Visual Comparison and Exploration of Natural History Collections

Martin Graham & Jessie Kennedy
School of Computing, Napier University
10 Colinton Road, Edinburgh, EH10 5DT
United Kingdom
++44 (131) 455 2749
{m.graham, j.kennedy}@napier.ac.uk

Laura Downey
Biology Department, University of New Mexico
1 University of New Mexico, Albuquerque
NM 87131-0001, United States
++1 (505) 277-3157
ldowney@lternet.edu

ABSTRACT
Natural history museum collections contain a wealth of specimen level data that is now opening up for digital access. However, current interfaces to access and manipulate this data are standard text-based query mechanisms, giving no leeway for exploratory investigation of the collections. By adapting previous work on multiple taxonomies we allow visual comparison of related museum collections to discover areas of overlap, naming errors, and unique sections of a collection, indicating areas of specialisation for individual collections and the complementarities of the set formed by the collections as a whole.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Graphical user interfaces, interaction techniques.*

General Terms
Human Factors.

Keywords
Multiple tree visualization, natural history collections, animation, taxonomy.

1. INTRODUCTION
Recently a strong focus in information visualization has emerged on the display and manipulation of biologically-sourced data, including genetic sequencing data [8], micro-array data [1], large scale taxonomies [3] and phylogenies [7]. Many museums also hold large amounts of biological data in the form of natural history collections of preserved specimens. These form a valuable resource for a wide spectrum of biology and ecology researchers - the specimens themselves are a primary source of data for taxonomists, whilst their recorded distributions over space and time in different collections can act as a record of past biodiversity states [12].

Interfaces to these information repositories are however not geared towards exploratory interaction; electronic portals that are available consist of query forms that require potential users to specify textually what they are looking for. Furthermore, despite efforts such as DiGIR [2], many other collections are unavailable through digital access and hence require manual searching [6]. Thus, it is currently difficult to get a feel of the coverage within and between natural history collections, to discover where one museum's strength lies and where collections complement or overlap each other.

We have previously demonstrated the utility of a visualization for comparing multiple taxonomies [3] for both large-scale comparisons (1 million plus nodes in total) and more involved scrutiny between smaller taxonomies that contain synonymy (the existence of various degrees of explicit matching between differently named nodes). By adapting our multiple taxonomy visualization to museum collection data we can reveal patterns in museum collection data such as areas of local expertise, complementary and overlapping areas of collections, misnamed taxa and under-represented areas in current collections.

For demonstration purposes we accessed data from the MANIS (Mammal Networked Information System) [5; 11] databank which holds electronic data for mammal collections in seventeen North American institutions plus the ITIS reference taxonomy [4]. Using the MANIS data portal we obtained thirteen sets of collections data (when we gathered the data, four were unavailable) plus the ITIS taxonomy. The collections are organized taxonomically – for the most part they classify specimens along the lines of the ITIS taxonomy – and contain at the bottom level a per-species count of the physical specimens held at the institution.

2. METHOD
The visualization displays each classification as a compressed top-down tree representation, using an abutment style of representing parent-child links similar to that found in Stasko and Zhang's radial tree visualization [10] and Sifer's web log visualization [9] as shown in Figure 1. Groups of leaves are organized in grid patterns to maximize use of space. This representation style is carried out for all the collections on display, with the screen space divided up between the classifications present.

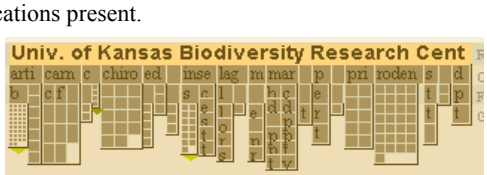

Figure 1. Individual collection representation.

AVI '06, May 23-26, 2006, Venezia, Italy.

A small toolbar tracks the mouse pointer across the collection representations, positioned in the upper left-hand corner of the current tree, and supplies options to move towards the root, jump directly to the root, hide/expand the entire tree, show unique nodes and set/unset the tree as the prime tree (which gives one collection more space to layout in). This floating toolbar device was adopted as having controls displayed permanently on a per-tree basis produced a large amount of screen clutter. Drill-down navigation and selection is performed by clicking over a node with the left and right mouse buttons respectively. The collection representations are linked by cross-highlighting - nodes that have been selected in one collection are highlighted where they occur in other, multiple, collections. A more detailed description of the interaction mechanisms and underlying data representation can be found in a previous paper [3].

