EcoPod: A Mobile Tool for Community Based Biodiversity Collection Building

YuanYuan Yu
Department of Computer Science
Stanford University
yuanyuan@stanford.edu

Jeannie A. Stamberger
Department of Biological Sciences
Stanford University
jeans@stanford.edu

Aswath Manoharan
Department of Computer Science
Stanford University
amanohar@stanford.edu

Andreas Paepcke
Department of Computer Science
Stanford University
paepcke@cs.stanford.edu

ABSTRACT
Biological studies rely heavily on large collections of species observations. All of these collections cannot be compiled by biology professionals alone. Skilled amateurs can assist by contributing observations they make in the field. The challenge with such contributions is their potentially questionable quality. We present our PDA-based application EcoPod, which replaces traditional paper field guides with a mobile computing platform. EcoPod aims both to increase the efficiency of the identification process and its reliability. The application solicits as little information from the user as possible. At the same time it places no restrictions on the sequencing of the identification process. This approach is to make our solution attractive to both skilled amateurs and professionals. The tool creates a record of the identification process, thereby providing an audit trail for quality assurance. EcoPod's user interface driver computes information gain over identification metadata to maximize screen utilization. The tool ingests SDD, an international standard for XML dataset that describes organisms.

Categories and Subject Descriptors
H.3.7 [Information Systems]: Digital Libraries - *Collection*, *Dissemination*, *Standards*.

General Terms
Design, Human Factors

Keywords
Biodiversity, PDA platform, data collection and dissemination, SDD, biology

1. INTRODUCTION
Biology researchers explore how the number and geographic ranges of organisms in our ecosystems change in response to shifting environmental conditions. One foundation of their efforts is the availability of observations made in the field across the globe. These observations include data such as species presence or absence, counts of individuals in a species, descriptions of individual specimens, time and place of observation, photographs, and environmental conditions under which the observations were made. Of particular importance are time series of repeated measures, taken in the same geographic area. Such collections sometimes span decades.

When researchers analyze these collections, they look for patterns of change in the abundance of particular species, or presence/absence of species that suggest geographic range shifts. For example, an investigator might discover a year-by-year decrease of a particular butterfly species in an area of study. More subtly, a change in average size, or delayed metamorphosis may indicate stress or species adaptations in progress. Such studies have been critical in documenting global climate change by revealing that globally organisms are shifting the timing of life events such as arrival of migrating birds, or flowering in plants, in response to climate change [1].

Traditionally, biologists have assembled their own collections of observations. Such individual efforts, however, are limiting as the need for observations that span large geographic areas grows urgent. Researchers rarely rely on amateur observations because of concerns of accurate species identification and observation bias in small datasets. Many global changes can only be evaluated by comparing findings in far flung, yet comparable places. Continuing the example above, an analysis of the same butterfly species in similar habitats elsewhere may offer insights into causes of the change: is that change a response to local urban development, or to more global climate change?

Individual efforts are limited not just in geographic coverage, but also simply in the number of data points that can be collected. Even a team of researchers can only spend so much time in the field.

The World-Wide Web has created the chance to include more data collectors for increasingly global efforts. The Calflora project, for example, has collected over 850,000 observations of more than 7,600 plants in California. The project is a community effort; observations are entered online by volunteer outdoors people. Community based data-collections have already yielded key scientific and conservation insights, including understanding the effect of environmental factors on population sizes and identifying critical habitat of the Monarch butterfly [2], which is considered an ‘endangered phenomenon’ because large numbers of individuals go on a long migration [3].

The difficulty with such community based efforts is that the collectors are not necessarily experts in identifying organisms. Many professional biologists are therefore wary of making the collections a basis for scientific study. Nonetheless, broadening the source of contributors to the body of observations is an exciting opportunity.

Several complementary approaches can combine to stabilize the quality of amateur and semi-professional collection efforts. One solution is to enlist the community in the quality assurance process. Online visitors to the collection could help correct possible misidentifications by inspecting the submitted photographs. Experts can then study the questionable entries in more detail.

Another approach is to use statistics across the growing collections to identify outliers. For example, one report of a rare sighting might automatically be set aside until more such

sightings are registered. The challenge with this approach is to tune the statistical threshold parameters to ensure that invasive species are quickly detected, while still safeguarding the collection's integrity. This tuning process can grow rather sophisticated once understanding of the biology is added to the statistical analysis. For example, thresholds might be lowered quickly when fast spreading species are beginning to be reported in new geographic areas, while observations of slower invading organisms could be treated more conservatively.

We are pursuing a third approach. We try to contribute to collection quality assurance by helping amateurs make the correct observations in the first place. In the context of our BioACT project we are constructing computing tools that support observers in the field.

Before we proceed to the details of our design and implementation, we explain how species identification is traditionally accomplished in the field. This description will expose the opportunities that computing support can provide.

1.1 How organisms are identified

Traditionally, species identifications are made on the basis of field guides in book format. Beyond general descriptions of how the plants or animals in question look and behave, these guides contain what are called "keys". In a computer science context these devices would be thought of as decision trees that ask the reader questions about an observed organism. Based on the answers, more questions are posed until the reader arrives at an identification. Field guides, which may not have readily apparent decision tree structure, still organize information in the book in a hierarchical manner reflective of decision trees (e.g. water birds are presented separately from land birds, which is similar to answering the decision tree question "Is the bird on land?") Figure 1 shows a schematic view of a key. Nodes in the tree are questions; the terminals are names of species, or "taxa" ("taxon" is the singular).

