

Quantitative Linnean Data Characteristics

Martin Graham

Abstract

The following is an analysis of typical data sets to be used in the course of developing and demonstrating a visualisation for multiple overlapping Linnean taxonomies. We find that such hierarchical data sets tend to distribute taxa internally in a roughly logarithmic fashion: a few parent taxa hold many child taxa, whilst many parent taxa hold only a single or few child taxa. No similar pattern is found for inter-classification concept relationships, with congruent relationships dominating the type of relationship that is declared.

Introduction

When attempting to develop visualisations for Linnean taxonomies one of the most vital tasks is to understand the characteristics of large taxonomic data sets.

In size terms, the larger taxonomies, such as those produced by ITIS [1], contain over a quarter of a million taxa, and so a series of six or seven such taxonomies can easily represent nearly two million node/classification pairings. In tree visualisation, research has attended to the size and depth of trees, resulting in interactive techniques such as drill-down and focus-and-context displays, but attention has not generally been paid to the branching structure of trees, with the exception of binary trees which have inspired a few dedicated visualisations.

The Hollow Curve

Linnean taxonomies have been observed as following a 'hollow curve' distribution [2] - simply put most taxa tend to contain very few sub-taxa, often having only one child taxon, whilst a few taxa contain many sub-taxa, sometimes numbering well into the hundreds. When plotted on a graph of taxa frequency against sub-taxa count this distribution forms a hollow curve shape, as shown in Figure 1, and when plotted on logarithmic scales produces for the most part a straight-line graph, though the exact form of these curves is an open question. Competing models attempt to fit the observed data [3; 4; 5], with power law or lognormal distributions the two most commonly cited matches, but either under or over-estimate the two tails of the distribution. [6] is a wide-ranging discussion of how this distribution pattern, and the subsequent power-law / lognormal argument, has occurred in many different research fields such as finance, network analysis, and computing.

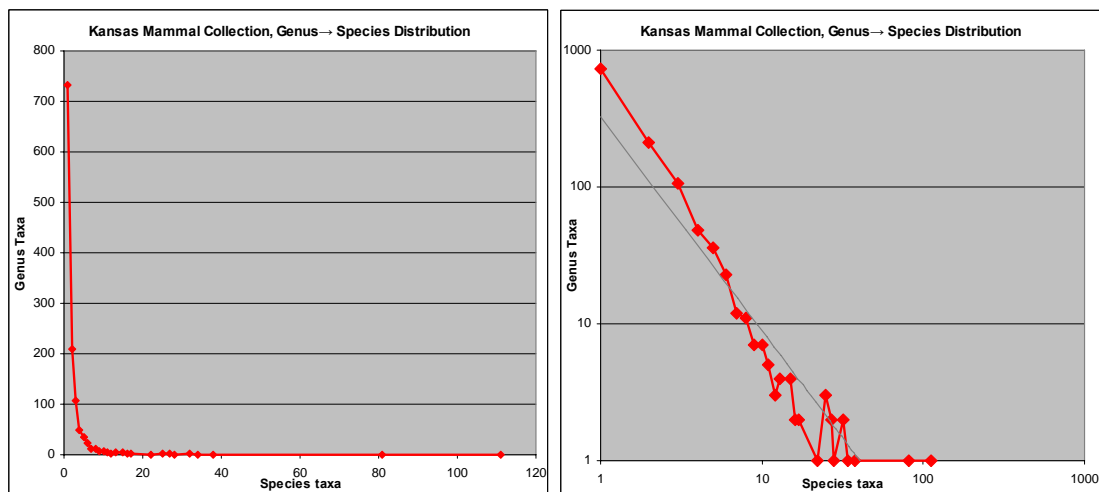


Figure 1. A typical 'hollow curve' distribution. The left-hand graph shows a plot on linear scales, right-hand graph uses log-log axes on the same data.

Various reasons have been proposed for the distribution being as it is; explanations which are grouped by two categories, evolutionary – the effect of evolution on the organisms being classified - and psychological – the effect of those doing the classifying.

In evolutionary arguments, [7] proposes an example of a ‘branching model’, whereby taxa distributions are explained by branching of species at different rates; more quickly branching areas of the tree grow at an exponential rate compared to slower growth areas. Similarly, [5] suggests evolutionary pressures differ throughout the tree of life, so some parts of the tree diversify more than others over a given period of time. Randomness is ruled out as they state that families that contain more than their fair share of genera contain an even more disproportionate number of species so the effect is propagated down the tree. [8] argues that monotypic taxa are old, dying branches that are heading towards extinction, whilst larger groups are examples of recent evolution. However, others argue that monotypic taxa are newer and larger taxa are older in the evolutionary timescale. [9] is representative of an argument that classifications are in fact fractal, driven by evolutionary processes.

