Growing Vocabularies for Plant Identification and Scientific Learning

Jane Greenberg
University of North Carolina at Chapel Hill
Tel: +01 919 9628066
Fax:+01 919 9628071
janeg@ils.unc.edu

Bryan Heidorn
University of Illinois at Urbana Champaign
Tel: +01 217 2447792
Fax:+01 217 2443302
pheidorn@uiuc.edu

Stephen Seiberling and Alan S. Weakley
University of North Carolina at Chapel Hill
Tel: +01 919 9626931
Fax:+01 919 9626930
sseiber@email.unc.edu
weakley@unc.edu

Abstract:

This paper reports on **U-PLanT's** (**U**niversity of North Carolina **P**lant **Lan**guage **T**eam) vocabulary solutions, including **Project OpenKey** vocabulary developments. The paper explores the meaning of vocabulary; discusses plant keys, plant taxonomy, and descriptive vocabulary used for plant identification; introduces U-PLanT's research and development activities and current inquiry. Vocabulary solutions presented include a suite of vocabulary tools, a preliminary process model with steps for the development of vocabulary tools, and guiding principles for the development of descriptive plant vocabulary. This work has been conducted to address the student/scientist vocabulary gap and facilitate student access to primary scientific resources found in education digital initiatives.

Keywords:

Plant vocabulary, plant taxonomy, descriptive vocabulary, science education digital libraries, primary resources, metadata.

1. Introduction

Digital initiatives providing access to the world's rich reservoir of primary scientific resources are among the most exciting developments supported by World Wide Web (Web) technology (Greenberg, et al, 2002). Examples include botanical and zoological organizations and partnerships such as the San Diego Zoo,[1] Royal Botanical Gardens, Kew,[2] and the Spanish and Portuguese Platform for Botanical Diversity Data Online project.[3] Institutions and partnerships of national and international stature are digitizing *scientific specimens* and targeting students at all levels of learning (academic to life-long learners). They want to reach audiences beyond the seasoned scientists and enrich curricula via access to scientific specimens. Web connectivity and digitization alone are not sufficient for student access to these primary resources. This is because the student's knowledge-base and thus working vocabulary differs greatly from the scientist's professional vocabulary. Student vocabulary and the vocabulary of any non-specialist is generally a mix of common terminology and newly learned scientific terms, whereas scientists communicate with a fine grained (descriptively rich) vocabulary and index specimens by scientific names.

The student/scientist vocabulary gap must be bridged in order for students to take advantage of digital initiatives containing scientific specimens. Enabling technologies such as eXtensible Markup Language (XML), and languages such as the World Wide Web Consortium's Web Ontology Language

[1] www.sandiegozoo.org/animalbytes/);

[2] www.rbgkew.org.uk/collections/herbcol.html

[3] www.gbif.org/Stories/STORY1063920331/#Title:_Spanish_and_Portuguese_Platform

(OWL) (), provide a *structural foundation* for integrating multiple vocabularies with different levels of detail (e.g., common to scientific vocabulary). In the sciences, there are also discipline specific developments such as the Structure for Descriptive Data (SDD), an XML-based standard for structuring taxonomic data in the biological sciences.[4] These developments are important, although they do not address the *intellectual tasks* of developing and growing vocabulary for student access to scientific specimens.

Researchers engaged in educational digital projects containing scientific specimens need to share experiences and identify the intellectual steps required for developing and growing vocabulary tools that bridge the student/scientist vocabulary gap. The **U**niversity of North Carolina **P**lant **Lan**guage **T**eam (U-PLanT) has been addressing this need over the last few years through a series of specimen digitization projects aimed at teaching students plant identification (Example 1 is a digital specimen). **U-PLanT** has identified and enhanced vocabulary tools for teaching students scientific terminology. Through **Project OpenKey**,[5] U-PLanT and **University of Illinois at Urbana Champaign** (UIUC), have also begun to study the complexity of taxonomic vocabulary for plants and explore methods for integrating student and scientist vocabulary.

This paper reports on U-PLanT's vocabulary solution implementations and current results of Project OpenKey. The rest of this paper is organized as follows: Section 2 discusses the meaning of vocabulary—noting its importance for science educational digital initiatives; Section 3 focuses on plant keys, plant taxonomy, and descriptive vocabulary used for plant identification; Section 4 introduces U-PLanT's activities and presents research and development questions motivating U-PLantT's partnerships, including Project OpenKey; Section 5 shares U-PLanT's research methods; Section 6 presents U-PLanT's current vocabulary solutions; and Section 7 includes a conclusion and identifies three directions for future research.

2. Vocabulary Defined

Providing a precise definition for *vocabulary* (as a concept) is difficult due to the varied applications of this term in different contexts and disciplines. Vocabulary can be represented by *communication symbols* (letters, numbers, and other symbols),

Example 1: Example of a Digitized Plant Specimena

gestures, and *words*. Examples include a set of *communication symbols* representing dance annotation or programmatic commands for computer processing; *gestures*, primarily hand motions comprising the universal sign language or a football team's repertoire of strategic plays; and *words* underlying a specific discipline (e.g., biology) or in a transaction (e.g., business transaction). The most common interpretation of vocabulary includes *words* used to express *concepts* in a language (e.g., the *Spanish language*) or a vernacular (e.g., *American-English with a Southern dialect*). Through further study, we understand that:

- *Vocabulary is a key component of language; vocabulary can also be a language.*

Vocabulary is the *main ingredient* of language (e.g., words are the main feature of language). Controlled vocabularies, authority lists, and classificatory systems are often referred to as *indexing languages*; these systems function as languages due to strict rules controlling name and word use. More recently, descriptive, structural, administrative, and

[4] 160.45.63.11/Projects/TDWG-SDD/ 5 http://www.isrl.uiuc.edu/~OpenKey/, and http://www.ibiblio.org/openkey/

other types of metadata standards demonstrate vocabularies as languages. Baker (2000) eloquently articulates this point when explaining that the Dublin Core metadata schema is akin to *pidgin*—a language comprised of simple patterned statements for tourists.