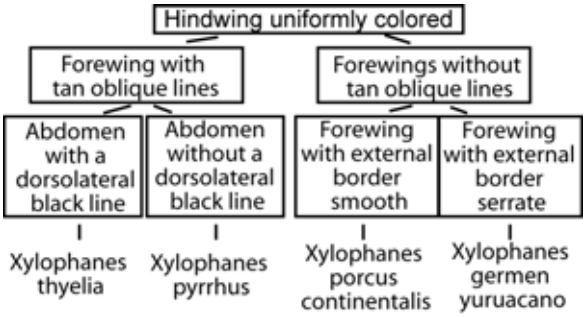

Figure 1: Excerpt from a dichotomous butterfly key

Identification is conducted by asking questions that determine whether the specimen has a particular attribute, such as wing color for birds, or leaf shape for plants. These attributes are called "characters". The values that characters may take are called "states". Thus the character "leaf color" might take the state "red". The process of identification, or "keying" is thus conceptually a decision tree traversal.

In practice, however, the process is messier. For one, some observations are more difficult to make in the field than others. Worse, the difficulty may vary from case to case. Fur length of a running squirrel, for example, is difficult to assess. On the other hand, if the squirrel is dead on the ground, that character state is easy to determine. Depending on the situation, progress towards an identification may therefore require random access to questions, rather than the clean top-down discipline implied by the decision tree view of identification.

States of characters within the same species can vary by geographic location, season, and life stage of the organism. Body color exemplifies this variation: butterflies in the northern extent of the species range tend to be darker than the southern extent of the range; fox fur color changes with seasons to camouflage with white snow in the winter and brown ground cover in the summer; and the spotted coat pattern of baby deer is lost in adults. Therefore, there can be tremendous variation in the state of a character within the same species, making identification complex.

Another source of difficulty in practice is that on a given specimen a particular character state may be ambiguous. For example, it may not be clear to the amateur whether the reddish-orange color of a butterfly wing falls into the state of "red" or "orange", and may require an expert who is familiar with the range of colors found in these butterflies (i.e. what is "red" and what is "orange"). Branching along the decision tree may therefore need to be deferred, or a tentative decision needs to be made. Observations may thus remain multi valued until final analysis at home where a local or remote expert can evaluate photos that document the specimen.

1.2 The Digital Advantage

This brief sketch of how observations are made suggests that computing support during the observation process can offer considerable improvements. Maybe the most important of these, beyond size and weight, is a computing solution's ability to adapt dynamically during the identification process in the field. Users can access questions about characters at random, while a physical guide book is more rigid.

Another aspect of this dynamism is a computer's context awareness. A computer can optimize its interface based on knowledge of the time and ─ if properly equipped ─ place where an observation is being recorded. The body color variation mentioned above exemplifies the advantage of context awareness that a computer can offer over a static book. Book format field guides list alternative states based on season or geographic location, but this generality makes book based keys cumbersome to use. If a particular observation is being made, say, in the fall, computers can exclude summer colors from the states that they offer to the user in the interface; books cannot adjust to such context.

Beyond dynamism, computers can well support the exploration and bookkeeping for alternative identification decisions. It is easy on a computer to maintain sets of alternative, tentative determinations of states. For example, an observer might use the computing device to take a photograph of a plant's leaf whose shape state he is unsure about. He would then tentatively enter "serrated" as the rim character's state and proceed entering the states that are more clearly distinguishable. The observer can then make maximum progress in the field and sort out remaining details later. Our user interface explicitly supports this process of generating identification hypotheses.

Finally, if it were possible to record not just the results of identifications, but also how the results were arrived at, experts might feel more confident in using data that was generated by

non-experts. Documentation of the process would be an audit trail that could then enable vetting.

While all these opportunities for digital support may be clear, their realization is challenging. We designed and implemented a framework that loads XML encoded character information about a set of species and automatically generates a dynamic user interface for the identification of these species. Our computing platform is personal digital assistants (PDAs).

In order to realize the digital advantages on the very limited screen real estate that PDAs have to offer we combined careful user interface design with an information theoretic engine. This underlying machinery must ensure that the most powerfully distinguishing characters are always close at hand on the screen, while still affording access to alternatives that the observer might wish to work with.

[4] has described a list of desired qualities of electronic field guides (EFGs) and software design issues involved. Though not using that list as an exclusive guide, we did address many of the things mentioned there in our PDA-based solution: EcoPod. In the following we describe the interface in detail. Its design is the result of many hours of interviewing biologists, both expert and semi-professional. We then explain the system's architecture and the mathematical process that drives it. We conclude with related work and our plans for the future.

2. EcoPod's User Interface

The EcoPod workflow supports three primary approaches to identification. A first approach has the observer successively specifying the states of characters. EcoPod responds each time by limiting and optimizing the layout of the remaining choices.

A second approach an observer might take is to work backwards, hypothesizing a taxon to be the correct identification and examining whether the evidence supports the hypothesis. A hypothesis may be arrived at, for example, by someone mentioning that species X is common in the observation area.

The third approach, finally, is to work primarily from pictures, rather than through the textual character-and-state methodology. Recognition, rather than analytic progression can sometimes lead to faster identification.

EcoPod allows observers to interleave all three workflows. After introducing the home screen of the interface, the following sections illustrate how EcoPod supports each workflow approach.

