Forest Ecology and Management 316 (2014) 21–31

Contents lists available at ScienceDirect

Forest Ecology and Management

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

Data and database standards for permanent forest plots in a global network

Richard Condit [a,*], Suzanne Lao [a], Anudeep Singh [a], Shameema Esufali [a], Steven Dolins [b]

[a] Center for Tropical Forest Science, Smithsonian Tropical Research Institute, Unit 0948, APO AA 34002, United States
[b] Department of Computer Science and Information Systems, Bradley University, Peoria, IL 61625, United States

ARTICLE INFO

Article history:
Available online 13 October 2013

Keywords:
Forest plot database
Integrity errors
Normalized data
Tree census

ABSTRACT

The Center for Tropical Forest Science established a network of 50-ha forest inventory plots in the 1980s, and now assists local scientists with field and database methods at 44 large-scale plots across boreal, temperate, and tropical forest biomes. We published detailed field methods over a decade ago, but at that time, data were maintained in spreadsheet-like formats, most harboring design flaws that resulted in frequent errors. We since established detailed database methods and a normalized data model for housing multiple censuses of large plots. Our largest databases include >2 million measurements, and each has a master version on a server where all collaborators can access and edit data. This paper focuses on the data requirements for handling tree census data and how to design databases to meet these requirements and to ensure data integrity. There are six key elements of a tree census which the database must reflect: (1) measurements, including individual trees (genetic units), stems within trees, and multiple measurements of stems; both field and data methods must assure that every tree, stem, and measurement is precisely identified and can be relocated easily; (2) coordinates, including quadrats within a plot, because field mapping is usually done by assigning *x*–*y* coordinates relative to local quadrat markers; (3) taxonomy, carrying a species identity for every tree with a history of individual re-identifications; (4) personnel, with records of the people who performed field and data work per quadrat; (5) assessment of field error via random re-measurements; and (6) a log of changes and a system of archiving so that errors can be tracked and past versions can be reconstructed and cited in publications. A well-designed database model reduces a variety of integrity errors and improves access to data tables in identical formats across many plots, allowing data analyses to be easily replicated and results to be compared. The principal disadvantage is that complexity of the database requires experienced data managers.

Published by Elsevier B.V.

1. Introduction

Trees are perfect subjects for population biology. They are large, easy to find, and do not move so can be relocated in future censuses. Whereas population studies of mammals, birds, and insects usually involve inference from subsamples and must account for failure to detect individuals, a *tree plot* is a complete count of all living and dead individuals over a predefined area (Table 1). The ease with which tree plots are censused coupled with the rigor they lend to population studies have fostered a proliferation of plot studies in every forest biome (Ayyappan and Parthasarathy, 1999; Phillips et al., 2003; Canham et al., 2006; Coomes et al., 2009; ter Steege et al., 2006). Some have been in place for decades (Crow, 1980; Whitney, 1984; Franklin and DeBell, 1988; Condit et al., 2012; Lilleleht et al., 2014), and tree plots are the basis of national forest inventories (Tomppo et al., 2010; Álvarez-González et al., 2014).

The greatest asset of a tree census comes from the recensus, when individual trees are relocated and remeasured, leading to a rigorous and repeatable count of the population size of all species present and measurements of individual growth and death rates. These measurements and remeasurements (Table 2) are the focus of data collection in the field and data maintenance in a modern database.

The measurements collected on a single tree are brief and straightforward, but most plots in the Center for Tropical Forest Science (hereafter CTFS) network, and in many national inventories, include hundreds of thousands of records, far more than can be proofread by eye. Prior to the year 2000, CTFS data were housed in spreadsheet-like tables, and we became familiar with a variety of oft-repeated errors caused by poor design. Particularly frustrating were integrity errors resulting from mismatched records, many of which occurred after data were transcribed into the computer (Table 3). Some of these errors are magnified from census to census

* Corresponding author. Tel.: +507 212 8045.
E-mail address: conditr@gmail.com (R. Condit).

0378-1127/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.foreco.2013.09.011

Table 1
Definitions of trees and tree plots.

Term	Definition
Tree	Woody plant, free-standing (i.e. excluding lianas), including all parts of one genetic unit
Stem	A single woody axis of a tree (i.e., one tree can have one or more stems)
Tree plot	Every individual tree stem in a predefined region, above a minimum size, is included 1. Coordinates estimated 2. Stem diameter measured 3. Species identity determined
Other vegetation surveys	Not qualifying as tree plots 1. Enumerations of individuals within plots, but no coordinates 2. Percent cover measured in a plot instead of enumeration of individuals 3. A subset of individuals or species enumerated, instead of every one

because species names and remeasurements are connected to the wrong trees.

We relate here our efforts to apply database theories and standard practices to tree plot data (Codd, 1971; Elmasri and Navathe, 2000; Date, 2004). The database model we designed to store Barro Colorado plot data, and subsequently data from other CTFS plots, follows basic theories of normalization and eliminates many common errors (Codd, 1971). Database normalization refers to a series of design rules for data tables whose intent is to minimize data redundancy, because redundant (i.e. repeated) data are prone to anomalies during updates. An anomaly is created when a repeated datum is changed in one place but not another; omitting repetition thus prevents anomalies.

This system, referred to here as the CTFS Data Model, has been designed and implemented in a database management system, MySQL, with interfaces written in HTML and PHP for online access (Hubbell et al., 2010). These interfaces are outside the scope of this presentation, as is any discussion of field methods, which Condit (1998) cover in detail. Our goal here is to describe the data standards adhered to in the CTFS plot network and enough details of the CTFS Data Model to explain how it upholds the standards.

2. Components of a tree census database

2.1. Tree Measurements and Re-measurements

The principal data from a tree census are repeated measurements of stems. This is straightforward at its most basic, with diameter of individual stems measured at a consistent position through time. The biggest single hindrance arises from multiple stems, where a single tree (a genetic unit) has several separate stems at breast height. Indeed, if trees always had single trunks, and never forked or sprouted from the base, the most frustrating difficulties with plot remeasurements would vanish. Yet multiple-stemmed trunks cannot be ignored: there are many tree species that routinely grow clonally, with many separate stems in a genetic unit. Examples include several palm species in the American tropics, Myrtaceae in Valdivian forests of Chile, and a variety of coppicing north temperate *Quercus*.

Assuring consistent measurements requires precise identification of individual stems and measurement positions, and this leads to three hierarchies in the measurement data: (1) Single trees have multiple stems. (2) Each stem has multiple measurements through time. (3) Each measurement may have associated with it several attributes, including breaks, death, swellings at the measurement point, etc.

2.1.1. Multiple stems

Reliable and permanent identification of multiple stems within a tree is crucial for ensuring that repeat measurements can be correctly linked. A routine method for tree plots should thus be to attach individually-numbered tags to every individual stem, just as traditional forest plots assume that every tree should be tagged. The preferred tagging system includes a principal sequence of numbers for individual trees and a subordinate sequence for stems; the two types of tag differ in form so are readily distinguished. So trees are numbered 1, 2, 3, ... through many thousands, and stems within trees have letters 'A', 'B', 'C', etc. In this system, there are both tree tags and stem tags. There is a commonly used alternative in which all stems get a single sequence and one tag type. The disadvantage to the latter method is that there is no tag defining the tree, and we consider the former the CTFS standard. In the former method, stems are readily associated with the tree they belong with, because each is identified by both a tree tag and a stem tag.

In the preferred method, a tree with a single trunk is not given a stem tag in the field, since it is not needed, but it is assigned stem tag 'A' in the database. In later censuses, if a trunk sprouts a second

Table 2
Standard stem measurements in CTFS tree plots. Alternatives indicate widely used variations on each measurement.

