SIXTY YEARS OF RESEARCH, 60 YEARS OF DATA: LONG-TERM U.S. FOREST SERVICE DATA MANAGEMENT ON THE PENOBSCOT EXPERIMENTAL FOREST

Matthew B. Russell, Spencer R. Meyer, John C. Brissette, and Laura S. Kenefic

Abstract.—The U.S. Department of Agriculture, Forest Service silvicultural experiment on the Penobscot Experimental Forest (PEF) in Maine represents 60 years of research in the northern conifer and mixedwood forests of the Acadian Forest Region. The objective of this data management effort, which began in 2008, was to compile, organize, and archive research data collected in the U.S. Forest Service silvicultural experiment and several auxiliary studies. Due to the hierarchical nature of these data, a relational database management system (RDMS) was used (Microsoft Office Access). The resulting data management system affords new opportunities for novel research through data mining and increased collaboration among researchers; many of the data have since been published online (Brissette et al. 2012a, 2012b). Data management efforts such as these bridge data collection and data analysis, and play an important role in preserving the integrity of long-term studies. The RDMS used in this project is contemporary and widely used, but data storage systems will continue to evolve. It is important that U.S. Forest Service data management efforts continue and that new systems are adopted as needed.

INTRODUCTION

The U.S. Department of Agriculture, Forest Service (USFS) silvicultural experiment on the Penobscot Experimental Forest (PEF) in Maine has generated 60 years of research in the northern conifer and mixedwood forests common to the Acadian Forest Region of Atlantic Canada and adjacent Maine (Braun 1950, Rowe 1972). Research began in the 1950s when the USFS initiated a study consisting of an array of silvicultural treatments applied to replicated experimental units (Sendak et al. 2003). Since then, many auxiliary studies have been implemented on the PEF, several of which are conducted by University of Maine faculty and students. Many of these studies are short-term; others include several years of measurements. These auxiliary studies were built upon the foundation of the long-term silvicultural experiment. Consequently, data from these studies are related and allow for synthesis and comprehensive analyses to address a range of intriguing questions.

Adequate management of data records is an essential component of any long-term research program (Burton 2006), but is often overlooked due to limited resources and the short-term nature of most projects.

Methods and data management practices for the USFS's PEF database have evolved tremendously over these 60 years. Punch cards were used during the 1970s, but were phased out in the early 1980s when data transferred to electronic formats (e.g., computer tapes). Data were maintained for a time on the University of Maine mainframe computer system using FORTRAN programs. In the 1990s, the USFS PEF data were converted to ASCII files. These methods were appropriate for their time, though they are now outdated and inefficient.

By the early 2000s, nearly 60 years of PEF data were stored in 3,605 ASCII data files in 255 folders, and contained 374 megabytes of information. As a result of

the numerous files and folders, data were not readily accessible to researchers. The USFS recognized that the size and complexity of the PEF database warranted organization in a relational database management system (RDMS).

Serving as a tool for understanding the dynamics of northern conifer forests, data management is crucial for maintaining the integrity and value of research conducted on the PEF. This report describes a project initiated in 2008 to archive research data collected in the USFS silvicultural experiment and auxiliary studies. Specific objectives were to (1) organize and compile existing data, (2) test the functionality of an RDMS for archiving these data and making them available to users, and (3) develop and document a process for new data to be appended to the database. Many of the data have subsequently been published and are available online through the USFS Research Data Archive (Brissette et al. 2012a, 2012b).

