
A Database Model for Integrating and Facilitating Collaborative Ethnomedicinal Research

Michael B. Thomas¹, Nan Lin² and Howard W. Beck³

¹Department of Botany, University of Florida, PO Box 118526, Gainesville, FL, USA; ²Agricultural and Biological Engineering Department, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL, USA;

³Agricultural and Biological Engineering Department, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL, USA

Abstract

A model for a database system that provides a standardized environment for submission, storage, and retrieval of ethnomedicinal data was developed. The model is based on object oriented database technology, and is suitable for not only storing data, digital images, sound and video, but also for modeling domain knowledge associated with plant-based medicinal preparations utilized in systems of traditional medicine. The model incorporates both linguistic and semantic elements. Terms in natural language are mapped to database objects that represent knowledge in various ethnomedicinal domains. The distributed object infrastructure permits integration with other authoritative taxonomic databases and includes an interface capable of supporting existing and emerging standards of data. The model provides a foundation for a globally current dynamic data resource that encourages comparative ethnomedicinal research through direct contributions by members of the research community. Examples of integrated domain models are presented incorporating medical terminology, plant systematics, ecology, and pharmacology.

Keywords: ethnomedicine, ethnopharmacology, medicinal plants, ethnobotany, database, object database, Java, internet.

Introduction

The study of the interactions of plants and people, including the influence of plants on human culture, is the focus of the interdisciplinary field of ethnobotany (Balick & Cox, 1996). Ethnobotanical research incorporates much interdisciplinary knowledge. In this paper, we present a data model based on

ethnobotanical fieldwork conducted in Brazil where medicinal plants and their properties were identified and knowledge gained through collaborative interviews with indigenous experts was included. The goal of this study is to demonstrate how knowledge from diverse disciplines can be integrated into a single dynamic database. A new database design is used to describe several interrelated areas of scientific knowledge and expertise. These areas include botany, systematics, ecology, plant geography, pharmacology, linguistics, and anthropology. Traditionally, each of these areas is highly specialized, and extensive training is required for individuals to achieve a level of proficiency in any one area. Yet in ethnobotanical research, interrelationships among these areas are also very important. In this paper, we present a data model capable of representing detailed knowledge in each area, as well as showing interrelationships across disciplines.

The technique of data modeling involves studying a concept and creating a symbolic representation of that concept; the symbolic representation captures the meaning of the concept. The symbolic representation is a data structure that can be stored and manipulated in a computer. Traditionally, databases are used to store raw field observations and other facts prior to analysis, and Web pages arbitrarily store much knowledge in unstructured HTML. We use a new database approach, based on object database technology (Barry, 1996), capable of storing concepts and knowledge as well as raw data. This new approach uses data structures called "objects" to represent concepts. In contrast to traditional established databases based on the relational model, in which data are viewed as records in tables, object databases provide a more natural description of the entities within a

Accepted: ■■

Address correspondence to: Michael B. Thomas, Department of Botany, University of Florida, PO Box 118526, Gainesville, FL 32611-8526, USA. Fax: (352) 392-3993, E-mail: mthomas@botany.ufl.edu

domain by examining objects, their properties, and interrelationships.

An example of a data object for a medicinal plant specimen is shown in Figure 1, which illustrates several characteristics of an object. First, an object has a name (Voucher MT543) that uniquely identifies the object. Objects also have properties, in this case plant characteristics such as height. Properties have values; in this case, the height is ca. 3.5 to 12 m. Objects have parts (leaves, flowers), and parts also have properties. In addition, objects have relationships with other objects. In this example, the specimen is related to more general classes taxonomically, and may have more specific species or cultivars, as indicated by “Superclasses” and “Subclasses”. The superclass/subclass relationships result in taxonomies of objects that can be quite large (thousands of objects arranged in many classes). Object database management systems (ODBMS) are databases used to store and manage large collections of objects. In general, an object has a unique name, properties, parts, associations with other objects, and general/specific relationships with other objects (a special kind of association). A more formal definition of the data modeling language used to define objects is given below.

As the object data modeling language is quite general, it can be applied to many different subjects. By using associa-

tion relationships, objects from within and between different disciplines can be interrelated. The study in this paper illustrates the application of object data modeling to ethnobotany. We demonstrate how object models can be built in each of several different disciplines related to ethnobotany, and show how these disciplines are integrated via object associations. One of the disciplines involved is medical terminology, in which the object database is used to describe a thesaurus and glossary of medical terms. Systematics is treated by using object classes to describe ordinal/family/genus/species relationships; the taxonomic structure of the object database lends itself well to this task. Several interesting problems in systematics, including the interrelation of different taxonomic conventions and the tracking of dynamic changes in the taxonomy, are also addressed. Ecology is treated by using object classes to describe ecological zones, which are characterized by climate among other properties. Chemical descriptions and relationships needed in ethnopharmacology are also handled using object descriptions of chemical properties and associations among chemicals.

The work presented here is intended for use by scientists working in ethnobotany, with the goal of developing and disseminating ethnobotanical data and knowledge among researchers. We have attempted to present the work in a fashion that does not require extensive training in data mod-

Voucher MT543
 Superclasses: Family Anacardiaceae
 Subclasses: a cultivar type
 Plant name - Genus/ Species/ Authority: *Anacardium occidentale* L.
 Synonym(s): *Anacardium occidentale* Linn.
 Common name(s): cashew, cashew apple, cashew nut
 Ethnic name(s): Cajueira, Cajú
 Locality: Pataxo Indigena Reserva
 County/municipality: Porto Seguro
 State: Bahia
 Country: Brazil
 Lat: 16° 51' S
 Long: 39° 09' W
 Elevation: 20 m absl
 Description: Along footpath leading to Posto. Open field.
 Habit: Spreading evergreen perennial tree
 Height: ca. 3.5 to 12 m tall
 Leaves: {Leaf Arrangement: leaves simple, alternate, obovate, glabrous Size: to 20 cm long, 15 cm wide, apically rounded or notched}
 Flowers: {Petal Arrangement: nearly radially symmetrical; Color: pinkish-yellow}
 Fruit: short oblong, rounded, with much swollen pedicel and apical curved fruit; red to yellow
 Cultivated: no
 Collector: M.B. Thomas
 Collection number: MT543
 Date Collected: 16 September 1998
 Photo(s)
 Entire Plant...
 Leaf
 Flower
 Fruit
 Voucher specimen
 Illustration(s)...

