Visual Exploration of Alternative Taxonomies through Concepts

Martin Graham & Jessie Kennedy

School of Computing

Napier University

Edinburgh

UK

{m.graham},{j.kennedy}@napier.ac.uk

Abstract

A graphical user interface is presented that allows users of taxonomic data to explore concept relationships between conflicting but related taxonomic classifications.

Ecological analyses that use taxonomic metadata depend on accurate naming of specimens and taxa, and if the metadata involves several taxonomies, care has to be taken to match concepts between them. To perform this accurately requires expert-defined concept relationships, which are more complex yet more representative than the simple one-to-one mappings found through simple name matching, and can accommodate nomenclatural changes and differences in classification technique (cf 'lumpers' versus 'splitters'). In the SEEK-Taxon (Scientific Environment for Ecological Knowledge) project we aim to help users of taxonomic datasets untangle and understand these relationships through a prototype visual interface which graphically displays these relationship structures, allowing users to comprehend such information and more accurately name their data.

Keywords

Concepts, taxonomy, visualisation, names

Introduction

SEEK (Michener et al., 2005) is a large-scale multi-disciplinary research project that aims to integrate disparate ecological, taxonomic and environmental data sources. Such a goal would allow end users to access these hitherto heterogeneous data sources, greatly enhancing the ability of interested scientists to not only access such data but also to draw inferences and findings from the relationships between such data sets.

The SEEK-Taxon subgroup is concerned with addressing the issues associated with matching such data sets on organism name. Their solution is to adopt a concept based approach, where a concept is a unique combination of a name plus an author and date – i.e. *Ranunculus occidentalis* in Benson 1948, as opposed to the traditional, but inaccurate, name-only methods for matching between overlapping taxonomies (Kennedy et al., 2005; Kennedy et al., 2006). A concept typically then has intra-classification parent and child relationships that define its positioning within the concept author's classification and other inter-classification relationships that define it with reference to taxa concepts in other classifications. There are two main scenarios that lead to the production of inter-classification relationships; either

relationships between concepts are asserted by the creator of a new taxonomic revision to accurately reference their work to past classifications or to incorporate parts of those classifications within their own work, or a qualified third-party taxonomist can assert relations between concepts in existing classifications. These relationships can be given varying types and strengths and are also annotated with the details of the taxonomist that introduced that relationship. A concrete implementation of this model is defined in the XML-based Taxonomic Concept Schema (TCS, 2006).

The relevance of this development to ecologists is that it provides a path for more accurately describing observed specimens in terms of a taxon name as described in a particular taxonomic publication. Names in taxonomy are in a constant state of flux; new discoveries and reinterpretations of existing specimens introduce new names into the mix - what is "*Aus aus*" in one publication may be equivalent to "*Aus bus*" in a second taxonomy and "*Xus aus*" in yet another. Trying to deduce this synonymy from a paper trail left by previous taxonomists is difficult, often requiring another taxonomist to do so properly, which leads at least one commentator to state that taxonomy is not made easy for its ultimate end-users – ecologists, conservationists, naturalists etc (Godfray, 2002). Alternatively, reliance on simple name-based matching would in many instances leave the ecologist unable to equate observations in one data set with those in another, or arguably worse make the wrong link between their observations, nor able to trace the taxonomic history of a species - a need recognised in community ecology by Gotelli (2004). Faced with intractable inter-relationships between taxonomies is a factor in many ecologists falling back on less accurate parataxonomy (Krell, 2004) or classifying specimens by higher-level taxa (Herman et al., 1988). In contrast, using the concept-based model, the relationships between taxa in related classifications are precisely defined such that an ecologist could discover that *Aus aus*, *Aus bus* and *Xus aus* are all names and descriptions of equivalent species.

