

Visualizations for taxonomic and phylogenetic trees

Cynthia Sims Parr, Human-Computer Interaction Lab, Univ. of Maryland, College Park, MD 20742*
and
Museum of Zoology, Univ. of Michigan, Ann Arbor, MI 48109

Bongshin Lee, Human-Computer Interaction Lab, Univ. of Maryland, College Park, MD 20742

Dana Campbell, Human-Computer Interaction Lab, Univ. of Maryland, College Park, MD 20742

Benjamin B. Bederson, Human-Computer Interaction Lab, Computer Science Department, Univ. of Maryland, College Park, MD 20742

*To whom correspondence should be addressed

RUNNING HEAD: Tree visualizations

ABSTRACT

Motivation. Despite substantial efforts to develop and populate the back-ends of biological databases, front-ends to these systems often rely on taxonomic expertise. This research applies techniques from human-computer interaction research to the biodiversity domain.

Results. We developed an interactive node-link tool, TaxonTree, illustrating the value of a carefully designed interaction model, animation, and integrated searching and browsing towards retrieval of biological names and other information. Users tested the tool using a new, large integrated dataset of animal names with phylogenetic-based and classification-based tree structures. These techniques also translated well for a tool, DoubleTree, to allow comparison of trees using coupled interaction. Our approaches will be useful not only for biological data but as general portal interfaces.

Availability. A working version of TaxonTree, with Tree A, and a video demonstration of DoubleTree, showing Trees A and C, are available at <http://www.cs.umd.edu/hcil/biodiversity>. Tree B is available in XML at <http://www.cs.umd.edu/hcil/iv03contest/index.shtml>.

Contact. cparr@umd.edu

INTRODUCTION

In recent years, vast amounts of biological information have been made available digitally. Examples include genomic and proteomic data, biodiversity and natural history data, and scientific publications. All of these resources require, at least in part, the ability to retrieve information about organisms, typically by the name of the organism. Users of this data vary in their knowledge of these organisms, so that effective interaction with these data sources often requires trial and error exploration until the correct taxonomic spellings and appropriate keywords are discovered that will allow a successful query. This problem is particularly acute for large databases for two reasons: 1) nobody can be a specialist on all parts of the tree of life; and 2) larger databases have more information, both names and attributes of names, with which to interact. Although substantial efforts are underway to develop and populate taxonomic and specimen-based databases (Bisby, 2000), and to improve distributed querying of such datasets (Mammal Networked Information Systems, MaNIS (<http://elib.cs.berkeley.edu/manis>), very little research has focused on the front-ends of such biological datasets.

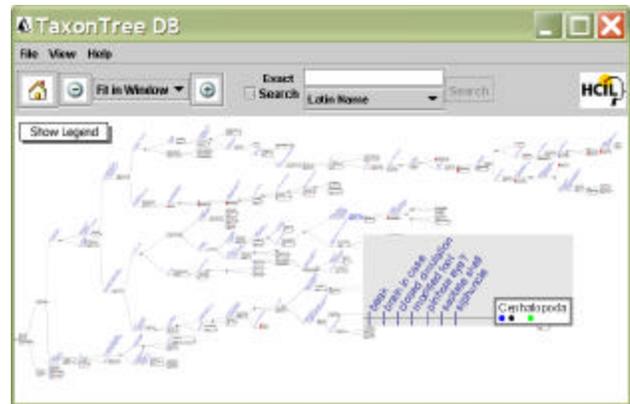


Figure 1. TaxonTree display of 182 node overview, with synapomorphies and tooltip magnification of a node

In part one of this paper, we describe our work on visualization techniques that allow effective interaction with vast trees of taxonomic names and associated attributes. Our goal has been to help users both find what they want and understand the biological context of what they have found. We visualize the tree of names necessary to access further information, thereby creating an innovative portal interface. Our tool provides examples of effective use of techniques such as animation and zooming of node-link diagrams and integrated searching and browsing. These techniques are well suited generally to tasks that require exploration of large hierarchically organized biological datasets.

In part two of the paper, we consider the challenges of visualizing and comparing multiple trees. We extend our work on TaxonTree by developing a new prototype, DoubleTree, which allows users to navigate two trees using coupled interaction. Such an interface not only provides a tool for exploring differences in two similar trees (for example, two conflicting scientific classifications or phylogenies), but can allow navigation in one tree (for example, a “folk” classification organized by a non-scientist) to assist information retrieval using another tree (for example, a scientific classification).

