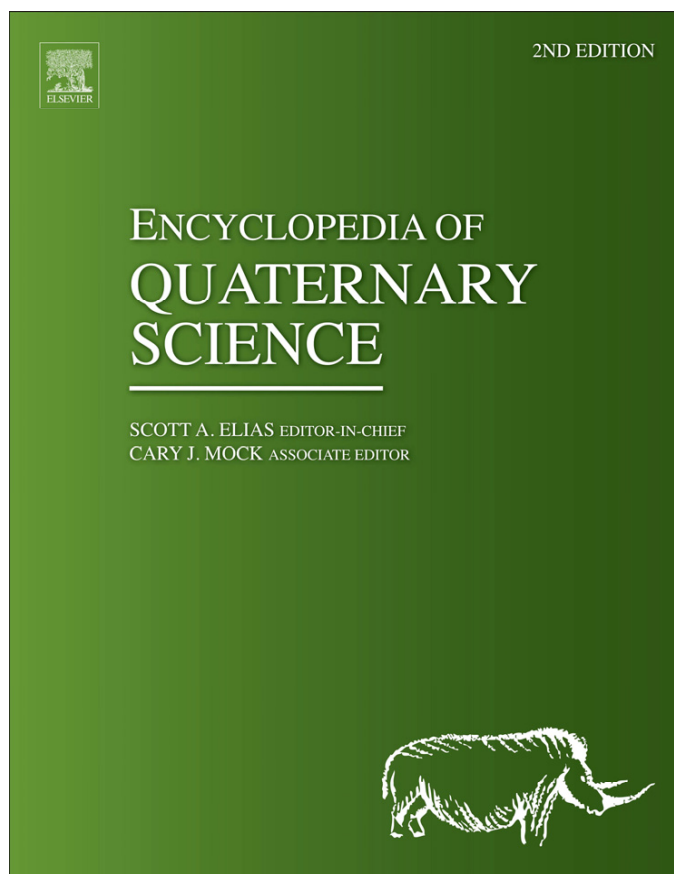


Provided for non-commercial research and educational use.  
Not for reproduction, distribution or commercial use.

This article was originally published in the *Encyclopedia of Quaternary Science* published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues who you know, and providing a copy to your institution's administrator.



All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>

Grimm E.C., Bradshaw R.H.W., Brewer S., Flantua S., Giesecke T., Lézine A.-M., Takahara H. and Williams J.W. (2013) Databases and Their Application. In: Elias S.A. (ed.) *The Encyclopedia of Quaternary Science*, vol. 3, pp. 831-838. Amsterdam: Elsevier.

© 2013 Elsevier Inc. All rights reserved.

## Databases and Their Application

**E C Grimm**, Illinois State Museum, Springfield, IL, USA

**R H W Bradshaw**, University of Liverpool, Liverpool, UK

**S Brewer**, University of Utah, Salt Lake City, UT, USA

**S Flantua**, University of Amsterdam, Amsterdam, The Netherlands

**T Giesecke**, University of Göttingen, Göttingen, Germany

**A-M Lézine**, Université Pierre et Marie Curie, Paris, France

**H Takahara**, Kyoto Prefectural University, Kyoto, Japan

**J W Williams**, University of Wisconsin, Madison, WI, USA

© 2013 Elsevier B.V. All rights reserved.

### Introduction

Palynological data from the recent geological past have been invaluable for understanding ecological dynamics at time-scales inaccessible to direct observation, including the reorganization of communities and ecosystems as species responded individually to climate change, historical influences on contemporary patterns of biodiversity, and biotic responses to climatic change, both gradual and abrupt. Understanding the dynamics of ecological systems requires ecological time series, but many ecological processes operate too slowly to be amenable to experimentation or direct observation. In addition to having ecological significance, fossil-pollen data are crucial for climate-model verification and are essential for elucidating climate–vegetation feedbacks that can amplify or mitigate climate variations.

Basic palynological research is site based, and palynologists have devoted innumerable hours to identifying and counting pollen grains from cores, sections, and excavations. These data are typically published in papers describing one or a few sites. Usually, the data are published graphically, as in a pollen diagram, often only showing a few of the identified taxa, and the actual raw data reside on the investigator's computer or in a file cabinet. These basic data are similar to museum collections, in that they are costly to replace, often irreplaceable, and their value does not diminish with time. Moreover, the raw data are the primary resource for new interpretations and analytical approaches. Also, similar to museum collections, they require cataloging and curation. Whereas physical specimens of large fossils, such as animal bones, are typically accessioned into museums, microfossils, such as pollen, are often not accessioned, and the digital data are the primary objects. Thus, the loss of pollen data is equivalent to losing valuable museum specimens.

Several pollen database cooperatives have assembled pollen data from various parts of the world and have archived these data in relational databases. These databases fulfill two critical needs: (1) they are a secure archive of invaluable paleobiological data and (2) they are essential infrastructure for paleoenvironmental research. Although data from individual sites are valuable in their own right, emergent properties can be investigated by marshalling data from geographic arrays of sites for synoptic, broad-scale ecosystem studies. Examples of such properties are continent-scale climate and vegetation reconstructions, which provide the historical context for

understanding biodiversity dynamics, including genetic diversity. Synoptic climate reconstructions are crucial for climate-model validation.

For the past 20 years, the managers of the various regional pollen databases have collaborated to deploy their data in identical database structures. Some regional pollen databases have had their data merged from the outset, as did, for example, the North American Pollen Database (NAPD), the Latin American Pollen Database (LAPD), and the Pollen Database for Siberia and the Russian Far East (PDSRFE). In recent years, additional efforts have been made to merge other regional databases into a Global Pollen Database (GPD). The logical next step has been to merge the pollen databases into an even broader paleo database containing a wide variety of fossil and related data, which will facilitate even broader studies of community and ecosystem dynamics. This objective is now being achieved with the development of the Neotoma Paleoecology Database.