Figure 1. Object representation of a medicinal plant collected in Brazil.

eling or other aspects of database design. To that end, we have developed some high-level data design and visualization tools that can be used by researchers to browse and edit the database. Using the object paradigm, researchers deal with objects that relate to their discipline, and therefore are easily recognizable. The visualization tools present the objects in a familiar form. These tools are accessible through a Web browser.

The data modeling language and visualization tools are described in the next section. Following that, we present detailed data models for several disciplines encompassed by ethnobotany and then describe the issues and problems addressed by the models. We present an overview of other related botanical databases, and outline an approach for building a global ethnomedicinal database.

Data modeling language and visualization tool

A data modeling language was developed to describe terminology, concepts, data, domain knowledge and multimedia in cross-disciplinary fields. This model captures the meanings of domain terminology as well as the semantic relationships between these terms. Knowledge in the database can be visualized and edited by users through the Internet. A Web-based database browser is being built for this purpose.

The data modeling language consists of two layers (Fig. 2), the linguistics layer and the semantics layer:

The linguistic layer organizes terms (words and phrases) for describing database objects. It is necessary because a given term can have multiple meanings (homonyms), or many terms can have the same meaning (synonyms). Furthermore, terms can play different syntactic rolls (parts of speech). For example, this is a common problem in taxonomic systems where many common names are used, and even scientific names may not uniquely refer to the same organism. We have initialized the linguistic layer with terms from WordNet, a general thesaurus of over 100,000 English terms (Fellbaum,

1998), with the intention of building in additional domain vocabularies specific to ethnomedicinal research. Following WordNet conventions, each term maps to one or more Synsets (synonym sets). Each Synset represents a different meaning of the term, and furthermore one or more synonymous terms are collected into the same Synset. A given term is first located within a dictionary, and one or more Synsets associated with that term are identified. Users can select which sense they intend (or contextual clues can sometimes be used to automate selection). Once a unique sense is identified, it is mapped to a unique concept in the Semantic layer of the database. The linguistics layer forms an interface between the user's natural language and the underlying semantic layer.

Consider for example the term "plant". The linguistic layer indicates that "plant" has four noun word senses:

1. buildings for carrying on industrial labor; "They built a large plant to manufacture automobiles." (plant, works, industrial plant)
2. a living organism lacking the power of locomotion (plant, flora, plant life)
3. something planted secretly for discovery by another; "The police used a plant to trick the thieves."; "He claimed that the evidence against him was a plant." (plant)
4. an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience (plant)

After each sense, the Synset (list of associated synonyms) is shown in parentheses. Each Synset is mapped to a unique Class in the semantic layer, which represents the meaning of the concept "Plant" for that sense.

The semantic layer provides a way to represent the meaning of concepts in a domain. A concept is modeled by a Class that describes its properties and relationships with other Classes. Furthermore we use "Instance" to describe a particular occurrence of a concept (the term "instance" is synonymous with "object"). That is, objects belonging to a certain class are Instances of that class. The elements of a Class are illustrated by an example shown in Figure 3.

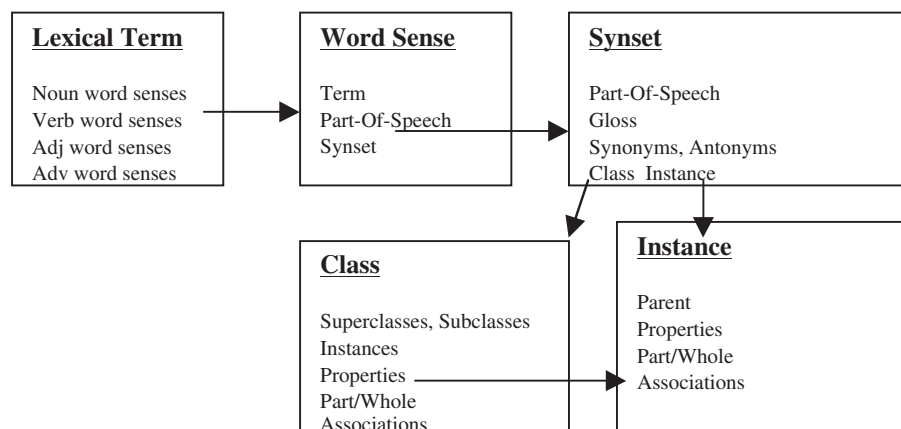


Figure 2. Data modeling language.

Edible Nut

Superclass: Angiosperms... - is an artifact, object and entity - a more general Class than itself.

Subclass: Cashew, Brasil Nut, Almond... - is a more specific Class than itself.

Properties: Common name(s), Ethnic name(s), Ecological Zone, Locality, County/municipality, State, Country, photographic images - is a collection of attributes of the Class. They can be strings, images, sounds, video clips or any other complex data type

Association: Tropical moist forest... - is a plant can be associated with some other objects such as tropical forest in an ecological relationship, relationships between another Class and itself.

Parts and PartOf: leaves, fruit, stem, roots... - is the part-whole relationship describing an entity and its composites.

Instances: MT602... - is a collection of specific objects belonging to this Class.

Figure 3. "Edible Nut" as a class.