Psychological and cognitive arguments from [4; 10; 11] suggest that early taxonomists tended to organise well-studied groups into taxa containing a happy medium of sub-taxa to ease learning, rather than the extremes of single or many sub-taxa. Higher taxa especially have no specific rules for their definition [4], thus they are more susceptible to being sized in line with a taxonomists idea of what a taxon should be composed of, rather than evolutionary effects. [10] argues that early taxonomists specifically tailored their classifications to hold groupings of certain sizes – for instance it can be seen in the Apiaceae graph of Figure 2 that some early classifications had no monotypic taxa whereas the later 1962 classification has many such taxa. It is proposed that these early taxonomies act as seeds for taxonomists, who tend to add new species to existing genera if closely related or place species by themselves if not [12] – this is analogous to the argument of older taxa being larger, except ‘older’ in this case is the date at which the taxa was first recognised rather than an evolutionary timescale. It is also suggested that taxonomists tend to make monotypic groups as it gives them an opportunity to delineate new genera, families etc.

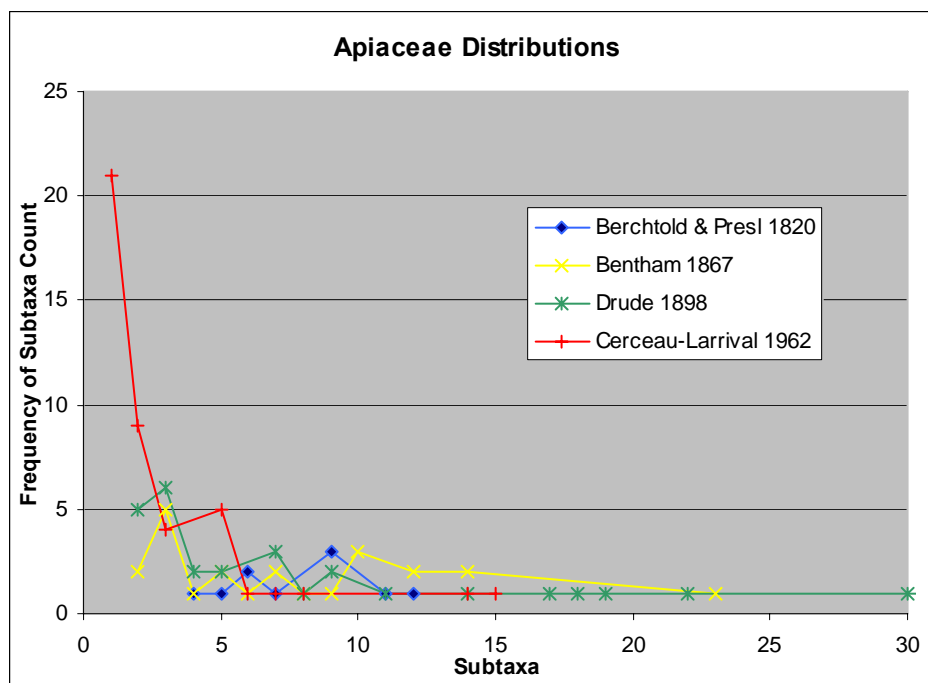


Figure 2. Distribution of subtaxa within taxa for several Apiaceae classifications.

Comparisons between different types of classification to reveal any specific patterns for biological taxonomy have also led to contradictory findings. [13] demonstrates both evolutionary and psychological effects on classification by comparing early taxonomists

classifications on biological and mineral samples. The biological classifications followed the hollow curve distribution whereas the mineral classifications didn't, showing an evolutionary effect on the classifications. Further, they compare all these early classifications with later, modern classifications which again expose a dearth of monotypic taxa in the early classifications. This is explained by the early taxonomists attempting to merge monotypic taxa into larger groups. [14] compares a biological classification with a non-biological soil taxonomy, and discovers that both follow a similar pattern of distribution, leading them state that such patterns are the result of innate cognitive processes used when classifying any group of objects.

[11; 15] both further note that the concept of what constitutes a group at any given rank differs between sub-fields of study in taxonomy and between individual taxonomists.

From this it can be seen that the reason *why* the distribution of subtaxa within taxa in classifications is still not entirely resolved, the fact that they do have this particular distribution is not in question.

Analysis of Current Data

While this distribution pattern has been shown for families such as birds, we are concerned with the integrated taxonomic classifications generated by ITIS et al, so the 2004 ITIS data set was analysed to see if a similar pattern emerged. Immediately from the Figure 3 below, it can be seen that the hollow curve distribution pattern holds for these larger classifications; when distributions are plotted on a log-log axis a straight line is formed (jitter at the bottom is due to differences in values nearer the log-scaled axes having greater spatial resolution.)