### Database History

Ever since Lennart von Post first introduced the science of paleopalynology and the first pollen diagram in 1916 ([von Post, 1967](#)), pollen data have accumulated as slides in desk drawers, count sheets in file cabinets, and files now stored on various electronic media. With the advent of radiocarbon dating in the 1950s, it became possible to compare temporally controlled pollen diagrams across vast areas to assess species migrations and expansions and synchronous and time-transgressive events. Such comparisons require a 'database,' whether it is a pile of pollen diagrams or digital data in a relational database.

Efforts to develop pollen databases began in North America and Europe in the 1970s. H.J.B. Birks and B. Huntley assembled data from sites throughout Europe to produce their massive atlas of isopoll maps at 500-year intervals beginning 13 000 <sup>14</sup>C year BP for 55 taxa ([Huntley and Birks, 1983](#)). For the most part, they extracted data from published pollen diagrams. However, this procedure had limitations, and Huntley was impelled to become one of the original organizers of the European Pollen Database (EPD; [Fyfe et al., 2009](#)). In North America, efforts of T. Webb III and colleagues at Brown University in the 1970s and 1980s led to the first organized pollen database as part of the Cooperative Holocene Mapping

Project (COHMAP Members, 1988; Wright et al., 1993). Although climate data-model comparison was the principal objective of the COHMAP project, the synoptic analyses of the pollen data, particularly maps showing the constantly shifting ranges of species in response to climate change, were revelatory and led to much ecological insight.

The COHMAP pollen 'database' consisted of a multiplicity of flat files with prescribed formats for data and chronologies. FORTRAN programs were written to read these files and to assemble data for particular analyses. T. Webb III managed the COHMAP pollen database at Brown University, but as the quantity of data increased, data management became increasingly cumbersome. Clearly, the data needed to be transferred to a relational database management system (RDBMS). Discussions with E.C. Grimm, who was based in a museum with a mission of data preservation, led to the initiation of the NAPD at the Illinois State Museum in 1990. At the same time in Europe, the International Geological Correlation Project (IGCP) 158b was conducting a major collaborative synthesis of paleoecological data, primarily of pollen, and the need for a pollen database became painfully obvious (Fyfe et al., 2009). A workshop to develop an EPD was held in Sweden in 1989, and the organizers of NAPD and EPD commenced a long-standing collaboration to develop completely compatible databases. NAPD and EPD held several joint workshops and developed the same data structure. Nevertheless, the two databases were independently established, partly because Internet and software capabilities were not yet sufficient to easily manage a merged database. The pollen databases were developed in Paradox®, which at the time was the most powerful RDBMS software readily available for the personal computer (PC) platform. NAPD and EPD established two important protocols: (1) the databases were relational and queryable and (2) they were publicly available. All data in NAPD are in the public domain. EPD contains data that are public domain as well as smaller set of restricted data that require permission from the contributor for use. In an effort to encourage public access and to diminish the number of restricted sites, the EPD Advisory Board has set a three-year limitation, after which restricted status may continue but must be reaffirmed by the contributor.

As the success of the NAPD-EPD partnership grew in terms of contributed datasets and publications utilizing the databases, working groups initiated pollen databases for other regions, including the LAPD in 1994 (Marchant et al., 2002), the PDSRFE in 1995, and the African Pollen Database (APD) in 1996. At its initial organizational workshop, LAPD opted to merge with NAPD, rather than develop a standalone database. PDSRFE also followed this model, and the combined database was called the GPD. APD developed independently, but uses the exact table structure of NAPD and EPD. Pollen database projects have also been initiated in other regions, and the GPD contains some of these data, including the Indo-Pacific Pollen Database (IPPD).

The growth of the Quaternary pollen databases has been paralleled by the development of databases for other fossil types, including the North American Plant Macrofossil Database, FAUNMAP (a database of Quaternary-Pliocene mammals from North America), the North American Non-Marine Ostracode Database, the Diatom Paleolimnology Data Cooperative, and the Strategic Environmental Archaeology

Database/Bugs Database. These databases are similar in that they essentially store lists of taxa with quantitative measurements from stratigraphic contexts, with similar metadata. All of these databases face long-term sustainability issues associated with the costs of database maintenance and development and continued data acquisition. Thus, to consolidate maintenance costs and to build a multiproxy paleo database even more useful for addressing emergent global-scale ecosystem questions than the individual databases alone, the Neotoma Paleocology Database ([www.neotomadb.org](http://www.neotomadb.org)) was established in 2007.

Named for the packrat, genus *Neotoma*, prodigious collectors of anything in their territories, the Neotoma database realizes the ultimate goal of a multiproxy paleodatabase with global coverage. Neotoma includes fossil data for the past 5 million years (the Pliocene, Pleistocene, and Holocene Epochs) and provides the underlying cyberinfrastructure for a variety of disciplinary database projects. Although Neotoma is a new structure, it was built on the foundation laid particularly by the GPD and FAUNMAP. Neotoma can accommodate virtually any type of fossil data. The data are accessible to anyone with an Internet connection. Neotoma database management is centralized but scientific oversight is distributed, thus providing domain scientists with quality control over their portions of the data. To this end, all data in Neotoma are stored in a single centralized database, but are conceptually organized into virtual *constituent* databases, for example, NAPD and EPD, each of which has its own governance charged with overseeing scientific issues such as taxonomy, data quality standards, data acquisition, and prioritizing tool development. Constituent database cooperatives may develop individualized and branded websites to display and distribute their portions of the database if they so desire. Constituent database projects can appoint 'data stewards' to remotely upload and update data. The primary reason for the centralized design is lower requisite IT cost for development and long-term maintenance – only one database is being maintained rather than a plethora of autonomous databases all requiring managers. Furthermore, the centralized database structure simplifies cross-disciplinary multiproxy analyses and common tool development.

The Neotoma project was initially funded by a 2-year grant from the U.S. National Science Foundation (NSF) Geoinformatics Program and was launched with a workshop at Pennsylvania State University in February 2007. The initial task was to develop a data model suitable for a wide range of paleo data, which was accomplished by merging the GPD and FAUNMAP, which represent a broad spectrum of paleo fossil data and depositional contexts, the former primarily from wet-sediment cores, the latter mainly from terrestrial excavations. Neotoma received a 5-year renewal from the NSF Geoinformatics program, which began in September 2010, when a second workshop was held in Madison, Wisconsin. The two workshops involved 50–60 participants, representing a large spectrum of paleo data, including representatives from several of the pollen database cooperatives. Other research projects utilizing Neotoma are contributing data as well.