- *Vocabulary is a means of communication—an ontology of sorts*

Vocabulary allows people to communicate and understand each other. Information-oriented disciplines with vested interest in the Web (e.g., library and information science and computer science) work daily with vocabularies, such as subject thesauri and metadata schemes supporting *person—to—machine*, *machine—to—person* and *machine—to—machine* communication. Vocabularies in this venue are community agreements and increasingly defined as *ontologies* (Jacob, 2003). Ontologies present a simplified view of a world (or domain) by documenting concepts and the relationship among concepts.

- *Vocabulary is fundamental to learning and scientific advancement.*

Researchers (e.g. Stahl, 1999) have demonstrated a strong correlation between reading comprehension and vocabulary knowledge. There are obviously intelligent people who gain knowledge through life experiences rather than reading and vocabulary development. For example, a farmer who knows how to interpret weather patterns is deemed intelligent regardless of her reading ability. Nevertheless, society dictates that one generally needs to understand domain vocabulary to grasp the *status quo* and advance discipline knowledge. This mode of being explains why scientists develop discipline specific language and constructs to advance knowledge (Somerville, 1998). Similarly, students studying a subject need to learn discipline language in order to further their knowledge. The relationship between learning discipline vocabulary and advancing knowledge is fundamental to the research and development activities presented in this paper.

- *Vocabulary is integral to the development and functionality of digital libraries.*[6]

Digital libraries, like traditional physical libraries, depend on vocabulary to facilitate basic organization and information retrieval operations. Advantages of controlled vocabulary are outlined in Rowley (1994). Examples include greater recall during an information retrieval operation, less searching burden for the user, and better collocation (the bringing together of related resources). Research on controlled vocabulary in information systems is reported on in Svenonius (1986). Known advantages of controlled vocabulary extend to name heading authority files, gazetteers, scientific taxonomies, and other standardized vocabulary systems; and any of these vocabularies may be implemented in a digital initiative, including plant keys used to facilitate plant identification.

3. Plant Identification: Methods and Vocabulary Needs

Plants are among the most advanced living organisms on earth, with a long evolutionary history dating back hundreds of millions of years. Plant identification is a common part of school curricula because of the historical and present significance of plants and their critical role in harnessing the sun's energy and making it available to support animal life. Applications known as *plant keys*, *plant taxonomy*, and *descriptive vocabulary* are all integral to plant identification.

3.1 Plant keys

The goal of plant identification is to arrive at the scientific name—that is the species taxon (*plant taxonomy* is discussed in section 3.2.1 below). Plant identification for students and amateur botanists is generally supported by a tool known as a plant key. There are two primary types of plant keys, a *fixed structure key* and a *polyclave key.*

With a fixed structure key, the species is identified by answering an ordered series of questions, typically in the form of "either/or" (dichotomous) pairs, one-at-a-time. For example: Are the leaves/buds on the tree branch *alternate* or *opposite* in their arrangement, and are they *toothed edged* or *smooth*? An important disadvantage of fixed structure keys is their inflexibility, requiring the user to answer predefined questions in a fixed order. When the user does not possess the particular information needed at each step in the key, the process must either be abandoned or demands following multiple pathways, a tedious, often frustrating, process. U-PLanT activities have been connected with the production of several fixed structure keys, such as the *Key to the Gymnosperms of the Southeastern U.S.*[7] (gymnosperms typically have leaves in the form of *needles* or *scales*, and include pines, hemlocks, and junipers—to name a few).

A polyclave key works by presenting the user the

[6] The term digital libraries is used broadly here to represent any type of repository with digital resources.

choice of many plant characters (*plant characters* are discussed in section 3.2.2 below) which may be selected in any order, and added to in an iterative process. The species name is arrived at by matching the character states selected with information contained in a database, and eliminating the species for which there are mismatches. Standalone polyclave applications include DELTA and Lucid. Web technology is ideal for interaction and the development polyclave keys to support plant identification, despite noted vocabulary and architectural challenges required for designing such system. For this reason, most polyclave plant keys are still in a nascent state. As part of Project OpenKey, U-PLanT has produced the *Common Trees of the North Carolina Piedmont* polyclave,[8] and UIUC has been working on Biological Information Browsing Environment (BIBE) (Heidorn, 2001), a polyclave key for the identification of both flora and fauna.

3.2 Plant Taxonomy and Descriptive Vocabulary

Plant identification involves two vocabularies: *plant taxonomy*, the scientific names for known plants;[9] and *descriptive vocabulary*, the terminology used to describe the characters of a plant to facilitate identification These two vocabulary systems are an essential component of digital plant keys and they are discussed in the next two sections of this paper.

3.2.1 Plant taxonomy

The object of plant identification is to determine the plant's scientific name—that is the appropriate species taxon, and, at times, all the other taxa levels in the process. Plant classification can be traced back more than 2,000 years to the Greek philosopher Aristotle (384-322 BC), who classified plants by characteristics such as shape and stem and by habitat. Aristotle's classification activities produced a *vocabulary* for studying and advancing knowledge about plants.