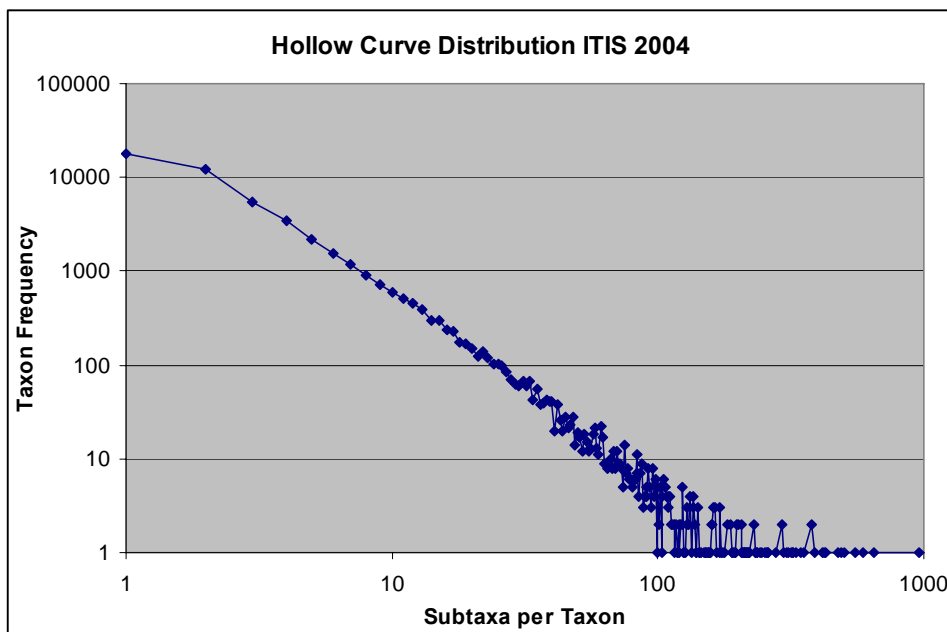


Figure 3. Taxa-subtaxa distribution in the 2004 ITIS classification.