METHODS

The Relational Database

Research institutions have increasingly relied on the RDMS model to archive experiment data in a hierarchical structure. Such systems can be customized to meet the needs and design requirements of the information being stored. The RDMS appeared to be an ideal tool for an experiment like that of the USFS on the PEF for several reasons. First, the RDMS allows various types of data to be related under a single framework. As an example, one data table may describe the silvicultural treatments, while another includes information about the experimental units to which each treatment is applied. In addition, each experimental unit contains a network of permanent sample plots, each of which has spatial data. With an RDMS, plots can be related to the experimental unit, and the experimental unit can be related to the silvicultural treatment (Fig. 1). Second, through

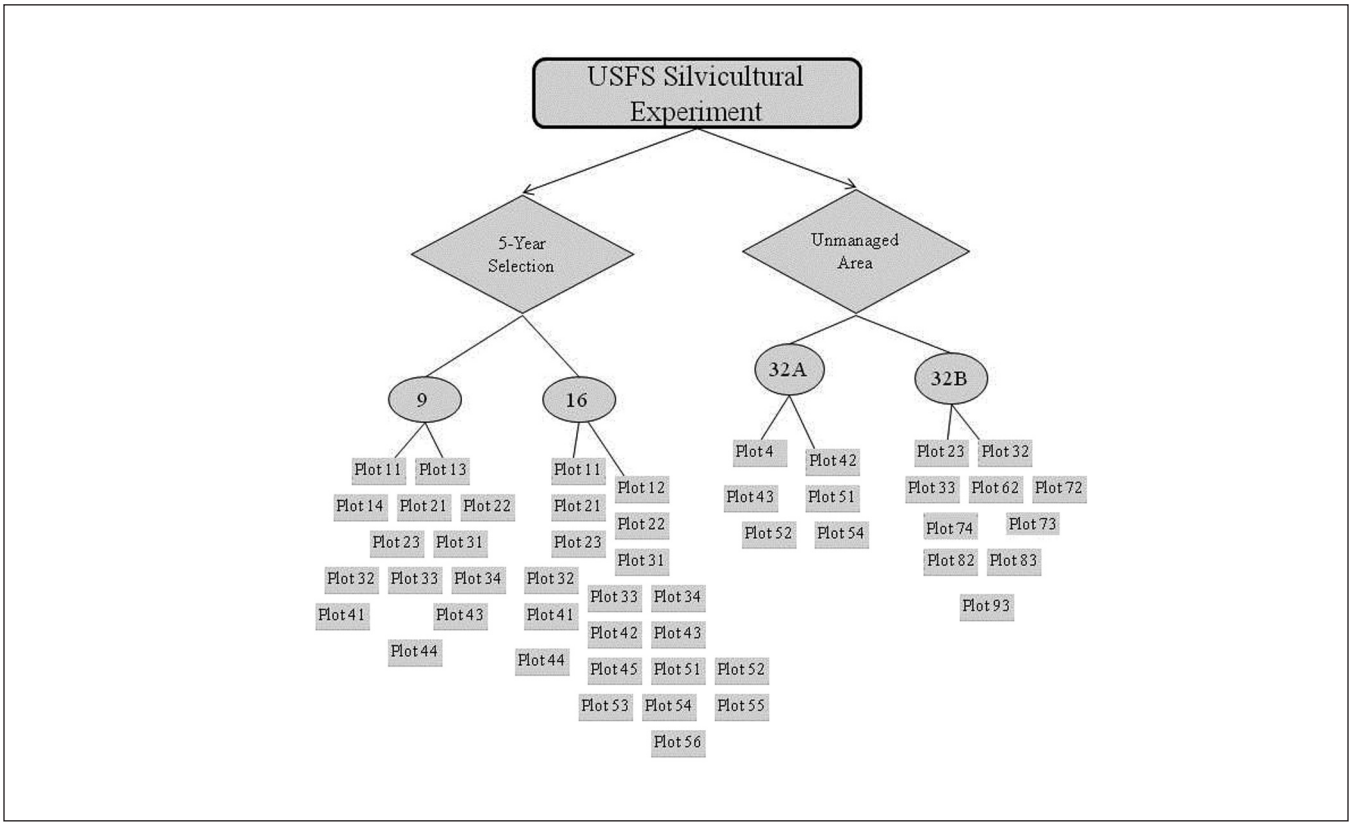

Figure 1.—Schematic displaying how U.S. Forest Service research on the Penobscot Experimental Forest fits into the structure of a relational database management system.