SYSTEM I: SINGLE TREE VISUALIZATION

In this project, two kinds of trees concerned us: biological classifications and phylogenies. By *biological classification* we mean the Linnaean system of scientific names grouped in a nested hierarchy. Each name receives a rank consistent with its level in the hierarchy, such as

Kingdom, Phylum, Class, Order, Family, Genus, and Species (in decreasing order of inclusiveness). The goal of a biological classification is to reflect the evolutionary relationships among the named organisms, but does not convey complete detail within named groups. In contrast, a *phylogeny* uses some form of scientific analysis of particular characters across members of the group, to support a tree topology that truly depicts what is known about evolutionary relationships among organisms (historical branching patterns). A well-resolved phylogenetic tree potentially has far more nodes, or branching points, than the classification. Moreover, some daughter groups are named and some are not. The concept of giving every named group a Linnaean rank to indicate the level of the hierarchy becomes unwieldy

At present, classifications are more readily available to the public and we focus primarily on those. Tools currently used for accessing data using or about taxonomic names rely on simple form fill-in interfaces or browsing indented lists of names. See, for example, the Integrated Taxonomic Information System (<http://www.itis.usda.org>) and the NCBI Taxonomy Browser (<http://www4.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html>). These interfaces work well only when a user can recall enough of a scientific name, or a recognized common name, to enter it into the search box, and if they can recognize their target in the search results. While fuzzy matching algorithms can assist the query input and clustering algorithms can structure the output, users typically must scan text results to first find and then understand the biological significance of the name they entered.

Despite a long history of general research in tree visualization (Herman *et al.*, 2000) these approaches have rarely been applied to or evaluated in the biodiversity domain. Visual approaches have been used in the Prometheus project (Graham *et al.*, 2000), in phylogenetic analysis software (e.g. Mesquite, and in the application of hyperbolic trees (Lamping *et al.*, 1995) to the phylogenetic domain. These interactive visualizations are aimed at highly trained taxonomists who already have a high level of familiarity with the taxonomic groups they are studying and who are likely to invest significant time learning how to use sophisticated tools.

IMPLEMENTATION

Our approach was to modify SpaceTree, a tool we had previously developed (Plaisant *et al.*, 2002). SpaceTree uses smooth animation as a user opens a node in a node-link diagram with a mouse-click. The system automatically closes nodes or subtrees distant from the user's focus to maximize screen usage. Search results are highlighted orange in the node link diagram with non-result branches hidden. The results set can then be a

starting point for more browsing in the tree structure. Zooming and panning are possible, but unlike hyperbolic trees, users have discrete control over which nodes are visible and apparently become less disoriented as a result (Plaisant *et al.*, 2002).

More specific details about the TaxonTree implementation are reported elsewhere (Lee *et al.*, in press). In that paper we describe use of TaxonTree by biology majors in an undergraduate course where repeated retrieval of information from our system would be useful. In this paper we expand on the nature of our dataset and on the biological implications of this work.

Test data set

We created a 200,000 node tree, Tree A, of animal names with several unique characteristics. Tree A resembled a classification at lower levels (usually below families) but a dichotomous phylogeny at higher levels. Our intent was to create the largest dataset possible with global, not local, geographic coverage, spanning as many levels of the taxonomic hierarchy as possible. We used the tree interface as a portal to several prominent websites (described below) so we needed to maximize the number of nodes that could link to these websites. This approach also allows us to consider the effectiveness of our approach for both kinds of trees.

Content was imported and merged from a variety of sources. We began with the classification of Kingdom Animalia from ITIS (<http://www.itis.usda.gov>), from which we took scientific names, ranks, and common names if available. The ITIS classification is a true Linnaean nested hierarchy. We replaced all names from Class Aves with the classification used by the University of Michigan Museum of Zoology. Because ITIS provides information only on North American members of Class Mammalia and Class Reptilia, we replaced those parts of the tree with classifications from the Smithsonian Institution's Mammal Species of the World (Wilson and Reeder, 1993) and EMBL (Uetz, 2003). Finally, to accommodate the needs of our test audience, who were required to know evolutionary relationships, we added unnamed nodes to reflect binary branching. Information for these nodes was consistent with the phylogeny used by our target audience, and was informed by the Tree of Life website (Maddison, 2003), the University of California Museum of Paleontology website (<http://www.ucmp.berkeley.edu>) and Hickman *et al.* (2003). Adding these additional nodes to reflect finer structure was not difficult, except in parts of the tree where classification conflicts with phylogenetic grouping. For example, Class Reptilia is paraphyletic, so it would be misleading as a named node in a phylogenetic tree, i.e. all members of Class Aves would be nested within Class Reptilia. We removed the label for Class Reptilia because our target audience was not required to use this name.