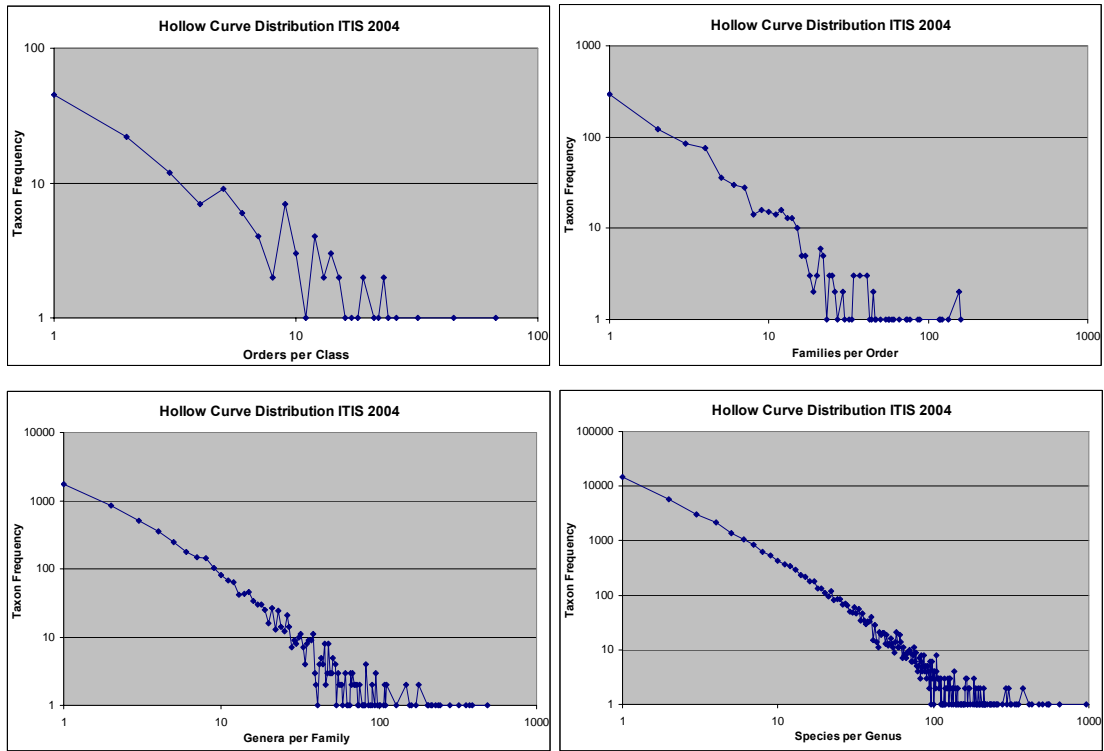


Figure 4. Rank by rank analysis of taxa distribution in ITIS 2004

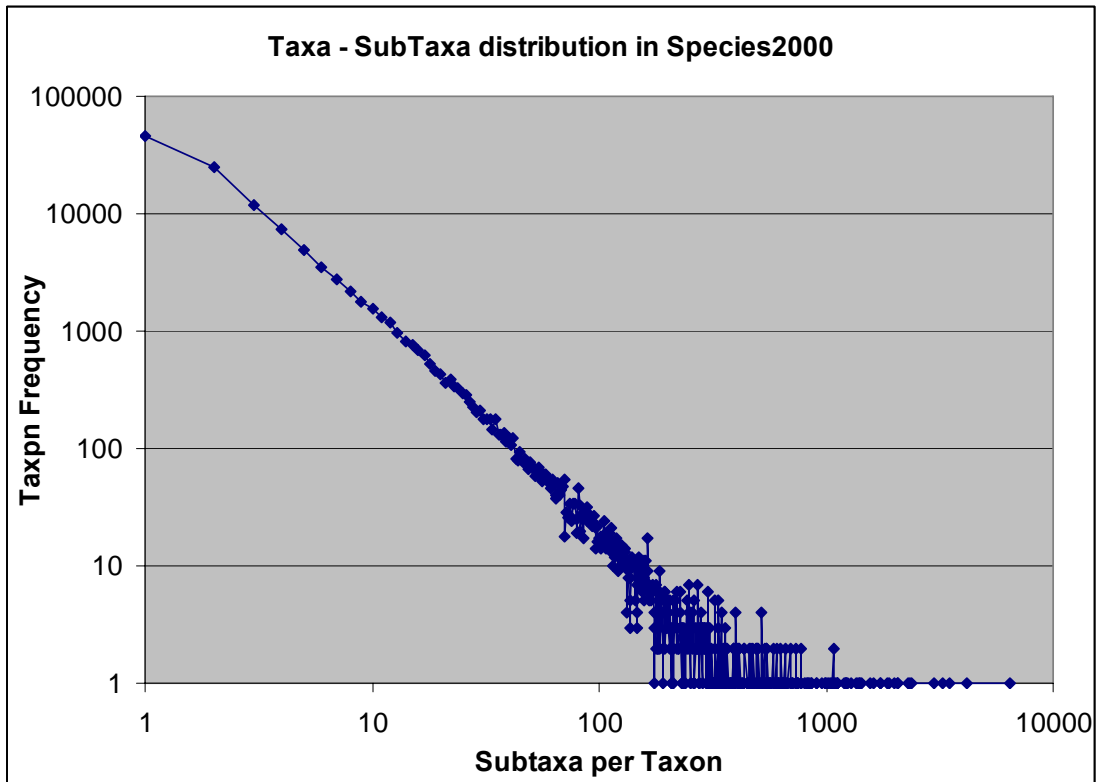


Figure 5. Taxa-subtaxa distribution for Species 2000, a classification of over a million taxa.

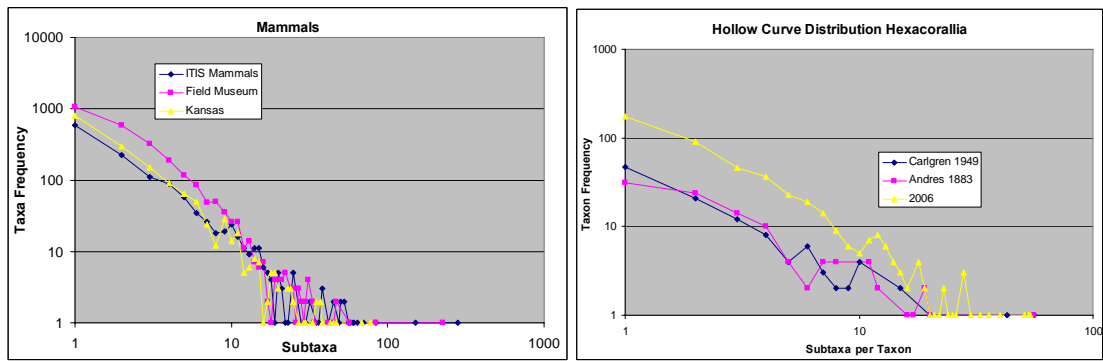


Figure 6. The same pattern emerges in various Mammal collections and Hexacorallia classifications.

Figure 4 shows the same pattern emerging in the ITIS data set for the distribution of species within genera, genera within families and so forth, the pattern becoming clearer as more taxa are considered – there are obviously more genera than orders in the data set. Figure 5 shows a taxa to subtaxa analysis on one Species 2000 dataset, a classification of over a million taxa – again the distribution follows the same pattern. Looking at smaller scales of data sets, Figure 6 shows the hollow curve for a) museum mammal collections and the ITIS mammal classification of a few thousand taxa and b) for various classifications of Hexacorallia consisting of a few hundred taxa.