2.1 The Home Screen

The application initially ingests an XML representation of the entire identification key, be it for a class of plants or animals. This representation, known as SDD (Structured Descriptive Data), is an international standard. EcoPod can thus display keys that were authored by experts anywhere (more on the SDD standard in section 3.1). Once the key with its factual information and images has been loaded, the home screen displays on the PDA.

The upper panel of the home screen is a tree view populated by all the characters that are used to identify the species encoded by this key. The lower panel is a list of all the species (a.k.a. taxa) that the current key covers (Figure 2).

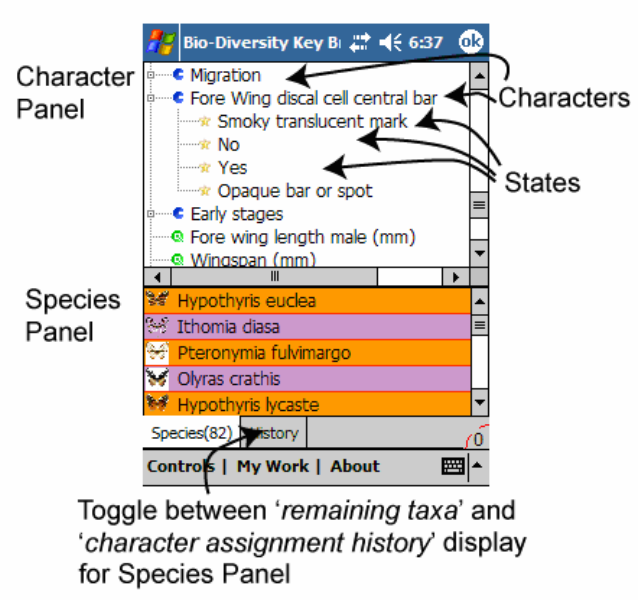

Figure 2: EcoPod's Home Screen

The observer may expand each of the characters in the character panel to expose that character's possible states. We distinguish between categorical and quantitative characters, which are natively defined in SDD. A categorical character is one that has a pre-defined set of states. The blue "C" icons in the home screen indicate categorical characters. A quantitative character is one that calls for a continuous value (green "Q" icons).

As identification continues, species are eliminated from the list in the species panel. Every taxon in that list has an accompanying thumbnail image displayed to its left. The lower left of the screen, finally, shows how many species remain in the possible result set (e.g. Species(82) in the screenshot).

2.2 Working Forward

When the user works towards an identification by specifying observed states for characters, a challenge for the user interface is to display the "best" characters in the visible portion of the character panel at all times.

2.2.1 Ordered character list

We drive the ordering of the character list by always minimizing the number of characters a user needs to specify to reach a final taxon. The characters at the top of the list, and therefore the ones visible without scrolling in the character panel, are thus the ones whose disposal affords the user the largest progress. EcoPod has this approach in common with a number of desktop-based identification programs [5]. The algorithm behind this ordering is based on the concept of information gain and classic ID3 decision tree learning algorithm. We will discuss this in more detail in section 3.4, including the mathematics involved, its limitation, and potential improvement.

Figure 3a shows the system recommending "Species range" as the character that guarantees maximal progress towards narrowing down to a final species. Species range is followed by "Abundance", "Color pattern", etc.

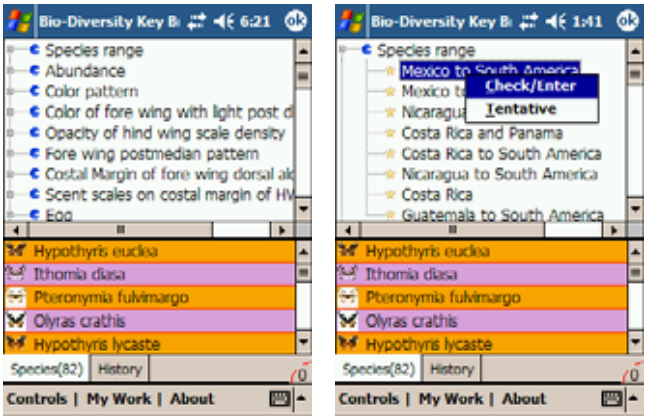

Figure 3a and b: "Best character" ordering and categorical entry

2.2.2 Entering a Character State

Figure 3b illustrates how a categorical character is entered. A tap-and-hold gesture exposes a dialog that offers the choice of committing firmly to the decision of entering the tapped state for the respective character, or making a tentative decision.

Once the state is entered, EcoPod will eliminate species that do not exhibit the specified state for the respective character from the list in the species panel.

The program will also re-compute the ordering of the character list. The new order will place at the top of the character panel the character that is the most distinctive among the remaining taxa.

Quantitative data is entered via a keypad. The keyboard buttons are large to minimize risk of erroneous entry (Figure 4a and b).

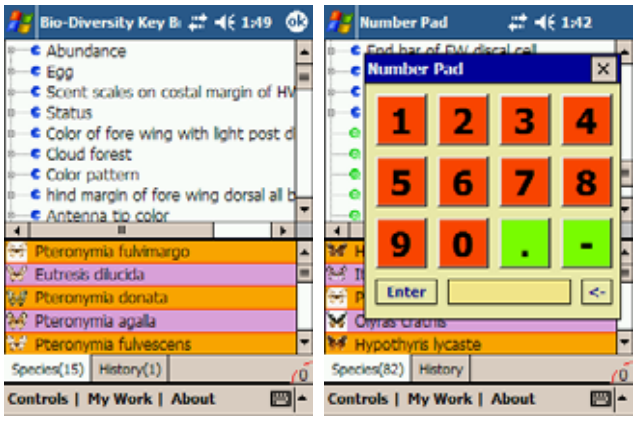

Figure 4a and b: Character list recomputed; Number pad

2.3 Working Backward from a Taxon

We now turn to the second workflow alternative, identification 'backward' from a hypothesis. In EcoPod the user simply clicks a taxon in the species list to obtain all the taxon's relevant information.