Table 1. Performance comparison of the two TaxonTree versions

Task	TaxonTree local	TaxonTree web
Browse to few children	62 ms	63 ms
Browse to many children	94 ms	96 ms
Exact search	1125 ms	1016 ms
Inexact search	2547 ms	2313 ms

These performance details indicate that our software handled this large dataset with little difficulty. This is primarily because the program shows only a subset of the tree and the other parts are hidden either by default or because a user has closed them. Also, the hierarchical organization of our database efficiently supports the core operation of “show this node’s children.”

User testing of the CD-ROM version of TaxonTree with 18 undergraduates is reported in (Lee *et al.*, in press) and selected results are summarized here in the biological research context.

Users took advantage of the node-link display. Despite our use of a non-traditional biological tree rendering style (linked rectangular, labeled nodes) and a non-traditional combination of phylogeny and classification, users had no difficulty navigating or understanding sister relationships. Users easily interpreted search results, quickly using the node-link structure to discover the quantity of biologically unique results. They often made more sophisticated inferences about relationships among the search results, which is nearly impossible by using a typical list of search results.

TaxonTree integrates searching and browsing seamlessly. After a search, users often carried out additional browsing, giving them a better sense of the search results. Users preferred to browse rather than search, but often employed both strategies for the same task, especially when they had partial knowledge of names or relationships – a situation likely to be common among biologists.

General user response to the style of interaction was positive. Users categorized as “high interest,” who may have had greater taxonomic expertise, were more likely to prefer manual opening and closing of nodes rather than automatic subtree closing when we offered them a choice. Almost all users responded positively to the 182 node overview and all used the interface to access external web pages.

SYSTEM II: VISUALIZING TWO TREES

While visualizing and interacting with single trees is useful, it is often more important to interact with multiple

trees. Biologists, particularly systematists, often compare trees during their phylogenetic analyses. Information managers need to understand differences in biological classification, a field of obvious utility in joining datasets. Non-biologists could also benefit from multiple tree visualizations. Cross-culturally, people build hierarchical mental biological trees (Medin and Atran, 1999). Such a tree may be at odds with a scientific tree, and can be less detailed, but allows people to organize their knowledge and draw appropriate inferences. A visualization that couples a folk tree with the scientific one could allow users to use their own knowledge to begin to understand the scientific tree.

There have been many efforts to visualize multiple trees. Prometheus (Graham *et al.*, 2000) uses a set-based approach combining elements of TreeMaps (Shneiderman, 1992) and ConeTrees (Robertson *et al.*, 1991). A recent contest at InfoVis 2003 spurred interest in this subject among both HCI researchers and biologists and provides benchmark datasets, including classifications provided by the present project (available at <http://www.cs.umd.edu/hcil/iv03contest/index.shtml>). Notable entries included TreeJuxtaposer, Zoomology, and TaxoNote. Each approach emphasizes a different feature of classification hierarchies. TreeJuxtaposer (Munzner *et al.*, 2003) focuses on tree topology differences: it displays two trees side-by-side as square-angled lines, with green lines indicating concordant topology and red indicating topology differences. Each full tree is always rendered, but a user can focus on a specific area of the tree which gives more space for the focus area and less space for the rest of the tree. As the user zooms into part of the hierarchy some labels are arbitrarily provided. Zoomology (Hong *et al.*, 2003) emphasizes rank and level by rendering the hierarchy as nested circles, color-coded by rank. The size of the circle indicates number of members, and white outlines indicate nodes absent in the other tree. TaxoNote (Morse *et al.*, 2003) emphasizes the names in the classification by providing Explorer-style hierarchies and panels listing missing and different taxa.