More familiar today is the hierarchical system of classification for living organisms developed in the mid-1700s by the Swedish naturalist Carl Von Linné, better known by Carolus Linnaeus—the Latinized version of his name. Linnaeus' system emphasized plant *morphology* basing his system on the organism's form and structure. He established a system of taxa (*pl*. of taxon) for grouping related organisms. Modern plant taxonomy has enhanced Linnaeus' original taxonomic hierarchy of five to seven top levels (Table 1, next page).

There are rules for constructing a plant's scientific name, as there are often rules for establishing any

Table 1: Plant Taxonomy

Taxonomic Level	Example for a Field Rose
Kingdom	Eucaryota
Phylum	Spermatophyta
Subphylum*	Magnoliophytina (Angiospermea)
Class	Magnoliopsida
Subclass*	Rosidae
Order	Rosales
Family	Rosacaeae
Genus	*Rosa*
Species	*Rosa avensis*

***Subphylum and subclass are subgroups.**

name (e.g., a person's full name is generated by combining their first name, middle name, and surname, and any titles in a specific order). Plant taxonomy combines the genus and specific epithet. For example, the scientific name for the plant commonly known as *white pine* is *Pinus strobus*, and is comprised of the genus "Pinus" and specific epithet "strobus." This system is

now encoded in the *International Code of Botanical Nomenclature* (the current iteration of which is also known as the St. Louis Code)[10] adopted by the Sixteenth International Botanical Congress St Louis, Missouri, July-August 1999. It will be replaced by the (Vienna CODE) after the International Botanical Congress in Vienna in 2005. In an information system, the scientific name can provide a key to the accumulated published knowledge about the species.

3.2.2 Descriptive vocabulary for plant identification

Plant identification requires descriptive vocabulary—a collection of terms that provide interpretative information representing *plant characters*, *character states*, and *character groups*.

A *plant character* is a property or attribute that can be observed, measured, counted or examined in some fashion. Examples include *growth habit*, *leaf shape*, *leaf margin* (the edge of the leaf), *leaf duration*, *stem type*, and *stem pith*. Davis and Heywood (1973), authors of *Principles of Angiosperm Taxonomy* , define plant characters as "abstract entities" that identify the "form, structure, or behavior of an organism for a particular purpose such as comparison or interpretation."

---

[7] www.ibiblio.org/pic/GymnospKey/index.html

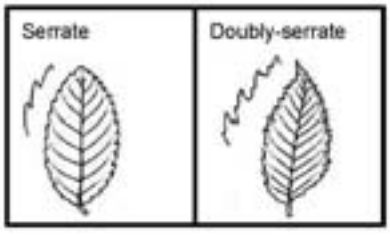
[8] www.ibiblio.org/openkey/intkey/index.htm

[9] Estimates of the total number of vascular plant species in the world today range from 250,000 to 500,000.

Table 2: Leaf Shape and Leaf Margin Character States

***Leaf shape* character states**	
acicular	obcorodate
awl-shaped	obdeltoid
cordate	oblanceolate
deltoid	oblong
elliptic	obovate
enisoform	orbiculate
falcate	oval
fan-shaped	ovate
filiform	pandurate
hastate	reniform
lanceolate	rhombic
linear	runcinate
lorate	sagittate
lyrate	scale-like
needle-like	spatulate
***Leaf margin* character states**	
ciliate	doubly serrate
crenate	entire
crenulate	erose
crisped	incised
deeply lobed	involute
denate	lacerate
denticulate	lacinate
divided	serrate

Example 2: Serrated Leaf Margin Examples



Character states are expressions or states of plant characters that help with plant identification: they are the possible values for plant characters. Examples of character states for the character's *leaf shape* and *leaf margin* are presented in Table 2. Theses character states document the fine grained terminology that botanists use to describe plants. Student terminology is more simplistic. A student, unaware of this detailed (granular) vocabulary, may simply describe leaf margin as being "smooth edged" or "toothed," and is likely unaware that there are different levels of serration as illustrated in Example 2.

[10] www.bgbm.fuberlin.de/iapt/nomenclature/code/SaintLouis/0001ICSLContents.htm)

It is difficult to teach students plant taxonomy without their understanding the fine distinctions among specimens that are documented in descriptive vocabulary. These vocabulary challenges have been central to U-PLanT's work over the last few years, driving the solutions given in this paper.

The most general category for descriptive vocabulary is *character groups*, which are artificial, abstract yet logical grouping of plant characters. For example, characters about *leaves*, *stems*, and *buds* may be grouped under the character group (a concept) *non-reproductive morphology*, and characters about *petals*, *sepals*, and *anthers* may be grouped under the character group of *flower*. *Flower petals* may have a number of characters as well, such as *petal color*, *petal texture,* and so forth. Character groups are defined only for logical grouping of characters for organization purposes.

4. U-PLanT: Research and Development

The U-PLanT (University of North Carolina Plant Language Team) is a partnership of UNC's Herbarium, UNC's School of Information and Library Science, and the North Carolina Botanical Garden. Members include biologists, vocabulary and metadata experts, and educators. The partnership has also recently extended to include the Graduate School of Library and Information Science, University of Illinois Urbana-Champaign. The U-PLanT team was formed to address vocabulary issues underlying educational use of digital specimens for plant identification. The three key projects catalyzing U-PLanT's formation are:

- BOTNET, a digital herbarium for plant specimens. (www.ibiblio.org/botnet/flora/indexstart.html).
- Plant Information Center (PIC) (www.ibiblio.org/pic), a digital learning center connecting students and the general public to primary research materials.
- Project OpenKey (www.ibiblio.org/openkey), a collaboration between UNC and UIUC providing access to botanical resources through polyclave plant keys that visually capture the way botanical experts identify species.