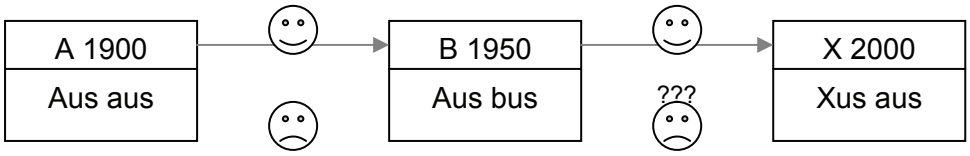


Figure 1. Concepts allow the finding of equivalent names between classifications

Having such concept information available is a step forward, but it also introduces its own complexities. For instance, a one-to-one mapping between taxonomies is preferable, but when mapping between taxonomies that have been defined at different levels of resolution (splitters versus lumpers), the relationships of a concept to those in another publication can be one-to-many in nature, with the types of relationship also diversifying, often signifying overlap or inclusion rather than a straight-forward match. Furthermore, the concept relationships themselves represent the opinions of a particular taxonomic expert, and in cases where a data set has been scrutinised by two or more taxonomists there may very well be disagreements between them in terms of which concepts are related and the strengths and types of those relationships.

The outcome can be that a dense set of inter-relationships is formed which are difficult to comprehend or analyse when presented in traditional paper or electronic text-based forms. To alleviate this we have developed a prototypical tool we term TaxVis for browsing and searching TCS-based data sets through name, and more importantly, concept-based relationships. TaxVis is one of a number of graphical visualisation tools developed within the SEEK project, where the main aim of such tools is to present complex information structures in a more readily perceived form than paper or standard interface metaphors such as lists or textboxes would allow.

These styles of visualisation, collectively known as Information Visualisation (Card et al., 1999), differ from the more commonly present cartographic and geographic visualisations in ecoinformatics interfaces as the data they display is not anchored in any direct sense to a representation of the real world. There are previous examples of information visualisations being developed for use with ecological data, amongst them Yoon et al's (2005) and Lee et al's (2006) network visualizations of food webs. Our work sits with these examples, as where specimen distributions, observations, field studies etc can be rooted at a particular coordinate and then correlated and overlaid onto map representations, a taxon or its associated classification are not fixed in physical space and thus their visual representation is less intuitive and straight-forward. Within the field of information visualisation work on graphically comparing multiple classifications has been of interest with previous systems by the authors (Graham et al., 2005) and comparing taxonomic name-based classifications, Munzner et al (2003) comparing phylogenies, and Sifer (2006) displaying multiple hierarchies over web log data.

Feedback from questionnaires after demonstrations of this tool to users of taxonomic and museum collection data has indicated that a graphical display of expert-defined relationships between taxonomic datasets is deemed more useful than that produced by naïve name matching – average usefulness of concept relationship-based operations was perceived to be higher than features that relied solely on names. These users cover many disciplines such as museum curators, ecologists, collection managers, taxonomists and taxonomic database administrators, but all have demonstrated a strong reaction when shown the ability to compare taxa across classification through concepts, as they are all too familiar with the problems caused by name matching or trying to reconcile unfamiliar taxa.

In the following sections we describe the application's layout and some of its operations, with an example of how a TCS-based data set can be explored using shared names or concept relationships, with an emphasis on the different results these two approaches can reveal, of which we claim the relationship-based outcome to be the most accurate of the pair. We then follow this with a description of some further requirements of interest to ecologists that were revealed when the tool was demonstrated to users of taxonomic data. Discussions of the advantages or difficulties involved in these proposed requirements, and of the visualisation in general, are then detailed.

Materials & Methods

TCS data sets can be loaded into the application via the menu bar at the top of the screen. Currently, the visualisation works with standalone XML files of TCS data sets that contain a fixed number of alternative taxonomies, such as the Koperski Moss data set (Koperski et al., 2000) or the annotated *Ranunculus* data set. Certain types of non-concept based data sets such as MaNIS (2001) (Mammal Networked Information System) data can also be uploaded, but as these do not contain concept relationships these are probably of little relevance here. Upon loading the classification display themselves and the visualisation is laid out as described below.