powerful querying capabilities, analysts can rapidly manipulate, summarize, or extract data of interest. By drawing from different tables, queries allow users to interpret data without changing the underlying data structure. Lastly, the RDMS has the ability to store large amounts of data, which reside as a single, easily replicated, and shared file on a personal computer or server. In contrast to typical spreadsheet or other "flat" data management systems, an RDMS reduces data storage capacity by extracting redundant information and making use of hierarchical relationships.

With large data sets such as the USFS's PEF experiment, data would be difficult to manage collectively as individual text files or spreadsheets. In the past, if researchers were interested in analyzing long-term data collected from a specific experimental unit, dozens of files from various inventories would first need to be compiled. Additional files that explained changing data collection methods and other associated metadata would similarly need to be compiled in order to interpret the data. This process of assembling data represented a cumbersome and time-consuming process for the analyst, and increased the probability of making errors. In addition, RDMS software is configured to work well with external software packages such as those used for statistical analyses by including fully featured input and output. RDMS software and database connection tools are available to work well with other operating systems such as Linux/UNIX and Apple operating systems (e.g., see R Development Core Team 2010).

Compiling and Archiving 60 Years of Information

Data collected on the PEF as part of the long-term USFS silvicultural experiment through the 2006 field season were used to develop the base structure and organization of the database. Data were aggregated into groups according to the type of study. Groups were organized according to the kinds of treatments applied in the experiment and the types of data collected. Data that were part of the long-term USFS experiment were classified in one group while auxiliary studies were grouped in another. Datasets were normalized when possible, meaning that data were arranged and restructured to meet the assumptions of conventional relational database design, thus reducing redundant storage of information. Management of data followed general guidelines established for ecological studies (Borer et al. 2009). Data were archived in a Microsoft Office Access database. Non-proprietary ASCII files of these data were also archived.

U.S. Forest Service Long-term Silvicultural Study

Data records for the long-term USFS experiment on the PEF were previously stored solely in ASCII files. One file existed for overstory tree data collected in each experimental unit (called a management unit, or MU) at each inventory. Tree species, diameter, and status were universally recorded in these files. Given that the same variables were collected in all inventories, these data were first grouped by management unit. For example, data from the 22 inventories that had occurred in MU 9 were previously stored in 22 separate files with related information. These were consolidated into one unified table in the database. After the files for each MU were aggregated, data were collapsed even further into a single table that contained all tree data collected on all MUs at all inventories. Tree regeneration data were organized in a similar manner as the overstory tree data.

Each MU contains an average of 15 permanent sample plots, totaling more than 600 plots in the USFS experiment on the PEF. These plots differed in terms of the inventory design used and the level of measurement detail (Fig. 2). Measurement protocols differ between "compartments" (replicated MUs in the long-term experiment), "units" (nonreplicated MUs used for other research), and the "management intensity demonstrations" (MUs managed for demonstration purposes). The measurement protocols for these areas evolved during the study (Table 1).

Other files, such as those containing spatial distribution and tree height and crown data for the