These three projects have focused on plant identification, specifically trees in the State of North Carolina; OpenKey, which was initiated at UIUC, has also focused on the identification of Prairie Plants in the State of Illinois. Primary research and development questions underlying these projects have been:

What tools can help students learn botanical

terminology?

What steps aid vocabulary development for plant keys, or other digital initiatives supporting plant identification?

What principles guide the development of descriptive vocabulary for plant identification?

5. Research Methods and Inquiry

A number of research projects, involving a combination of research methods, have been conducted during BOTNET, PIC, and OpenKey. Although several of the research projects addressed vocabulary issues more intimately than others, all the research has helped U-PLanT to better understand vocabulary issues central to specimen identification and for science education digital initiatives containing primary resources. Research methods and activities have included the following:

- Informal expert interviews and vocabulary tool analyses were conducted during the BOTNET project in order to identify vocabulary tools that might help students learn discipline-specific and scientific vocabulary.
- A pilot project BotNetDC (Botanical Information Network Dublin Core) (Buch, et al, 1999) was conducted testing the application of Dublin Core for plant description (e.g., specimen family, genus, and species descriptions). BotNetDC was extended to BotDC (Botanical-Dublin Core), which included the development of an XML DTD and a schema specification for describing botanical resources. Both projects identified plant description controlled vocabularies requirements.
- Three PIC usability studies (Dopke & Carlson, 2001; Hall, 200111) were conduced. The usability studies gathered data on the usefulness of PIC's plant database, linked vocabulary tools (see Solution 1 below in Section 6.1), and the overall Website.
- Two experiments were conducted with PIC Advisory Board members. One experiment was conducted at each annual Advisory Board meeting. The first experiment studied the topical classification of PIC's Website resources, and the second experiment studied metadata creation for personal botanical images.
- An analysis of botany-related frequently asked questions (FAQ) was conducted by Williams (2000). Williams' research resulted in the development of a FAQ taxonomy (reported in Greenberg, 2001), and informed the development of a FAQ module for the PIC Website (Warmoth, 2002). This project captured vocabulary that students and the general public use to ask botanical queries.
- Two metadata creation studies were conducted (Harmes, 2001; Hanrath, 2002). Metadata creators included students, life-long learners, and amateur botanists. These studies provide data on how students' might use vocabulary when searching for specimens.[12]
- Preliminary quantitative analysis of the character groups, characters, and character states has been conducted throughout the OpenKey project to get a sense of the extent of the vocabulary needed for description of both North Carolina Trees and Illinois Prairie Plants.

6. Development Solutions

U-PLanT has implemented three solutions to help address the vocabulary challenges discussed in the first parts of this paper. The solutions have been informed by research reviewed directly above and by team members having expertise in the areas of biology, vocabulary and metadata, and education. The three development solutions include: 1.) a suite of vocabulary tools, 2.) a preliminary process model outlining vocabulary development steps, and 3.) the identification of guiding principles for vocabulary development.

6.1 Solution 1: Develop a Suite of Vocabulary Tools

U-PLanT's efforts have led to the development of five vocabulary tools, all of which have been implemented and integrated into the three digital plant projects (BOTNET, PIC, and OpenKey). Vocabulary terms and definitions presented in these tools have been obtained from close to twenty bibliographic sources and from U-PLanT members with expertise in botany. Bibliographic citations are linked to terms in cases where a definition is taken verbatim or almost exactly from a published source. All tools in the U-PLanT vocabulary suite are available via the Web. Development of these vocabulary tools has been incremental. That is, each vocabulary tool is successive and borrows from the preceding tools. Each tool is described below in consecutive order of development.

1. **Technical Plant Glossary** (www.ibiblio.org/botnet/glossary/index.html) is a highly technical plant glossary linking to a digital version of *Vascular Plant Systematics* (Radford, Dickison, Massey & Bell, 1976). This

[11] Results from the third usability study have not been published.

is a comprehensive text of plant morphology terms. The terms are not organized in one complete list, but rather are grouped according to specific plant parts (stem, leaves, root, etc.). U-PLanT does not have the ability to edit or add terms to this glossary, but has borrowed from it for additional tools in the U-PLanT suite.

2. **Student Botanical Dictionary** (www.ibiblio.org/pic/student_glossary.htm) is a glossary that includes 157 terms fundamental to studying botany. The dictionary was designed to include basic botany terms that are useful for students or the general public with little or no botanical knowledge. Each dictionary entry includes a term followed by a definition. Example 3 shows entries for the terms "botany" and "pollen." The initials "GT" following the definition for "botany" links to the citation for this source, which is presented underneath in this example. All bibliographic references are hyperlinked and appear on the last page of the online *Student Dictionary*.

Example 3: Student Botanical Dictionary Term Entries and Bibliographic Citation

Botany: The scientific study of plants.

- Pollen: Minute grains, usually powdery, containing the male sex cells of gymnosperms and flowering plants. [GT]

*[GT] noted the bibliographic source: Greenway, Theresa. (2000). The Plant Kingdom: A Guide to Plant Classification and Biodiversity. Texas: Steck-Vaughn Company.