The visualization had to be adapted in a number of ways to accommodate the collections data and the operations we wished to perform on the data. Firstly, though specimens can be distinguished individually in MANIS, we decided to present them as aggregated totals. As our previous visualization [3] displays overlaps between taxonomies/trees/classifications, specimen-level data across collections would be superfluous as the same physical specimen cannot be kept in two different places at once; thus the specimens themselves would not be active elements when comparing collections. As we were not representing individual specimens, we incorporated a 'count' value into leaf nodes that would reflect the number of specimens the leaf node ultimately acted as a classifier for (leaf nodes were usually the taxa at the species and sub-species ranks), and extended a visual sort functionality to allow ordering on these values.

Secondly, we incorporated a function to directly find unique elements within a classification, available from the toolbar (see top left-hand corner of Figure 2). Further, we introduced a control to toggle the presence of trees within the visualization. This is not a purely visual control for releasing screen space (there is an expand/collapse control in the floating toolbar for that purpose), rather this control determined whether trees were involved in searches for unique elements. Thus, removing or adding collections from the visualization with this control enabled us to find nodes that are unique to a collection given a particular subset of other collections.

2.1 Unique aspects of collections

One function that individual museum collections carry out is to act as centres for species found within their geographical 'catchment'. By introducing the facility to find unique elements within a collection, we are able to reveal to what extent a particular collection performs this function. Selecting the 'find unique' function from the toolbar highlights all the taxa in a collection that are unique to that institution, either through incorrect spelling, or more interestingly, because they comprise a specifically local set of taxa or were brought back from a foreign field trip. Obviously no cross-highlighting occurs in the other collection hierarchies, which confirms the taxa as a unique set to the collection. As an example, in Figure 2, the California Academy of Sciences collection has been highlighted with respect to taxa that do not occur in the other collections present. Brushing the highlighted nodes reveals misspellings, but also intriguing examples such as that shown, where the *Lepus Californicus* species hold several sub-species that do not occur elsewhere. Such unique taxa are important; they are not anomalies but areas where the collection forms a niche resource in that particular species or genus. For instance, as this collection is held at a Californian institution it indicates a degree of local specialism for this species. It could be argued that it would be more cost effective for museums to concentrate on these particular branches and forge themselves areas of expertise rather than replicating holdings that are covered in several other institutions.

2.2 Comparison against a reference taxonomy

Another function is to compare the ITIS reference taxonomy against the museum collections. When part of or the whole reference taxonomy is highlighted by selecting one of its taxa, the consequent cross-highlighting reveals the coverage or omission of the reference taxonomy in the museum collections' structure. Figure 3 shows that many of the collections contain 50% or more of taxa which do not correlate to the ITIS taxonomy. Most of this discrepancy is revealed to be at the sub-species rank, as ITIS stops at the species level and does not differentiate further. However, the highlighted taxa at the species rank or above reveal misspellings, local collections and re-structuring that differs from the reference taxonomy. Figure 3 reveals that the Bishop Museum holdings, amongst others, contain the order *Marsupialia*, whose highlighted contents are scattered across several different orders in ITIS. This can be explained by ITIS being a North-American centred organization, and *Marsupialia*, being indigenous to