IMPLEMENTATION

Our approach is to focus on the mode of interaction of two trees, assuming that this would scale up to N trees. One might also argue that cognitively, users are best able to compare trees pairwise even if they are ultimately comparing across N trees. We do not in this research test algorithms for determining correspondences or computing topographical differences among multiple trees; this vitally important work is being addressed by other researchers, e.g. (Day, 1985; Munzner *et al.*, 2003), and complements ours.

Test data sets

To provide another scientific tree to compare with Tree A, described above, we created Tree B by importing unmodified data from the Integrated Taxonomy Information System for Kingdom Animalia. Comparing Tree A with Tree B allows one to see the modifications we made to ITIS when preparing Tree A. We can also compare interaction with a tree with unlabeled dichotomous nodes (e.g. the phylogeny part at high levels of the Tree A), with a similar tree that is strictly a classification (Tree B). Both are very large trees, but names are largely similar and a simple correspondence between organisms with the same name can be assumed. Some groups have additions of names, others have different topologies, such as Reptilia as explained above.

A second example we explored was the idea of allowing navigation in a folk tree to be coupled with navigation in a scientific tree. Here the topologies are not only largely different, with the folk tree expected to have fewer hierarchical levels, but the mapping of nodes is likely to be one (in the folk tree) to many (in the scientific tree). We created Tree C by taking a sample of 165 organisms found in suburban and rural areas on southeastern Michigan, taken from a publicly available website (Espinosa *et al.*, 2002). We chose this dataset as an example of a set of organisms with relevance to non-experts. A non-biologist on our project sorted the leaf nodes into simple groups and created higher level names to apply to these groups. A biologist on our project then examined the tree and created appropriate mappings to nodes in the vastly larger Tree A. The three trees used in these visualizations are summarized in Table 2.

Table 2. Characteristics of datasets used in the study

Dataset	Source(s)	Number of nodes	Number of levels	Number of children per node
Tree A	Multiple	193786	32	Max: 965
				Avg: 5.6
Tree B	ITIS	190266	15	Max: 965
				Avg: 5.4
Tree C	BioKIDS, folk	165	5	Max: 17
				Avg: 4.1

Features

DoubleTree displays two trees in upper and lower frames of a window (Figure 3). It preserves the basic interaction of SpaceTree and TaxonTree because to view the children of a node, one clicks on the node. The corresponding node in the other tree is opened and marked in rosy purple. If a node in Tree A has no corresponding node in Tree B, the focus node in Tree A remains at the least inclusive node that does map onto Tree B. If one node maps to several

nodes in the other tree, we outline those other nodes in red, rather than have multiple focus nodes in the same tree.

Users can browse either tree and watch coupled interaction with the other. Because this application was more exploratory, we did not implement all features available in TaxonTree. Branches automatically close in both trees as focus shifts to other branches, an option available in but not the default in TaxonTree. Searching is currently available only for exact Latin names in the top pane, but illustrates how efficient the comparison of a specific part of a tree can be.

DoubleTree does not calculate or indicate where tree topology differs. As in TaxonTree, DoubleTree labels all open nodes, and does not label any closed nodes to avoid overwhelming the user. Similarly, DoubleTree does not draw lines between corresponding parts of trees due to difficulty of distinguishing these lines from those linking nodes in the hierarchy.

RESULTS

DoubleTree offers interactive tree exploration and comparison for users with a need to examine details in trees already known or suspected to be different. The simple, intuitive nature of the interface provides quick answers to basic tree topology questions.

Table 3. Performance comparison of TaxonTree and DoubleTree

Task	TaxonTree local	DoubleTree
Browse to few children	62 ms	125 ms
Browse to many children	94 ms	203 ms
Exact search	1125 ms	359 ms

Table 3 shows the performance differences between the TaxonTree local version and DoubleTree comparing Trees A and B. The browsing time almost doubles compared to TaxonTree, because queries are run on two different databases, but DoubleTree's performance remains reasonable. TaxonTree has a longer exact search time because more time is spent calculating the layout and zooming factor of results so that multiple results fit to screen in the most readable way.