2.3.1 Displaying a Taxon's Characters

EcoPod clusters all of the selected taxon's relevant characters and pushes them to the beginning of the character list. The list is again in 'most-powerful-first' order. A red "S" icon indicates relevant categorical characters, a red star icon highlights relevant states, and a green "S" icon indicates a relevant quantitative character. Characters that don't apply to the selected taxon are appended to the end of the list. Figure 5 shows how the characters of one particular butterfly are displayed.

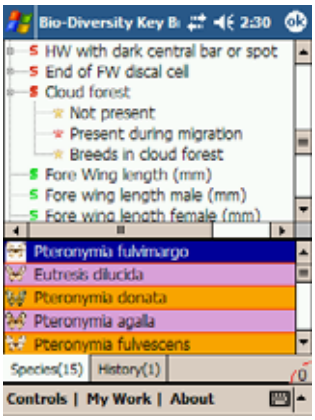

Figure 5: One taxon's characters

Once this display is available, the user may work down the list of the taxon's characters and ensure that all the states of the observed specimen match the key's prescription. Discrepancies indicate that the chosen taxon may not be what the observer has spotted in the field.

2.4 The Image Alternative

It is often easier to work with pictorial than textual information. SDD allows media resources such as images to be associated with taxa. When users select a taxon to view in EcoPod, the corresponding images are displayed (Figure 6).

Figure 6: Image display

Due to space constraints traditional paper-based keys typically include no more than 1-2 images for each taxon. Multiple images may, however, be required to convey different sexes, life stage and characters that vary by season, behaviors, etc.

In addition, field guides must choose between artist illustrations that highlight the diagnostic features of a species but may not look realistic, and photographs of specimens that are realistic, but may be confusing. The reason for this confusion is that it is often unclear which of an individual's particulars are idiosyncratic variations, and which are representative of the species as a whole.

SDD can associate a virtually unlimited number of image resources with each taxon, and EcoPod is constrained only by memory, which is consistently increasing in modern PDAs. Note that compared to, for example, photo album browsers on PDAs, organizing this many images in the interface is much less of an issue in EcoPod. This is because the context of the identification task and the extensive metadata that constitutes the key are excellent browsing constraints. This is in contrast to collections of photographs without context. Those are difficult to organize and randomly access.

2.4.1 *Image Gallery Quick View*

Beyond viewing images of a selected taxon, EcoPod provides an image gallery. This facility allows an observer to browse through all of a loaded key's images. The image gallery is therefore a pure browsing alternative.

However, in the context of EcoPod's application, domain browsing also requires the ability to investigate alternatives in side-by-side comparisons. Figure 7 shows how these close-up comparisons are accomplished in the EcoPod interface.

Figure 7: Image gallery quick view

Available through a popup menu, the image gallery quick view displays a thumbnail strip along the left side of the display. Clicking on a thumbnail enlarges the image and displays it in the right panel, allowing the user to examine the details of the image. Clicking on a second thumbnail generates the dual view comparison in Figure 7. Clicking on the large images allows the user to locate the species directly in the species list.

The Image gallery quick view is particularly useful to speed up the identification process when there are only a few remaining species to choose from. In that case, the user may view images of the remaining species and quickly select the one that matches the real object. Again, the context of the user's task makes the availability of many images useful, rather than just overwhelming.

2.5 Recovering from Mistakes

The uncertainties of identifying a particular specimen may at times lead to 'dead ends.' The symptom of a dead end in identification work is often that none of the states in a character that is required for all remaining species fits the observed specimen. For example, after specifying a number of states for a series of characters one character in the remaining list might call for a decision on whether “Leaves are flattened” or “Leaves are not flattened and needle-shaped”. If the leaves of the specimen at hand are neither flattened nor needle-shaped, then a dead end has been reached, and the user needs to backtrack. Our tool to handle this situation is the History Panel.

2.5.1 *History Panel*

The display area that usually houses the species panel may instead be switched to show a history of actions. The switch is made by clicking on the “History” tab at the bottom of the screen. Once in History mode, the panel shows a list of characters, somewhat like the list in the character panel in the top portion of the screen. However, each character's entry in the history also shows the state that the observer has entered for that character during the course of the identification process so far.

When the user selects an item in the history list, a combo box on the right is populated by all the possible states for that character, and the current user-specified state is displayed (Figure 8).

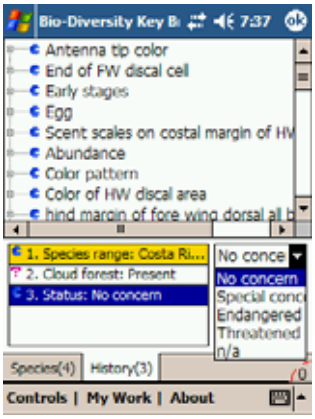

Figure 8: History panel

2.5.2 *Revising Previous Inputs*

There are two ways of changing previous user input.

1) If the user believes he has selected the correct character but simply entered a wrong state for it, he may modify the state in the combo box on the right. In response, EcoPod will re-filter the species list and the ordering of the list in the character panel. These computations are all synchronized by the click on the combo box.