The data from the GPD and other paleoecological databases are now available from the Neotoma website. Users can search for individual sites or datasets on a map or by several different

search criteria. Datasets can be downloaded or copied and pasted into a spreadsheet for use with other applications. In addition, the entire database can be downloaded in Microsoft SQL Server or Access format. The database and website are hosted at the Center for Environmental Informatics at Pennsylvania State University.

Neotoma currently contains data from the NAPD, EPD, LAPD, PDSRE, and IPPD. The APD is currently independent, but is slated for incorporation into Neotoma. The LAPD and Japanese Pollen Database (JPD) have also inventoried data for future incorporation into Neotoma (Figure 1).

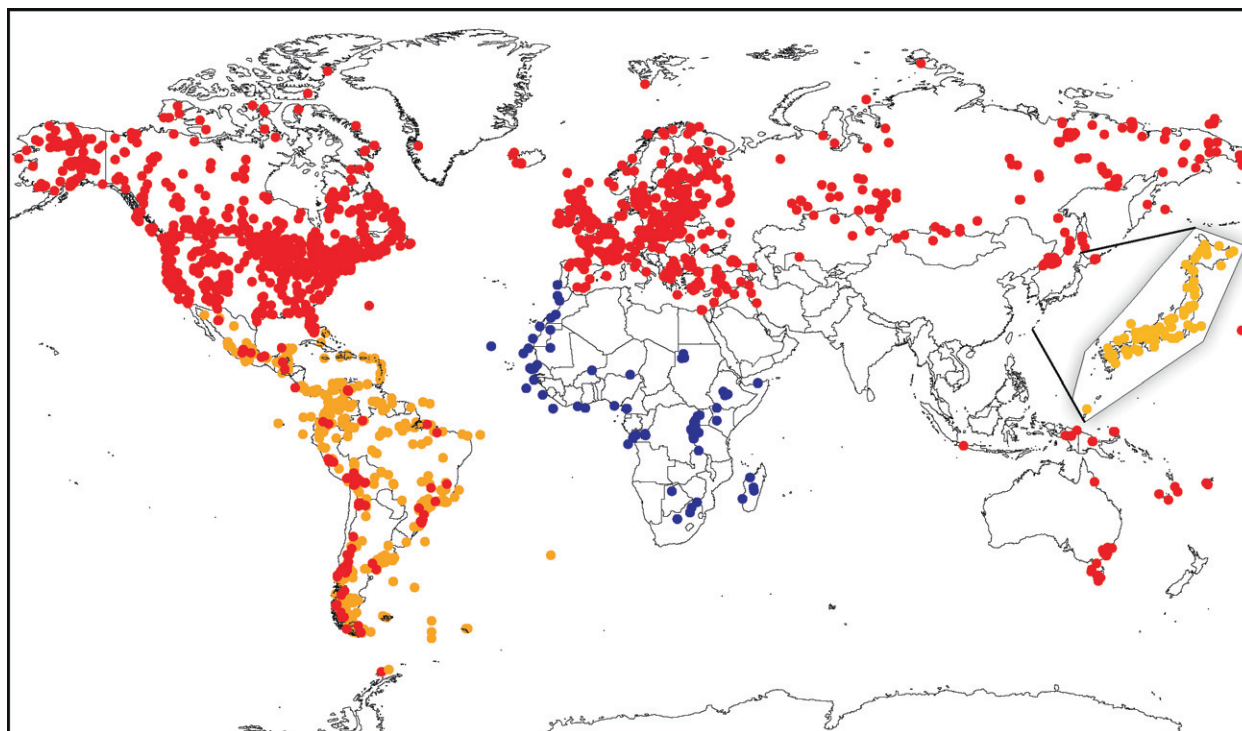
APD and LAPD contend with much larger and less well-known floras than do the pollen databases for temperate regions. APD has made a major effort at synonymization of pollen taxa and has produced a taxonomy list, which is available from its website. Images of pollen grains are also available from the APD website. Marchant et al. (2002) have published a major compendium of the ecology and distribution of sporophyte parent-taxa in LAPD.

The pollen databases make available data published in many languages, from many countries, and from literature not always widely available. Of particular note, PDSRFE assembled data from more than 175 sites, most of which were either unpublished or published in Russian in publications not widely available outside of Russia; and they published a compendium of these sites in Russian and English, with the original reference, a short description of each site, radiocarbon dates, short interpretation, and pollen diagram (Anderson and Lozhkin, 2002).

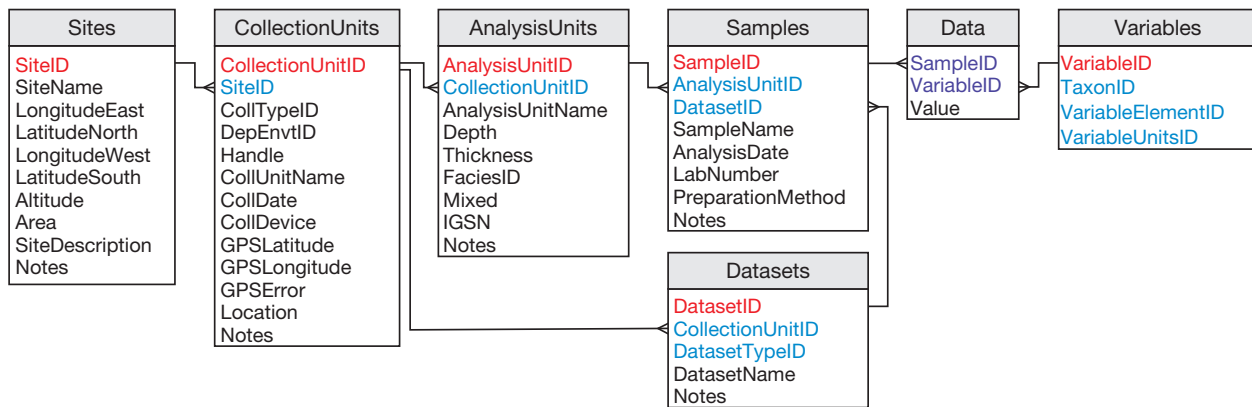
## Database Structure

Neotoma consists of a large number of tables and fields for data and metadata. Central tables to the database are (Figure 2)