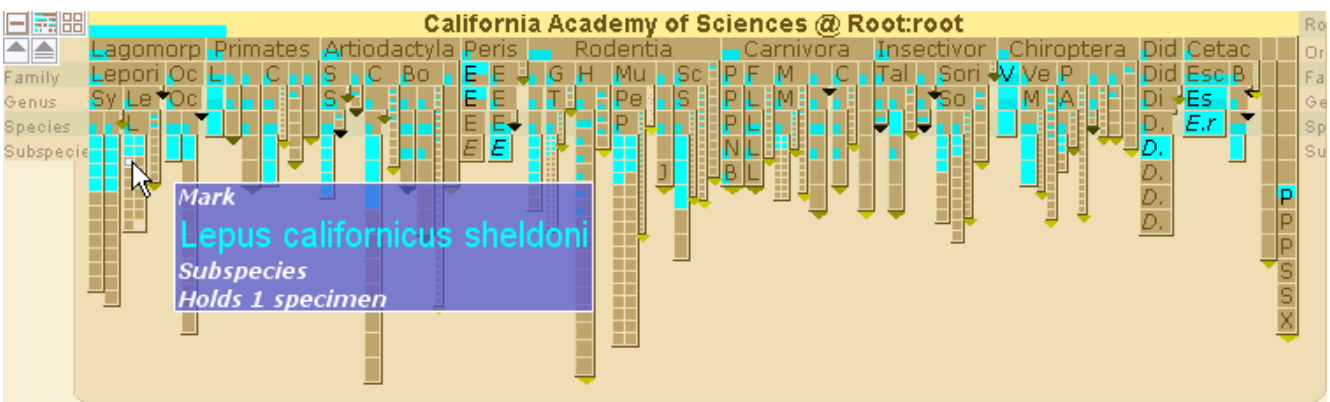

Figure 2. Unique taxa in the California Academy of Sciences collection indicates some local specialisms.