DoubleTree also illustrates a novel approach to providing public access to scientific resources. Navigation in a folk tree coupled with a scientific tree that is a portal to more information can allow users to 1) find information more quickly, even without scientific training, and 2) gain a better understanding of the scientific organization of the data. Even limited interaction with both trees reveals areas of strong concordance (e.g. birds are appropriately

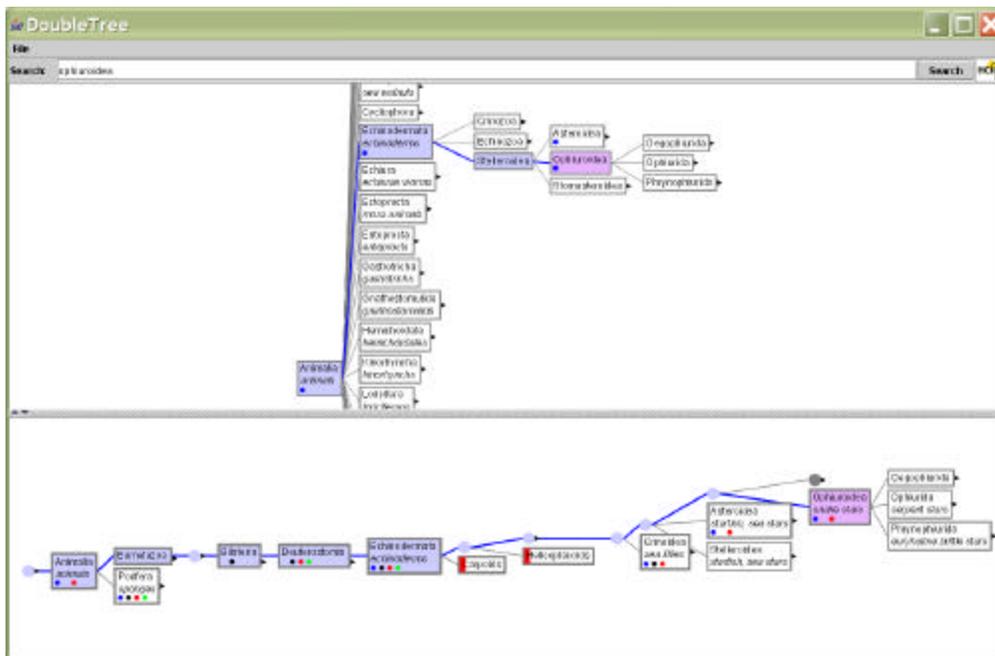


Figure 3 DoubleTree comparing Tree B (from ITIS, above) with Tree A (below)