- **Sites:** Name and location of site.
- **Collection units:** Typically cores, sections, or excavation units. Sites may have multiple collection units, which have spatial coordinates.
- **Analysis units:** Individual units analyzed, typically identified by depth and thickness within a collection unit. Collection units may have multiple analysis units.
- **Samples:** A data type tabulated from an analysis unit, for example, a pollen sample, a plant macrofossil sample. An analysis unit may have multiple samples.
- **Datasets:** Samples of a particular data type from a collection unit comprise a dataset, for example, a pollen dataset. A collection unit may have multiple datasets (e.g., pollen, plant macrofossils, ostracodes), and a dataset may have multiple samples of the same data type.
- **Variables:** A variable consists of a taxon, element, and measurement unit. For plants, the element is the plant organ, for example, pollen, spore, seed, needle. Thus, *Picea* pollen and *Picea* needles are different variables. For pollen, possible measurement units are number of identified specimens, percent, or present/absent. Lookup tables exist for taxon names, element types, and measurement units.



**Figure 1** Pollen sites (cores and sections). Red dots – Neotoma Paleocology Database, blue dots – African Pollen Database, yellow dots – inventoried sites for inclusion.



**Figure 2** Central tables in the Neotoma Paleoecology Database. Table names are in the gray boxes. Red fields are primary keys; blue fields are foreign keys; violet fields are compound primary keys. Relationships shown by connection lines. IGSN, International geo sample number.

In addition, tables exist for:

- *Contacts*: principal investigators, collectors, analysts, data contributors;
- *Geochronology*: radiocarbon and other radiometric dates;
- *Chronologies*: chronology controls (e.g., calibrated  $^{14}\text{C}$  ages, tephras, biostratigraphic ages), age models (algorithms for assigning sample ages based on chronology controls), and ages assigned to individual samples;
- *Lithology*: core or section descriptions;
- *Publications*: bibliographic references for data and metadata;
- *Synonymy*: accepted names and synonyms, original identifications.

## Quality Control

The database cooperatives make every effort to ensure data accuracy. Quality control is maintained by having data stewards with palynological training enter data and by a software interface, currently under development, between the data input files and the database files. This interface checks for many kinds of common errors and inconsistencies. Probably, the thorniest quality control issues have been nomenclature and taxonomic synonymization. On a basic level, misspellings are corrected, but, in addition, the database does synonymize some names. The database coordinators must be cognizant of taxonomic resolution. Although some datasets may not contain the taxonomic resolution of others, they may nevertheless be suitable for synoptic studies. Whereas the quality of the data is ultimately the responsibility of the contributor, checking the appropriateness of any dataset is the responsibility of the user.

## Taxonomy and Nomenclature

A protocol of the pollen databases is that contributors' names for pollen types are not changed in their meaning. Thus, a taxonomic identification is not questioned; although, of course, the database coordinators may choose not to include datasets with obvious misidentifications. However, some correction and synonymization is desirable and necessary to make the database more functional. Plant taxonomy of higher plants

is largely based on the diploid sporophyte generation, and taxonomic synonymization for pollen involves complications in addition to the usual issues of nomenclatural synonymy. An important problem arises from the fact that sporophyte names are used for pollen types, but pollen types may derive from more than one sporophyte taxon. As a consequence, names are combined, many pollen types have the suffix-type appended to them, and the usage of *cf.* to indicate uncertainty is common.

## Taxonomic Hierarchy

The pollen database contains a flexible taxonomic hierarchy, which stores the maximum taxonomic precision made by the analyst but allows users to aggregate taxa into higher level categories. Whereas high taxonomic precision might be desirable for certain ecological studies, lumping is necessary for synoptic studies. The essence of the hierarchy is that each pollen variable has a field that points to the next higher variable in the hierarchy. For example, *Fraxinus nigra* points to *Fraxinus nigra*-type, which points to *Fraxinus*. The hierarchy then facilitates the retrieval of all instances of *Fraxinus* pollen. Changing botanical definitions and taxonomic synonymies are a persistent problem for databases of species occurrences.

## Synonymy and Naming Conventions

Three general types of synonymy exist: nomenclatural, syntactic, and pollen morphological. The database synonymizes for nomenclature and syntax, but not for morphology. Nomenclatural synonymy refers to botanical nomenclature. The regional pollen databases follow major works relevant for their area, such as *Flora of North America* (Flora of North America Editorial Committee, 1993 et seq.), *Flora Europae* (Tutin et al., 1964–1993, 1993), *Vascular Plants of Russia and Adjacent States* (Czerepanov, 1995), *Énumération des plantes à fleurs d'Afrique Tropicale* (Lebrun and Stork, 1991–1997), and the *Australian Plant Name Index* (Chapman, 1991). Inevitably, with a global database such as Neotoma, nomenclatural differences occasionally occur. In general, the newer authority will be followed, or if one name varies from recommendations of the *International Code of Botanical Nomenclature* (2006), then the name