3. **Conceptual Table of Descriptive Vocabulary** () is a matrix of approximately 200 characters and over 2000+ character states for both North Carolina Trees and Illinois Prairie Plants. The conceptual table was developed for the OpenKey project. The vocabulary was structured for the development of an interoperable polyclave plant keys. The vocabulary in the conceptual table can be described as an ontology because the "plant groups" and "plant characters" act as nodes or classes, which are part of a *superclass*, and they have multiple means of expression via *subclasses* in the form of "characters" and "character states."
4. **Botanical Dictionary** (www.ibiblio.org/pic/botanical_dictionary.htm). This dictionary is a comprehensive list of vocabulary words for botany; it contains approximately 1,600 terms. Entries include a word followed by a definition; both the vocabulary and definitions are generally more technical than the *Student Dictionary*, although there is some overlap. This dictionary consists of the complete set of terms and definitions from *Vascular Plant Systematics* (Radford, Dickison, Massey & Bell, 1976) arranged in alphabetical order.
5. **UNC-OpenKey Glossary of Botanical Terms** (www.ibiblio.org/openkey/searchform.php) is a searchable glossary with 510 terms and definitions developed in conjunction with the *Conceptual Table*. This glossary defines vocabulary used in the *Common Trees of the North Carolina Piedmont*—UNC's polyclave produced for the OpenKey project. The glossary is stored in a MySQL database and uses a combination of X/HTML and PHP (PHP Hypertext Preprocessor) to support searching, retrieval, and presentation of terminological entries. Entries follow a standard format shown in Example 4. The *UNC-OpenKey Glossary* is also accessible as Microsoft WORD and Adobe PDF documents. This glossary provides a model for developing a UIUC—OpenKey glossary.

Example 4: Entry example from OpenKey Glossary of Botanical Terms

Term- [Botanical structure] {category of application} Definition. (Terms for comparison.) [source reference]

Acicular- [Leaflets, Leaves] {shape} Very long and slender, gradually tapering to a point, like a needle; needle-shaped. (Compare with awl-shaped, filiform and linear.) [modified from K&P, p. 11]

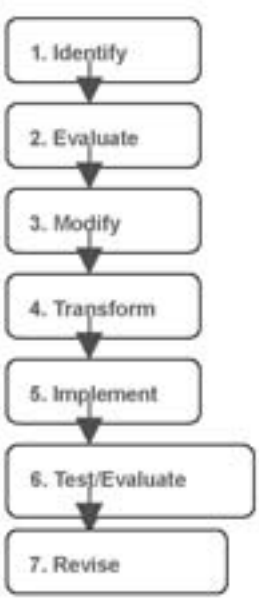
6.2 Solution 2: Preliminary Process Model

U-PLanT's vocabulary development activities have been extensive, and also expedient in an effort to complete project deliverables. U-PLanT has identified seven general steps underlying its vocabulary development, which form the basis of a *preliminary process model* (Example 5). These steps continue to guide the iterative development of U-PLanT vocabularies as digitization extends to new sets of taxa. These steps may also be useful to other digitization efforts focusing on student access to primary scientific resources and facing similar vocabulary challenges.

Descriptions of the seven steps in the preliminary process model follow below:

1. **Identify**. Identify any existing vocabularies (controlled vocabulary, name authority files,

Example 5: Process Model for Vocabulary Development



metadata schemas, ontologies, and so forth) that are useful to project goals. The digitization of the *Vascular Plant Systematics* (Radford, Dickison, Massey and Bell, 1976) for the *Technical Plant Glossary* is demonstrative of this first step. The goal is to work with existing tools, rather than expend often limited resources where work has already been done. The cliché here is *why reinvent the wheel*, and it is very applicable to vocabulary development.

2. **Evaluate**. Evaluate existing vocabularies. Use practical/economic measures to identify what *is useful* and *what is not useful* in existing vocabulary tools. Determine which vocabulary requires limited revision and will serve as an excellent source for project needs, and which vocabulary requires too much revision such that it is more efficient to build a new tool. In general, useful vocabulary sources are available and can provide a base vocabulary; even a noted monograph's index can be a starting point.
3. **Modify**. Enhance, extend, and delete vocabulary in existing tools to meet project needs. *Vascular Plant Systematics* (Radford, Dickison, Massey and Bell, 1976) provided the base vocabulary for U-PLanT's activities. It was identified (step 1), evaluated (step 2), and then modified (step 3) to *grow* vocabulary for all of UNC's plant digitization projects. Existing vocabulary tools will likely require modification to fit project goals, particularly when dealing with primary resources. This is because the availability of standardized discipline-specific vocabularies from primary resources is limited, having not yet been published. Even so, various vocabularies combined and modified contribute greatly to the growing vocabulary.
4. **Transform**. Make the vocabulary suitable for the access environment. *Vascular Plant Systematics*' vocabulary was first simply scanned and made accessible via HTML. Latter, as this vocabulary was modified. More sophisticated technological applications were applied due to new vocabulary needs and increased demand for access to vocabulary tools. The UNC—OpenKey Glossary, the latest vocabulary tool to be developed, is the most sophisticated, accessible through a MySQL database with a corresponding XML schema. An example of a plant description following the *Conceptual Table's* XML schema is found in Appendix A.
5. **Implement**. Make the vocabulary tool operational after transformation to the desirable format(s).
6. **Test/evaluate**. Once implemented, vocabulary functionality should be evaluated. The research and development activities outlined in Section 5 note a couple of usability studies that gathered data on the use of PIC's vocabulary tools during a plant identification exercise. More formal analyses are needed to assess the overall usefulness of these tools and inform the following step (step 7). Among evaluation methods that could be used are transaction log analyses, user surveys, and plant identification exercises that specifically measure the use of the existing vocabulary tools.
7. **Revise.** Vocabulary development is organic and revisions are required due to collection growth. Significant events, such as the discovery of a new species will also impact standard vocabularies. The process of growing vocabularies can also be circular. In such cases, the revision step (step 7) might require vocabulary developers to start by identifying new vocabulary sources (step 1) to fill system gaps, and then proceed through the steps outlined in this preliminary model to grow a vocabulary.