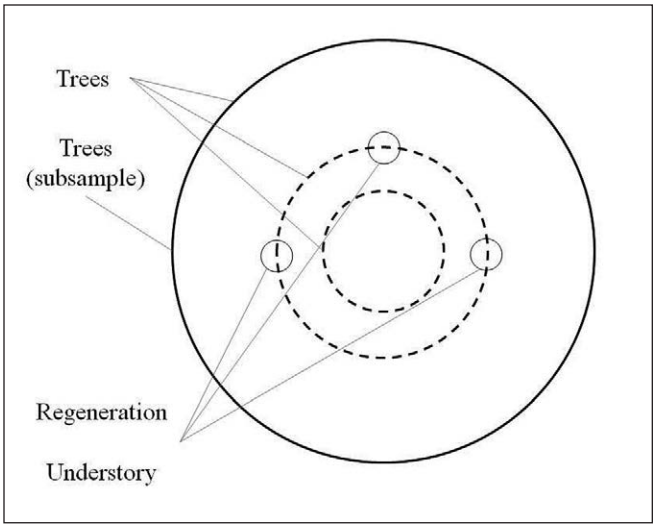

Figure 2.—U.S. Forest Service permanent sample plot schematic displaying 1/5th-, 1/20th-, 1/50th-, and 1/1000th-acre nested plots and the types of data associated with each as archived in the USFS's Penobscot Experimental Forest database (key information archived in these tables is described in Table 2).

compartments resided in separate Microsoft Office Excel spreadsheets. Similar to the tree data, these data existed in separate files for each MU. Given that identical variables were measured across the MUs, files of these types could be merged. Altogether, several key data tables were archived, and serve as the basis of the USFS silvicultural experiment on the PEF (Table 2).

Much of the supplementary information for the silvicultural experiment was obtained from scanned historical documents. Descriptions of silvicultural treatments, information on plot sizes, and tree species codes are examples of information obtained this way.

After data through the 2006 field season were compiled and archived in Access, field data collected in 2007 and subsequent years were used to test the functionality of the database in terms of checking and appending subsequent remeasurement data.

Auxiliary Studies

A wealth of information existed in the USFS archives concerning auxiliary studies, i.e., those expanding upon the foundation of the long-term experiment but not directly a part of it. For example, complete metadata and tree-level data from a precommercial

Table 1.—Overview of general historical data collection methods for trees in compartments, units, and the management intensity demonstration areas (MIDs) in the U.S. Forest Service's long-term silvicultural study on the Penobscot Experimental Forest (1950-2010)

	Plot size (ac)		Minimum diameter at breast height measured (in)	
	Before 2000	2000 and after	Before 2000	2000 and after
Compartments	1/5, 1/20	1/5, 1/20, 1/50	0.5	0.5
Units	1/5, 1/20	1/5	1-in class	5-in class
MIDs	varied	1/5, 1/20, 1/50	0.5 or 1.0	0.5

Table 2.—Key data tables for the U.S. Forest Service's long-term silvicultural experiment, as archived in the Penobscot Experimental Forest Microsoft Office Access database (local-use only)

Data table	Key information
Management units	MU ID, silvicultural treatment, acreage, status
Plots	MU ID, plot ID, spatial coordinates, depth to water table
Trees	MU ID, plot ID, tree ID, species, diameter at breast height, status
Trees (subsample)	MU ID, plot ID, tree ID, height, height to crown, crown width, spatial location
Regeneration	plot ID, species, count of stems by height class
Understory	plot ID, ground cover class percentages

thinning study (Brissette et al. 1999) were archived and well documented, as were data from a study of tree age. Some studies had limited data or metadata (e.g., for a study of growth efficiency on Study 58 plots and a logging technology study in MUs 2A and 2B). Auxiliary data that were well documented were imported directly into the database.

Additional data from other auxiliary studies were obtained from individual researchers. Examples included studies of tree leaf area (Kenefic and Seymour 1999, Maguire et al. 1998) and additional tree size measurement data sets (Saunders and Wagner 2008) (Table 3). Many of these data sets were archived following the overall database design used for the long-term silvicultural experiment; data manipulation was minor and done only to ensure consistency across all data tables archived within the database.

Table 3.—Data sets for completed and ongoing studies included in the U.S. Forest Service's Penobscot Experimental Forest local-use Microsoft Office Access database at the time of the 60th anniversary in 2010. Data from USFS and S58 have since been published online (Brissette et al. 2012a, 2012b).