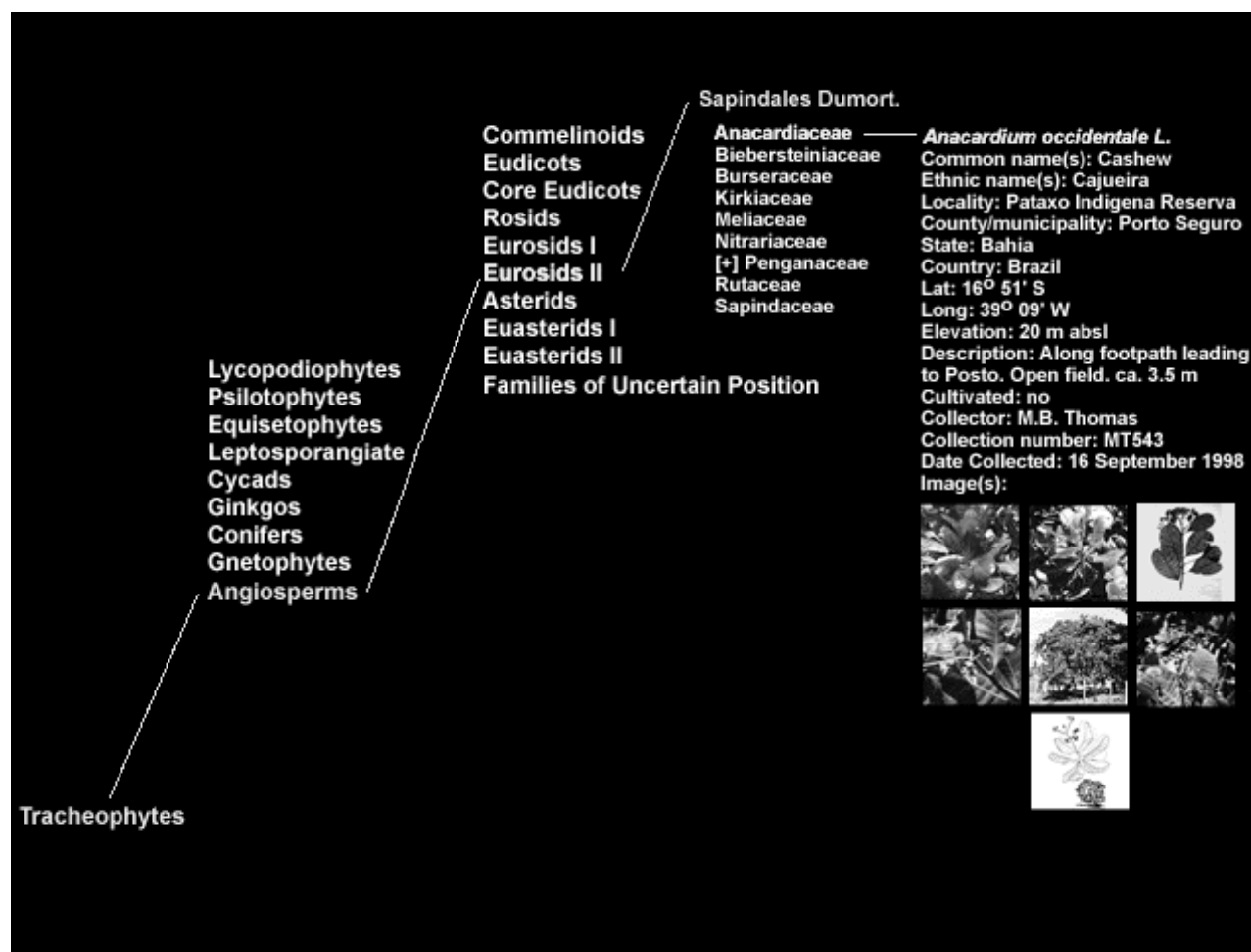


Figure 4. Web-based data visualization tool.

Since Classes capture the semantic structure of a concept, they can be used to automatically classify new concepts through a structure matching process. When a new Class or Instance is added, its structure is compared against those of existing Classes. This process identifies a location in the taxonomy to which the new concept belongs.

A graphical visualization tool is being created for browsing and editing the database (Fig. 4). This database browser is a Java applet that runs inside standard Web browsers. It

displays the database graphically as a node and link diagram where nodes represent concepts and links represent relationships between concepts. The details of the underlying data modeling language are hidden. To start navigation of the database, users can select a term and the term will be displayed as a node. By clicking on a node, users can inspect relationships with other nodes. Users can also view the properties of a node from an opened window attached to the node. Authorized users can edit the database through the browser

and the changes will be visible for all users. This tool can be used for on-line ethnomedicinal information acquisition, retrieval and decision making.

Integrating multidisciplinary ethnobotany studies

Next, we present examples of using the data model to describe four different, yet interrelated disciplines within Ethnobotany. We describe models for medical terminology, systematics, ecological zones, and pharmacology. These are by no means the only or most important areas in ethnobotany, but these examples illustrate the basic approach involved. We are currently working to incorporate additional domains within the overall database design.

Medical terminology glossary

The use of medical terminology to describe medicinal plant specimens occurs at several levels. For example, during ethnobotanical interviews with medicinal plant experts, a particular specimen may be described as, “relieving pain or ache in the head”. There is a general problem of unambiguously mapping commonly used medical terms to specific medical concepts. There is a distinction between common and scientific terms, as well as between folk or indigenous uses of terms compared with international usage. Furthermore, there are many different languages (such as English {headache} verses Portuguese {dor de cabeça}) which must also be addressed.

As a framework, the data model handles mapping between terms used at various levels, through the linguistic layer, and maps terms to medical concepts represented in the semantic layer. Consider the term, “headache”. This term has two senses in WordNet, one being the more common “pain in the head”, technically described as “caused by the dilation of cerebral arteries”, the other being a psychological state of anxiety. Someone describing a medicinal plant as used to combat “headache” might be referring to either sense. Synonyms for the first sense would include “cephalalgia”, and synonyms for the second sense would include “concern”, “worry”, and “vexation”. Thus, the term entry for “headache” would map to two different Synsets, and each of these Synsets would map two different, unique concepts in the Semantic layer.

The unique concept for the first sense of “headache” (Fig. 5) describes the “pain in the head” as belonging to the more general class of aches (toothache, backache, stomachache, earache . . .), and having more specific kinds of headaches (migraine headache, tension headache, sinus headache, cluster headache). Properties of headaches can also be described such as intensity (steady, mild, moderate, severe), occurrence (chronic vs. rare) or location (entire head, left side, right side).