The preliminary process model outlined here can be further developed over time to help guide vocabulary development, which is an activity that is becoming essential to the development of educational science digital initiatives and other digital projects.

6.3 Solution 3: Guiding Principles for Descriptive Vocabulary Development

U-PLanT in collaboration with UIUC has

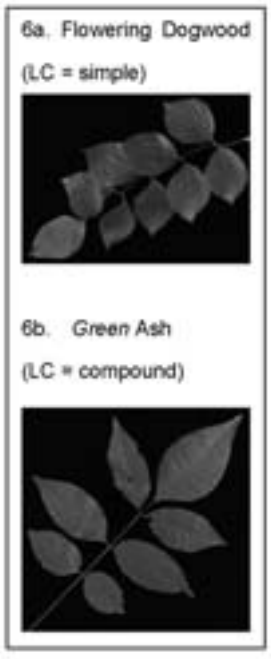
identified three principles that drive the creation of descriptive vocabulary recorded in the *Conceptual Table*. These include the need for *understandability*, *uniqueness* and *consistency*. These principles are important because of vocabulary challenges stemming from the need to support multiple user communities (students, life-long learners, etc.) and the need to support distributed development of vocabulary—in OpenKey's case distributed development is between UIUC and UNC.

In terms of *understandability*, OpenKey users need to "understand" or "comprehend" the terminology used to identify plant characters. Consider leaf complexity, which may be *simple* or *compound*. Each Flowering Dogwood *(Cornus florida)* leaf (Example 6a) comes from a separate bud, so the leaf complexity is simple. Example 6b is *a single* Green Ash (*Fraxinus pennsylvanica*) leaf that has developed from a single bud; it is considered compound because the specimen appears to have multiple leaves. The two character states of *simple* and *compound* are understandable to the botanists, but not so obvious to the amateur botanists. In folk taxonomy, they'd both be compound. Understandability—as a guiding principle, requires that all characters are coded using the most specific scientific terminology available. For the descriptions of trees and prairie plants there are several hundred characters used to describe each species. To make these understandable each character and each character state is tied to both a text definition and where possible to an image.

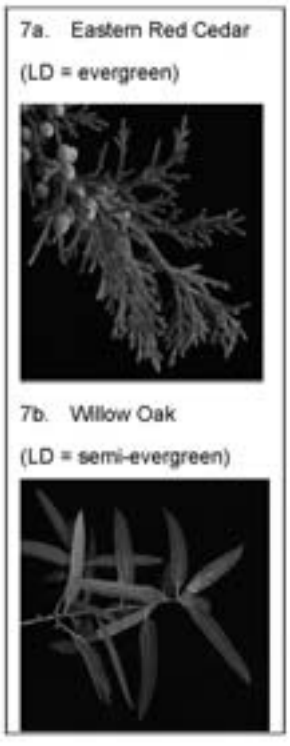
Each vocabulary entry must also be *unbiguous and discreet*. By unambiguous we mean that there can be only one meaning to any particular word. Discreet means that concepts should have well defined boundaries so that a particular item can belong to a category (character state) or not. Despite this measure, there are continuous character states, such as "height," which should be clearly coded. In some cases the plant may possess more than one character state value. For examples, the flowers may be either "red" or "blue" or both "red" and "blue." The definitions of "red" and "blue" should be clear so that these values are unambiguous and discreet. A final example discreetness is leaf duration (the length of time a leave lives upon a tree):

- *Deciduous plants such as the flowering* dogwood (Example 6a) have leaves that fall off the tree during the winter season.
- *Evergreen* plants such as the Eastern Red Cedar (*Chamaecyparis thyoides*) (Example 7a), have leaves tho not fall off the tree during the winter season regardless of weather and climatic region.
- *Semi-evergreen* plants such as the Willow Oak (*Quercus phellos*) (Example 7b) have leaves that fall off the tree during the winter season due to certain weather conditions and when this plant is found in certain climatic regions.

Example 6: Leaf Complexity (LC)



Example 7: Leaf Duration (LD)



The last guiding principle is that vocabulary must have *consistency*—that is the same term must have the same meaning each time it is used. U-PLanT and UIUC, together, have been able to accomplish this through a process of continuous refinement and negotiation. These practices are particularly important when new vocabulary items were added to any of the vocabulary tools, but specifically the *Conceptual Table*, which is a shared tool with UNC and UIUC, and will likely gain other partners in the near future. ust have the same meaning each time it is used. U-PLanT and UIUC, together, have been able to accomplish this through a process of continuous refinement and negotiation. These practices are particularly important when new vocabulary items were added to any of the vocabulary tools, but specifically the *Conceptual Table*, which is a shared tool with UNC and UIUC, and will likely gain other partners in the near future.