2) If the user needs to discard a character completely, he can simply tap-and-hold on an entry in the history list and select “Remove Item”: An important point: users can remove steps in arbitrary order. The process is not constrained in the classic “Undo” stack discipline, which cannot undo step 2 unless step 3 is undone first. Our interviews with both professional and amateur biologists documented consistently that such 'random access' is essential.

Notice in Figure 8 that the second item has a purple question mark as its icon. This is how EcoPod reminds the user where he indicated uncertainty when making state decisions previously. Such decisions may be good first choices for backtracking when hitting a dead end.

2.6 Documenting the Identification

We stressed the importance of documenting how observers, particularly skilled amateurs, arrived at their identifications. Such documentation may ensure that the identification work will be useful to professionals. We include a number of facilities in this vein.

All identifications are saved in persistent store. Saving may be requested at any time; when identification is complete, or to save an unfinished piece of work for continuation at a later time.

The history panel described earlier in the context of backtracking is, of course, a tool for documenting decisions. Even once the user decides on a taxon as the final identification, the history panel documents which characters the user explicitly specified during the identification. Recall that the user may, for example, pick from a photograph, rather than running through all characters.

Another documentation tool we built into EcoPod is notes. A popup menu can raise a memo window into which users can jot notes about the ongoing identification. Embedded OCR facilities provide translation to ASCII text, or the note may simply be retained in ink format (Figure 9). Alternatively, users may record short audio messages and play them back later on.

A third documentation tool is digital images. The photographs will be taken with the camera that is built into the PDA, which will be enabled in future user trials. Digital images can be used to record a character of questionable state, or to document unusual variation in organism morphology or behavior. Images, however, may not be enough to identify a specimen, since the states of all characters may not be show in the image, and images, of course, cannot capture states of diagnostic characters such as movement and vocalizations.

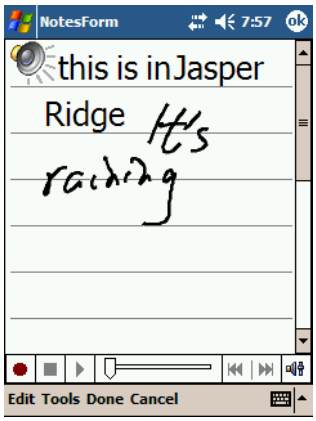

Figure 9: Notes help document the ID process

An off-screen pile facility allows users additional means for saving supporting evidence, or taxa that they were considering during identification. Off-screen piles provide the illusion to the user that material flung with the stylus towards a corner of the screen end up on a pile beyond the edge of the screen. A small arc in the corner of the screen provides both a reminder of the pile's existence, and a handle for recalling the piled information to the display. Previous research works [6][7][8] have been done on exploring and verifying the feasibility of this "virtual space visualizing" technique.

We currently have one pile implemented to contain taxa entries that the user has flung from the species panel to the lower right of the screen. In the future we will include another off-screen pile for photographs taken to identify steps in the identification.

3. Architecture and Implementation

EcoPod is implemented on an HP iPAQ rx3715 PDA with a 400MHz CPU and 64MB of RAM. We wrote on top of Windows Mobile 2003 2ed, using the Microsoft .NET compact framework. Note that battery life is of major concern in this application.

Figure 10 shows how EcoPod is put together. The front-end user interface communicates with both the piles and the Global Cache. The Cache uses the Load Module to ingest one of any available XML encoded keys (SDD standard) into memory. The Cache also maintains the history trace and the set of characters that are (still) relevant at any given moment in the identification. The latter 'in-set' is computed by the Character Set Computation Engine. The presentation order of this in-set is re-computed by the Character Ordering Engine as required by the user's manipulations on the display. We next present selected elements of this architecture.

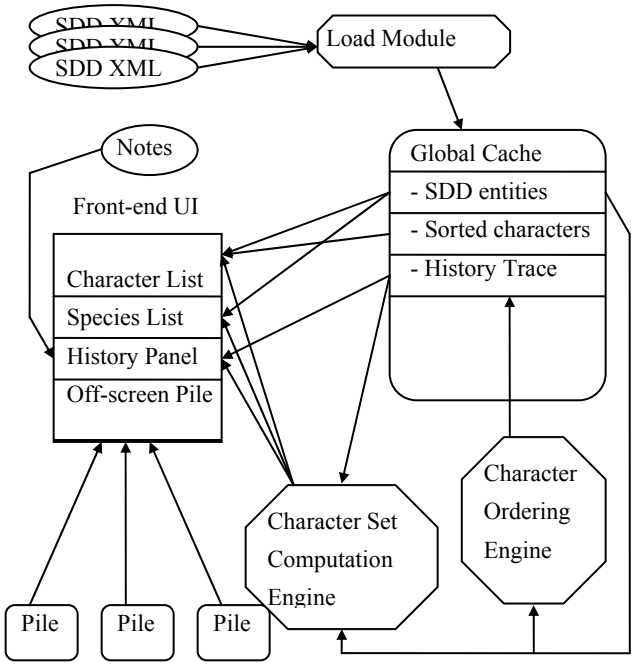

Figure 10: Architecture of EcoPod

3.1 The SDD Representation of Keys

SDD (Structured Descriptive Data) is an international XML-based standard for capturing and managing descriptive data for organisms. It is actively developed and maintained by TDWG (Taxonomic Databases Working Group, http://www.tdwg.org) [9][10]. Using an XML format, SDD supports descriptions of taxa and specimens both in natural language and as "CodedDescriptions". These descriptions conform to a community constrained vocabulary ("SDD Terminology") for describing characters and states. A number of systems have been created for authoring and managing SDD descriptions [11][12]. Desktop-based application that uses SDD exist as well (e.g. Lucid, http://www.lucidcentral.com). We are not aware of a PDA-based software application that utilizes SDD as the underlying representation and presents users with an efficient identification workflow.