Using the database to describe a simple term such as “headache” has several advantages. For one, the use of the term observed in the field can be put into context of other

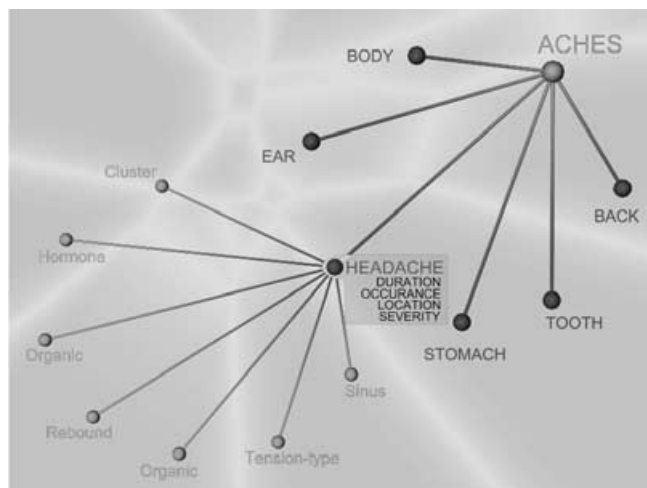


Figure 5. Glossary of medical terms (headache).

usage standards. This can both help to clarify the use of the term locally, as well as place the use into context of existing usage. Furthermore, the term is mapped to medical concepts and beyond that to other domains in the database. Ultimately the term “headache” can be cross-referenced to specific ethnopharmacological classes or phytochemical properties/actions of a plant found in the pharmacology section of the database.

In addition to WordNet, other linguistic sources can be incorporated into the data model, such as the Medical Subject Headings (MeSH) developed by the National Library of Medicine (Medical Subject Headings, 2000). The MeSH thesaurus is a carefully constructed sets of terms often connected by “broader-than”, “narrower-than”, and “related” links. Thus, it maps relatively easily onto our data modeling language. These links show the relationship between related terms or subject headings and provide both an alphabetic and a hierarchical structure that permits searching at various levels of specificity from broad to narrow. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. The MeSH vocabulary is continually updated by subject specialists in various areas. Each year hundreds of new concepts are added and thousands of modifications are made. 2000 MeSH includes more than 19,000 main headings, 110,000 Supplementary Concept Records (formerly Supplementary Chemical Records), and an entry vocabulary of over 300,000 terms.

Systematics – reflecting current flowering plant phylogeny

A similar ambiguity problem arises in mapping names of the taxonomic groupings of species into genera, families, and orders. The taxonomic structure of the object database inherently lends itself well to this task. Additionally, the data model needs to be globally current and reflect the current

classification or phylogeny of flowering plants. Flowering plant classification systems from the early 1980s such as those by Cronquist (1981) and Takhtajan (1980), although still in frequent use, have become outdated as new kinds of molecular data and new methods of analyzing conventional data have become firmly established (Stevens, 1986). Existing online medicinal plant databases such as the Worldwide Ethnobotany Database (Beckstrom-Sternberg et al., 1994) and the Native American Ethnobotany Database (Native American Ethnobotany Database, 2000) do not support changes in current angiosperm phylogeny. These databases were not designed to serve as dynamic resources and therefore have remained static data models unable to reflect taxonomic changes. Our database addresses this problem by providing a dynamic model reflecting the current phylogeny of flowering plants.

Two issues are of particular interest. The first is the problem of handling synonyms and homonyms occurring in both scientific and common names used for naming plants (Bisby, 2000). A number of synonymic relationships are used by systematists, including homotropic and heterotropic synonyms, pro parte synonyms, and orthographic variants of names. The linguistic component of the database handles this directly by mapping synonymous terms to the same database class, and using multiple word senses to describe a homonymic term. The second problem is inter-operability among different taxonomic classification systems. This is addressed through the use of “association” links between different taxonomies. Figure 6 illustrates associations between two systems where certain classes can be merged, split, added, deleted, or remain uncertain. The association between a class in the taxonomy of one system (Cronquist) is mapped to zero or more classes in the second system (Angiosperm Phylogeny Group). This technique can also be used to map dynamic changes between old and new versions of the same taxonomy.

Although our data model was designed to adhere to the revised suprafamilial classification of 462 flowering plant families and forty orders by the Angiosperm Phylogeny Group (APG), it can be integrated with emerging authoritative taxonomic systems such as the International Plant Name Index (INPI). Therefore, as new scientific names, or amendments to existing ordinal or familial relationships become available through the published literature or automatically updated through integration with INPI, these changes can be reflected by the database. There would be little need to manually update the database to reflect the changed circumscription of the orders recognized by the APG except for the inclusion of yet unassigned families of unknown systematic position and transfer of misplaced families. Additional recognized orders could be added as the phylogenetic relationships of families that are not yet placed are clarified. In the current APG classification, many families are listed without assignment to orders; these families are known to belong within major groups under which they are listed but their ordinal position remains uncertain. In addition, the database

can leverage existing taxonomic classification systems, including Cronquist’s and Takhtajan’s, by cross-referencing taxonomic synonyms against the AGP classification (Fig. 6).

Modeling ecological life zones

One of the primary requirements for the data model is to enable ethnobotanists to record plant biodiversity using current classification systems such as ecological climate-vegetation models. As an example, we incorporate the well-established and widely used Holdridge Life Zone classification (Holdridge et al., 1967) developed in 1947 by using object classes to describe ecological biomes, which are characterized by climate and other properties. The life zones are defined first according to a climatic variable – degrees mean annual biotemperature (and not according to degrees latitude or meters of elevation). The broad climatically defined life zones are further subdivided into associations on the basis of local environmental conditions and actual vegetation cover or land use.

The Holdridge life-zone system provides a logical basis for defining local ecosystems in a globally comparable framework. In this classification scheme, all terrestrial ecosystems can be uniquely defined in terms of three climate parameters for which data are widely available. Holdridge was able to define boundaries between these major vegetation units according to 1) logarithmic increases in mean annual biotemperature, 2) logarithmic increases in total annual precipitation, and 3) the ratio of mean annual potential evapotranspiration to mean total annual precipitation.

In the data model (Fig. 7), Holdridge life zones are represented using Classes. The three climate parameters are class properties. Each class has these three properties, but with different allowable ranges of definitive values. The classes and instances in the life zone classification system naturally form a taxonomy based on these three parameters.