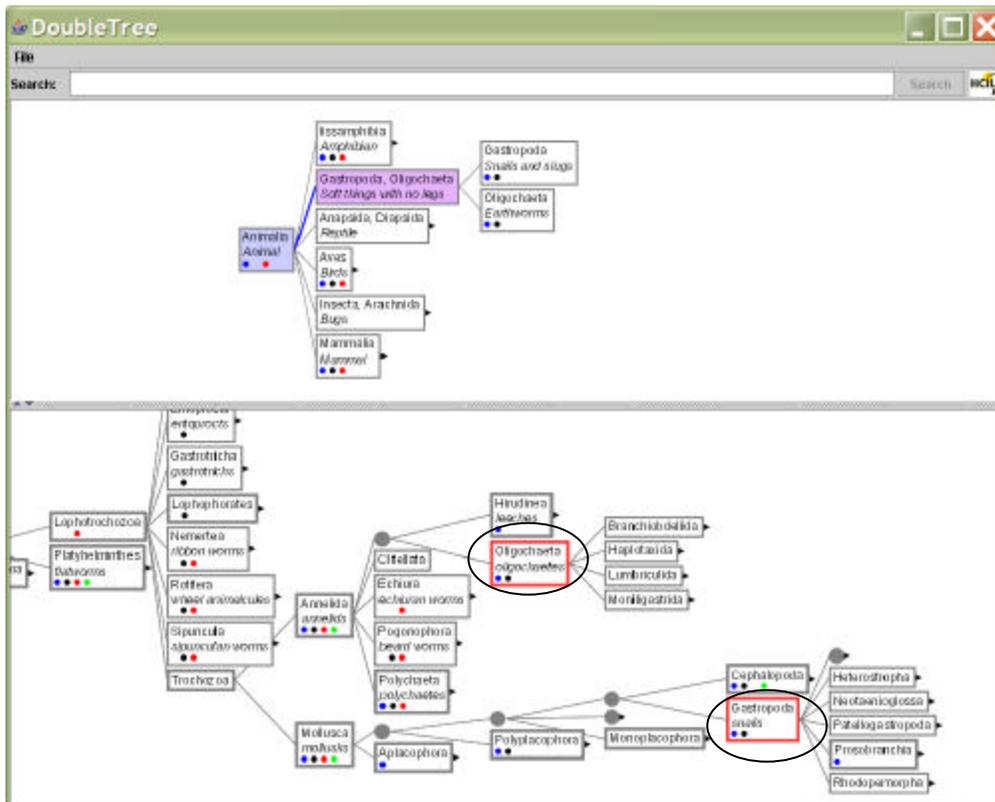


Figure 4. DoubleTree shows differences in tree topology between Tree C (top, the folk tree) and Tree A (below, a composite scientific tree). “Earthworms and slugs” group in Tree C maps onto two distant groups in Tree A (circled).

grouped in the folk tree) and areas of distinct difference in the topologies (Figure 4).

Coupling the interaction of two trees is difficult when node names differ, or when there is a one to many mapping. We have proposed one set of solutions to these problems. These solutions work well for the comparison of Tree A and Tree B, but work less well for Tree C because so many of its nodes map onto multiple nodes.

Setting aside the conceptual, biological differences in Tree A and Tree B, a comparison of these trees (Figure 3) illustrates the logistical trade-offs inherent in visualizing phylogenies versus classifications. As a strict classification, Tree B is bushy near the root of the tree, meaning it has many children per node, making it a broad and shallow tree. This requires large amounts of vertical space if names are rendered as a readable, vertical list. Scrolling or zooming and panning were immediately necessary to view all nodes. Tree A's bifurcating nodes, on the other hand, allow it to be less bushy but deeper. Scrolling or zooming is not immediately necessary, but viewing lower levels requires large amounts of horizontal space if depth is portrayed on that axis. Furthermore, tree level is more obvious in a visualization of Tree B (upper pane of Figure 3), and not apparent in Tree A (lower pane of Figure 3).

DISCUSSION

Value of general approach

Our case study demonstrates an approach to biological database visualization that would be particularly valuable for publicly available database and portal front-ends. Previous controlled studies with SpaceTree indicated that users were better able to revisit already visited nodes and preferred its interaction over that of hyperbolic trees and Microsoft Windows Explorer interfaces on the same data. (Plaisant *et al.*, 2002). Controlled studies are needed to extend the present case study to see if users similarly prefer a TaxonTree-style interface to online resources over the current standard of form-fill in pull-down menu interfaces.

Combined phylogeny and classification

In the scientific realm, we expect that biologists may be primarily concerned with interacting with either phylogenetic trees or classifications. However, several factors combine to argue that the future may hold elements of both in general use. First, the exponential increase in genetic data and its analysis mean that phylogenetic hypotheses can be, and already are, generated in abundance. Second, the trend in biological education is to encourage "tree-thinking," so that students and scientists can take full advantage of the predictive power of evolutionary relationships (Freeman, 2001). Following from these first two factors is a third: the increase of named groups and emphasis on tree thinking

rather than the Linnean hierarchies have led to a decreased emphasis on rank relationships. In fact, because ranks have no equivalent meaning across taxa, they are not always very useful, and often misleading in that they imply parallels where none exists. Yet, we argue that it can be useful to use a classification as a placeholder for an area of the tree of life for which the phylogeny is uncertain.

Our work illustrates that a combined approach can be successful. We are able to take steps towards integrating knowledge across the whole tree of life, which would not be possible without the classification. Because it is relatively comprehensive as a classification and contains some phylogenetic information, Tree A now serves as the backbone for the Animal Diversity Web and is updated there (Myers, 2003).

Folk taxonomies and common names

The DoubleTree approach of showing coupled navigation in folk and scientific trees may be useful in educational contexts where students need to move from an intuitive understanding of biological relationships to a scientifically-constructed hierarchy. It could be argued that users with no need to know the scientific hierarchy would only need the folk tree. However, it is unclear if many biologists would encourage the use of representations of organismal classification that are non-scientific.

Both TaxonTree and DoubleTree serve as effective illustrations to the general public that common names and ranks, while often useful, are not predictable indicators of biological significance.