3.2 Key Loading and Hashing

EcoPod's operation requires non-trivial processing (especially for a PDA platform, which has limited processing power compared to

desktop PCs), and SDD structures are complex. Efficiency of the internal data representation is therefore of concern.

When loading an SDD key, EcoPod undergoes a process called "flattening the XML", which extracts the nested information items from the XML file and populates them into hash tables. For instance, here is a character encoded in SDD:

```
<CategoricalCharacter id="332">
  <Label xml:lang="en-us" audience="en">
    <Text>Fore wing postmedian pattern</Text>
  </Label>
  <States>
    <StateReference ref="326" id="333" />
    <StateReference ref="328" id="334" />
    <StateReference ref="330" id="335" />
  </States>
</CategoricalCharacter>
```

When "flattened" into the hash table, the structure turns into a key/value pair, where the key is the "id" (332), and the value is a memory reference pointer to the XML node. This kind of hash table is widely used in EcoPod, including the taxon table, the character table, the coded description table, the media resource table, the state definition table, and others. These data structures are very efficient (O(1)) for the runtime cross-referencing and lookup that we need to perform.

3.3 Character 'In-Set' Management

One of EcoPod's user interface strengths is its dynamic adaptation as the user makes progress. The success of this adaptation requires fast response times. For large keys the management of the character in-set, the set of characters that are relevant at any given time, requires care. Here is an example and the solution we chose in our Character Set Computation engine for implementing backtracking.

Consider the scenario that the user specifies the character C (with possible states X, Y, and Z) has state X. This eliminates all the species having character C but taking state Y or Z. Once this character decision is made, EcoPod remembers which species are being eliminated. These species are placed into the RemovedList of that particular character C.

Later, if the user decides to remove the assignment of X to C, instead of processing the entire species list to identify those with character C taking state Y or Z, the program only needs to consult the RemovedList and restore the species in that list.

Computation at this complexity level poses no problem for desktop PCs at all. But on a PDA, without this type of care, such a search may cause significant delay in UI response time. The situation can grow more complex for this optimization.

Suppose this time the user specifies that character C has state X, and character D (with possible states P, Q, and R) takes state P. Then what happens to a taxon T which has state Y for character C and state Q for character D? It is eliminated by both of the user's assignments (the assignment to C and the one to D). Thus, removing the assigned state X to character C from the history list should not result in restoring T, as it is still eliminated by D=P. For this situation the Character Manager maintains a RemoveCount associated with each taxon. Only when the RemoveCount reaches 0 will a taxon be present in the remaining species list.

3.4 Ordering of Characters

As pointed out, the limited screen real estate of PDAs requires that 'good' questions about characters need to be pushed up into the visible area of the screen where they do not require scrolling. In the context of species identification there are two measures of 'good'. By one measure we wish EcoPod to ask as few questions as possible before arriving at a conclusion. Relying on this measure alone, however, would be naive in EcoPod's application domain.

A second, sometimes competing, measure for the goodness of character ordering is rooted in biology. EcoPod should also prefer questions that are feasible to answer from simple observation. For example, the shape of an animal's heart valves might be an excellent question when the goal is to optimize the minimality measure. But when the observer is out in nature, most of the time he will not be able to answer the question. EcoPod's ability to provide random access to any of the characters at any time helps the user optimize the feasibility measure. He can use a greedy approach for a time, entering states for characters that are obvious to observe in his current situation.

Optimizing the minimality measure is where EcoPod can help. Again, the way such support manifests in the interface is that best characters are favored to be pushed up into the visible area of the character panel.

In order to understand how the system accomplishes minimality optimization, it is helpful to consider a different view of the identification process. In the introduction we explained a decision tree view of the procedure. Alternatively, we can think of identification in the context of a matrix in which each row is a taxon and each column is a character. The cells are the states that the respective taxon takes. Figure 11 shows the matrix view of a hypothetical butterfly key.

	Wing Color	**Head Wider than Thorax**	**Wing Spots**
King	*Green*	*No*	*No*
Queen	*Red*	*Yes*	*No*
Duke	*Green*	*Yes*	*Yes*
Bishop	*Red*	*No*	*No*

Figure 11: Matrix view

Four possible taxa can be identified: a King butterfly, a Queen, a Duke, and a Bishop. Three characters in combination determine the identification: Wing color, whether the head is wider than the thorax of the insect, and whether the wings have spots.

The user interface problem for EcoPod is to order the questions such that many rows are eliminated quickly. Figures 12 and 13 show two possible organizations of the question tree.