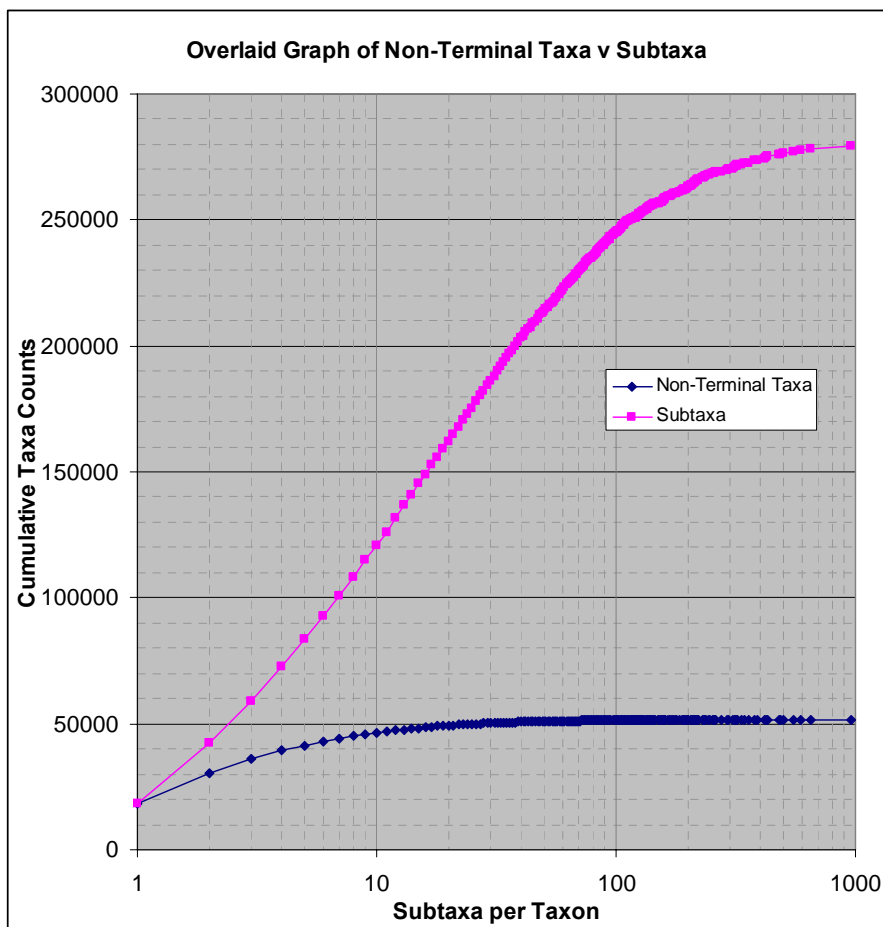


Figure 7. Cumulative graph of sub-taxa distributions, from both the perspective of a parent taxon and a child taxon.

Percentile	ITIS 2004 Subtaxa	ITIS 2004 Non-terminal taxa	ITIS 2004 Genus->Species subtaxa	ITIS 2004 Genus->Species Non-terminal taxa	ITIS Mammals Subtaxa	ITIS Mammals non-terminal taxa
5%	1	1	1	1	1	1
10%	2	1	2	1	2	1
20%	3	1	4	1	3	1
50%	14	2	19	2	11	2
80%	62	5	75	6	45	5
90%	123	11	144	12	64	10
95%	217	19	244	22	151	15
100%	965	965	965	965	281	281

Table 1. Size distributions of taxa child groups and subtaxa sibling groups.

Adding together all the taxa distribution values for the large ITIS dataset supplies a cumulative distribution graph as seen in Figure 7 with selected percentiles in Table 1. Analysing the non-terminal taxa (those that contain sub-taxa), the majority (>50%) contain only 1 or 2 sub-taxa, and 90% contain 11 or fewer sub-taxa. However, looking at the data from the perspective of a sub-taxon, we see that the average modal sub-taxon shares a parent taxa with approximately 13-14 other sub-taxa and the 90th percentile lies at groups of 123 sub-taxa. The difference is accountable due to groups with many sub-taxa being less frequent but obviously containing more subtaxa per occurrence.

Figure 8 shows a similar analysis of the distribution of files in directories on a computer hard disk. [16] analysed many more file repositories to demonstrate file sizes follow a log-normal distribution, and the same distribution as for taxonomies generally holds with our example, but a strange feature is seen in the cumulative graph (better seen here due to the linearly-scaled y axis) where a cluster of directories with roughly 1000 files each is present, accounting for roughly 20,000 files. These turned out to be temporary internet browser file directories. Such anomalies from the hollow curve pattern are not seen in large taxonomic classifications of a similar size (100,000+), though smaller taxonomies or analyses may well throw up more ‘jitter’ in the curve due to small discrepancies having a larger effect on a smaller data set. As a demonstration of this, it can be seen in Figure 4 that the plotted line gets progressively smoother as more taxa are considered as the analysis proceeds down the ranks from order to genus.