Measure	CTFS standard	Alternative
Height of diameter measure (*HOM*)	'Breast height', or 1.3 m	Some sites use 1.37 m as standard. *HOM* must be above buttresses. When buttresses grow upward, additional *HOM* needed
Minimum size limit	1 cm	10 cm
Death	Criteria are seldom specified but usually include 1. Leafless state in non-deciduous species 2. Trunk broken 3. Rotten wood encircling trunk 4. Trunk entirely fallen or vanished	
Species	Standard botanical taxonomy	Morphospecies (i.e., unidentified), Subspecies
Local coordinates	*X*, *Y* distance to a precisely surveyed grid post	Polar coordinates to a grid post
Stem diameter	Small stems: Largest axis (with calipers) Large stems: Circumference (with dbh tape)	Dbh tape for all sizes

Table 3
Common integrity errors that are avoided with normalized tables, i.e. errors that cause mis-matching or undefined records. Not included are errors of measurement or identification that cannot be prevented by any database system. By *typos*, we include any transcription error, either in the field or later when copying data into computer tables.

Error	Cause
Tag number duplicated in one plot	Typos, or sometimes tags really are duplicated
Multiple spellings of one species' name	One typo among 1000 records for one species creates a 'new' species
Species codes assigned to trees but missing in the species table	Typos
Stems within a tree at widely different locations	Typos in tree tag
Stems within a tree with >1 species' name	Typos in tree tag
Measurement attribute codes without definition	Typo in attribute code

Table 4
Definition of types of individual stems on a single tree (i.e., within a genetic unit).

Type	Description
Main	Largest stem emerging from the ground
Secondary	Any additional stem emerging separately from the ground but evidently linked underground to the main stem
Fork	Division off the main stem within 30° of vertical (i.e., close to upright)
Branch	Division off the main stem >30° vertical (i.e., close to horizontal)

stem, it becomes necessary to put an 'A' tag on the first trunk and a 'B' tag on the second.

In CTFS plots, four stem types are defined (Table 4). In the Barro Colorado census, branches below breast height, even horizontal branches, are considered additional stems, with diameters recorded. But this is not typical, and many forest censuses do not include the branch category as stems. This is an important reason to record stem type, so that branches can be excluded in comparisons with plots where they are not measured.

Sprouting of new stems off existing trunks is one of the most confusing aspects of a recensus. There are two distinct sprouting scenarios.

First, a tree already having two or more stems loses some stems while others remain alive. In this case, all stems received tags in the earlier census, making it clear that one stem has been lost. We use 'lost_stem' with some care. If one stem dies while others on the same tree remain alive, it must be noted differently from tree death (see more on tree death below).

The second scenario is more complicated. A single stem on a tree breaks off (or otherwise dies), but sprouts at the base to produce a new stem. We wish to assign the new stem a distinct tag, but also want to associate it with the original tree. The new sprout retains the same tree tag, but is assigned the letter 'B', since the first, now deceased stem, was 'A'. This scenario is where the alternative tagging system, in which a single sequence of tags is applied to all stems, is at its worst, because the sprout must get a new number distant in sequence from the number on the deceased stem, while the original number must be removed in the field but remain in the database.

2.1.2. Consistent HOM

Given that stems are tagged and can be relocated, the next crucial step is identifying the *height-of-measure* (*HOM*) in consecutive censuses. At its simplest, this means relocating breast height (Table 2) and remeasuring the diameter there. But some trees cannot be measured at this standard height, such as when buttresses, irregularities, or lianas make it an inappropriate diameter (Condit, 1998, provides details). At Barro Colorado, paint is used to mark the trunk at the point-of-measure whenever it is not 1.3 m, but otherwise not; there are many plots, however, where every single stem is painted. Another way to ensure consistent *HOM*s is to use the nail that bears the tag as an indicator, so that the measurement is always made at the nail, or offset a consistent distance from the nail.

Given the importance of the height of measure, it needs to be part of the data. Whenever the normal height is used, we do not record it in the field, but it is later assigned automatically. Whenever the *HOM* is not 1.3 m, it must be recorded in the field.

2.1.3. Measurement attributes

Attributes associated with a dbh measurement are a few discrete states, most importantly whether a stem is broken or the entire tree is dead, both of which are required information in a tree plot. Other states relevant to the diameter, such as stem irregularities or other difficulties that might be taken into consideration when interpreting stem growth, are required within the CTFS standard. Additional site-specific attributes can be included.

First, death needs to be recorded clearly. We have encountered plot data where field notes for an individual are simply blank when the tree is dead, because there is no diameter to measure. The most confusion we have encountered, though, harks from the difference between an entire tree dying and individual stems dying. These two events must be recorded differently, otherwise the death of a tree has to be deduced based on the records for all its stems. In the CTFS Data Model, we have adopted definitions to standardize this difference: the attribute 'Death' is only applied when an entire tree is dead. When a stem dies or disappears while the tree remains alive, we use the attribute 'Stem lost'. For brevity, discrete attributes are recorded in the field with one- or two-letter codes, such as 'D' for dead or 'R' for broken stem, but codes can be defined differently at different sites.

The most important single goal of a recensus is to account for every tree and stem already tagged. Every stem should have a diameter measurement, or a code indicating it is lost (i.e. broken or vanished). A stem with neither diameter nor the code 'lost' is anomalous. Cases where living stems cannot be measured do arise, such as the presence of stinging wasps, and these require an additional code indicating the situation. For trees, it must be true (1) every one of its stems is accounted for, or (2) the entire tree is dead. The straightforward way to assure that every stem is accounted for during a recensus is to have field workers carrying a

Table 5
Sample tree plot data from two censuses in spreadsheet format. Columns for measurement attributes and stem tags are omitted. In this format, columns must be added to accommodate new censuses or stems, and species names tend to get misspelled on some rows but not others (Table 3). Data normalization (Table 6) overcomes these and additional problems (see main text).

Tag	Species	*X*	*Y*	DBH1	DBH2	Code1	Code2	Date1	Date2	Stem 2.1	Stem 2.2
260277	Alchornea costaricensis	15.1	34.5	161	171	.		01/04/00	01/02/05		
255640	Alchornea costaricensis	29.3	61.5	112	122	M		01/13/00	01/11/05		
255714	Alchornea costaricensis	35.2	60.2	119		.	R, X, TS	01/13/00	01/11/05		
7917	Alchornea costaricensis	71.6	164.5	348	358	.		01/20/00	01/18/05		
6982	Alchornea costaricensis	183.8	28.5	300	310	.		01/19/00	01/17/05		
6980	Alchornea costaricensis	186.7	22.5	250		.	D	01/19/00	01/17/05		
230712	Alchornea latifolia	123.2	90.6	207	217	.		01/24/00	01/22/05		
260174	Alseis blackiana	1.7	35.4	146	156	.		01/04/00	01/02/05		
260790	Alseis blackiana	1.9	114.3	228	238	.		01/19/00	01/17/05		
261167	Alseis blackiana	10.5	158.6	116		.	D	01/25/00	01/23/05		
261227	Alseis blackiana	0.2	166.3	127	137	.		01/28/00	01/26/05		
261281	Alseis blackiana	6.4	167.9	210	220	.		01/28/00	01/26/05		
8410	Alseis blackiana	17.6	186.1	250	260	.		01/31/00	01/29/05		
255322	Alseis blackiana	32.2	33.8	150	160	.		01/07/00	01/05/05		
8346	Alseis blackiana	23.1	45.3	257	267	.		01/12/00	01/10/05		
8342	Alseis blackiana	24.4	59.3	364	374	M, B		07/24/00	07/23/05	12	
8321	Alseis blackiana	38.3	118.8	440	450	B		07/25/00	07/24/05		
8311	Alseis blackiana	23.5	135.3	450	460	.		01/19/00	01/17/05		
8309	Alseis blackiana	30.3	138.3	277	287	B		07/25/00	07/24/05		
8315	Alseis blackiana	33.9	128.6	457	467	B		07/25/00	07/24/05		
8305	Alseis blackiana	30.8	151.0	313	323	.		01/19/00	01/17/05		
220455	Andira inermis	173.9	53.5	102	112	.		01/13/00	01/11/05		
255464	Annona spraguei	29.8	54.8	153	163	Q		01/12/00	01/10/05		
260639	Apeiba membranacea	1.3	91.1	109	119	.		01/18/00	01/16/05		
260793	Apeiba membranacea	3.9	112.6	142	152	.		01/19/00	01/17/05		

Table 6
Database terminology. A single table has columns called *attributes*. Normalization is accomplished by assuring tables satisfy the normal forms (Codd, 1971; Elmasri and Navathe, 2000; Date, 2004). Fourth and fifth normal forms are omitted, since they are rarely relevant.