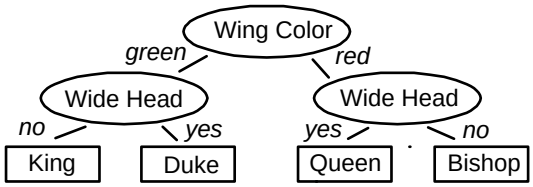

Figure 12: Tree of fewest questions

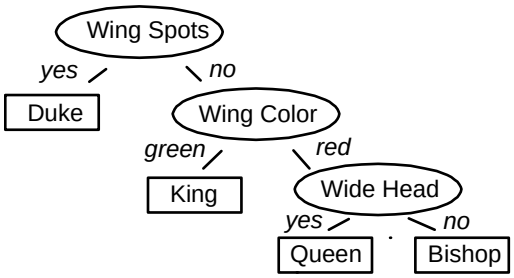

Figure 13: Suboptimal question tree (assuming uniform probability distribution)

In evaluating the minimality measure of the first tree, we compute the average number of questions that EcoPod needs to ask in order to come to an identification. In Figure 12, two questions suffice for any specimen that the observer might encounter.

Now examine Figure 13. Following this tree, if the specimen is a Duke butterfly, one question will make the identification. For Kings, two questions will be required. If the specimen turns out to be a Queen or a Bishop, the user will need to answer three questions. On average, EcoPod would pose 2.25 questions (1+2+3+3/4). The system would therefore choose the tree of Figure 12 over the one in Figure 13 to drive its interface.

Note that this computation assumes equal probabilities among the four taxa: it is assumed that any observed specimen is equally likely to be a King, Queen, Duke, or Bishop. If it were known that in the geographic area where the observer is currently working Duke butterflies are vastly more frequent than the other species, then proceeding as per Figure 13 would be the better choice. In that case EcoPod would only need to ask one question most of the time. We are working on integrating such leverage of historic data into our interface driver.

Figure 12 illustrates another opportunity for systems like EcoPod. Note that this tree never asks about wing spots. The identification is unique without this character being specified. In the interest of quality assurance we could choose to ask the wing spot question anyway. For example, when the system is about to conclude King or Duke, it could ask the user about wing spots. For King the user's answer should be 'no'; for Duke a 'yes' answer is expected. If this safeguard triggers, the system could ask the observer to reconsider some earlier choices. We are working to introduce such double checking as well.

The derivation of optimal answer trees is a well known problem in Artificial Intelligence. The notion of "Information Gain" is prominent in the related solutions (the ID3 decision tree learning algorithm). We refer to [13] for details. Information gain, as the term suggests is the amount of discrimination EcoPod achieves by asking for the state of a particular character. Unless we 'luck out', asking about wing spots first often still leaves 3/4 of the taxa as possibilities for the observed specimen. Asking for wing color or head width instead narrows the choices to 1/2. Visually speaking, maximizing information gain works by balancing the question tree to minimize its depth.

The advantage of using the mathematical machinery behind information gain for EcoPod was that it naturally includes the case of non-uniform probabilities, which will in turn make it easy to include historic observation data in the user interface controller.

4. Related Work

Our current research has been motivated by the relative paucity of species identification tools for mobile/handheld devices. Most of the efforts in the Computer Science community have been directed towards the desktop (large screen) where screen real estate is not an issue. Among the well-known systems is Delta (DEscription Language for TAxonomy), developed by Mike Dallwitz [14]. Delta comprises a suite of tools for generation and organization of keys [15]. It consists of an interactive identification program called Intkey. Intkey offers a variety of features such as random access, encoding uncertainty in observations, and backtracking. It also computes an ordering of the best characters using a variant of the standard equations from Information Theory [16]. EcoPod incorporates the most important set of Intkey's features on a small device.

Another well-known system is the Lucid System developed at the Center for Biological Information Technology (CBIT), University of Queensland in Australia (http://www.lucidcentral.com). Lucid is a commercial program and uses its own proprietary data format, but also supports SDD. Lucid again consists of tools to help in the authoring and publishing of keys and also in the identification of species.

There have been some efforts to develop technologies and applications for handheld devices. Campbell Web and a team from Dartmouth College have been working on porting a Delta-Intkey style application onto the Palm Pilot, running under Palm OS (http://www.phylodiversity.net/cwebb/software/jfilekey.html). As far as we can tell, this effort seems to be work in progress. Two other tools are called SLIKS (Stinger's Lightweight Interactive Key Software) (http://www.stingersplace.com/SLIKS), and Sasol eWildlife (http://www.pdasolutions.co.za). These are simple tools which do not exhibit many of the features necessary for a powerful tool in the field (ordering of best characters, tracking hypotheses, error-checking).

CyberTracker (http://www.cybertracker.co.za/Help/Index_v3.htm) is another PDA-based identification program. But it is handles limited pictorial information and simplifies the user interface to an extent that even non-literate people can use it.

5. Future Work

We are exploring several approaches to further improve EcoPod's character ordering machinery. In particular, we are preparing an experiment to test whether the inclusion of historic observation data in a given geographic region can be used to bias character ordering. The intuition is that characters of species that are known to be common in the area where an identification is being made should be favored when deciding which characters to move up into the visible screen real estate. The danger of this approach is

that rare, or only currently invading species will be missed and thereby under-observed.

A related effort will examine whether context such as temperature measured at the time of the observation, or altitude, can gainfully be applied to the effort of optimizing screen usage. The same cautions apply as to the history based context approach above.

Recall that EcoPod records all characters used in identifications. As EcoPod is deployed in the field, character ordering could in the future be further optimized by favoring the most often used characters. This approach will automatically capture the characters whose states are easiest to identify in field conditions, but are not explicitly identified as such in the key.

Another area of investigation will be error checking. In the interest of quality assurance, one might imagine not eliminating all irrelevant characters as we currently do in EcoPod. Instead, one would retain characters of species that are easily confused with some of the species that remain at a given moment in the identification process. The hope would be to catch errors very early.