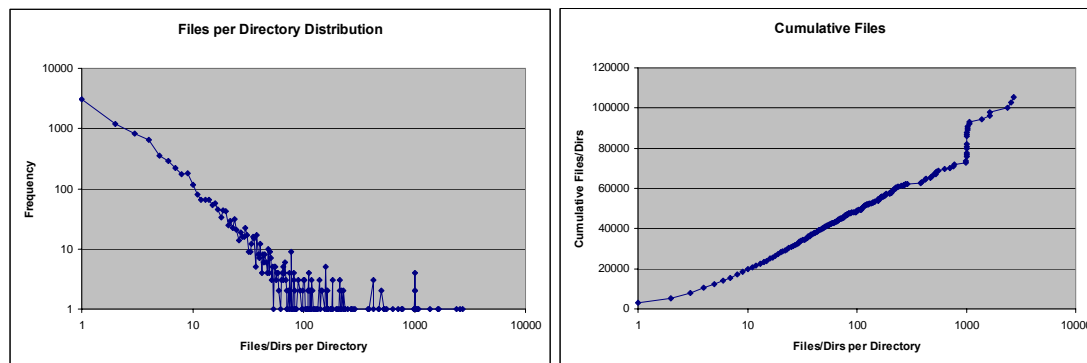


Figure 8. The organisation of files in directories on hard disks shows a similar pattern, but this has an unexpected bulge as seen in the cumulative graph.

In summary, we see that the taxonomic hierarchies to be considered for visualisation have a distinct character, the trees are unbalanced with a near-logarithmic distribution in subtree(-taxa) sizing. For ITIS 2004 we find that 80% of non-terminal taxa hold 5 or less subtaxa, but this number increases exponentially (+ more) towards the end of the distribution. When

considering all taxa, including terminal taxa, in the distribution, it is found the average modal taxon shares a parent taxon with roughly 14 other taxa. The incongruence with the first result is due to a few large groups holding a relatively much larger share of the overall taxon population.

Concept Relations

Concept relations are explicit taxonomist-defined relationships between taxa in different classifications, used when simple name or id matching does not capture the complexity of the relationships between classifications. However due to the labour-intensive nature of defining such relationships, concepts are only fully or partially mapped for small data sets – here we present a breakdown of the concept relationships present in two such data sets, the Koperski German Moss data set [17] and Ranunculus.

The Koperski moss classification defines many thousands of relationships to taxa in 13 other moss classifications. From 1,588 accepted taxa in Koperski emanate 7,921 relationships of four types – *congruent*, *includes*, *included_in*, and *overlaps* (it should of course be remembered that ‘is parent’ and ‘is child’ are also concept relationships that tend to be treated as special cases – they do not occur between classifications but within them, arguably their presence defines and separates classifications). Of those taxa that have relationships defined, most have around 5 relationships, which can be single relationships to multiple classifications, multiple relationships to the same classification, or a mixture of both. The relationships are mostly of the congruent type – they account for over 75% of the relationships in the data set.