ID	Experiment name
USFS	Silvicultural Experiment
REGEN	Regeneration Study
BRYCE	Understory Vegetation and Cover
CORE	Tree Core Analysis
MOORE	Light/Seedling Experiment (Spruce, Fir, Hemlock)
SAUND	Tree Measurements
WEAV	Seedling/Downed Woody Debris Study
S58	Study 58
PHLPS	Study 58 Stem Analysis Measurements
REHAB	Rehabilitation Study
LEAP	Land Use Effects on Amphibian Populations[a]
KZELL	White Pine Study
GAP	Expanding Gap Silvicultural Study[a]
CZELL	White Pine Study
CTRN	Maine Commercial Thinning Research Network[a]
WPINE	White Pine Quality under Varying Silviculture
LKFOL	Leaf Area of Eastern Hemlock
DMFOL	Growth Efficiency of Red Spruce
WEATH	PEF Weather Data/Weather Station
DAMAG	PEF Harvest Damage Survey
2020	Agenda 2020 Vegetation Competition Study[a]

[a] Study overview only

RESULTS

At the time of the PEF's 60th anniversary in 2010, the USFS PEF data were archived and resided as a fully integrated Microsoft Office Access database of 80 megabytes in size. An additional 120 megabytes of supplementary information was linked to this database. This information included references to external files, such as maps (including maps of management units, plots, and soils) and key publications of PEF research. This database is for local use by researchers on the PEF; Russell and Meyer (2009)[1] serves as a guide for researchers using the database and details procedures for documenting future data. Many of the data have since been published and are publicly available via the Web (Brissette et al. 2012a, 2012b). The local-use database laid the groundwork for a smooth, timely transition between the multitude of ASCII files and full online access. It is a valuable resource for researchers and staff on the PEF, and allows management of data prior to publication. Those seeking to obtain data from this database should follow the appropriate procedures for acquiring data through the Northern Research Station.

Twenty tables were initially archived within the local-use database. These tables included data collected as part of the silvicultural experiment, as well as auxiliary datasets (Table 3). In 2009, the "trees" data table contained more than 900,000 records with information on the species, diameter, and status of trees measured on permanent sample plots since the early 1950s.

Several reference tables were included in the database to aid in interpreting and analyzing data. Examples of these tables include comprehensive tables of species codes used for all the experiments, coefficients for estimating tree volume (Honer 1967), and a list of all inventory and harvest dates for the MUs.

[1] Russell, M.B.; Meyer, S.R. 2009. Penobscot Experimental Forest: a guide for data management and the Microsoft Access database. 61 p. Internal report available by request from M.B. Russell, University of Minnesota, College of Food, Agricultural and Natural Resource Sciences, Department of Forest Resources, 1530 Cleveland Ave. N., St. Paul, MN 55108.

Relationships were identified to associate data types in one table to similar types of data in another (Fig. 3). This step had important implications for using the database for querying tables and interpreting data. These data relationships form the backbone of the RDMS and allow the powerful query language to summarize similar data.

Action queries proved to be an effective and efficient tool for the database in several ways. Data collected after 2006, for example, were appended to existing data tables through action querying to take advantage of the structure of the database. In addition, several queries were designed to summarize stand-level statistics by using the underlying trees data table. These statistics included total and species-specific number of trees, basal area, and volume per acre for each of the inventories in the silvicultural experiment, as well as diameter distributions.

DISCUSSION

The RDMS proved to be an effective tool for archiving and managing 60 years of research data collected in the USFS long-term experiment on the PEF. The RDMS structure allows data in one table to be associated with data in another, and is an ideal instrument for archiving forest inventory information. In the example of the USFS silvicultural experiment on the PEF, the "management units" table contains a list of areas of land that are managed in the different USFS experiments on the PEF, while the "plots" table is a list of measurement plots used in each MU. The ability of the RDMS to associate data of different types in a hierarchical fashion makes it well-suited to managing long-term forest inventory data sets such as those of the PEF.