Beyond hierarchies

The simplicity of the TaxonTree and DoubleTree approach appears well-suited for navigation and understanding of taxonomic hierarchies. Would a similar approach be useful for allowing users to search, browse, and understand biological data with reticulated or network relationships? A significant amount of research has been conducted in graph visualization (reviewed in Herman *et al.*, 2000), but good interactive solutions remain a challenge. Our next steps are to draw on our experiences with TaxonTree and DoubleTree to develop and test new interaction techniques for graphs and ontologies.

ACKNOWLEDGEMENTS

We thank Catherine Plaisant and Jesse Grosjean for their assistance related to our transformations of SpaceTree. We thank Jeff Jensen for content advice and access to his undergraduate course. This work was supported in part by NSF #0219492 (PI Bederson). Preparation of Tree A was also supported by the Interagency Education Research Initiative (IERI) grant REC-0089283 (PI's Songer and Myers).

REFERENCES

- Bisby,F.A. (2000) The Quiet Revolution: Biodiversity Informatics and the Internet. *Science*, **289**, 2309-2312.
- Day,W.H.E. (1985) Optimal algorithms for comparing trees with labeled leaves. *J. Classif.*, **2**, 7-28.
- Espinosa,R., Dewey,T., Hammond,G., Parr,C.S., Jones,T., Pappas,J., and Myers,P. (2002). BioKIDS Critter Catalog. <http://www.biokids.umich.edu>.
- Freeman,S. (2001) *Biological Sciences*. Prentice Hall.
- Graham,M., Kennedy,J., and Hand,C. (2000). A comparison of set-based and graph-based visualisations of overlapping classification hierarchies. *Proc. Advanced Visual Interfaces (AVI 2000)*, ACM Press, 41-50.
- Herman,I., Melançon,G., and Marshall,M.S. (2000) Graph visualization and navigation in information visualization: a survey. *IEEE Transactions on Visualization and Computer Graphics*, **6**, 24-43.
- Hong,J.Y., D'Andries,J., Richman,M., and Westfall,M. (2003). Zoomology: Comparing two large hierarchical trees. *Proc. Poster Compendium of the IEEE Symposium of Information Visualization (InfoVis 2003)*, 120-121.
- Lamping,J., Rao,R., and Pirolli,P. (95). A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. *Proc. Human Factors in Computing Systems (CHI 95)*, ACM Press, 401-408.
- Lee,B., Parr,C.S., Campbell,D., and Bederson,B. (in press). How Users Interact with Biodiversity Information Using TaxonTree. *Proc. Advanced Visual Interfaces (AVI 2004)*, ACM Press.
- Maddison,D.R. (ed.) (2003). Tree of Life. <http://www.tolweb.org>.
- Medin,D.L., Atran,S. (1999) *Folkbiology*. MIT Press, Cambridge, MA.
- Morse,D.R., Ytow,N., Roberts,D.M., and Sato,A. (2003). Comparison of multiple taxonomic hierarchies using TaxoNote. *Poster Compendium of the IEEE Symposium of Information Visualization (InfoVis 2003)*, 126-127.
- Munzner,T., Guimbretière,F., Tasiran,S., Zhang,L., and Zhou,Y. (2003). TreeJuxtaposer: Scalable tree comparison using focus+context with guaranteed visibility. *Proc. 30th annual conference on computer graphics and interactive techniques. (SIGGRAPH 03)* ACM Press, 453-462.
- Myers,P. (ed.) (2003). Animal Diversity Web. <http://www.animaldiversity.org>.
- Plaisant,C., Grosjean,J., and Bederson,B.B. (2002). SpaceTree: supporting exploration in large node link tree, design evolution and empirical evaluation. *Proc. Information Visualization Symposium (InfoVis 2002)* New York: IEEE, 57-64.
- Robertson,G.G., Mackinlay,J.D., and Card,S.K. (91). Cone Trees: Animated 3D Visualizations of Hierarchical Information. *Proc. Human Factors in Computing Systems (CHI 91)* ACM Press, 189-194.
- Shneiderman,B. (1992) Tree Visualization with Treemaps: A 2-D Space-Filling Approach. *ACM Transactions on Graphics*, **11**,92-99.
- Uetz,P. (ed.) (2003). European Molecular Biology Laboratory Reptile Database. <http://www.reptile-database.org>.
- Wilson,D.E. and Reeder,D.M. (1993) *Mammal Species of the World*. Washington, DC, Smithsonian Institution Press.