Comparison of multiple taxonomic hierarchies using TaxoNote

David R. Morse[1]
The Open University, United Kingdom

Nozomi Ytow [2]
University of Tsukuba, Japan

David McL. Roberts[3]
The Natural History Museum, United Kingdom

Akira Sato[4]
University of Tsukuba, Japan

Abstract

In this paper we describe TaxoNote Comparator, a tool for visualising and comparing multiple classification hierarchies. In order to align the hierarchies, the Comparator creates an integrated hierarchy containing all the taxa in the hierarchies to be compared, so that alignment of the hierarchies can be maintained. A table of assignments reports the taxonomic names that are common to all hierarchies and the differences between them, which facilitates structural comparisons between the hierarchies.

CR Categories: I.3.6 [Computer Graphics]: Methodology and Techniques – Graphics data structures and data types; J.3 [Life and medical sciences]: Biology and genetics

Keywords: taxonomy, nomenclature, visualisation, rough set theory, formal concept analysis.

1. Introduction

Recent work on modelling taxonomic names and their relationships has highlighted the need to capture the multiple names and hierarchies that exist in nomenclature. A number of projects have considered this problem, including Nomencurator [Ytow et al. 2001] and Prometheus [Pullan et al. 2000]. Data models incorporating multiple hierarchies are crucial in facilitating the effective integration of biodiversity data from diverse sources, since multiple and overlapping taxonomic concepts must be tracked, as well as the names that have been applied to these concepts. Equally important are visualisations which permit the comparison and exploration of several hierarchies simultaneously.

In this paper we will describe an extension to our previous work on the Nomencurator data model [Ytow et al. 2001] by giving an overview of the visualisation and comparison tools within TaxoNote. TaxoNote (short for **Taxo**nomist's **Note**book) is a graphical user interface to the Nomencurator data structures.

2. Hierarchy visualisation and comparison

The TaxoNote Comparator hierarchy visualisation and comparison tool is shown in Figure 1. The display is divided into three:

- A **Query** panel can be used to search the displayed hierarchies for particular taxonomic names, by text entry.
- A **Hierarchy Comparison** panel shows the two hierarchies that are being compared (centre and right) and an 'integrated view' (left) where the hierarchies have been merged into one, composite, hierarchy. An additional pane would be added for each hierarchy being compared by the application. The hierarchy comparison panel provides a list of siblings and children of a taxon. It also captures the parent taxon and the path to the hierarchical root. These may not be displayed if there are many siblings or children of a node, in which case a **Pop-up panel** gives a short summary of the path to the root.
- An **Assignment Table** at the bottom shows various alternative views of where names that appear in the hierarchies are assigned. It contains information on the parent taxon and potential equivalence of taxon concepts depending on its modes. While the **Hierarchy Comparison** panel gives a top-down oriented view, the **Assignment Table** gives a bottom-up oriented view.

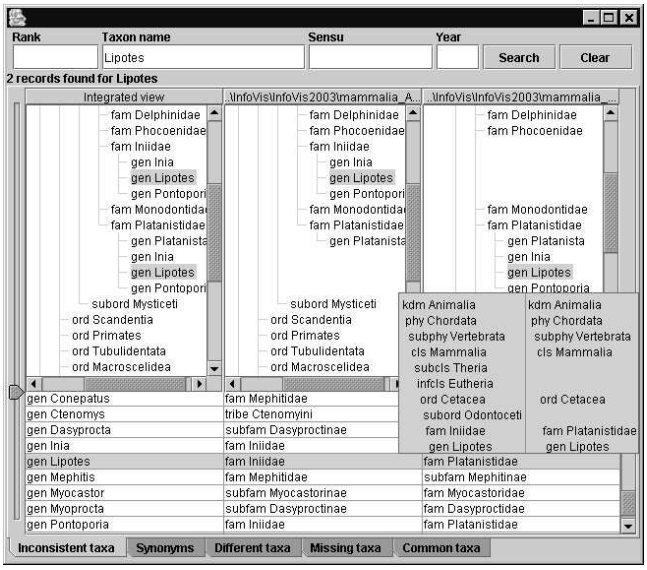

Figure 1. The TaxoNote Comparator hierarchy visualisation and comparison tool.

2.1. The Query Panel

In large data sets, efficient search tools are necessary to focus the display and the user's attention on the area of interest. Additional fields to the taxon Name are included as potential query fields in order to refine the search. These search fields are metadata which are important in modelling multiple taxonomic hierarchies, since they allow you to compare, distinguish between and reconcile different taxonomic opinions of the taxon concepts that are linked to the same taxonomic name.

2.2. The Hierarchy Comparison Panel

In Figure 1, we prefixed all names with an abbreviated form of the taxonomic rank as an aid to navigation and comparison. We chose

[1]e-mail: d.r.morse@open.ac.uk [2]e-mail: nozomi@biol.tsukuba.ac.jp
[3]e-mail: dmr@nhm.ac.uk [4]e-mail: akira@cc.tsukuba.ac.jp

an indented representation for the hierarchies because this is familiar to taxonomists and to most computer users through applications such as Microsoft Explorer. As with that interface, additional levels of the hierarchy can be expanded and contracted at will. While other representations such as Hyperbolic Trees and TreeMaps [Bederson et al. 2002; Graham and Kennedy 2001] may have a higher information density, it is important that the names retain their visibility and readability at all times. The hierarchies and integrated view can be scrolled in concert by holding down the middle mouse button while any of the hierarchy display panes is scrolled. This facilitates the search for a particular taxon and the structural comparison of the different hierarchies.