Term	Definition	CTFS model example
Primary key	One or more columns whose values together uniquely define every row	Arbitrary integer in every table, i.e., *TreeID* in *Tree* table
Composite primary key	Primary key requiring two or more attributes	Stems are identified by tree tag and stem tag
Foreign key	A column in one table whose value is the primary key of another table	*TreeID* in the *Stem* table
Functional dependency	Relationship between two attributes in which each value of the first is associated with only one value of the second	*TreeTag* determines *SpeciesID*
Normalization	Organization of database tables to minimize redundancy, simplify insertions and updates, and assure valid joins	Dbh measurement information divided into *Species*, *Tree*, *Stem*, *Dbh* and other tables
First normal form	Every row has the same attributes, all single-valued, and there is a primary key	A table of trees should have one dbh column having a single dbh
Second normal form	If primary key is composite, every attribute depends on the full key (full functional dependency)	Table of stems only has information about stems, not trees
Third normal form	Every attribute depends on the primary key, not on non-key attributes	Table of dbhs does not have information about the plot or species
Join	Reassemble data in two or more tables in a single table based on foreign keys	Join *Species, Tree, Stem*, and *Measurement* tables into a single display

data form listing every stem, with its stem tag; next to every one of which something should be noted.

2.2. Database issues covering measurements

A simple format for handling tree measurement data from a forest plot is a "spreadsheet", using a single row to represent a single tree and its associated remeasurements and attributes. This format requires extra columns for secondary stems on a tree, and new columns whenever additional measurements are made in a new census (Table 5). The simple spreadsheet format, however, is prone to several types of errors, examples of which are covered in the following sections. Database normalization is a series of rules for data formatting whose purpose is to avoid these routine errors, while also simplifying updates and queries (Codd, 1971). The CTFS Data Model was designed to meet the three most important aspects of normalization (Table 6).

2.2.1. Database issues and multiple stems

Consider first trees and multiple stems within trees. Attributes of a tree, such as an individual's tag and species name, can be represented in a *Tree* table (Fig. 1). A tree, however, can have multiple stems, and new stems may be added during any census. If this *Tree* table also holds measurements of stem(s), with new columns added as needed, as in spreadsheet format described above, then diameters would comprise a *multivalued attribute*, with many values per individual tree. This design is inflexible: every tree must have the same number of columns for stems and measurements, and the *Tree* table would have to accommodate the maximum number observed. Many trees would have fewer than the maximum number of stems, so many rows would have columns with null values. Null values complicate data manipulation and indicates poor database design.

Thus, the first normal form means that tables do not have multi-valued attributes (Table 6). To achieve this, we use a *Stem* table to store each individual tree's stem information, including the

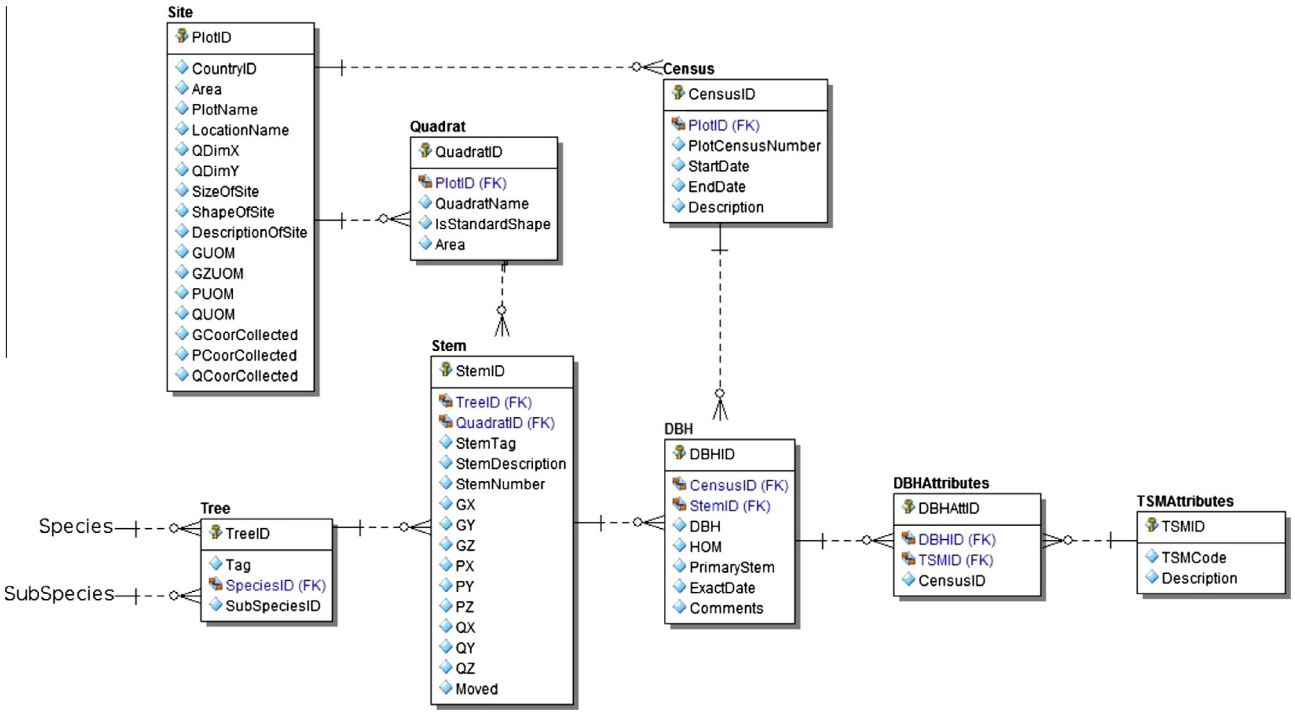

Fig. 1. Tables handling tree and stem measurements in the CTFS Data Model. Each box has a list of all attributes in a table; above the line is the primary key, always a single auto-incremented integer. FK = foreign key. Lines between tables indicate relationships in which an offspring table has an FK which equals the primary key of a parent table. Each relationship is one:many, with the many side of the relationship indicated with crow's feet (for example, one tree to many stems). See Fig. 2 for diagrams of *Species* and *SubSpecies* tables. Detailed definitions of all attributes as well as a diagram of all tables can be found in an online Appendix (http://si-pddr.si.edu/dspace/bitstream/10088/20863/1/CTFSDataModelDocumentation.html).

stem's tag and location (Fig. 1). Each row in the *Stem* table represents an individual stem; some trees have a single row in the *Stem* table, while others have two or more rows (one tree in the Barro Colorado 50-ha plot once had 105 living stems). This solves the problem of not knowing in advance how many stems there will be per tree, and simultaneously rids the table of null values.

The *cardinality* between the *Tree* and *Stem* tables is one-to-many, because one tree can have many stems while each stem is associated with only one tree. *Tree* is called the *parent* table and *Stem* the *child* table. For one-to-many relationships, the primary key of the parent table (*TreeID*) is added in the child table as a foreign key (Fig. 1, Table 6). The foreign key in the *Stem* table preserves referential integrity, i.e. a row in *Stem* has to be associated with a tree or a row that exists in the *Tree* table. Relational database management systems enforce referential integrity by forbidding updates that would create *orphaned* stems, ones for which there is no associated tree.