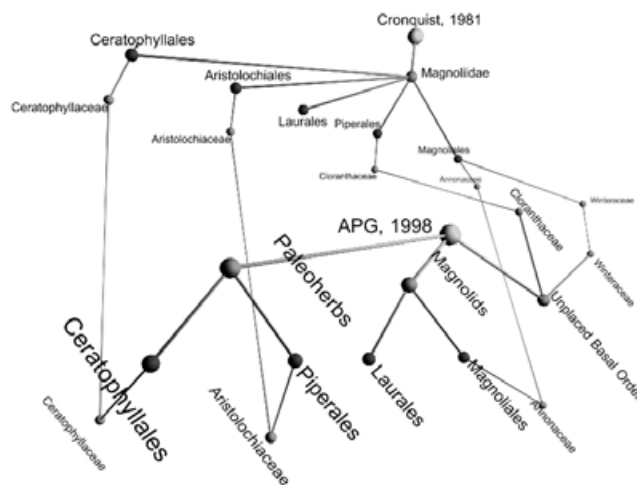


Figure 6. Phylogenetic relationships of Cronquist (1981) and the Angiosperm Phylogeny Group classification system (1999).

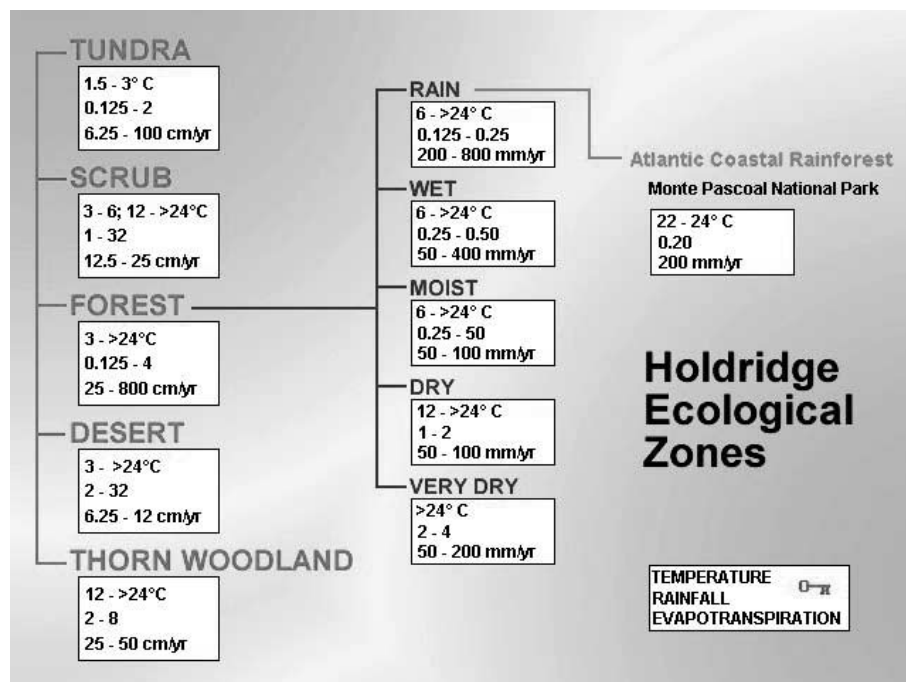


Figure 7. Holdridge Ecological Life Zones.

The automatic classification feature of the database (McGuinness & P. F. Patel-Schneider, 1997) enables new classes and new instances to be inserted into the taxonomy; their position in the taxonomy can be determined automatically. The basic rule is, class A subsumes class B (A is above B in the hierarchy) if every instance of B is also an instance of A. This can be computed based on properties. Class A subsumes class B if the properties of B logically imply the properties of A. Thus, the taxonomy in Figure 7 can be built incrementally and automatically by inserting new life zone class descriptions one at a time. For example, “Forest” subsumes “Rainforest” since the rainfall, biotemperature, and evapotranspiration for “Rainforest” fall within the ranges required for “Forest” (and furthermore, every rainforest is also a forest).

In this example, an Instance would be a particular habitat, such as the Monte Pascoal National Park in the Atlantic Coastal Rainforest region of Brazil. Since an instance would have a particular combination of values for the three climate parameters, it can also be automatically classified within the taxonomy. Its correct position below “Moist Forest” is shown in Figure 7.

This example demonstrates incorporation of an ecosystem classification system within the database. Even though the three climate parameters are sufficient for classification purposes, additional properties associated with ecological biomes can also be incorporated. For example, edaphic factors or associated flowering plant families, such as Apocynaceae and Rubiaceae (which may be found in association within a specific Holdridge life-zone system) can also be included.

Automatic classification is one of the tools that utilizes inference over object structure. Such tools can assist in structuring the database automatically, or assist users in locating information. Inference is based on a graph-matching technique that compares the structure of two objects to see how they are similar or different. Another use of this technique is in query processing (Beck et al., 1989), where the user’s query is stated as an object which is then matched against objects in the database to find exact and also approximate matches.

Uniting pharmacology with traditional knowledge

During the last decade, great efforts have been made to archive ethnomedicinal knowledge and to utilize this data in the search for novel plant-derived pharmaceuticals. Two different approaches, random and targeted, were utilized and the targeted selection programs were generally of three types. In phylogenetic surveys, the close relatives of plants known to produce useful compounds are collected. In ecological surveys, plants that live in particular habitats or have certain characteristics are selected. And in ethnobotanical investigations, plants used by indigenous peoples in traditional medicine are collected for analysis. Although ethnobotanical approaches to drug discovery are of significant historical importance, many pharmaceutical companies now believe that new approaches, incorporating molecular biology and combinatorial chemistry, supersede traditional knowledge as a potential source of new pharmaceuticals. As a result of this often deeply rooted divide, the concept of developing comparative data models has not been explored. Consequently,

random and targeted pharmacological data have remained distributed.

Currently, no data model incorporates all three primary methods of targeted selection including phylogenetic, ecological, and ethnobotanical sampling approaches. Our data model, however, does provide for such a comparative mechanism. By using object descriptions of plant species, genera, or family and including information about the plant's chemical properties, that plant's ecology, chemical description, and relationships can be handled. Comparative associations among related chemical structures can be represented including multimedia three-dimensional virtual models. It is clear that there is considerable collaborative potential in developing such a comparative model since less than $\frac{1}{2}$ to 1% of the species of flowering plants have been exhaustively studied to determine their chemical composition and medical potential (Balick & Cox, 1996). Data derived from the combination of all three of these targeted sampling techniques might approach an analogy to human bioassay data, and provide ever-increasing unified evidence of efficacy or acute toxicity.