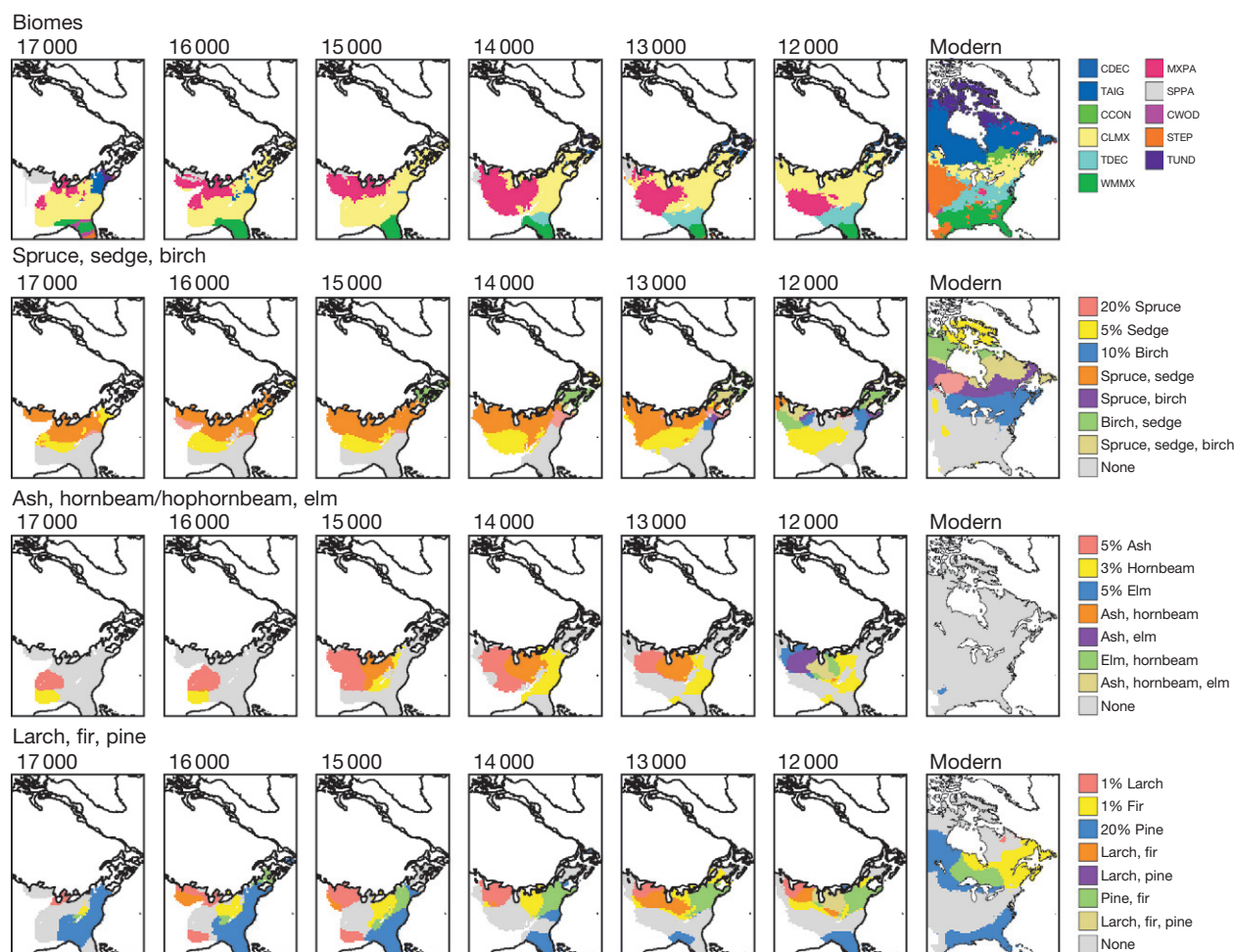


following the Code will be used. Thus, for example, families with the 'aceae' endings such as Poaceae and Asteraceae are preferred over Gramineae and Compositae.

In Neotoma, Angiosperm family names and higher ranks follow the phylogenetic classification of the Angiosperm Phylogeny Group (APG; [Angiosperm Phylogeny Group, 2003, 2009; Stevens, 2001 et seq.](#)). Some of the APG revisions based on phylogenetics change long-standing taxonomic classifications, but are consistent with pollen morphology. For example, the genus *Celtis* has been moved from the Ulmaceae to the Cannabaceae. The scabrate, triporate pollen grains of *Celtis* are more consistent with the Cannabaceae than with the rugulate, stephanoporate pollen grains of *Ulmus* and other genera remaining in the Ulmaceae. Pollen from the traditional families Chenopodiaceae and Amaranthaceae are commonly grouped together as Chenopodiaceae/Amaranthaceae (or often informally as 'ChenoAms'). Conveniently from the

palynological standpoint, APG has synonymized Chenopodiaceae with Amaranthaceae, thus the Chenopodiaceae/Amaranthaceae simply becomes Amaranthaceae (as does the former superfamily Chenopodiineae).

Nomenclatural synonymization should not change the range of pollen morphotypes intended by the original analyst. For example, APG has synonymized the Agavaceae with the Asparagaceae, a family that includes many more taxa than the former Agavaceae. However, the genera of the former Agavaceae now comprise the subfamily Agavoideae, and the name is thus changed to Asparagaceae subf. Agavoideae. Neotoma has tables for synonyms and for original identifications. Thus, an original identification of Agavaceae can be retained. In this example, the circumscription of potential taxa has not changed. However, synonymization that involves splitting or lumping potentially does change the circumscription of morphotypes included in the analyst's original identification. For example, an



**Figure 3** Representations of late Pleistocene and modern plant associations as biomes (top row) and as multitaxon isopoll maps (rows 2–4). Numbers indicate cal year BP. Biomes are: CDEC, cold deciduous forest; TAIG, Taiga; CCON, cool conifer forest; CLMX, cool mixed forest; TDEC, temperate deciduous forest; WMMX, warm mixed forest; MXPA, mixed parkland; SPPA, spruce parkland; CWOD, conifer woodland; STEP, steppe; TUND, tundra. Three taxa are mapped in each multitaxon map, with a single isopoll shown for each plant taxon. Primary colors (red, blue, cyan) indicate regions where only one of the three taxa is present in appreciable numbers. Secondary colors (orange, purple, green) denote associations between pairs of taxa. Beige indicates where all three taxa overlap. Scientific names: spruce, *Picea*; sedge, Cyperaceae; birch, *Betula*; ash, *Fraxinus*; hornbeam, *Ostrya/Carpinus*; elm, *Ulmus*; larch, *Larix*; fir, *Abies*; pine, *Pinus*. Reproduced from [Williams JW, Shuman BN, and Webb III T \(2001\)](#) Dissimilarity analyses of late-Quaternary vegetation and climate in eastern North America. *Ecology* 82: 3346–3362, with permission.

identification of Cannabaceae prior to APG would have included the genera *Cannabis* and *Humulus*, but probably not *Celtis*. Thus, in Neotoma these earlier identifications of Cannabaceae have been changed to Cannabaceae *sensu stricto*.

