oxford University Press, New York, NY, US, and Oxford, UK.
- Swenson, N.G. & Enquist, B.J. 2007. Ecological and evolutionary determinants of a key plant functional trait: wood density and its community-wide variation across latitude and elevation. *American Journal of Botany* 94: 451–459.
- Swenson, N.G., Enquist, B.J., Thompson, J. & Zimmerman, J.K. 2007. The influence of spatial and size scale on phylogenetic relatedness in tropical forest communities. *Ecology* 88: 1770–1780, doi: 10.1890/06-1499.1.
- Toronto International Data Release Workshop. 2009. Prepublication data sharing. *Nature* 461: 168–170.
- U.S. Federal Geographic Data Committee (FGDC), Vegetation Subcommittee. 2008. National Vegetation Classification Standard (Version 2). FGDC-STD-005-2008. Available at: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/vegetation/NVCS_V2_FINAL_2008-02.pdf.

- van der Maarel, E. 1979. Transformation of cover-abundance values in phytosociology and its effects on community similarity. *Vegetatio* 39: 97–114.
- Weber, H.E., Moravec, J. & Theurillat, J.-P. 2000. International code of phytosociological nomenclature. 3rd ed. *Journal of Vegetation Science* 1: 739–768.
- Weiser, M.D., Enquist, B.J., Boyle, B., Killeen, T.J., Jørgensen, P.M., Fonseca, G., Jennings, M.D., Kerkhoff, A.J., Lacher, T.E. Jr, Monteagudo, A., Núñez Vargas, M.P., Phillips, O.L., Swenson, N.G. & Vásquez Martínez, R. 2007. Latitudinal patterns of range size and species richness of New World woody plants. *Global Ecology and Biogeography* 16: 679–688.
- Whitlock, M.C., McPeck, M.A., Rausher, M.D., Rieseberg, L. & Moore, A.J. 2010. Data archiving. *American Naturalist* 175: 145–146.
- Whittaker, R.H. 1975. *Communities and ecosystems*. 2nd ed. Macmillan, New York.
- Wiser, S.K., Bellingham, P.J. & Burrows, L. 2001. Managing biodiversity information: development of the National Vegetation Survey Databank. *New Zealand Journal of Ecology* 25: 1–17.
- Zimmerman, A.S. 2008. New knowledge from old data: the role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values* 33: 631–652.