We are particularly interested in evaluating whether EcoPod's interface is flexible enough to appeal to both experts and amateurs alike. A user study towards insights in that regard will also test a number of field related issues, such as whether or not glare from sunlight makes the screen difficult to see, whether battery life is suitable to data collection, etc.

The tool is designed to be further extended towards uploading observations to large databases, such as eBirds, a bird observation database from a collaboration of Cornell Lab of Ornithology and the National Audubon Society which is supported by the National Science Foundation (http://www.ebird.org/content/), or the North American Butterfly Association's "Butterflies I've seen" database (http://www.nababis.org/servlets/Sightings), eliminating tedious transcription of observations into cumbersome and repetitive Web-based data entry formats. These extensions are yet to be realized.

6. Conclusion

We described EcoPod, a PDA based tool for identifying plants and animals in the field. The intention is to enable the construction of large biodiversity collections by non-professionals. Global changes in the ecosystem can only be analyzed effectively when large and geographically diverse data sets of observations are available. Such datasets cannot be collected by highly trained professionals alone. On the other hand, for scientific conclusions to be correct, the underlying observations must be correct.

EcoPod helps semi-skilled amateurs participate in the collection of this essential biodiversity data. The system provides 3 methods that can be interwoven for users to navigate dichotomous keys that EcoPod reads from internationally standardized SDD encodings:

1) Follow the "best character" ordering, or randomly access any characters in the key, 2) Examine all relevant characters for a particular species (hypothesis driven identification), 3) Use the image gallery quick view.

While our future work section points to a number of steps yet to be taken, EcoPod can help ensure that community based biodiversity collections reach their highest possible standards.

7. ACKNOWLEDGMENTS

We thank a number of biologists and naturalists at Stanford's Jasper Ridge Biological Preserve, who explained their process of identification. Stan Blum of the California Academy of Sciences pointed us towards a number of existing efforts. Bob Morris of University of Massachusetts Boston helped our understanding of SDD.

8. REFERENCES

[1] Root, T. L., et al., *Fingerprints of global warming on wild animals and plants*. Nature, 2003. 421(6918): p.57-60.

[2] Meitner, C.J., L.P. Brower, and A.K. Davis, *Migration patterns and environmental effects on stopover of monarch butterflies (Lepidoptera, Nymphalidae) at Peninsula Point, Michigan.* Environmental Entomology, 2004. 33(2): p. 249-256.

[3] Brower, L.P. and S.B. Malcolm, *Animal Migrations Endangered Phenomena.* American Zoologist, 1991. 31(1): p. 265-276

[4] Stevenson, R.D., W.A. Haber, and R.A. Morris, *Electronic field guides and user communities in the eco-informatics revolution.* Conservation Ecology, 2003. **7**(1): p. 3.

[5] Dallwitz, M.J. 1992. *A comparison of matrix-based taxonomic identification systems with rule-based systems*. In Proceedings of IFAC workshop on expert systems in agriculture, pp. 215–8.

[6] Baudisch, P., and Rosenholtz, R. *Halo: a Technique for Visualizing Off-Screen Location*. In Proceedings of the Conference on Human Factors in Computing Systems (CHI'03), p. 418-488, 2003

[7] Hsieh, Tony; Wang, QuianYing; Paepcke, Andreas. *Piles Across Space: Breaking the Real-Estate Barrier on PDAs*, Submitted for Publication, http://dbpubs.stanford.edu/pub/showDoc.Fulltext?lang=en&doc=2005-8&format=pdf&compression=&name=2005-8.pdf

[8] A. Cockburn and B. McKenzie. Evaluating the effectiveness of spatial memory in 2d and 3d physical and virtual environments. In Proceedings of the Conference on Human Factors in Computing Systems (CHI'03), p. 203-210, 2003.

[9] Gregor Hagedorn, Robert Morris, Kevin Thiele, P. Bryan Heidorn. *Introduction to SDD (Structured Descriptive Data)*, TDWG 2004, http://www.tdwg.org/2004meet/paperabstracts/TDWG_2004_Papers_Hagedorn_4.htm.

[10] Gregor Hagedorn, Robert Morris, Kevin Thiele, P. Bryan Heidorn. *Structured Descriptive Data (SDD) version 1.0*, TDWG 2005, http://www.tdwg.org/2005meet/TDWG2005_Abstract_26.htm

[11] Jacob K. Asiedu & Robert Morris, *Experience exporting and importing SDD 1.0*, TDWG 2005, http://www.tdwg.org/2005meet/TDWG2005_Abstract_1.htm

[12] Kevin R. Thiele, *SDD and the Key to Life*, TDWG 2005, http://www.tdwg.org/2005meet/TDWG2005_Abstract_71.htm

[13] Stuart Russell, Peter Norvig, Artificial Intelligence - A Modern Approach 2nd ed, p. 653-664, Prentice Hall, 2003

[14] Dallwitz, M. J., Paine, T. A., and Zurcher, E. J. 1993 onwards. *User's guide to the DELTA System: a general system for processing taxonomic descriptions. 4th edition*. http://delta-intkey.com

[15] Dalllwitz, M. J. 1980. *A general system for coding taxonomic descriptions*. Taxon 29: p. 41-46.

[16] Dallwitz, M. J. 1974. *A flexible computer program for generating identification keys*. Systematic Zooloogy. 23: 50–7.