Syntax refers to elements of pollen-type names that are not parts of the Linnean system (i.e., species, genus, family...). Neotoma has a set of rules for pollen syntax; for example, for types the convention is name-type with a hyphen, for example, *Ambrosia*-type, not *Ambrosia* type. For a global database, types are a problem because they often are not precisely defined, or they are defined differently in different regions. Researchers need to be aware of these issues of taxonomic precision when utilizing the database.

Pollen types with different names may be morphologically identical, but the GPD does not attempt to synonymize these for several reasons. The principal reason is that, unlike floras for the sporophyte life-history stage, complete palynological floras and monographs do not exist. The database coordinators, although experienced palynologists, have no published authority on which to base their morphological naming conventions, nor do they have the expertise to determine the full realm of morphological synonyms. They could possibly reject fine morphological distinctions that, in fact, a given palynologist has made with confidence. Although, the GPD does not synonymize based on morphology, some regional pollen databases that contribute to the GPD may synonymize based on intimate knowledge of the palynology of their region. Another important argument against morphological synonymization is that the pollen-type name may indicate the most probable derivative sporophyte taxon based on biogeographic or ecological considerations.

One other database protocol that produces name changes is that, for any dataset, pollen-type names must be mutually exclusive, which is necessary for lumping algorithms to work properly. For example, if a dataset includes *Potentilla* and Rosaceae, the latter is changed to Rosaceae undiff., because Rosaceae includes *Potentilla*. Some of these names can be somewhat cumbersome. For example, many contributed datasets include *Artemisia* and Asteroideae (or the obsolete Tubuliflorae). The latter is changed to Asteraceae subf. Asteroideae undiff.

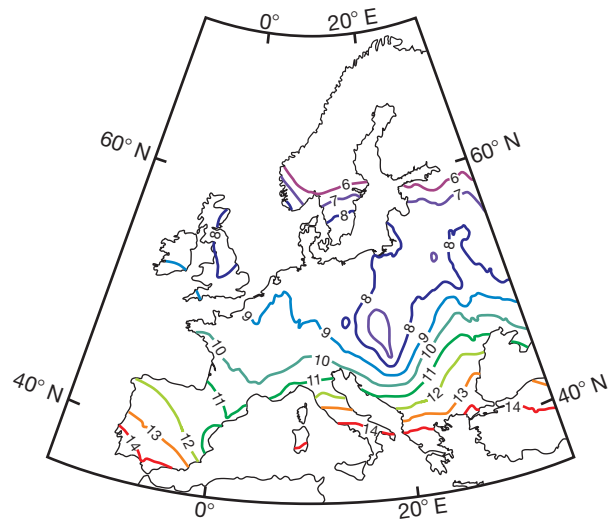
## Database Applications

The pollen databases have been widely used. Applications range from seeking to place a new record into a regional context (e.g., Nelson and Hu, 2008; Schauffler and Jacobson, 2002) to synoptic mapping at continental to global scale (e.g., Allen et al., 2010) and paleoclimatic reconstructions for evaluation of global climate models (e.g., Bartlein et al., 2011). Applications range from evolutionary biology and macroecology (e.g., DiMichele et al., 2004; Giesecke and Bennett, 2004; Petit et al., 2004) to natural resource management (e.g., Cole et al., 1998; Heard and Valente, 2009). Given the diversity of studies utilizing the pollen database, selection of a few representative examples is difficult. Nonetheless, the following two illustrate some of the more powerful synoptic applications.

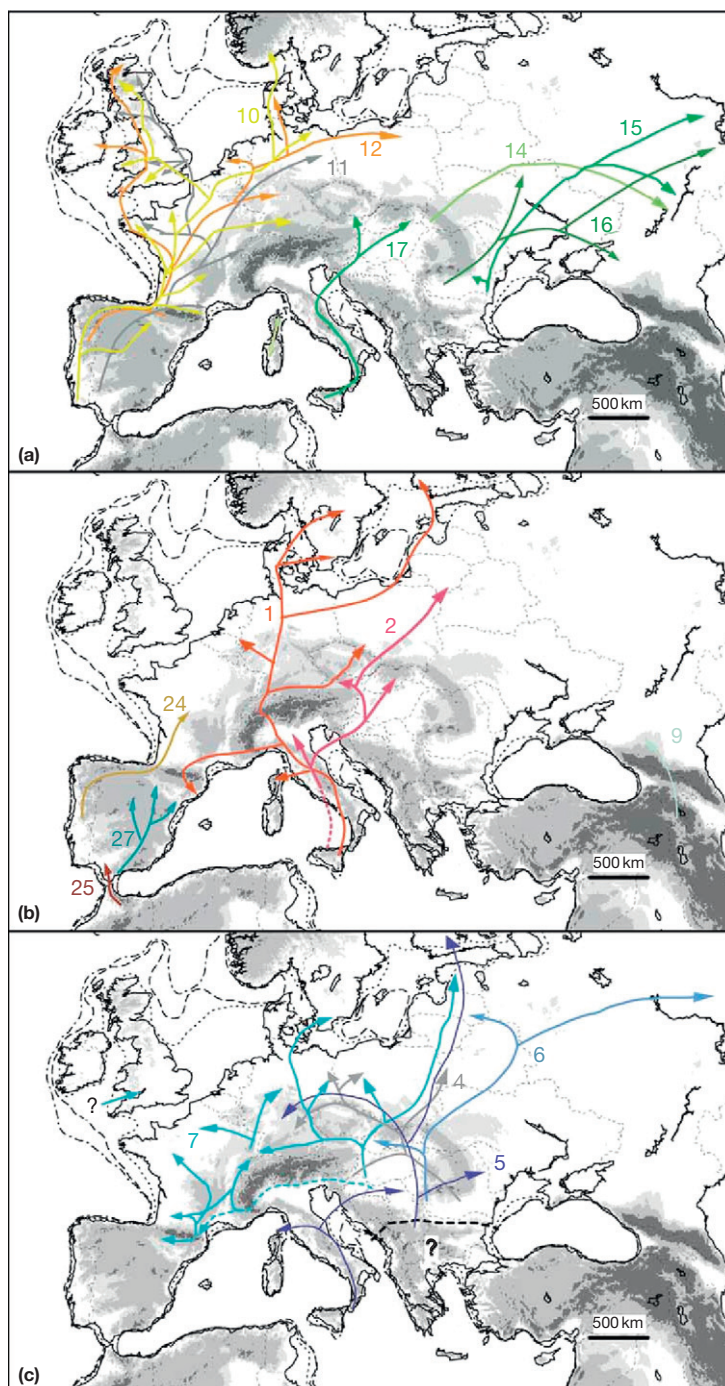
Biomization is a particularly powerful technique for deriving biomes from pollen assemblages. Rather than using actual species, biomization utilizes plant functional types (PFTs),