The stem tag consists of a simple character ('A', 'B', 'C'; sometimes '1', '2', '3'), and stem tag 'A' is found on many different stems. Thus both *TreeID* and stem tag are needed to identify a stem. For the sake of simplicity and efficiency, we created a new, alternate primary key called *StemID* which is a unique number for every stem. This is a common database practice, since a single attribute uniquely identifying rows means better performance and easier querying.

Why not use tree tag number as the primary key of the tree table? Tag numbers are sometimes repeated, often in different plots within a database, or just by mistake. Using the tag as primary key would thus violate a fundamental design principle: the primary key must uniquely identify rows. We resolved the problem caused by duplicates with the arbitrary key *TreeID*. These primary keys, both *TreeID* and *StemID*, are generated by the database to assure integrity, but they are never used by field workers, since long, arbitrary integers could easily cause confusion.

2.2.2. Database issues and multiple measurements

Given a separate table for individual stems, we still need to store multiple dbh measurements for each stem, certainly during a second census, but also because a second *HOM* is required when growing buttresses reach the previous *HOM* (Table 2). Storing several dbh values in a single row per stem would be another violation of first normal form, leading to a third table, *DBH* (Fig. 1). Each row is a single measurement of one stem in one census at one *HOM*. Since the primary key must be *StemID-Census-HOM* together, a *composite key* (Table 6), the table cannot include information about any one of those three attributes, otherwise it would violate second normal form (Table 6). The remedy for putting a table in second normal form is to create separate tables for each *partial functional dependency*, in particular a *Census* table is needed to handle information about each census (census number, start date, end date).

There is a one-to-many relationship from *Stem* to *DBH* and also a one-to-many relationship from *Census* to *DBH*. For both of these relationships, *DBH* is the child table, while *Stem* and *Census* are the parent tables (Fig. 1). Therefore, *StemID* and *CensusID* (primary key of the census table) are part of the primary key as well as foreign keys in *DBH*. Since there may be diameter measurements at two different heights on a single stem during a census, we add *HOM* to the primary key of the *DBH* table. Although *StemID*, *CensusID*, and *HOM* makeup a unique key for the *DBH* table, the efficient design is to define a single arbitrary primary key, *DBHID*, which automatically increments with every new record (Fig. 1).

2.2.3. Database issues and measurement attributes

Since there may be several attributes associated with a stem's measurement during one census, for example, 'I' for irregular and 'L' for leaning might be assigned to a single measurement. Another table is thus needed in order to avoid violation of first normal form, the *DBHAttributes* table. *DBH* is the parent table, *DBHAttributes* is the child table, and the primary key of *DBH* (*DBHID*) is a foreign

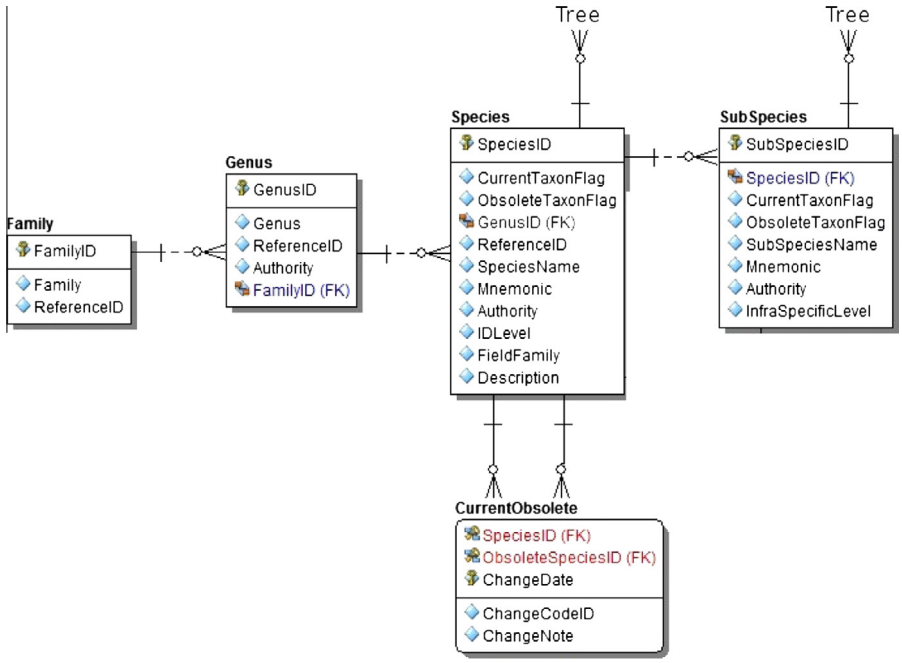

Fig. 2. Tables handling taxonomy in the CTFS Data Model. Each box has a list of all attributes in a table; above the line is the primary key, always a single auto-incremented integer. FK = foreign key. Lines between tables indicate relationships in which an offspring table has an FK which equals the primary key of a parent table. Each relationship is one:many, with the many side of the relationship indicated with crow's feet (for example, one tree to many stems). See Fig. 1 for a diagram of the *Tree* table. See http://si-pddr.si.edu/dspace/bitstream/10088/20863/1/CTFSDataModelDocumentation.html.

key in *DBHAttributes*. An additional small table called *TSMAttributes* holds definitions of the codes used in the field (such as 'I', 'L'). *DBHAttributes* is an associative table, linking *DBH* and *TSMAttributes*, necessary because the two latter tables have a many-to-many relationship. *DBHAttributes* has a one-to-many relationship with both *DBH* and *TSMAttributes*.

The *DBHAttributes* table has facts about *stem-census-HOM* combinations: every attribute is linked to a single *HOM* of one stem in one census. One crucial attribute – death of an entire tree – is instead a fact about a *tree-census* combination, and this requires a second attribute table, *TreeAttributes*. The codes used for *DBHAttributes* and *TreeAttributes* are both defined in *TSMAttributes*.

We do not record any attribute *alive* for trees or stems. It is derived from the other attributes: every tree which does not have the attribute *dead* is considered alive, though living trees may not have any stems to measure (if only the trunk base is alive). Dead stems (i.e. broken off) on living trees should carry *DBHAttribute* = 'lost', otherwise they are assumed alive and should have a diameter.

2.3. Coordinates and mapping

Horizontal positions of individual stems measured to an accuracy of 1 meter or better are referred to as *x*–*y* coordinates, and are obviously essential to a mapped plot. In standard CTFS methods, permanent stakes are surveyed to an exact 20-m grid before the tree census begins (Condit, 1998), and we consider placement of permanent stakes a crucial standard for any tree plot. The stakes define a checkerboard of quadrats, each labeled with a number, whose coordinates (multiples of 20 m relative to one corner of the plot) are surveyed with precision (Fig. 3). The value of the permanent stake and quadrat system is that consistency of locations through time is assured. Here we assume these stakes are in place (Condit, 1998) and consider mapping trees relative to them. We leave aside calculations of elevation, as they follow easily from the survey.

2.3.1. Local coordinates

Stem positions relative to the 20-m quadrats are referred to as *local coordinates* or *quadrat coordinates*. Most often, the coordinates are Cartesian *x* and *y* distances relative to two sides of the quadrat, so we refer to them as *qx*, *qy*. Plots are not always laid out in cardinal directions, so *qx* and *qy* need not be east-west and north-south axes. Also, the coordinates collected in the field could just as easily be polar (qr, $q\theta$) instead of Cartesian.