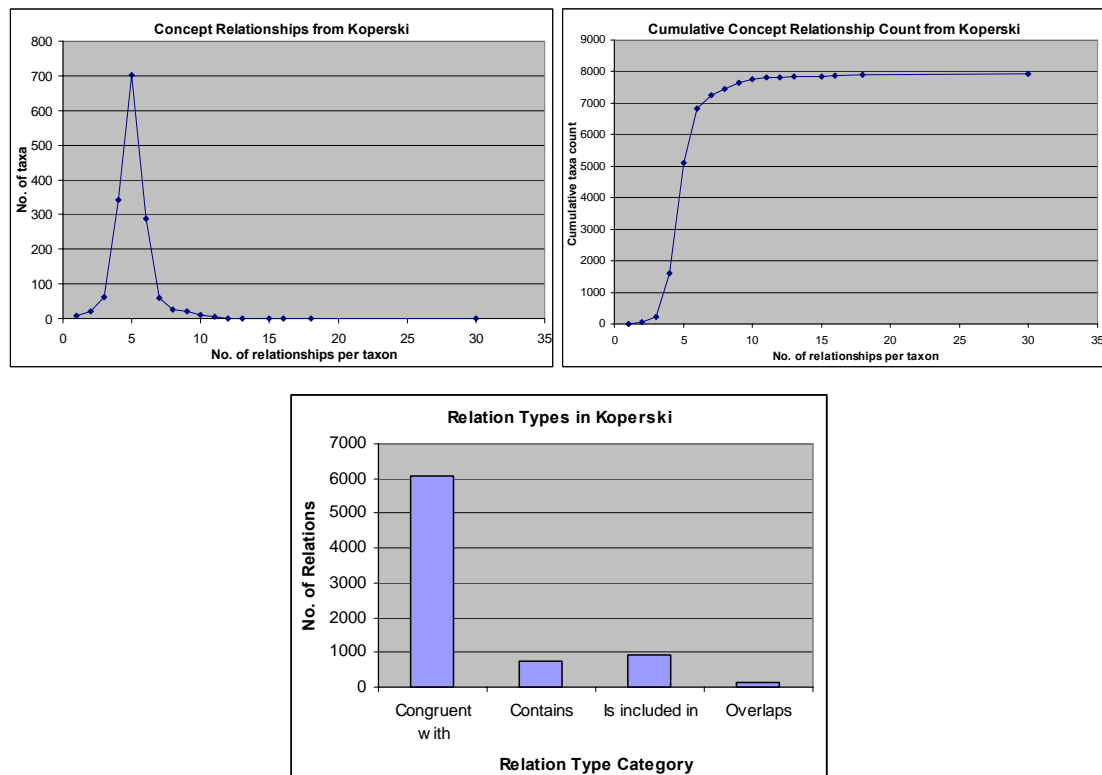


Figure 9. Concept relation counts for Koperski's moss classification.

This is in theory a simple data set as all the relationships are defined by one authority (Koperski) from their classification only to the other taxonomies. In practice, multiple taxonomists can define authors and in some cases these opinions will be contradictory to some extent.

The Ranunculus data set contains seven classifications with relationships defined from two classifications to the other classifications, with two authors now involved in defining separate

relationships (Kartesz & Peet). Since there are fewer classifications in the data set the average number of relationships is smaller. Again, the dominant relationship type is congruence – two-thirds of the relationships are of this type.

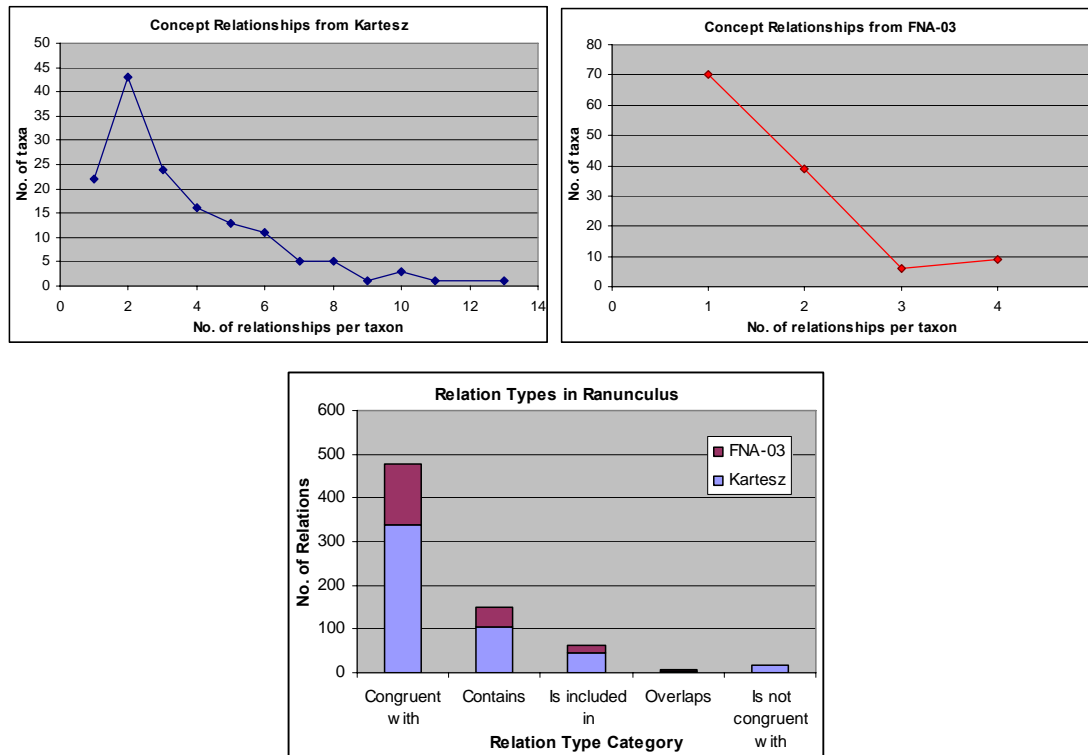


Figure 10. Relationship distributions and types defined from two Ranunculus classifications.

It is difficult to put a limit on the number of relationships that could be defined over a set of classifications. The number of possible inter-classification relationships is a product of the size and number of the classification. This is however compounded when multiple authors can add relationships between the same taxa, in effect producing a multigraph over the taxa set.