which are based on life form, physiology, phenology, leaf form, and climatic tolerances. The method entails assigning PFTs to biomes globally, assigning pollen taxa to PFTs regionally, and using these assignments to assign regional pollen taxa to biomes. For fossil samples, 'affinity scores' are calculated between pollen samples and biomes. The biome with the highest affinity score is assigned to the fossil-pollen sample, subject to a tie-breaking rule (Prentice et al., 1992, 1996; Williams et al., 1998, 2000, 2001). Advantages of biomization are that nonanalog biomes are definable and that no reliance is placed on comparison with modern samples, which may be biased by anthropogenic activities. Williams et al. (2001) mapped biomes and major taxa for the Late Glacial interval (17 000–12 000 cal year BP) in eastern North America (Figure 3). Mapped are two nonanalog biomes – mixed parkland and spruce parkland – that no longer exist. The maps clearly show the changing biomes and taxon distributions through time. This study also showed that the peak interval of dissimilarity from the present preceded the time of maximum climate change. Thus, nonanalog vegetation does not appear to be a nonequilibrium transient state, but, in fact, is in equilibrium with nonanalog climate conditions.

Maps of fossil-pollen data show species migrations from glacial ranges or 'refugia' to their modern ranges. Modern genetic data also contain clues to the past migration history of taxa. In Europe, collaboration between palynologists and geneticists has produced a detailed history of the migration of deciduous *Quercus* from glacial refugia (Brewer et al., 2002; Petit et al., 2002). Because of frequent introgression, the various *Quercus* species share chloroplast DNA (cpDNA) haplotypes regionally, and haplotype lineages cross species lines. Consequently, the taxonomic resolution of the cpDNA lineages is comparable to the identification of pollen to the genus level. Palynological data (Figure 4) and the geographic distribution of haplotypes (Figure 5) show the existence of refugia



**Figure 4** Isopoll map showing time of first arrival of deciduous *Quercus*. Adapted from Brewer S, Cheddadi R, de Beaulieu J-L, Reille M, and Data contributors (2002) The spread of deciduous *Quercus* throughout Europe since the last glacial period. *Forest Ecology and Management* 156: 27–48.



**Figure 5** Maps showing inferred postglacial colonization routes for frequent cpDNA haplotypes of deciduous *Quercus*. Colors indicate different lineages, and numbers indicate haplotypes. Maps (a), (b), and (c) show different haplotypes moving from the same refugia. Gray shading indicates altitude. Reproduced from Petit RJ, Brewer S, Bordács S, et al. (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management* 156: 49–74.

in the Iberian Peninsula, Apennine Peninsula, and Balkan Peninsula; and isopoll maps and the distributions of haplotypes show migration routes north from these southern peninsulas to northern Europe and into central Asia. Pollen and genetic data are in agreement about general migration trends. Synoptic mapping of *Quercus* pollen shows where refugia were and when migrations occurred. Genetic survey adds detail to

the migration history; for example, DNA data show that *Quercus* in the British Isles is related to the Iberian Peninsula populations, rather than to those from the Apennine refugium. Genetic data can also suggest where more palynological work is warranted. For example, a unique haplotype in Corsica and Sardinia suggests a refugium there, although no pollen data are available to confirm it.



## Conclusions

The quantity of data in the pollen databases and the hundreds of papers that have utilized these databases demonstrate their great success. The databases are public and access is free. The databases have fostered collaboration between scientists in different disciplines and from different countries, and they are critical to global change research.

The pollen data were some of the first paleo data to be organized into a relational database structure. Many lessons concerning scientific database organization have been learned. To be successful, the database must satisfy needs and wishes of both data contributors and data consumers. Some scientists had initial reluctance to contribute data. However, willingness to participate has increased with the maturation of the database and the increasing evidence of its value as a scientific resource, combined with the broader informatics revolution. A large database effort must involve a number of data cooperatives. Unfunded data contributors will send data 'as is,' but typically they will not spend much time entering or reformatting data to fit database standards.

The incorporation of the pollen databases into the multi-proxy Neotoma Paleocology Database attends to several long-standing needs. The common cyberinfrastructure should reduce long-term maintenance costs, enhance data preservation, and guard against data loss, while the distributed governance permits expert review and oversight for particular data types. The development of a software interface allowing data stewards from the various pollen database cooperatives to upload and manage data remotely should permit these cooperatives to focus on data entry rather than IT needs. Neotoma will facilitate new multi-proxy studies of Quaternary environments.

**See also:** Pollen Records, Last Interglacial of Europe. **Plant Macrofossil Methods and Studies:** Validation of Pollen Studies. **Pollen Methods and Studies:** Archaeological Applications. **Pollen Records, Late Pleistocene:** Africa; Australasia; Middle and Late Pleistocene in Southern Europe; Northern Asia; Northern North America; South America; Western North America. **Pollen Records, Postglacial:** Africa; Australia and New Zealand; Northeastern North America; Northern Asia; Northern Europe; Northwestern North America; South America; Southeastern North America; Southern Europe; Southwestern North America.