2.3.2. Stake, plot, and global coordinates

The surveyed coordinates for the southwest corner stake of each quadrat are defined as *sx*, *sy*, where *s* refers to a stake. To precisely locate a stem in the plot, its local coordinates *qx* and *qy* are added to the *sx* and *sy* for the stem's quadrat. Thus, $px = qx + sx$ and $py = qy + sy$, where *px*, *py* are Cartesian coordinates within the entire plot and are the basis of maps of trees within a plot (Condit, 1998). A single set of coordinates for one corner of the plot, (base coordinates, *bx*, *by*, either latitude-longitude or UTM), is taken with GPS or off a map. Then a tree's global coordinates are $gx = bx + px$ and $gy = by + py$.

Plot and global coordinates are not recorded (or known) at the time of data collection. Field workers only record local coordinates, or mark stems on a map; the rest are calculations carried out after data are assembled.

2.3.3. Stem versus tree coordinates

Our currently recommended standard is to collect coordinates for every individual stem, including separate locations for separate stems of the same tree when their positions differ. This is an update relative to the methods we first used at Barro Colorado, when only tree coordinates were recorded. The earlier method makes sense for stems connected above ground (stem forks and branches), but secondary stems require separate coordinates (Table 4). Field workers should only collect what is necessary — local coordinates for each trunk base. Later, when the database is

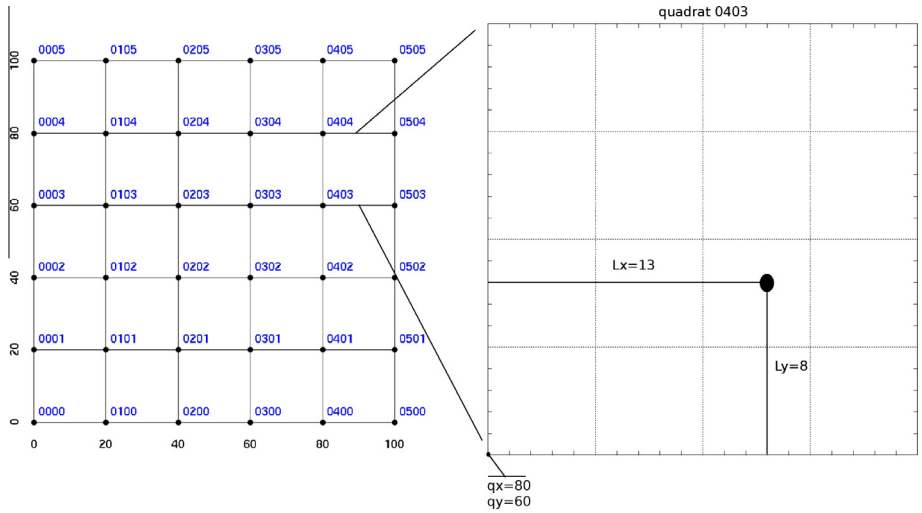

Fig. 3. Diagram of quadrats defined by a 20×20 m grid in a 100×100 m plot, with the quadrat naming convention used in the Barro Colorado 50-ha plot. The first two digits in a quadrat name indicate a column number, and the second two row number, for example 0403 means column 4, row 3, where columns are rows are 20-m wide. The enlarged section to the right is a single quadrat, showing coordinates of 0403's post referred to in this diagram as (qx, qy) and the local coordinates of one tree labelled here as (Lx,Ly).

assembled, local, plot, and global coordinates are calculated for every stem.

2.3.4. Modern mapping technology

Methods for recording stem coordinates have seen more changes due to technological advances than any other component of the tree census. Prior to GPS and laser-range finders, the *qx* and *qy* distances were collected with tape measure or marked on a paper map in the field. In the latter case, local coordinates were captured by digitizing positions off the map (and this is still our method at Barro Colorado). With laser range finders, a pair of coordinates can be automatically digitized by pointing a device from any stake to any tree; indeed, many devices skip quadrat and plot coordinates altogether and immediately produce global coordinates.

2.4. Database issues covering coordinates

Information about the plot and quadrats must be stored in tables for those purposes. The *Site* table holds a record for the plot, including name, size, shape, and a description of the plot; with a single plot, the table has just one record. The *Quadrat* table has one record per quadrat, which in the BCI 50-ha plot means 1250 rows, with a name for each quadrat (as labeled in the forest), plus the size and shape. We have moved the base coordinates for the plot (*bx*,*by*) and the base coordinates for each quadrat (*sx*,*sy*) into a separate *Coordinates* table; the reason for this will only be evident when we mention irregular geometries below. The CTFS database schema stores the quadrat number plus local stem coordinates (*qx*,*qy*) in the *Stem* table.

In an early version of the CTFS Data Model, used for BCI data alone, the coordinate data stored in the database were the plot base (*bx*,*by*), quadrat corners (*sx*,*sy*), and local stem coordinates (*qx*,*qy*). Plot and global coordinates were calculated from those as needed. This arrangement is not generalizable, however. Modern surveying tools now allow direct calculation of global coordinates in the field, and in sites where quadrats are irregular, the calculations of tree coordinates are much more complicated. For these reasons, we decided to include three sets of coordinates in both *Coordinates* and *Stem* tables: (1) local, (2) plot, and (3) global. At BCI, (1) and (2) are used to calculate (3), but at some sites, (3) and (2) are used to calculate (1).

2.4.1. Irregular geometry of plots

Having coordinates in a table separate from the *Quadrat* table allows for irregular shapes. Instead of a single pair *qx* and *qy* at one quadrat corner, coordinates can be a polygon outlying any shape desired, and the same goes for plots. Both *Site* and *Quadrat* tables include an attribute for area, which is calculated once and stored. This arrangement allows, for example, plots with a boundary set by a lake. The calculation of plot coordinates, however, no longer matches the simple rectangular system described above. There are several CTFS plots where the calculations are site-specific and rely on independent algorithms to fill the *Stem* table with each individual's position.

2.4.2. Multiple plots

To this point, we have described data in which all trees are in a single plot, but the database structure accommodates additional plots, and the BCI database now includes 63 separate plots in Panama. The *Site* table requires one row per plot, and the *Coordinates* table has the base location of each. The *Quadrat* table has additional rows for as many quadrats as there are in each plot. The name of each, *QuadratName*, can be repeated in different plots, as are plot coordinates, but global coordinates would differ.

2.5. Taxonomy

In tropical plots names of species are the source of frequent mistakes and confusion, owing to the large number of individuals that are not identified and continuous correction and reidentification. Many of the mistakes, though, amount to trivial misspellings and can be overcome with straightforward data tools. A particular focus of those tools must be the *indets* — the unidentified tree species so frequent in tropical forests.

2.5.1. A history of name changes

Many tropical plots have hundreds to thousands of species within a few hectares, and botanists cannot identify every individual during a single pass. Large numbers of leaf specimens are collected and returned to herbaria for sorting and subsequent study, and some individuals must be revisited for further study, especially to search for flowers. Unidentified specimens are gradually identified through the additional work, and many trees are given one name, then another, and sometimes another and more. In all CTFS plots, the names assigned in the field are abbreviated codes,

or *mnemonics*, usually 4–8 letters indicating the genus and species; at Barro Colorado, mnemonics are six letters: the first four of the genus plus the first two of the species.

We recommend that all specimens, no matter whether identified or not, be assigned names using mnemonics that reflect what is known. In Panama, when the genus is known, we use the 4-letter genus code plus a number to represent an unknown species; when the family is unknown, mnemonics are '*unid1*', etc. The mnemonic is never left blank, and comments such as 'tag 101 is the same as tag 55' cannot be used as identification. If the two are deemed the same species, then they should be assigned the same mnemonic in the field. Full species names for indets can be descriptive, such as *Nectandra* 'fuzzy', with quotes included.