Interface Layout

The visualisation is laid out as shown in Figure 2; sub-divided into four main sections, numbered i-iv in the figure. The largest part (i) of the interface is devoted to displaying classifications visually and is itself sub-divided according to the number of active classifications in the current data set. Each classification is drawn 'top-down' as in Figure 3, with lower rank taxa displayed underneath their parent taxa. Each taxon is drawn as a rectangular box which contains information such as name and author.

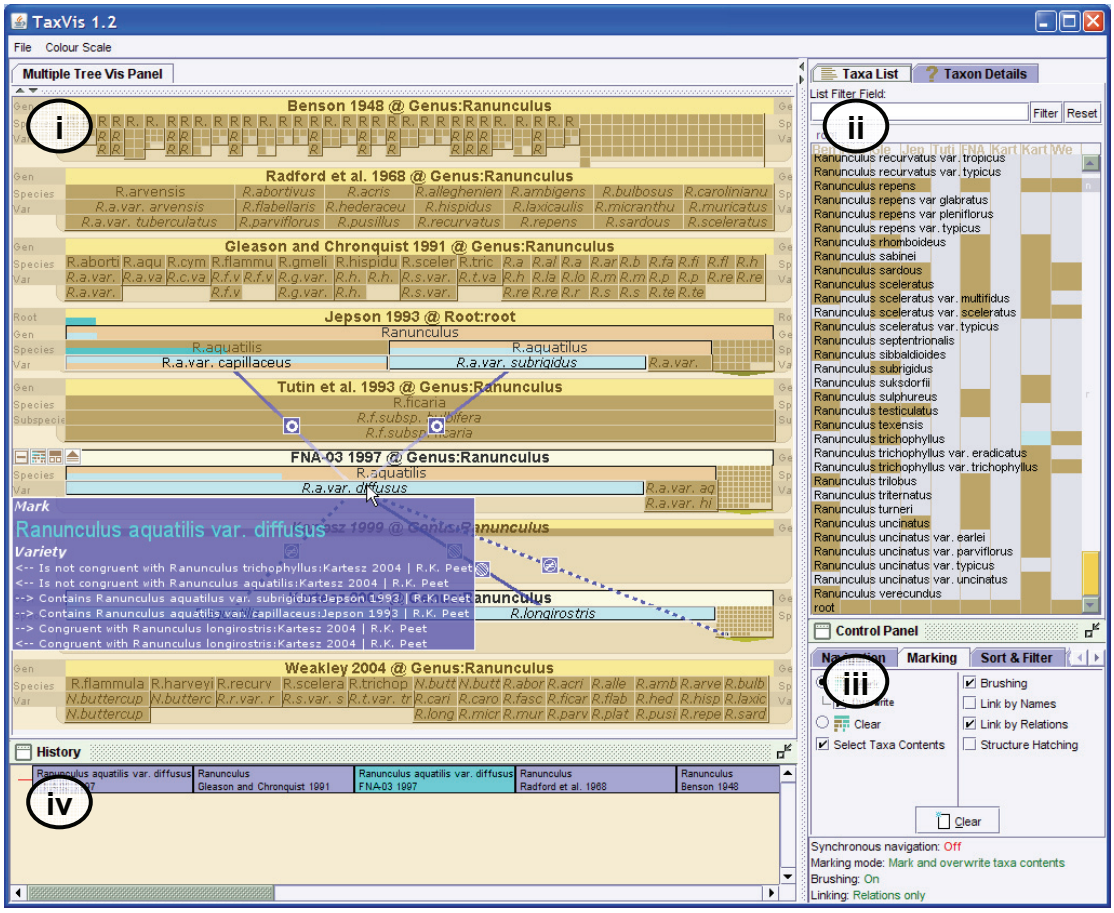


Figure 2. Screenshot of TaxVis application.