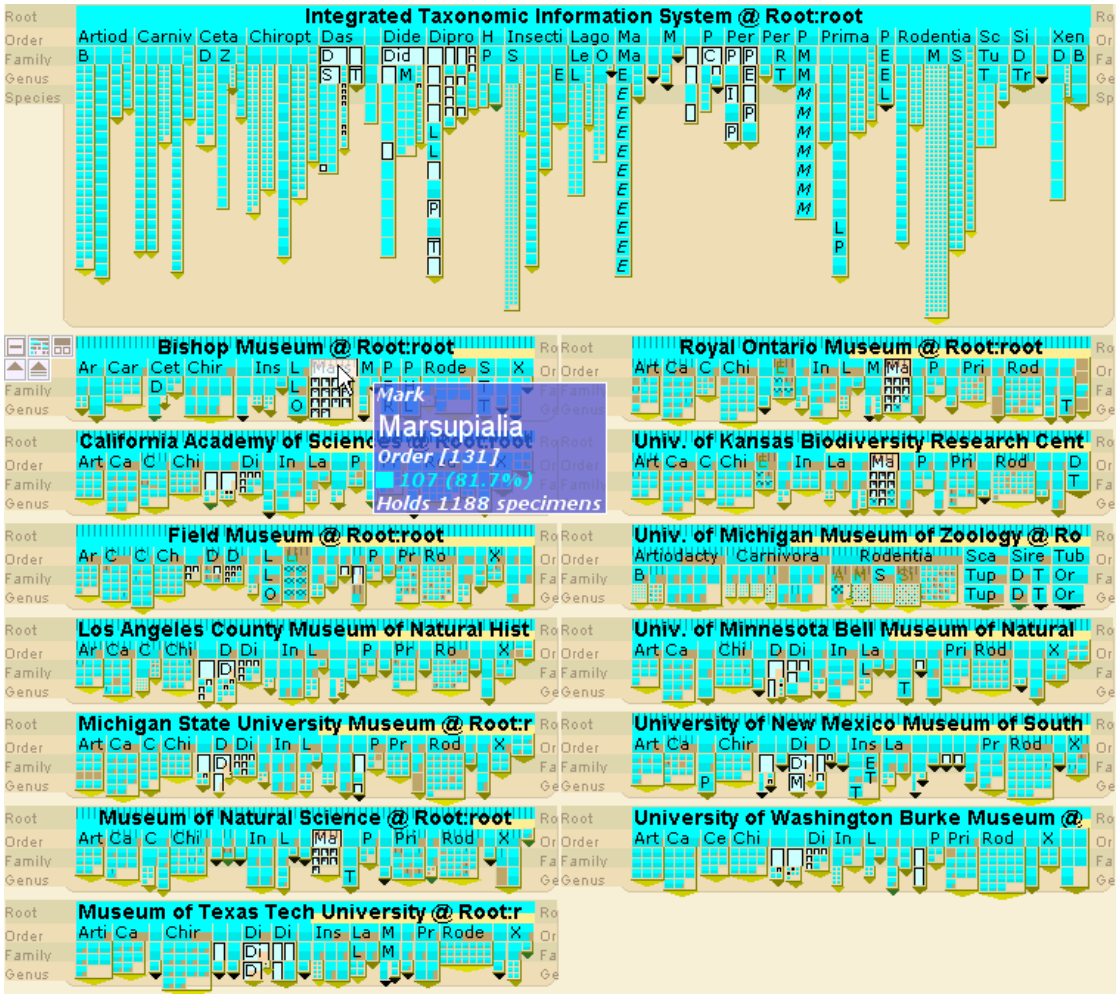

Figure 3. ITIS reference taxonomy distribution across collections.

Australasia, has not been included in the ITIS taxonomy. This indicates that several collections have decided that they have enough material in *Marsupialia* that following the ITIS taxonomy is no longer appropriate.

Texturing is applied to nodes where structural changes have occurred compared to the originally selected classification. This explains the patterning in the top half of the Marsupialia node, the vertical lines indicating a change within it compared to the ITIS taxonomy. The hatching on the child nodes of Marsupialia mark the points of the actual changes, indicating these taxa are now classified under a different parent taxon to those in ITIS. What stands out when there are many changes present are the areas without texturing, indicating no change. In Figure 3 there are three collections that contain no hatching (Michigan State, Museum of Texas Tech and University of Washington), indicating that the ITIS structure they have followed has not been re-arranged or fused with parts of other classifications. This provides useful feedback to the collections managers as to where they need to revise the naming of their index taxonomies.

The reciprocal comparison, that of a collection against ITIS, can show where a collection is missing parts of the reference taxonomy and thus gaps in their specimen collection that they may wish to consider holding.

2.3 Operating on a subset of collections

If a query is concerned with only a few of the collections available then the unnecessary collections can be removed from consideration (and other collections brought back into frame).

For example, to discover the overlap in the *Lepus californicus* species between the two Californian institutions - the California Academy of Sciences and the Los Angeles County Museum of Natural History - we remove the other trees from the display and drill down to the *Lepus* taxa. We then highlight the unique nodes within *californicus* for both collections, as shown in Figure 4. This shows that the Academy of Sciences contains seven sub-species that Los Angeles lacks while the reciprocal display shows the Los Angeles holds specimens for only one subspecies that the Academy does not. Following the earlier discovery that the Academy contained several unique *Lepus californicus* sub-species compared to all the collections, any *Lepus* researchers could be advised to head there rather than to Los Angeles to find interesting specimens.

2.4 Animating & Sorting

Animation was introduced to ease disorientation effects when drilling down and zooming out in larger trees. The methodology, as used in various graph and tree visualizations, was to fade out

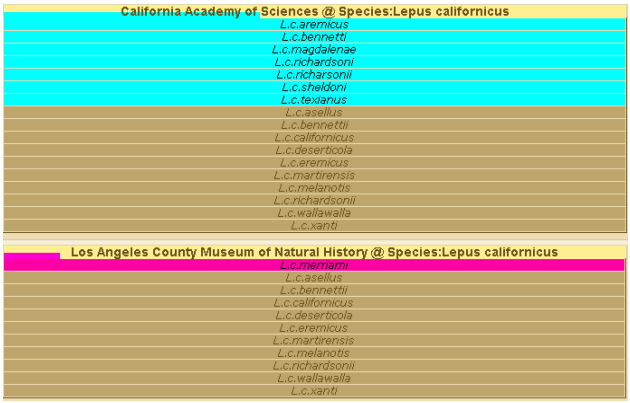

Figure 4. Finding unique elements based on a subset of the whole collection set.

elements that would not be present in the new layout, fade in those that will be newly introduced and move elements present in both the old and new views between their respective positions.

This also allowed animation of the visual sort. To make the display more informative the classification representations can be rearranged internally by a number of sort metrics to allow ordering by taxa size, percentage of taxa selected etc, with the default being alphabetical by name. Animating this operation allows the type and degree of difference between alternative orderings to be observed. For instance, switching the order of "taxa size" and "specimen total" would reveal the extent of correlation between the total of taxa and specimens contained in subtrees. In most cases animation shows clearly that the number of taxa and number of specimens under a given taxon tend to be tightly correlated, so individual species or genera with extreme numbers of specimens will move sharply against a mostly static background.