A table of mnemonics and their associated full names, whether valid Latin binomials or not, must be maintained and expanded as needed. A field *IDlevel* is included to indicate the taxonomic depth to which identification is certain. Fully identified species have *IDlevel* = *species*, while those known to genus have *IDlevel* = *genus*. Other possibilities are *family*, *none* (completely unidentified), *subspecies*, or *mixed*. The last is important, describing cases where a mnemonic includes a mixture of unknown species, often arising with difficult genera where separating species must await special methods. There have been cases in CTFS plots where temporary names like *Zanthoxylum brillante* appear like valid Latin names, and the *IDlevel* is needed to indicate one way or the other. There is no species *Z. brillante*: they were specimens where the leaves looked shiny, and the name was intended as a descriptor while the valid Latin name was sought. Thus, the correct *IDLevel* was *genus*; it is also useful to write temporary names like this with quotes, i.e. *Zanthoxylum* 'brillante'.

Combining the need to revisit trees and revise species identification with careful use of partial identification means updating the identities of many individuals. Updating the table of species means adding new rows when specimens are identified, but keeping the old rows even if they are no longer needed. Many times, there is a comments field indicating that specimens with one name were changed to a new name. A crucial and complicated aspect of a plot database is the need to keep track of name changes, which we consider in detail below in the section on the taxonomy database.

2.5.2. Taxonomic hierarchy: subspecies, genus, family

The expanding table of species names includes genus and family names for every mnemonic. Field botanists will always use these, and they are important for partial identification of indets. Some botanists are also keen to assign infraspecific names, such as varieties or subspecies, and these names should be stored in their own column (not appended to the species name). When two different subspecies of the same species are found within a plot, they get separate mnemonics.

2.6. Database issues covering taxonomy

The working table described above, holding all mnemonics and species names, is temporary, though it is close to the final database table named *Species*. To avoid repetition of genus and family names (which would violate third normal form), separate *Genus* and *Family* tables are required (Fig. 2). We take the further step of establishing these tables in advance, *Genus* with 21240 known genus names and *Family* with 549 family names, both intended to include all angiosperm, gymnosperm, plus tree fern taxa. The hierarchy assures that every species is in a valid and correctly spelled genus and family (assigning indets to a genus 'Unidentified' with family NULL). The genus-family taxonomy is taken from the Angiosperm Phylogeny Group II (Angiosperm Phylogeny Group, 2009; Stevens, 2012), so that all CTFS plots are uniform in this regard. We also move subspecies into a separate table, otherwise, a single species would require two or more rows in the *Species* table (one per subspecies).

There is a one-to-many mapping from *Species* to *Tree,* so the child table *Tree* table includes a foreign key, *SpeciesID*, which is the primary key of the *Species* table. This enforces referential integrity so a tree can only be associated with a single species in the Species table. Since *subspecies* are in a separate table, there must be a second foreign key, *SubSpeciesID* in the *Tree* table. Separating species and subspecies like this ensures straightforward queries at the species level or the subspecies level.

The important and challenging component of the taxonomy database is maintaining a history of name changes. Without it, old publications and old analyses carry names that could no longer be traced. To manage changes, two columns were added to the *Species* table: *CurrentTaxonFlag* and *ObsoleteTaxonFlag*. If the *CurrentTaxonFlag* is set, then the species is currently in use. To store the relationships between current and obsolete species names, a mapping table *CurrentObsolete* is introduced. There are two one-to-many identifying relationships between *Species* and *CurrentObsolete*; one relationship for the old, obsolete species and one for the new, valid species.

Some name changes do not involve *CurrentObsolete*. If a single tree is reidentified due to an error, it is not a taxonomic change. Only the *Tree* table is updated for the one record, reflecting a different *SpeciesID*. Taxonomic changes are those where one name is replaced by another, often for example when a group of unidentified trees finally gets a name. This requires multiple changes to the *Tree* table and new records in *Species* and *CurrentObsolete*. For example, consider a species known by *unid1*, which includes several hundred individuals, that is finally identified as *Nectandra globosa*. The new name, *nectgl*, must be added to the *Species* table with *CurrentTaxonFlag* = 1, while the old name, *unid1* remains but is updated so *CurrentTaxonFlag* = 0 and *ObsoleteTaxonFlag* = 1. A new record is added to *CurrentObsolete* showing that *unid1* is the obsolete name for *nectgl* (Fig. 2). After a species change, hundreds of records in the *Tree* table must be assigned the *SpeciesID* of the new *nectgl*. To simplify the process, we built software to accomplish multiple changes like this easily (Dolins et al., 2008).

2.7. Personnel

The CTFS data standard is to store names of field and data personnel and to keep track of who censused which quadrats. These records lead to quantitative measures of progress and labor required. Moreover, by identifying who measured each tree, the source of errors can be explored more fully. Though neither of these is crucial to a successful census, the time it takes to store names is trivial relative to the full data entry, and once a database is built to handle names, there is no further cost.

Data are collected on a quadrat basis, and names of personnel are recorded once at the top of a form where the date and quadrat name are typed. Multiple names can be typed, with a record of what aspect of work each person performed. The names are saved, providing a new table with quadrat name, date, personnel, and type of work. One person can have several roles and several people might work in one quadrat in the same census, so four tables are required: (1) *Personnel*, simply a list of all people who have worked on the plot; (2) *RoleReference*, a list of chores (such as field technician, data entry technician, and supervisor); (3) *PersonnelRole*, a mapping table including a record for every person-chore combination; and (4) *DataCollection*, a mapping table from *PersonnelRole* to *Quadrat* and *Census*. Therefore, the one-to-many mapping between these tables means that the primary keys of *Census*, *Quadrat*, and *PersonnelRole* become foreign keys in *DataCollection*.

2.8. Precision and accuracy

The CTFS data standard includes a quantitative estimate of measurement error. Diameter measurements and species identification are repeated in a random selection of trees. Error rates at Barro Colorado were documented in Condit (1998) and diameter error is now incorporated in growth models (Rüger et al., 2011). The CTFS Data Model includes a table, *Remeasurement*, whose sole purpose is to store independent repeat measurements for quality control purposes. A companion table, *RemeasAttrib*, holds attributes associated with the remeasurements, just as *DBHAttributes* does for the *DBH* table.

2.9. A realized view

The census data used in most analyses requires a query joining *Family*, *Genus*, *Species*, *Tree*, *Stem*, *DBH*, *DBHAttributes*, and *TreeAttributes* tables, most often for just one census and one measurement per stem (in case there is an extra measurement at a second height). A single row of the result has one stem's species name, coordinates, and diameter (or status 'dead'), and is a denormalized reassembly of the normalized tables. Because many CTFS databases have >100,000 measurements per census, we store the table as a *realized view* rather than repeating the query every time it is needed. That is, this table is a denormalized representation of the data, and it can used for analytical and reporting purposes only. The view is updated with every change to the core tables. A smaller realized view of all species names joined to *Genus* and *Family* tables is also stored.

2.10. History of changes

There is a frequent need during data analysis and data screening to understand why past queries do not produce the expected result on a recent version of a database. This leads to specific questions about which records were changed when. The CTFS Data Model maintains records of most changes with a single history table that logs updates, inserts, and deletions, covering all tables and all columns.

2.11. Updates and archives

A principal advantage of carefully organized data is the ease with which an entire database — millions of measurements — can be backed up and stored in a single computer file. Our standard at BCI is to maintain a complete plot database from Panama unchanged for one year. At the end of the year, after updates are made, the new stable version is immediately backed up and archived, thus allowing us to restore any prior version of the database. Every five years, the stable version is published with a DOI (*digital object identifier*) at the Smithsonian Institution digital repository (A. Hutchinson, Smithsonian Library, pers. comm.). This permanently identifies every published version of the database, and past versions remain available indefinitely, allowing literature citations to point to the exact version used in a paper.

3. Discussion

Moving long-term tree census data into a carefully designed data model has had three principle advantages. Two are resolutions to well-known problems, but the third is a benefit to the standardized data that we did not clearly articulate at the outset. We also address here several disadvantages to a complex data model.