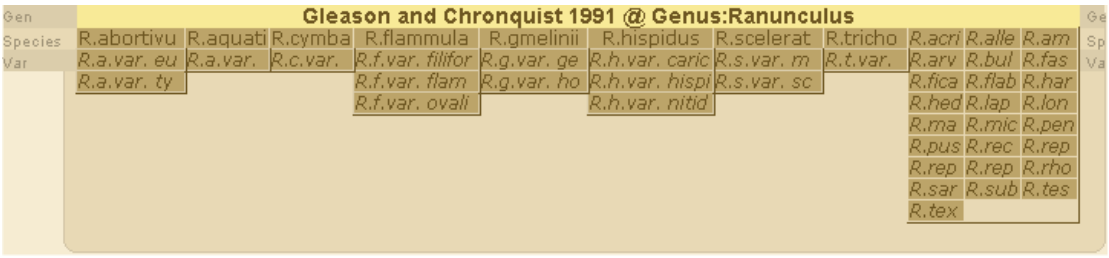


Figure 3. A single classification in close-up. Taxa are laid out top-down, underneath their parents.

The upper right-hand portion (ii) contains two tabbed panels which are each in effect alternative views of the data presented in the main display. The first tab displays a list of all taxa names present in the classifications under examination, with a horizontal pattern of shading to indicate that name's occurrence or not in individual classifications – each classification represented as a narrow column within the list. In effect, each name in the list represents a number of possible concepts, one per classification - though names tend not to have been used within all classifications, so the shading pattern behind the name reflects which classifications it is present in. Selections performed in either the list or the main visualisation panel are reciprocated between the two views in terms of navigation and the colouring applied to both displays. Direct searching by type-ahead keyboard input or by simple regular expression substring matching can either locate known names quickly or narrow down the range of the list considerably.