7. Conclusion and Future Research Directions

U-PLanT was formed to address the student/scientist vocabulary gap and to facilitate student access to digital plant specimens. U-PLanT's vocabulary activities are directly connected to UNC's BOTNET, PIC, and OpenKey projects, and the OpenKey project included UIUC as a partner. Vocabulary development is integral to functionality and ultimately the success of each of these projects. On a larger scale, vocabulary is essential to any educational digital initiative working with scientific specimens. This paper focuses on common vocabulary challenges in this area and presents a series of solutions.

The research and development activities presented in this paper examined the types of tools that can help students learn botanical terminology. Work has resulted in steps that can guide vocabulary development for plant keys, or other digital initiatives facilitating plant identification; and principles that can inform descriptive vocabulary development. The work conducted via BOTNET, PIC, and OpenKey has results in three solutions. These include a suite of vocabulary tools, a preliminary process model to guide the development and growth of plant vocabulary, and guiding principles to inform the development of descriptive terminology for plants as recorded in a conceptual table. Although the current vocabulary solutions are limited in that they are specific to UNC's efforts and their collaboration with UIUC, they may aid other science education digital initiatives facing wanting to provide access to primary specimens and facing similar vocabulary challenges.

Although the research and development activities presented in this paper address only a few selected vocabulary challenges, they are po imrtant to furthering UNC's and UIUC's goals to provide student access to primary scientific resources, to teach them plant identification, and ultimately engage them in the process of scientific discovery. Another goal is to use technology to teach students about and connect them to the natural world. We believe these experiences are invaluable and can greatly enrich science education.

Research and development activities shared in this paper provide a baseline that is useful for conducting more research on the vocabulary challenges connected to science education digital initiatives. We have identified several directions for future research:

- Study the discipline of botany through selected *laws of distribution* (e.g.,Zipf's law and Bradford's law) to better understand the domain and characters of North Carolina Trees and with OpenKey also Illinois prairie plants.
- Conduct user studies using BOTNET, PIC, and OpenKey to test the usefulness of all the tools in the vocabulary suite.
- Analyze extensibility of the Conceptual Table for additional families (both flora and fauna).

As Wilson (2003) states, people are easingly calling for a single Web resource escribe all life on earth. In conclusion, if we want to move in this direction and inform people about the natural world on a global scale, we need to continue to study vocabulary challenges and share knowledge.

References

1. Baker, T. (2000). A Grammar of Dublin Core. *D-Lib Magazine*, 6(10): org/dlib/october00/baker/10bak www.dliber.html.
2. Buch, P., Martin, K., and Silbajoris, C. (1999). BotNetDC. ils.unc.edu/~buchp/botnet/web/BotNetDC.html
3. Davis, P. and V. Heywood. (1973). *Principles of Angiosperm Taxonomy*. Robert Krieger Publ. Co., NY.
4. Dopke, J. and Carlson A. (2001). Evaluating Children's Electronic Educational Systems: A Case Study of the ings: Plant Information Center (PIC) Website. *WebNet 2001 Conference Proceedings:* 317-322.
5. Greenberg, J. (2001). Metadata Applications for the Plant Information Center (PIC): A Web-based Scientific Learning Center. *Interactive Learning Environments*, 9(3): 291-313.
6. Greenberg, J. Bullard, K., James, M. LDaniel, E., and White, P. (2002). Student Comprehension of Classification Applications in a Science

Education Digital Library. In Maristella Agosti, Constantino Thanos (Eds.). *ResearAdvanced Technology for Digital Libr6th European Conference, ECDL 2002*, Rome, Italy, September 16-18, Proceedings. Lecture Notes in Computer Science 2458 Springer (ISBN 3-540-44178-6), pp. 560-567

7. Hall E. P. (2001). The Plant Information Center Database Interface. MSIS, ry Science, Master Paper for completion of School of Information and LibraUniversity of North Carolina at Chapel Hill: http://ils.unc.edu/MSpapers/2687.pdf.

8. Hanrath, R. S. (2002). A Usability Study of a Tool for Contributor-supplied hapel Metadata Creation: the Use of Metadata Element Definitions and Examples in Online Help. Master Paper for completion of MSIS, School of Information and Library Science, University of North Carolina at Chapel Hill: http://ils.unc.edu/MSpapers/2793.p.

9. Harmes, H. (2001). Development of an Input Form to Capture Author-Generated Metadata for a Botanical Image Collection. Master Paper for completion of MSIS, School of Information and Library Science, University of North Carolina at Chapel Hill: http://ils.unc.edu/MSpapers/2674.pdf.

10. Heidorn, P. B. (2001). A Tool for Multipu rpose Use of Online Flora and Fauna: The Biological Information Browsing Environment (BIBE). *First Monday,* 6(2): www.firstmonday.dk/issues/issue6_2/heidorn/index.html

11. Jacob, E. K. (20011. 3). Ontologies and the Science and Semantic Web. *Bulletin of the American Society for Information Technology*, 29(4): 19-22. www.asis.org/Bulletin/Apr-03/BulletinAprMay03.pdf

12. Radford, A. E., W. C. Dickison, J. R. Massey, C. R. Bell. (1976). *Vascular Plant Systematics*. Harper and Row, New York.

13. Rowley, J. (1994). The Controlled Versus Natural Indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science* 20: 108-119.

14. Somerville, M. A. (1998). UNESCO Transdisciplinary Programmes, nary. Preliminary Insights. Transdisciplinary Paris: UNESCO, Division of Philosophyand Ethics

15. Stahl, S. A. (1999). *Vocabulary Development*. Cambridge: Brookline Books.

16. Svenonius, E. (1986). Unanswered Questions in the Design of Controlled Vocabularies. *Journal of the American Society for Information Science*, 37: 331-340.