One of the principal requirements for the data model is to permit scientists to record ethnomedicinal data using current classification systems derived from a diverse range of related disciplines including Economic Botany, Ecology, Plant Systematics, and Pharmacology to name just a few. In the data model (Fig. 8), these four disciplines and their related taxonomies are represented using superclasses and subclasses. In this example, the Instance is a particular voucher specimen of a medicinal plant native to Brazil, *Pilocarpus jaborandi* Holmes. The instance has a particular combination of parameters (properties), which could be automatically classified within the taxonomy of the related disciplines. For example, the specimens correct habitat position below "Moist Rain Forest" is shown in the left Ecology section. Its correct taxonomic position in the order Sapindales, family Rutaceae is shown in the upper Plant Systematics section. Its use as a medicinal is identified to the right in the Economic Botany section. Finally, its use as a treatment in Ear, Eye, Nose and Throat therapeutic category more specifically for treating glaucoma is shown in the lower Pharmacology section. This example demonstrates the collaborative approach and ability of the database to demonstrate a comparative analysis of various classification systems within the database. It should be noted that the database is designed to be extensible meaning the database supports the addition of other taxonomies such as linguistics or specific indigenous taxonomic classification systems.

Contributions mechanism

The database described above includes a contribution mechanism for researchers to submit and retrieve ethnomedicinal data and other associated data to an object database on a server. Data can be submitted by one of two methods, either directly online via a Java applet or through a downloadable Java application (Figs. 4 and 9). Ethnomedicinal data, which

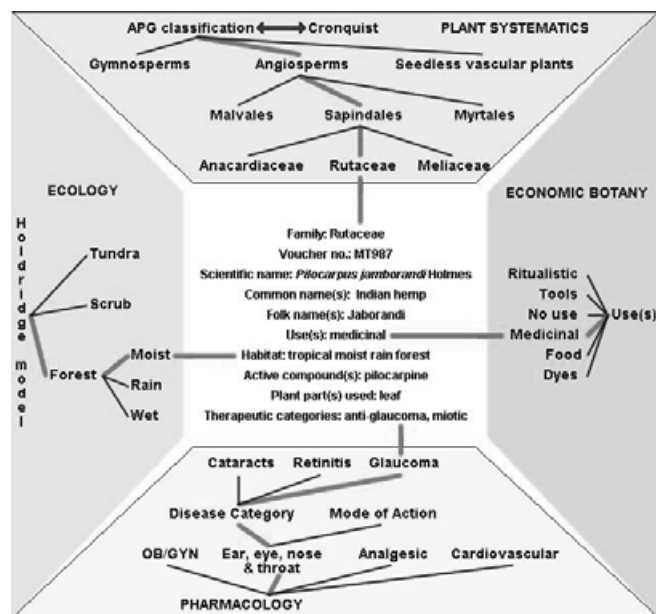


Figure 8. Medicinal plant integrated with multiple domains, including pharmacology.

can include text, figures, tables, and multimedia objects such as photographs, video, or sound, can be stored directly into the object database via the Internet. The system utilizes a standardized methodology and glossary of terms for archiving data. As a consequence, the data model effectively creates a single archived digital library of multimedia objects that can be used in comparative analysis, educational training, and data mining.

To summarize, the following layers exist in the software architecture proposed with our model (Fig. 10). The top layer is the one most visible to the users, and provides graphic displays (as illustrated in several of the figures of this paper) and other devices for interacting with the database. The underlying data model, as described in this paper, acts as the language for structuring database information, but is largely hidden from most users. The object database management system provides a storage layer where physical data are stored and manipulated. Finally, the bottom layer is an interface for exchanging data with other systems. CORBA (Common Object Request Broker Architecture) and RMI (Remote Method Invocation) provide high-level interfaces for direct program-to-program exchange of objects (Beck & Xin, 1998). XML (2000) is a convenient standard for non-database exchange of complex structured data.

Current databases and internet Resources for ethnobotany

Currently, the infrastructure for collaborative interdisciplinary scientific research through the Internet and Web is growing and becoming rich with power and promise (Lucky, 2000). However, many Web pages delivering medicinal plant

The screenshot shows a Netscape browser window with the address bar displaying 'http://geirs.ifas.ufl.edu/dataentry/'. The main content area contains a data entry form with the following sections:

- User:** DataEntry | Search
- Collaborator:** Taxonomic | Collections | Multimedia | Mgmt | MedicalUses | Literature Referenced
- Herbarium Voucher:**
 - Collection no: MT543
 - Collection date: 08 | 15 | 1998
 - Collector: M. B. Thomas
 - Collected with: A. Salvador
 - Determined by: M. B. Thomas
 - Location description: Along footpath leading to Posto. Open field.
 - Site name: Pataxo Indigena Reserva
 - Locality/Municipality: Porto Seguro
 - State: Bahia
 - Country: Brazil
 - Altitude: 20 m absl
 - GPS: 16o51'S 39o09oW
 - Form: tree
 - Description: Spreading evergreen perennial tree ca. 3-5 to 12 m tall
 - Reproduction status: flowering
 - Deposited at: CEPEC, NYBG, FLAS
- Bioassay Collection:**
 - Sample collected: None (dropdown) Type: []
 - Collection no. []
 - Sample date [] [] []
 - Amount []
 - Deposited at []
 - Notes: No collection made.

Figure 9. Java data entry display.

Data Visualization and Data Entry
Data Model
Object Database (ODBMS)
CORBA/RMI/XML

Figure 10. Software architecture.

information contain questionable data and research material and often do not include the detailed references necessary for sound scientific research. Even though scientists are, as a group, comfortable with computers and networks, they generally have been slow to provide online access to their research. Nonetheless, with Internet traffic doubling every six months, and wireless capacity doubling every nine months, the value of using current information technology for both storing and retrieving biological research data for a variety of purposes has become increasingly apparent.