3. EVALUATION

We conducted initial evaluations with collections managers in order to gauge the usefulness and applicability of the collections-based visualisation. Each participant was given a profiling instrument to complete, after which the visualisation tool was demonstrated via a series of likely tasks of interest. Next, participants were asked to assign a perceived overall usefulness rating of the visualisation tool (average: 3.4/5.0). The final step in the evaluation process was a facilitated discussion of user interface issues, clarifying questions and feature enhancements. Analysis of this qualitative feedback provided input for user interface improvements and a list of useful feature enhancements. Amongst others, this discussion elicited perceived benefits on the ability to identify incorrect data and that it would be easier to explore the data through the visualisation than with a regular text-based search engine. The most desirable enhancements voiced were for a geography-based index to correlate chosen specimen groups against and access to specimen details such as collectors' names. Future plans include regular usability testing to measure participant effectiveness and satisfaction whilst directly interacting with the collections data.

4. CONCLUSIONS

We have shown that a visualization of museum collection data can show areas of overlap and complementary coverage between multiple collections. The ability to quickly find resources unique to a collection reveals specialist knowledge within the holding institution, usually either locally-sourced specimens or those from a specific field trip. Unique elements can also be due to misnomers and comparison to a reference taxonomy can indicate where a collection's nomenclature may need updated. Such operations are not feasible using a textual query interface, where misnamed data will simply not match and be thought of as non-existent, and data unique to a particular collection must be either hunted down through exhaustive search or known about in advance to be directly queried.

5. ACKNOWLEDGMENTS

This work was funded by the SEEK project (Science Environment for Ecological Knowledge) - NSF Grant award 0225676. Thanks to Jim Beach for feedback and suggestions in requirements and testing.

6. REFERENCES

[1] Craig, P., Kennedy, J. and Cumming, A. Animated Interval Scatter-plot Views for the Exploratory Analysis of Large Scale Microarray Time-course Data. *Information Visualization*, *4*, 3 (Sept. 2005), 149-163.

[2] DiGIR. Distributed Generic Information Retrieval. Retrieved 11 Jan, 2006, from http://digir.net

[3] Graham, M. and Kennedy, J. Extending taxonomic visualisation to incorporate synonymy and structural markers. *Information Visualization*, *4*, 3 (Sept. 2005), 206-223.

[4] ITIS. Integrated Taxonomic Information System. Retrieved 11 Jan, 2006, from http://www.itis.usda.gov/index.html

[5] MANIS. Mammal Networked Information System. Retrieved 12 Jan, 2006, from http://manisnet.org/

[6] O'Connell Jr., A. F., Gilbert, A. T. and Hatfield, J. S. Contribution of Natural History Collection Data to Biodiversity Assessment in National Parks. *Conservation Biology*, *18*, 5 (Oct. 2004), 1254-1261.

[7] Parr, C. S., Lee, B., Campbell, D. and Bederson, B. B. Visualizations for taxonomic and phylogenetic trees. *Bioinformatics*, *20*, 17 (Nov. 22, 2004), 2997-3004.

[8] Sairaya, P., Lee, P. and North, C. Visualization of Graphs with Associated Timeseries Data. In *Proc. of IEEE InfoVis* (Minneapolis, Minnesota, USA, October 23-25, 2005), IEEE Computer Society Press, 225-232.

[9] Sifer, M. Exploring Web Site Log Data with a Multi-Classification Interface. In *Proc. of IEEE Conference on Information Visualisation* (London, UK, 16-18 July, 2003), IEEE Computer Society Press, 94-101.

[10] Stasko, J., Catrambone, R., Guzdial, M. and McDonald, K. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human-Computer Studies*, *53*, 5 (Nov. 2000), 663-694.

[11] Stein, B. and Wieczorek, J. Mammals of the World: MaNIS an example of data integration in a distributed network environment. *Biodiversity Informatics*, *1*, 1 (2004), 14-22.

[12] Suarez, A. V. and Tsutsui, N. D. The Value of Museum Collections for Research and Society. *BioScience*, *54*, 1 (Jan. 2004), 66-74.