17. Warmoth, E. K. (2002). The UNC PlantInformation Center's "Ask the Expert" module: A Usability Study. Master Paper for completion of MSIS, School of Information and Library Science, University of North Carolina at Chapel Hill: http://ils.unc.edu/MSpapers/2771.pdf.

18. Williams, J. (2000, *unpublished*). Creatingan XML Document Type or es and n and Definition for "Frequently Asked Questions" WebPages. Final project fINLS 210-92, Metadata ArchitecturApplications, School of InformatioLibrary Science, University of North Carolina, Chapel Hill.

19. Wilson, E. O. (2003). Trends in Ecology and Evolution. *Encyclopedia of Life*, 18(2): 77-80.

Acknowledgements

Wewould also like to thank Evelyn Daniel, Professor, SILS/UNC; Peter White, Director of the North Carolina Botanical Garden and Professor, UNC/Biology; and Kenneth R. Robertson, Administrative Curator of the Herbarium, Illinois Natural History Survey for their contributions to this paper and vocabulary research and development.

Appendix A

Example 1: Description for
Tulip Tree (Liriodendron tulipifera)

```
<Plant_Description>  <Global_ID_Number>18086
 <Ran
 <Name> </Name>
 <Authority>
 <Vernacular>tulip-tree</Vernacula
 <Vernacular>yellow poplar</
 <Vernacular>tulip poplar</Vernacular>
 <Synonym>
  <Name>Liriodendron procera</Name>
  <Authority>Salisbury</Authority>
```

```
    </Synonym>
    <Synonym>
     <Name>Tulipifera liriodendron</Name>
     <Authority>P.
    </Synonym>
   </Taxon>
   <Special_Diagnostic_Characters></Special_Diagn
ostic_C
   ers>
    <Habitat>forest</H
    <Habitat>mixed edge</Habitat>
     <Life_Span>perennial</Life_
     <Woodiness>woody </Woodines
     <Growth_Habit>tree</Growth_H
     <Growth_Form>single upright stem</Growth_
                <Plant_Height_When_Mature>30-
40</Plant_Height_
   ure>
     <Nutrition>autotrophic</Nutrition>
     <Carnivory>not carnivorous</Carnivory>
    <Stems>
     <Stem_Types>aerial stem</Stem_Types>
    <Stem_Surface_F
     <Pith>continuous</Pith>
     <Leaves>
     <Leaf_Duration>deciduous</Leaf
     <Leaf_Len
     <Leaflet_Le
     <Leaf_Width>6-20</Leaf_Width>
     <Leaflet_Width></Leaflet_Width>
   <Leaf_Arrangement>alternate</Leaf_Arrangement
>
     <Leaf_Complexity>simple</Leaf_Complex
     <Leaf_Shape>orbicular</Leaf_Shape>
     <Leaflet_Shape></Leaflet_Shape>
     <Leaf_Veins>pinnate</Leaf_Veins>
     <Leaflet_Veins></Leaflet_Veins>
     <Leaf_Margin>entire</Leaf_Margin>
     <Leaflet_Margin></Leaflet_Margin>
     <Leaf_Lobing>pinnately lobed</Leaf_Lobi
     <Leaflet_Lobing></Leaflet_Lobing>
     <Leaf_Base>cordate</Leaf_Base>
     <Leaf_Base>truncate</Leaf_Base>
     <Leaflet_Base></Leaflet_Base>
     <Leaf_Attachment>petiolate</Leaf_Attachm
     <Leaflet_Attachment></Leaflet_Attachment>
     <Leaf_Apex>emarginate</Leaf_Apex>
     <Leaf_Apex>truncate</Leaf_Apex>
     <Leaflet_Apex></Leaflet_Apex>
     <Stipules>present-deciduous</Stipules>
   <Leaf_Upper_Surface>glabrous</Leaf_Upper_Sur
fa
    <Leaf_Lower_Surface>glabrous</Leaf_Lower_Su
     <Petiole_Surface>glabrous</Petiole_Surface>
     <Rachis_Surface>glabrous</Rachis_Surface>
</Leaves>
    <Flowers_Cones>
     <Bloom_Time>March-June</Bloom_Time>
     <Flower_Attac
     <Inflorescence_Position
     <Inflorescence_Type>flowers
   so</Inflorescence_Type>
                <Breeding_System>flowers
perfect</Breeding_System
        <Flower_Symmetry>regular</Flower_S
<Petal_Color>g
   base</Petal_Color>
     <Petal_Number>6</Petal_Number>
     <Petal_Fusion>separate</Petal_Fusion>   <Petal
     <Sepal_Color>green</Sepal_Color>
     <Sepal_Number>3</Sepal_Number>
     <Sepal_Fusion>separate</Sepal_Fusion>
     <Sepal_Length>3-6</Sepal_Length>
     <Position_of_Ovary>superior</Position_of_Ov
     <Stamen_Number>numerous</Stamen_Number>
     <Pistil_Number>numerous</Pistil_Number>
     <Styles_per_Pistil>1</Styles_per_Pistil>
     <Placentation>marginal</Placentation>
    </Flowers_Cones>
    <Fruits_Cones_Seeds>
     <Fruit_Type>Samara</Fruit_Type>
     <Seeds></Seeds>
    </Fruits_Cones_Seeds>
    <Root_Type></Root_Type>
   </Plant_Description>
```