Ethnomedicinal information is increasingly being recorded in databases and delivered via the Internet. However, the term "database" has been used broadly to

describe many kinds of collections. These databases vary from facsimile replications of Gerard's Herbal (Gerard, 1633) to Herculean efforts such as HerbWeb.com (HerbWeb, 2000) developed by compiling published medicinal plant research data using conventional database technology. Existing databases delivered via the Internet can be roughly categorized into (1) facsimile reproductions of printed materials, (2) a listing of medicinal plants and some of their properties as HTML-based web pages, similar to Ethan Russo's Plants of the Machiguenga website (Plants of the Machiguenga, 2000), (3) downloadable spreadsheets, (4) downloadable relational database tables often using Microsoft Access or Filemaker Pro files, such as Roy Ellen's downloadable Brunei Dusun database (Roy Ellen's Brunei Project – Databases, 2000), (5) on-line relational databases that can be searched by a Web interface, such as the Native American Ethnobotany database by Moreman (Beckstrom-Sternberg et al., 1995) and the Worldwide Ethnobotany database developed by James A. Duke and Stephen M. Beckstrom-Sternberg (Beckstrom-Sternberg et al., 1994), and (6) object databases such as the one described in this paper.

Additionally, the notion of web scientific portals has

emerged and is becoming popular. These are single web sites that serve as entrances to integrate material in a particular field similar to popular portals such as Yahoo or Excite. A good example of this concept is EcoPort (EcoPort, 2000), which aims to be a one-stop encyclopedia of information on every known plant and animal. EcoPort was developed to bring together information that is currently scattered across the Internet on individual researcher or university Web sites and stores all its data in one database that is accessed through one site. Similarly, NAPRALERT (Gyllenhaal et al., 1993) compiles information on natural products (chemistry, medicinal folklore, and biological activities) from many sources. Both EcoPort and NAPRALERT use centralized relational databases.

In contrast to these centralized approaches, meta-database search engines such as The Gatherer (The Gatherer, 2000) or Species Analyst (The Species Analyst, 2000) search and gather data from other databases. These use various techniques, including full text search, Z39.50 common database interface standards (Z39.50, 2000), and XML (XML, 2000) for data exchange. They work either with unstructured text data or minimal models of the record structure of underlying databases that they search.

One of the oldest and largest international collections of plant and animal species information, Species 2000 (Bisby et al., 1993, Bisby, 2000, Species 2000), focuses on the problem of developing taxons for the world's collections of biodiversity information. Species 2000 also acts as a meta-database by pointing to other, locally maintained databases built within the Species 2000 framework. Collaborators worldwide integrate their information into Species 2000. The key to this collaboration consists of dividing global taxonomic classifications vertically so that individual projects don't conflict as much as they would across horizontal, flora-by-flora studies. The work in building an ethnobotanical database partially overlaps with Species 2000, and some sharing can occur, especially in the area of plant systematics. Unfortunately, the interdisciplinary nature of ethnobotanical research will make it difficult to apply some of the simplifying assumptions used successfully in Species 2000.

It has been recognized that the standardization of terms and a unified system to describe and record medicinal plant uses would be of enormous benefit to researchers, especially where exchanges of data sets are involved (Cook, 1995). One such standard is the Economic Botany Standard developed by the International Working Group on Taxonomic Databases for Plant Sciences (TDWG) (Economic Botany Data Collection Standard, 2000), which has been adopted as a standard by the International Union of Biological Sciences (IUBS) Taxonomic Databases Working Group. This standard provides a mechanism whereby uses of plants (in their cultural context) can be described, using standardized descriptors and terms, and attached to taxonomic data sets. However, attempts to utilize a standardized schema have not been widely implemented nor fully realized for several reasons. First, the Economic Botany Data Collection Standard is not

very flexible because it does not allow users to extend the range of terms beyond what is defined, and it also contains inconsistencies (TDWG Economic Botany Subgroup Report, 1999). Second, implementing the standard into a relational format is an awkward step (Economic Botany Data Collection Standard – Data Model, 2000). Finally, it was designed not as a single unifying database but rather as a tool for potential users to consult in development of independent economic botany databases. The TDWG Economic Botany Subgroup plans to evaluate the extent to which the standard meets the needs of users and what revisions and additional data standards would be useful.

In contrast to the Economic Botany Data Collection Standard, the data model we have developed is more dynamic and the glossary can be extended to include specific terms identified or suggested by individual researchers. Consequently, researchers have the ability to contribute their expertise and experience to improve the design of the data model. Yet it is possible to use standards such as the Economic Botany Standard to bootstrap a glossary of terms in a particular domain. The Economic Botany Standard can readily be incorporated into the Linguistic portion of the data model as a starting point for covering economic botany terminology. The taxonomic structure of the object data model easily supports the existing structure of this standard without having to worry about mapping into relational tables.

With ethnobotanical research data becoming increasingly distributed, it becomes more difficult to conduct collaborative analysis. The proposed data model and associated contribution mechanism provides researchers the ability to utilize shared data and to initiate more collaborative and comparative analysis. Currently, almost nothing is known about possible patterns of medicinal plant selection by humans across cultures, regions, or hemispheres (Moerman et al., 1999) because ethnobotanical knowledge is often inaccessible and no standard system of archiving data exists. Having the ability to recognize such patterns, supported by the proposed data model, may increase the efficiency of contemporary bioprospecting programs for useful natural products and at the same time provide increased support for documenting global patterns of human knowledge (Moerman et al., 1999).

Our approach uses a common data model and provides tools for ethnobotanical researchers to develop a global database. In this way it is similar to EcoPort (EcoPort, 2000) but with a more complex data model. The approach is logically centralized because a common data model is used. But it can be physically distributed, because it is possible to have many subsets of the database existing in different places. In contrast, many systems (The Species 2000, Species Analyst, 2000, The Gatherer, 2000) assume that researchers will develop their own local databases, and provide a standard interface to higher-level meta-databases so that their data can be searched globally. There are pros and cons to both approaches. A compromise is to use the object data model to describe the contents of existing databases (such as legacy

systems) and thus it can both as a central repository and a meta-database, but with a richer modeling language.

What is needed is a starting point for building these databases. The data model described here is one contribution to this process. But construction of a collaborative global database requires international standards that can only come from organizing societies within the ethnobotany community. Such standards should support existing local databases within a larger framework (a rigid, centralized standard should be avoided). The object database model provides a framework for achieving this goal. Unfortunately, most researchers will be slow to adopt the object model, and will tend to use established relational database software instead. This is unfortunate since the object data model provides a richer environment. Nevertheless, existing relational databases can be integrated within the object data model framework.