2.2.1 Alignment of taxonomic names

Core to the alignment problem is establishing the BCN (Best Corresponding Node, see [Munzner et al. 2003]). Ideally, corresponding nodes would represent equivalent taxonomic concepts. Unfortunately the taxonomic concept itself is extremely difficult to pin down [Ytow et al. 2001] and is approximated in one of two ways, either by consideration of the objects (taxa or specimens) included in the concept [Pullan et al. 2000; Munzner et al. 2003] or by analysis of the attributes of the taxon, i.e. the shared characters of the group. The former method is very sensitive to the contained set being incomplete for any reason, and data for the latter method are rarely available. Other proxy measures of the taxon concept have to be combined to establish the BCN, which include the hierarchical position (parent list), the included objects (the child list), but interpreted in a flexible manner, where positive matching counts for more than missing data and absence of conflict counts in favour, conflict against. This set of relationships is subtle and is currently being explored using rough set approximations and formal concept analysis [Yao et al. 1997].

In order to align the two hierarchies and to maintain their alignment while the display panels are scrolled, a consensus hierarchy is constructed from the source hierarchies that are being compared. This is shown in the left hand pane in Figure 1, as the Integrated View. In the Hierarchy Comparison panel, rows which are aligned have the same names in the same hierarchical position in both hierarchies (e.g. family Phocoenidae in Figure 1). Rows which are not aligned are indicative of names missing from one hierarchy, perhaps because they are newly created (e.g. family Iniidae) or names whose hierarchical position has changed from one hierarchy to the other (e.g. genus *Lipotes*). The necessary inclusion of duplicates of a name has the potential to be a way of indicating regions of difference between trees. Indeed, an estimate of the number of incompatible views can be obtained by simply counting the number of duplicate names in the Integrated view.

Construction of the consensus hierarchy requires the establishment of the BCN for each taxon in the Integrated View. Hierarchies proposed by different taxonomists are likely to embrace different taxon concepts that may or may not have the same name. Therefore, establishing node equivalence is not trivial and we are still working on algorithms for constructing the composite hierarchy that is shown in the Integrated View.

2.3. The Assignment Table

The bottom panel contains the Assignment table which consists of a number of organised lists whose purpose is to allow the user to explore the differences and commonalities between taxon concepts in the hierarchies. The table is structured into columns, one for each hierarchy pane. The primary taxon is given on the left, underneath the integrated view while the parent taxon is listed underneath the appropriate hierarchical pane. Tabs at the bottom of the Assignment Table allow the user to see those taxa which are missing from one set or the other ('Missing taxa' tab), while those taxa with different positions are summarised under the 'Different taxa' tab. Other forms of difference are given on the 'Inconsistent taxa' and 'Synonyms' tabs. Finally, those nodes in common are listed under the 'Common taxa' tab.

One use of the Assignment Table is illustrated under the 'Missing taxa' tab by the species *Acomys cineraseus* (in Mammals A) and *Acomys cinerasceus* (in Mammals B), that looks like a spelling error either in the original publication or in the data preparation.

3. The InfoVis 2003 Contest Data Sets

It is our contention that no one tool can solve all visualisations of hierarchical data problems. We have chosen to address one particular type of data – classification hierarchies – which may be characterised as being non-quantitative data. Our approach would need significant additions in order for it to perform well at visualising hierarchically arranged quantitative data; data which is often well suited to visualisations using TreeMaps [Bederson et al. 2002]. Such additions to our system could include colour-coded glyphs or bars alongside, or in place of the text labels.

Classification hierarchies are unusual in that the names in the hierarchies should be unique. The appearance of the same name in different places is indicative of homonymy and is of interest to taxonomists as an area that requires taxonomic revision. In contrast, file system hierarchies are replete with duplicated names. Files called 'index.html' abound in websites – the file logs_A_03-02-01.xml records 3356 occurrences of this file, for example.

In classification hierarchies, the name is just that because of the assumption that taxonomic names in a hierarchy are unique. The position of the name in the hierarchy – the rank – gives extra information about the name. In contrast, in a file system hierarchy, the name consists of the path to the file in addition to the actual file name. While components of the path may give additional information about the file, this interpretation is not as strong as the rank in taxonomy. Clearly very different visualisation techniques are required in order to navigate and compare hierarchies with such different properties.

References

BEDERSON, B. B., SHNEIDERMAN, B. AND WATTENBERG, M. 2002. Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies, *ACM Transactions on Graphics 21*, 4, 833 - 854.

GANTER, B. AND WILLE, R., 1999. *Formal Concept Analysis: Mathematical Foundations*, Springer-Verlag.

GRAHAM, M. AND KENNEDY, J. 2001. Combining linking & focusing techniques for a multiple hierarchy visualisation. In *Fifth International Conference on Information Visualisation*, IEEE Computer Society Press. 425-432.

MUNZNER, T., GUIMBRETIÈRE, F., TASIRAN, S., ZHANG, L. AND ZHOU, Y. 2003. TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility. In *ACM SIGGRAPH*, ACM Press.

PULLAN, M. R., WATSON, M. F., KENNEDY, J. B., RAGUENAUD, C. AND HYAM, R. 2000. The Prometheus Taxonomic Model: a practical approach to representing multiple classifications, *Taxon 49*, 1, 55-75.

YTOW, N., MORSE, D. R. AND ROBERTS, D. M. 2001. Nomencurator: a nomenclatural history model to handle multiple taxonomic views, *Biological Journal of the Linnean Society 73*, 1, 81-98.