Our primary goal with the CTFS Data Model was to eliminate the anomaly errors that routinely occur in spreadsheets. Once our datasets were normalized, we could proceed with analyses knowing that integrity errors (Table 3) were not a concern. A second primary goal of the CTFS Data Model from the outset was to assure that every project's dataset had a clearly designated master version housed on a server that all collaborating scientists can access. Designating a master has nothing to do with normalizing the data – it is just careful governance – but it was easier after a plethora of errors had been removed in building the CTFS version. Moreover, the standard for modern database systems such as MySQL is to aggregate many tables under one umbrella, and this also fosters good governance. With multiple, independent tables, it is common for users to exchange some tables but not others, leading to inappropriate mixtures, such as old species tables joined with updated measurement tables. Moreover, a single MySQL database fosters back-ups and archives by assuring that all relevant information in its most recent state can be easily copied.

Associated with the designation of a master version is the importance of distributing data via online portals where downloads can be tracked. When we started in the late 1980s, this option was not available, but the CTFS data system now includes an online version of the master that all participating scientists can access. This reduces the tendency for several scientists to maintain diverging versions, though it does not prevent it. Indeed, the desire of individual scientists to maintain separate versions continues to plague our attempts at data consistency.

We discovered a third advantage to the standardized data model after several forest plots in the CTFS network began to use it. Having multiple separate sets of data in the same format has greatly fostered comparative data analysis. Given data in a standard format, scientists can now build one piece of software which probes multiple datasets. A variety of basic analyses of diversity, growth, mortality, and spatial patterns written in the programming language *R* have been tailored to the standard CTFS Data Model and are disseminated freely (Condit, 2012); the Barro Colorado and Panama Forest Plot data are also available freely in the format required by the *R* programs (Condit, 2012). Because all plots have data in the same format, querying and analyzing across many CTFS sites is simplified. Numerous cross-site analyses are now published (Condit et al., 2005; Condit et al., 2006; Chave et al., 2008; De Cáceres et al., 2012).

A separate advantage to the standard format has been that it allows experts with experience on one project to quickly screen data and recommend corrections on other datasets. The same step is tedious and time-consuming when data are in multiple formats.

There are, however, disadvantages to standardizing multiple forest plot datasets in the common CTFS Data Model. The greatest is the complexity inherent with multiple tables. Certain updates, such as inserting (or deleting) a tree, with all its stems, measurement, and measurement attributes, require simultaneous changes to many tables. Moreover, the computer language required, *SQL* (Structured Query Language) is unfamiliar to many scientists. To assure cohesive changes, it is essential to devise software tools that enforce the correct set of updates across several tables and simultaneously allow less experienced programmers to make changes. Our forest plot data sit under a software interface in the languages *PHP* and *HTML*, but this has required database professionals on our staff. The difficulty of working with the full CTFS Data Model has hindered its widespread use, meaning that some collaborating projects still use poorly designed formats.

Another difficulty we have encountered has to do with broadening a data model to accept a wider range of plot methods. A major goal of the CTFS network has been to standardize methods at many sites (Condit, 1998), but the projects are independent and precise repetition of methods has not been possible. We had to add some

complex features to the CTFS Data Model in order to encompass alternate methods. We mention several important ways this has happened within our network.

Stems tags are one. Some sites do not put tags on individual stems, and at Barro Colorado in Panama, we started the census using tree tags but not stem tags (i.e., one tag per tree). One stem is still one clear-cut entity about which facts are collected in one census, but in a subsequent census, we cannot always identify the same stem, if for example two stems of similar size on the same tree do not have tags. We accommodate this in the Data Model by adding records to the stem table – those two stems must become four stem records after the second census. This is peculiar, but there is no other obvious way to do it.

Another odd discrepancy is caused by moving stems. We assumed at the outset that the location of a stem is a permanent attribute, and thus belongs in the stem table. But there are sites where landslides cause stems to move. We now keep track of this by storing extra records in the stem table – old and new coordinates for one stem. The alternative – separate coordinates in every census for every stem – would mean unnecessary repetition of data at most sites, where movements do not occur.

The design of data models to encompass a variety of field methods is illustrated by *VegBank* (Peet et al., 2012; EcologicalSocietyofAmerica, 2013), a database of North American vegetation surveys, and *VegX*, a general data exchange standard for plots (Wiser et al., 2011). Both must be able to handle a wide variety of information, and as a result, neither explicitly handles stems but instead hold generic facts about plots and trees. One great strength of the complexity of *VegBank*, though, is its ability to host multiple taxonomic concepts simultaneously, for example according to several scientists across different time periods.

Two databases that target tree plots specifically are *ForestPlots.net* for tropical plots of the RAINFOR network (Malhi et al., 2002; Peacock et al., 2007; Lopez-Gonzalez et al., 2011; Lopez-Gonzalez et al., 2012) and the United States Forest Inventory and Analysis, or FIA (Woudenberg et al., 2010). Both models include tables for tree measurements, taxonomy, and plots, but neither has separate stem tables so cannot link multiple stems within a tree. Nor does either handle unidentified species and keep track of their later identification, a rare concern in temperate forests but a universal need in the tropics.

Integrating tree plot records from many projects such as CTFS, RAINFOR, and the FIA will greatly improve our understanding of forest communities and how they change through time. But integrating is only possible when the individual databases are clearly and efficiently organized, with consistent definitions of individual facts in normalized tables. The CTFS Data Model was created to ensure consistency and thus allow comparative analyses, allowing a network of plots to be greater than the sum of its parts.

Acknowledgments

We thank the many principal investigators, field technicians, and database managers of all CTFS plots for their input on field and data methods, especially R. Foster and S. Hubbell for pioneering the idea of a large plots with small diameter limit at Barro Colorado Island. Funding for the plot has been provided by the National Science Foundation (DEB-0640386, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042, DEB-8206992, DEB-7922197), the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Small World Institute Fund, and numerous small grants from other donors. M. Overholt, M. Campos, and H-C. Su assisted with database design. The Bradley University Department of Computer Science & Information Systems and Smithsonian Tropical Research Institute SIGeo program provided funding, and we thank S. Davies and E. Bermingham for their support.

References

Álvarez-González, J.G., Cañellas, I., Alberdi, I., von Gadow, K., Ruiz-González, A.D., 2014. Forest observational studies in Spain: applications to forest modeling. Forest Ecol. Manag. 316, 54–64.

Angiosperm Phylogeny Group, 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot. J. Lin. Soc. 161, 105–121.

Ayyappan, N., Parthasarathy, N., 1999. Biodiversity inventory of trees in a large-scale permanent plot of tropical evergreen forest at Varagalaiar, Anamalais, Western Ghats, India. Biodiversity Conserv. 8, 1533–1554.

Canham, C.D., Papaik, M., Uriarte, M., McWilliams, W., Jenkins, J.C., Twery, M., 2006. Neighborhood analyses of canopy tree competition along environmental gradients in New England forests. Ecol. Appl. 16, 540–554.

Chave, J., Condit, R., Muller-Landau, H.C., Thomas, S.C., Ashton, P.S., Bunyavejchewin, S., Co, L.L., Dattaraja, H.S., Davies, S.J., Esufali, S., Ewango, C.E.N., Feeley, K.J., Foster, R.B., Gunatilleke, N., Gunatilleke, S., Hall, P., Hart, T.B., Hernández, C., Hubbell, S.P., Itoh, A., Kiratiprayoon, S., Lafrankie, J.V., de Lao, S.L., Makana, J.R., Noor, M.N.S., Kassim, A.R., Samper, C., Sukumar, R., Suresh, H.S., Tan, S., Thompson, J., Tongco, M.D.C., Valencia, R., Vallejo, M., Villa, G., Yamakura, T., Zimmerman, J.K., Losos, E.C., 2008. Assessing evidence for a pervasive alteration in tropical tree communities. PLoS Biol. 6, e45.

Codd, E., 1971. Further normalization of the database relational model. In: Rustin, R. (Ed.), Database Systems, Courant Computer Science Symposium, vol. 6. Prentice-Hall, pp. 33–64.