Conclusion

Equipped with new scientific tools from molecular biology, analytical chemistry, mechanical engineering, and medical anthropology, modern ethnobotanists are asking a broad array of new questions while shedding new light on older questions. However, the lack of a unified approach and standard data model has led to a paucity of comparative ethnobotanical studies – studies that not only examine different uses of the same type of plants in different cultures, but also compare the ways plants figure in different world views (Balick & Cox, 1996). Few papers published by leading ethnobotanical journals address comparative analysis of possible patterns of medicinal plant selection by humans across cultures, regions or hemispheres. Indeed the absence of readily accessible comparative sources of ethnomedicinal data has been recognized as a serious hindrance.

While the amalgamation of many different backgrounds and disciplines has enriched the field of ethnobotany, the lack of a clear consensus on such basic issues as disciplinary goals and archiving methodologies has hampered the development of a unified and collaborative approach (Balick & Cox, 1996). The described data model not only satisfies the need for a tool for archiving traditional knowledge but may also provide the foundation for a globally current data resource. A dynamic resource would encourage a unified approach by facilitating a greater opportunity for comparative ethnomedicinal research through direct contributions by members of the scientific research community.

In this paper, we have described a new data model for ethnobotanical information based on the object database. We have shown how it can be applied to several areas (medical terminology, systematics, ecosystem zones, and pharmacology) resulting in an integrated, cross-disciplinary database. We have developed tools to assist researchers in browsing and editing the database. Next steps would include expanding the data model to cover additional disciplines, and

accessing information from existing databases and resources either by importing data or creating meta-data descriptions of these resources. We suggest that an international standards committee would ultimately need to be formed within the ethnobotany community, as has been done successfully for other global databases, but the framework for a global ethnobotany database can started based on the model presented in this paper.

References

- Balick MJ, Cox PA (1996): *Plants, people, and culture: the science of ethnobotany*. (Scientific American Library Series, No. 60). New York, W.H. Freeman & CO.
- Barry DK (1996): *Object Database Handbook: How to Select, Implement, and Use*. Object-Oriented Databases. John Wiley & Sons.
- Beck HW, Gala SK, Navathe SB (1989): Classification as a query processing technique in the CANDIDE semantic data model. *Proc. Fifth International Conference on Data Engineering*. IEEE. Los Angeles, CA, pp. 572–581.
- Beck HW, Xin JN (1998): Using Java, CORBA, and ODBMS to develop agricultural databases. *Proc. 7th International Conference on Computers in Agriculture*. Orlando, Florida. ASAE. St. Joseph, MI.
- Beckstrom-Sternberg SM, Moerman DE, Duke JA (1995): The Medicinal Plants of Native America Database. <http://ars-genome.cornell.edu/cgi-bin/WebAce/webace?db=mpnadb>. (Data version June 1995).
- Beckstrom-Sternberg SM, Duke JA, Wain KK (1994): The Ethnobotany Database. <http://ars-genome.cornell.edu/cgi-bin/WebAce/webace?db=ethnobotdb>. (Data version July 1994).
- Bisby FA (2000): The quiet revolution: Biodiversity informatics and the internet. *Science* 289: 2309–2312.
- Bisby FA, Russel GF, Pankhurst RJ (eds.) (1993): *Designs for a Global Plant Species Information System*. Oxford, Oxford Science Publications.
- Cook FEM (1995): *Economic Botany Data Collection Standard*. Prepared for the International Working Group on Taxonomic Databases for Plant Sciences (TDWG). Kew, Royal Botanic Gardens pp. ■■–146.
- Cronquist A (1981): *An Integrated System of Classification of Flowering Plants*. New York, Columbia Univ. Press.
- Economic Botany Data Collection Standard. 2000. <http://www.rbgekew.org.uk/tdwguses/index.htm>.
- Economic Botany Data Collection Standard Data Model. 2000. <http://www.rbgekew.org.uk/tdwguses/datastruct.htm>.
- EcoPort. 2000. <http://www.ecoport.org>.
- Fellbaum C (ed.) (1998): *WordNet®: An Electronic Lexical Database*. Boston, MIT Press.
- Gerard J (1633): *The Herball, or, Generall Historie of Plantes*. London: A. Islip, J. Norton, and R. Whitakers.
- Gyllenhaal C, Quinn ML, Soejarto DD, Farnsworth NR (1993): NAPRALERT: Problems and achievements in the field of natural products. In: Bisby FA, Russell GF, Pankhurst RJ,

- eds., *Designs for a Global Plant Species Information System*.
- Herbweb. 2000. Global Botanical Exchange. <http://www.herbweb.com>.
- Holdridge LR (1967): Life zone ecology. Revised ed. San José, Costa Rica: Tropical Science Center. p. 206
- Lucky R (2000): The quickening of science communication. *Science* 289: 259–264.
- McGuinness DL, Patel-Schneider PF (1997): Usability Issues in Description Logic Systems. *Proceedings of the 1997 International Workshop on Description Logics*, Gif sur Yvette (Paris), France.
- Medical Subject Headings. 2000. <http://www.nlm.nih.gov/mesh/meshhome.html>.
- Moerman DE, Pemberton RW, Kiefer D, Berlin B (1999): A comparative analysis of five medicinal florals. *J Ethnobiol* 19: 49–67.
- Native American Ethnobotany Database. 2000. <http://www/umd.umich.edu/cgi-bin/herb>.
- Plants of the Machiguenga. 2000. <http://www/montana.com/manu>.
- Roy Ellen's Brunei Project – Databases. 2000. <http://lucy.ukc.ac.uk/brunei.html>.
- Species 2000. <http://www/sp2000.org>.
- Stevens PF (1986): Evolutionary classification in botany. 1960–1985. *J Arnold Arbor* 67: 313–339.
- Takhtajan A (1980): Outline of the classification of flowering plants (Magnoliophyta). *Bot Rev* 46: 225–359.
- TDWG Economic Botany Subgroup Report, October 1999. <http://www/tdwg.org/1999rep3.html>.
- The Gatherer. 2000. Plant Use Multiple Database Search Engine. <http://www/kippewagardens.com/cgi-bin/Gatherer.pl>.
- The Species Analyst. 2000. <http://habanero.nhm.ukans.edu>.
- XML. 2000. <http://www.w3.org/XML>.
- Z39.50. 2000. <http://lcweb.loc.gov/z3950/agency/>