Querying functions allow users to interpret and analyze data found in the underlying data tables. By using the relationships defined among the different data tables, queries can be built that pull data from different source tables; this process allows users to summarize data sets quickly and repeatedly. For example, the "trees" table in the local-use database for the long-term silvicultural experiment can be queried to compute stand-level basal area, volume, and tree

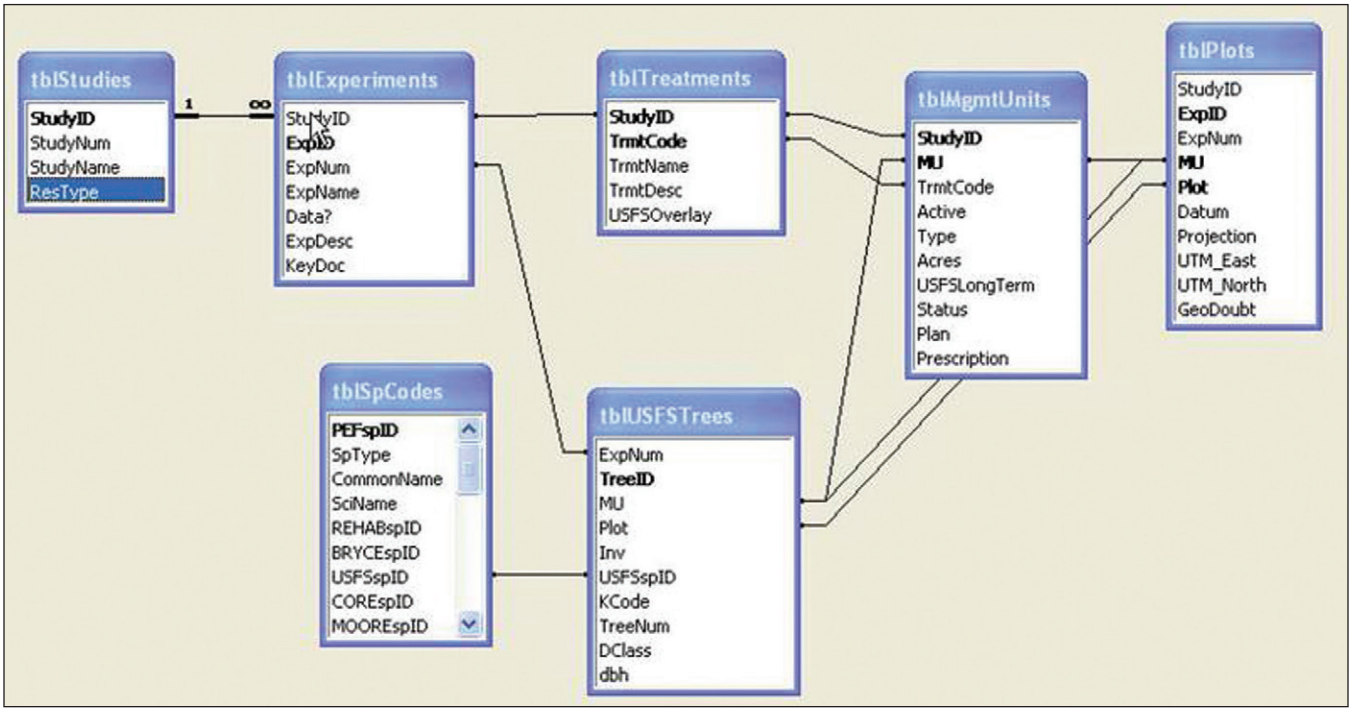

Figure 3.—Relationships window displaying associations of data tables within the U.S. Forest Service's Penobscot Experimental Forest local-use database.

diameter distributions across MUs, years, treatments, or other parameters. Users can readily graph stand development patterns throughout the duration of the study. A researcher may wish to routinely analyze trends in basal area among the differing selection system stands (Fig. 4). Once the analyst designs a query, it is instantaneously updated as new data are included in the database. This feature greatly facilitates analyses that are conducted annually.