Condit, R., 1998. Tropical Forest Census Plots: Methods and Results from Barro Colorado Island. Panama and a Comparison with Other Plots. Springer-Verlag, Berlin.

Condit, R., 2012. CTFS R Package. <http://ctfs.arnarb.harvard.edu/Public/CTFSRPackage>.

Condit, R., Ashton, P., Balslev, H., Brokaw, N., Bunyavejchewin, S., Chuyong, G., Co, L., Dattaraja, H., Davies, S., Esufali, S., Ewango, C., Foster, R., Gunatilleke, S., Gunatilleke, N., Hernandez, C., Hubbell, S.P., John, R., Kenfack, D., Kiratiprayoon, S., Hall, P., Hart, T., Itoh, A., LaFrankie, J., Liengola, I., Lagunzad, D., Lao, S., Losos, E., Magard, E., Makana, J., Manokaran, N., Navarrete, H., Mohammed Nur, S., Okhubo, T., Pérez, R., Samper, C., Seng, L.H., Sukumar, R., Svenning, J., Tan, S., Thomas, D., Thomspon, J., Vallejo, M., Villa Muñoz, G., Valencia, R., Yamakura, T., Zimmerman, J., 2005. Tropical tree α-diversity: results from a worldwide network of large plots. Biol. Skr. 55, 565–582.

Condit, R., Ashton, P., Bunyavejchewin, S., Dattaraja, H.S., Davies, S., Esufali, S., Ewango, C., Foster, R., Gunatilleke, I.A.U.N., Gunatilleke, C.V.S., Hall, P., Harms, K.E., Hart, T., Hernandez, C., Hubbell, S., Itoh, A., Kiratiprayoon, S., Lafrankie, J., de Lao, S.L., Makana, J.R., Noor, M.N.S., Kassim, A.R., Russo, S., Sukumar, R., Samper, C., Suresh, H.S., Tan, S., Thomas, S., Valencia, R., Vallejo, M., Villa, G., Zillio, T., 2006. The importance of demographic niches to tree diversity. Science 313, 98–101.

Condit, R., Chisholm, R.A., Hubbell, S.P., 2012. Thirty years of forest census at Barro Colorado and the importance of immigration in maintaining diversity. PLoS ONE 7, e49826.

Coomes, D.A., Kunstler, G., Canham, C.D., Wright, E., 2009. A greater range of shade-tolerance niches in nutrient-rich forests: an explanation for positive richness-productivity relationships? J. Ecol. 97, 705–717.

Crow, T.R., 1980. A rainforest chronicle: a 30-year record of change in structure and composition at El Verde, Puerto Rico. Biotropica 12, 42–55.

Date, C.J., 2004. An Introduction to Database Systems, eighth ed. Addison-Wesley, Longman, Boston USA.

De Cáceres, M., Legendre, P., Valencia, R., Cao, M., Chang, L., Chuyong, G., Condit, R., Hao, Z., Hsieh, C., Hubbell, S., Kenfack, D., Ma, K., Mi, X., Supardi, N.N., Kassim, A., Ren, H., Su, S., Sun, I., Thomas, D., Ye, W., He, F., 2012. The variation of tree beta diversity across a global network of forest plots. Global Ecol. Biogeogr. 21, 1191–1202.

Dolins, S., Condit, R., Su, H., Lao, S., 2008. Can AI techniques be applied to forest science data integration problems? Symposium on Semantic Scientific Knowledge Integration. AAAI Technical Report SS-08-05, pp. 21–23.

Ecological Society of America, 2013. Vegbank. <http://vegbank.org/vegbank/general/faq.html>.

Elmasri, R., Navathe, S., 2000. Fundamentals of Database Systems, sixth ed. Addison Wesley.

Franklin, J., DeBell, D., 1988. Thirty-six years of tree population change in an old-growth Pseudotsuga-Tsuga forest. Can. J. For. Res. 18, 633–639.

Hubbell, S.P., Condit, R., Foster, R.B., 2010. Barro Colorado forest census plot data. <http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci>.

Lilleleht, A., Sims, A., Pommerening, A., 2014. Spatial forest structure reconstruction as a strategy for mitigating edge-bias. Forest Ecol. Manag. 316, 47–53.

Lopez-Gonzalez, G., Lewis, S.L., Burkitt, M., Phillips, O.L., 2011. ForestPlots.net: a web application and research tool to manage and analyse tropical forest plot data. J. Veg. Sci. 22, 610–613.

Lopez-Gonzalez, G., Burkitt, M., Lewis, S., Phillips, O., 2012. ForestPlots.net – managing permanent plot information across the tropics. In: Biodiversity and Ecology Special, Vegetation Databases for the 21st Century, vol. 4, pp. 95–103.

Malhi, Y., Phillips, O., Lloyd, J., Baker, T., Wright, J., Almeida, S., Arroyo, L., Frederiksen, T., Grace, J., Higuchi, N., Killeen, T., Laurance, W., Leaño, C., Lewis, S., Meir, P., Monteagudo, A., Neill, D., Núñez Vargas, P., Panfil, S., Patiño, S., Pitman, N., Quesada, C., Rudas-Ll, A., Salomão, R., Saleska, S., Silva, N., Silveira, M., Sombroek, W., Valencia, R., Vásquez Martínez, R., Vieira, I., Vinceti, B., 2002. An international network to monitor the structure, composition and dynamics of amazonian forests (RAINFOR). J. Veg. Sci. 13, 439–450.

Peacock, J., Baker, T., Lewis, S., Lopez-Gonzalez, G., Phillips, O., 2007. The RAINFOR database: monitoring forest biomass and dynamics. J. Veg. Sci. 18, 535–542.

Peet, R.K., Lee, M.T., Jennings, M.D., Faber-Langendoen, D., 2012. Vegbank: a permanent, open-access archive for vegetation plot data. In: Biodiversity and Ecology Special, Vegetation Databases for the 21st Century, vol. 4, pp. 233–242.

Phillips, O.L., Vargas, P.N., Monteagudo, A.L., Cruz, A.P., Zans, M.E.C., Sánchez, W.G., Yli-Halla, M., Rose, S., 2003. Habitat association among Amazonian tree species: a landscape-scale approach. J. Ecol. 91, 757–775.

Rüger, N., Berger, U., Hubbell, S.P., Vieilledent, G., Condit, R., 2011. Growth strategies of tropical tree species: disentangling light and size effects. PLoS ONE 6, e25330.

Stevens, P.F., 2012. Angiosperm Phylogeny Website, Version 12. Missouri Botanical Garden. St. Louis, USA. <http://www.mobot.org/MOBOT/research/APweb/>.

ter Steege, H., Pitman, N.C.A., Phillips, O.L., Chave, J., Sabatier, D., Duque, A., Molino, J.F., Prevost, M.F., Spichiger, R., Castellanos, H., von Hildebrand, P., Vasquez, R., 2006. Continental-scale patterns of canopy tree composition and function across Amazonia. Nature 443, 444–447.

Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R., 2010. National Forest Inventories: Pathways for Common Reporting. Springer, Dordrecht.

Whitney, G.G., 1984. Fifty years of change in the arboreal vegetation of Heart's Content, an old-growth hemlock-white pine-northern hardwood stand. Ecol. 65, 403–408.

Wiser, S.K., Spencer, N., De Cáceres, M., Kleikamp, M., Boyle, B., Peet, R.K., 2011. Veg-X – an exchange standard for plot-based vegetation data. J. Veg. Sci. 22, 598–609.

Woudenberg, S.W., Conkling, B.L., O'Connell, B.M., LaPoint, E.B., Turner, J.A., Waddell, K.L., 2010. The Forest Inventory and Analysis Database: Database Description and Users Manual Version 4.0 for Phase 2. General Technical Report RMRS-GTR-245. United States Department of Agriculture Forest Service.