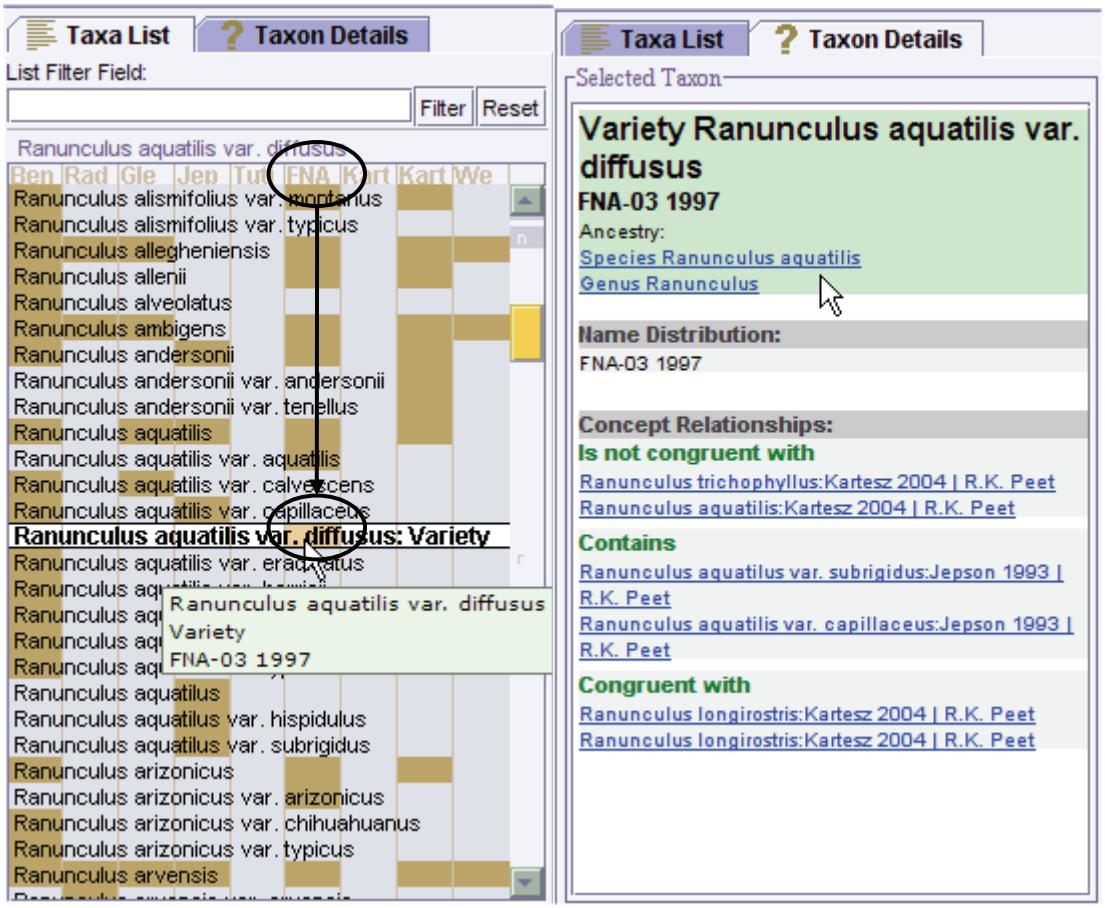


Figure 4. The left-hand side of the above figure shows the taxa list tab, an ordered row list of names divided into columns according to currently active classifications. Colouring occurs at the intersections where a concept using that name occurs in a classification, i.e. marked in the figure is a colouring on the name *R. aquatilis var. diffusus* where it intersects the FNA column indicating a concept using that name occurs in the FNA classification. The right hand part of the figure shows the details tab for the concept currently under investigation, including other classifications where the name also occurs and details of concept relationships that have been defined over it in the TCS data set.

The second tab holds information on the last navigated or selected taxon concept, including its parent taxa in the classification it was probed in, plus other classifications where the taxon name occurs, along with details of their child taxa if present. Finally a list of concept relationships is given for this taxon concept if they exist. All this information is hyperlinked so the various relationships can be navigated with simple mouse-clicks.

The lower right-hand portion of the application window (iii) contains a control panel for adjusting various properties of the display and interaction, such as should matching occur by name or concept (or both), should navigation be synchronised between the classifications or not, and sorting mechanisms for ordering displayed groups of taxa besides the default alphabetical ordering.

The bottom strip of the application window (iv) contains a history bar, which records all navigations and selections of taxa made in the other components. These are displayed as buttons containing the name and classification of the selected concept and selecting one of these buttons will repeat the operation it represents. Finally, a menu bar along the top of the application window allows TCS files to be loaded into the visualisation, along with an option for choosing the colour scale used to mark out selections. This option presents a number of perceptually linear colour spectra as

defined by Levkowitz and Herman (1992), in which colours are discernable in human vision by a degree equivalent to their separation in the scale.

In the main panel of the visualisation the classifications are drawn according to the available space. More space means more depth can be drawn per classification. Horizontal space is usually the limiting factor for drawing tree-based structures, the exponential increase in size of each consecutive rank quickly reducing available space per taxon to below a single pixel in width. Solutions such as drawing the smallest displayable taxa in a grid formation are used, along with increasing the allocation of display space to user-selected taxa, as these are obviously of more current relevance to the user. Also, one classification can be selected as the 'prime' classification, which simply sets aside more space for it in a separate sub-panel above the other classifications.

Interaction

Navigation within the main classification panel is performed by clicking the left-hand mouse button when hovering over a taxon concept representation; this moves the selected taxon to the top of the displayed portion of the classification, in effect displaying only the sub-classification underneath that taxon and thus able to display those taxa in greater detail. An exception is when selecting the taxon that is currently placed at the root of the displayed sub-classification, in which case the taxon moves down to reveal its parent taxon. Hence, navigation can easily be made up and down a classification. Animation of these movements is performed to make the user aware of the direction of the navigation and helps to reconcile the new layout with the old configuration.