Storing data in non-proprietary formats is an effective data management practice (Borer et al. 2009) and was accomplished as part of these efforts. Whereas Microsoft Office Access is proprietary software that (1) is subject to continuous updates, (2) could potentially become unavailable in the future, and (3) could be replaced by other newer and improved types of software, ASCII, or text files, can always be read. Similarly, analysis scripts have been maintained that allow a user to import USFS PEF data into statistical packages such as R (R Development Core Team 2010), MATLAB® (MathWorks, Inc., Natick, MA), and SAS (SAS Institute, Inc., Cary, NC). To preserve the relationships between different data tables in the RDMS, structured query language (SQL) scripts have been maintained of essential queries for importing data into other database systems, such as MySQL or PostgreSQL. Although currently the data are primarily managed within the Access RDMS, connectivity with other operating systems is offered. The open-source R statistical package is recommended for users seeking to use these coded scripts because of its compatibility with multiple operating systems and well-developed database connectivity packages (R Development Core Team 2010). Online USFS PEF data are in an Oracle® database; both raw data and summary statistics can be downloaded.

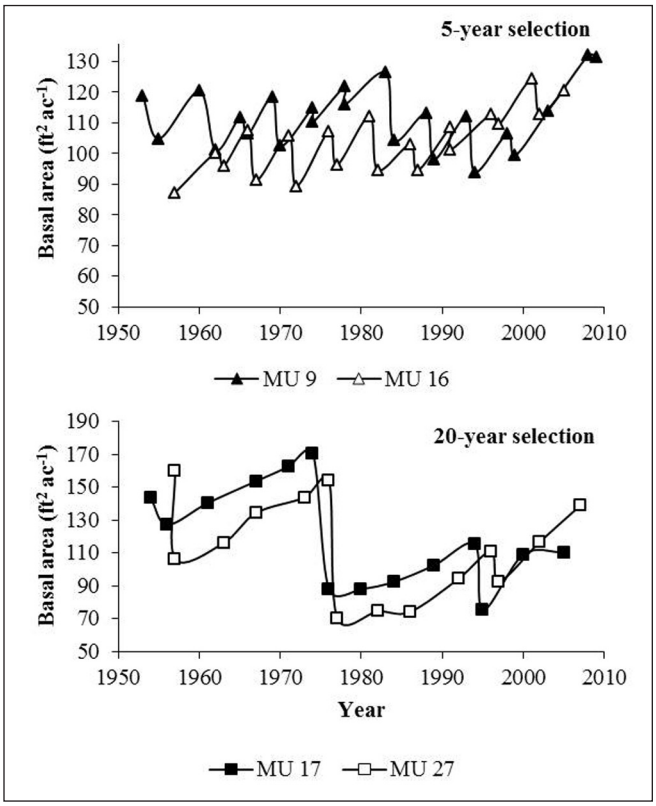

Figure 4.—Trends in basal area per acre (trees ≥0.5 in diameter at breast height) for management units treated with the selection system on 5- and 20-year cutting cycles, as archived in the U.S. Forest Service's Penobscot Experimental Forest database.

The ability to append data has tremendous value to the USFS long-term experiment. Data that are cohesively managed with a consistent structure provide a data format that can be easily maintained. For new data types and data from auxiliary experiments, new data tables can be created and incorporated into the existing database structure.