## References

- Allen JRM, Hickler T, Singarayer JS, Sykes MT, Valdes PJ, and Huntley B (2010) Late glacial vegetation of northern Eurasia. *Quaternary Science Reviews* 29: 2604–2618.
- Anderson PM and Lozhkin AV (2002) *Late Quaternary Vegetation and Climate of Siberia and the Russian Far East (Palynological and Radiocarbon Database)*. Magadan: North East Science Center, Far East Branch, Russian Academy of Sciences.
- Angiosperm Phylogeny Group (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141: 399–436.
- Angiosperm Phylogeny Group (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* 161: 105–121.
- Bartlein PJ, Harrison SP, Brewer S, et al. (2011) Pollen-based continental climate reconstructions at 6 and 21 ka: A global synthesis. *Climate Dynamics* 37: 775–802.
- Brewer S, Cheddadi R, de Beaulieu J-L, and Reille M (2002) Data contributors: The spread of deciduous *Quercus* throughout Europe since the last glacial period. *Forest Ecology and Management* 156: 27–48.
- Chapman AD (1991) *Australian Plant Name Index. Australian Flora and Fauna Series 12*, 4 vols. Canberra: Australian Government Publishing Service.
- COHMAP Members (1988) Climatic changes of the last 18,000 years: Observations and model simulations. *Science* 241: 1043–1052.
- Cole KL, Davis MB, Stearns F, Walker K, and Guntenspergen G (1998) Historical landcover changes in the Great Lakes region. In: Sisk TD (ed.) *Perspectives on Land Use History of North America: A Context for Understanding Our Changing Environment*. United States Geological Survey, Biological Resources Division, Biological Science Report USGS/BRD/BSR 1998-0003, pp. 43–50.
- Czerepanov SK (1995) *Vascular Plants of Russia and Adjacent States (the Former USSR)*. Cambridge: Cambridge University Press.
- DiMichele WA, Behrensmeyer AK, Olszewski TD, et al. (2004) Long-term stasis in ecological assemblages: Evidence from the fossil record. *Annual Review of Ecology, Evolution, and Systematics* 35: 285–322.
- Flora of North America Editorial Committee (1993 et seq.) *Flora of North America North of Mexico*, 14+ vols. New York: Oxford University Press.
- Fyfe RM, de Beaulieu J-L, Binney H, et al. (2009) The European Pollen Database: Past efforts and current activities. *Vegetation History and Archaeobotany* 18: 417–424.
- Giesecke T and Bennett KD (2004) The Holocene spread of *Picea abies* (L.) Karst. in Fennoscandia and adjacent areas. *Journal of Biogeography* 31: 1523–1548.
- Heard MJ and Valente MJ (2009) Fossil pollen records forecast response of forests to hemlock woolly adelgid invasion. *Ecography* 32: 881–887.
- Huntley B and Birks HJB (1983) *An Atlas of Past and Present Pollen Maps for Europe: 0–13000 Years Ago*. Cambridge: Cambridge University Press.
- International Code of Botanical Nomenclature (Vienna Code) (2006) *Regnum Vegetabile*, vol. 146. Ruggell: A.R.G. Gantner Verlag KG.
- Lebrun J-P and Stork AL (1991–1997) *Énumération des plantes à fleurs d'Afrique Tropicale*, 4 vols. Geneva: Conservatoire et Jardins Botaniques de la Ville de Genève.
- Marchant R, Almeida L, Behling H, et al. (2002) Distribution and ecology of parent taxa of pollen lodged within the Latin American Pollen Database. *Review of Palaeobotany and Palynology* 121: 1–75.
- Nelson DM and Hu FS (2008) Patterns and drivers of Holocene vegetational change near the prairie-forest ecotone in Minnesota: Revisiting McAndrews' transect. *New Phytologist* 179: 449–459.
- Petit RJ, Bialozyt R, Garnier-Géré P, and Hampe A (2004) Ecology and genetics of tree invasions: From recent introductions to Quaternary migrations. *Forest Ecology and Management* 197: 117–137.
- Petit RJ, Brewer S, Bordács S, et al. (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management* 156: 49–74.
- Prentice IC, Cramer W, Harrison SP, Leemans R, Monserud RA, and Solomon AM (1992) A global biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography* 19: 117–134.
- Prentice IC, Guiot J, Huntley B, Jolly D, and Cheddadi R (1996) Reconstructing biomes from palaeoecological data: A general method and its application to European pollen data at 0 and 6 ka. *Climate Dynamics* 12: 185–194.
- Schauffler M and Jacobson GL Jr. (2002) Persistence of coastal spruce refugia during the Holocene in northern New England, USA, detected by stand-scale pollen stratigraphies. *Journal of Ecology* 90: 235–250.
- Stevens PF (2001 et seq.) Angiosperm Phylogeny Website. Version 9, June 2008. <http://www.mobot.org/MOBOT/research/APweb/>.
- Tutin TG, Heywood VH, and Burges NA, et al. (eds.) (1964–1993) In: *Flora Europaea*, 5 vols. Cambridge: Cambridge University Press.
- Tutin TG, Burges NA, Chater AO, et al. (1993) *Psilotaceae to Platanaceae, Flora Europaea*, 2nd edn., vol. 1. Cambridge: Cambridge University Press.
- von Post L (1967) Forest tree pollen in south Swedish peat bog sediments. *Pollen et Spores* 9: 375–401.
- Williams JW, Shuman BN, and Webb T III (2001) Dissimilarity analyses of late-Quaternary vegetation and climate in eastern North America. *Ecology* 82: 3346–3362.
- Williams JW, Summers RL, and Webb T III (1998) Applying plant functional types to construct biome maps from eastern North American pollen data: Comparisons with model results. *Quaternary Science Reviews* 17: 607–627.
- Williams JW, Webb T III, Richard PH, and Newby P (2000) Late Quaternary biomes of Canada and the eastern United States. *Journal of Biogeography* 27: 585–607.
- Wright HE Jr., Kutzbach JE, Webb T III, Ruddiman WF, Street-Perrott FA, and Bartlein PJ (eds.) (1993) *Global Climates Since the Last Glacial Maximum*. Minneapolis: University of Minnesota Press.