The graph formed by the relationships itself cannot be said to be directed, relationships such as ‘include in’ will tend to point upwards towards taxa of higher rank in other classifications, whilst ‘includes’ will thus tend to point in the opposite direction. ‘Congruent’ will generally match taxa at the same rank, though that is not always the case - some subspecies may be equivalent to species in different classifications. ‘Overlaps’, ‘Not congruent’ and ‘dissimilar’ could be applied to a far wider spectrum of taxa; logically everything that is not a ‘congruent’ relationship is ready for recognition as a ‘not congruent’ relationship – in practice though these are usually reserved to differentiate nodes with either closely related names or sibling, parent and child taxa with similar names in different classifications.

Conclusion

We find that the data sets we are faced with follow a distinct distribution pattern, roughly logarithmic in nature and similar to the distribution patterns found in other areas such as income analysis and scale-free networks, with many taxa containing only a single sub-taxon, and a few taxa containing many subtaxa.

The taxa holding many subtaxa may themselves be rarities in the classification but the sum total of the subtaxa they contain account for a significant percentage of the overall taxa in a classification. Conversely, whilst a high percentage of taxa exist that hold only one other taxa, in total these subtaxa compose a smaller percentage than would be expected of the overall classification. Such a distribution must be borne in mind when constructing a visualisation to

accommodate the analysis and manipulation of such structures, indeed it might even be possible to take advantage of such a distribution, though that remains to be investigated.

Concept relationships as of yet are found to follow no such particular distribution; data sets containing them are still few and so no universal rules on their distribution or optimal size given a classification size can be produced. These relationships are not given to any hierarchical structure themselves, forming a general graph over the hierarchically organised taxa present in the classification set, and unlike straight name-name relationships they are attributed according to the type of relationship that is being defined. Given that multiple authors could place relationships over the same set of classifications, the effect is that of constructing a multigraph and as such there is no effective upper limit to the number of relationships that could be envisaged on a classification set of a given size.

References

- [1] ITIS, *International Taxonomic Information System*, <http://www.itis.usda.gov/>, 2006.
- [2] E. Mayr, "The Number of Species of Birds," *The Auk*, vol. 63: 64-69, 1946.
- [3] W.D. Clayton, "The logarithmic distribution of angiosperm families," *Kew Bulletin*, vol. 29: 271-279, 1974.
- [4] R.W. Scotland and M.J. Sanderson, "The Significance of Few Versus Many in the Tree of Life," *Science*, vol. 303: 643, 2004.
- [5] K.P. Dial and J.M. Marzluff, "Nonrandom Diversification within Taxonomic Assemblages," *Systematic Zoology*, vol. 38(1): 26-37, 1989.
- [6] M. Mitzenmacher, "A Brief History of Generative Models for Power Law and Lognormal Distributions," *Internet Mathematics*, vol. 1(2): 226-251, 2004.
- [7] J. Chu and C. Adami, "A simple explanation for taxon abundance patterns," *Proceedings of the National Academy of Sciences of the USA*, vol. 96(26): 15017-15019, 1999.
- [8] Q.C.B. Cronk, "Measurement of Biological and Historical Influences on Plant Classifications," *Taxon*, vol. 38(3): 357-370, 1989.
- [9] A. Minelli, G. Fusco and S. Sartori, "Self-similarity in biological classifications.," *Biosystems*, vol. 26(2): 89-97, 1991.
- [10] P.F. Stevens, "Mind, memory and history: How classifications are shaped by and through time, and some consequences," *Zoologica Scripta*, vol. 26(4): 293-301, 1997.
- [11] P.F. Stevens, "How to interpret botanical classifications - suggestions from history," *BioScience*, vol. 47(4): 243-250, 1997.
- [12] S.M. Walters, "The Name of the Rose: A Review of Ideas on the European Bias in Angiosperm Classification," *New Phytologist*, vol. 104(4): 527-546, 1986.
- [13] E.W. Holman, "Evolutionary and psychological effects in pre-evolutionary classifications," *Journal of Classification*, vol. 2(1): 29-39, 1985.
- [14] J.J. Ibáñez and M. Ruiz-Ramos, "A mathematical comparison of classification structures: The case of the USDA Soil Taxonomy," *Eurasian Soil Science*, vol. 39(7): 712-719, 2006.
- [15] W.J. Bock and J. Farrand Jr., "The Number of Species and Genera of Recent Birds: A Contribution to Comparative Systematics," *American Museum Novitates*(2703): 1-29, 1980.
- [16] A.B. Downey, "The Structural Cause of File Size Distributions," *Proc. IEEE Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, Cincinnati, Ohio, USA, 2001, pp. 361-370.
- [17] M. Koperski, M. Sauer, W. Braun and S.R. Gradstein, *Referenzliste der Moose Deutschlands*, LV Druck im Landwirtschaftsverlag GmbH, Münster-Hiltrup, 2000, p. 519.