Opportunities

As an artifact of today's technological age, computer technologies change and data management software continually evolves. Employing contemporary software used by scientists and managers is central to the research integrity of the USFS's PEF data sets. In future years, the design and structure of the database should be evaluated to determine whether or not it is effectively meeting users' needs. Similarly, new avenues of research and data management have arisen for the USFS's PEF database. First, there are opportunities for spatially explicit data summary and analysis. Spatial technologies and geographic information systems software are now widely used, and technologies for bridging observed tree data with spatial data sets are available. Second, the database adds value to the long-term experiment by

providing a unified data set that can be readily used by researchers for forest growth and yield modeling, threat assessment, and other associated areas. Finally, opportunities for collaborations with new researchers have been established through Web-based PEF data-sharing sites (e.g., http://www.fs.usda.gov/rds/archive/Product/RDS-2012-0008 and http://www.fs.usda.gov/rds/archive/Product/RDS-2012-0009). Such sites showcase the USFS's data from the PEF and increase the real and perceived value of the experiments.

CONCLUSIONS

Sixty years of research data from USFS experiments on the PEF have been compiled, archived, and made available to researchers. The relational database model proved effective given the design of the long-term silvicultural experiment and associated auxiliary studies. New opportunities and collaborations continue to arise as a result of these data management efforts.

LITERATURE CITED

Borer, E.T.; Seabloom, E.W.; Jones, E.B.; Schildhauer, M. 2009. **Some simple guidelines for effective data management.** Bulletin of the Ecological Society of America. 90(2): 205-214.

Braun, E.L. 1950. **Deciduous forests of eastern North America.** New York: Hafner. 596 p.

Brissette, J.C.; Frank, R.M.; Stone, T.L.; Skratt, T. 1999. **Precommercial thinning results in a northern conifer stand: 18-year results.** The Forestry Chronicle. 75(6): 967-972.

Brissette, J.C.; Kenefic, L.S.; Russell, M.B. 2012a. **Precommercial thinning x fertilization study data from the Penobscot Experimental Forest.** Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station. Available at http://dx.doi.org/10.2737/RDS-2012-0009. (Accessed June 17, 2013).

Brissette, J.C.; Kenefic, L.S.; Russell, M.B.; Puhlick, J.J. 2012b. **Overstory tree and regeneration data from the "Silvicultural Effects on Composition, Structure, and Growth" study at Penobscot Experimental Forest.** Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station. Available at http://dx.doi.org/10.2737/RDS-2012-0008. (Accessed June 17, 2013).

Burton, P.J. 2006. **The need for long-term research installations and datasets.** In: Irland, L.C.; Camp, A.E.; Brissette, J.C.; Donohew, Z.R., eds. Long-term silvicultural and ecological studies: results for science and management. GISF Res. Pap. 005. New Haven, CT: Yale University: 136-138.

Honer, T.G. 1967. **Standard volume tables and merchantable conversion factors for the commercial tree species of central and eastern Canada.** Ottawa, ON: Canadian Department of Forestry and Rural Development, Forest Management Research and Services Institute. Info. Rep. FMR-X-5.

Kenefic, L.S.; Seymour, R.S. 1999. **Leaf area prediction models for *Tsuga canadensis* in Maine.** Canadian Journal of Forest Research. 29: 1574-1582.

Maguire, D.A.; Brissette, J.C.; Gu, L. 1998. **Crown structure and growth efficiency of red spruce in uneven-aged, mixed-species stands in Maine.** Canadian Journal of Forest Research. 28: 1233-1240.

R Development Core Team. 2010. **R data import/export.** Vienna, Austria: R Foundation for Statistical Computing. Available at http://cran.r-project.org/doc/manuals/R-data.html. (Accessed October 25, 2010).

Rowe, J.S. 1972. **Forest regions of Canada.** Publ. 1300. Ottawa, ON: Canadian Forest Service, Department of the Environment. 172 p.

Saunders, M.R.; Wagner, R.G. 2008. **Height-diameter models with random coefficients and site variables for tree species of Central Maine.** Annals of Forest Science. 65(2): 1-10.

Sendak, P.E.; Brissette, J.C.; Frank, R.M. 2003. **Silviculture affects composition, growth, and, yield in mixed northern conifers: 40-year results from the Penobscot Experimental Forest.** Canadian Journal of Forest Research. 33(11): 2116-2128.