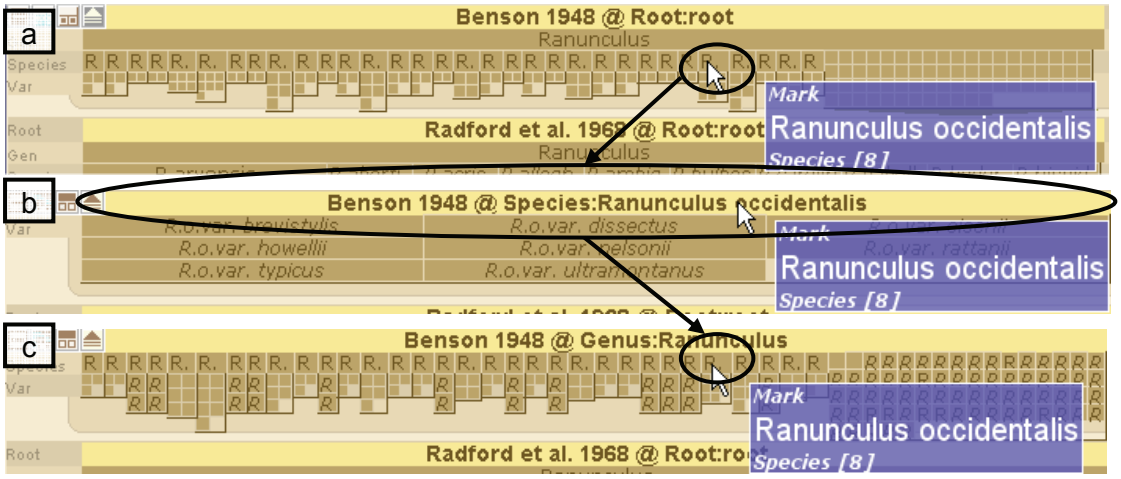
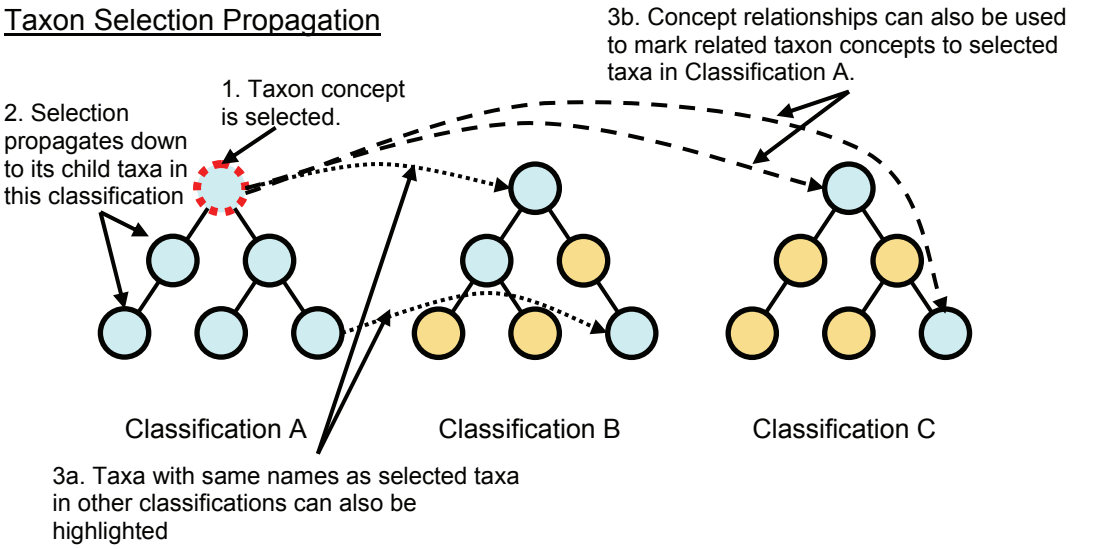


Figure 5. Pressing the left mouse button when over *Ranunculus occident*alis in sub-figure a displays that taxon and its contents in more detail, as seen in sub-figure b. Subsequently, the same operation on *R. occidentalis*, now it is the root of the currently displayed information, will show its parent taxon and its immediate contents – in this case the *Ranunculus* genus itself as shown in sub-figure c.

Similarly, selections are made by pressing the right mouse button when the pointer is over a concept representation. The concept and its descendants in the classification in question are highlighted in colour. Dependant on mode settings, highlighting is also performed on concepts in the alternative classifications that share the same names as the selected concept set, and/or on concepts in other classifications that have explicitly defined relationships to this set. Figure 6 demonstrates diagrammatically and through a visualisation screenshot what the results are of selecting a given taxon concept in a classification.



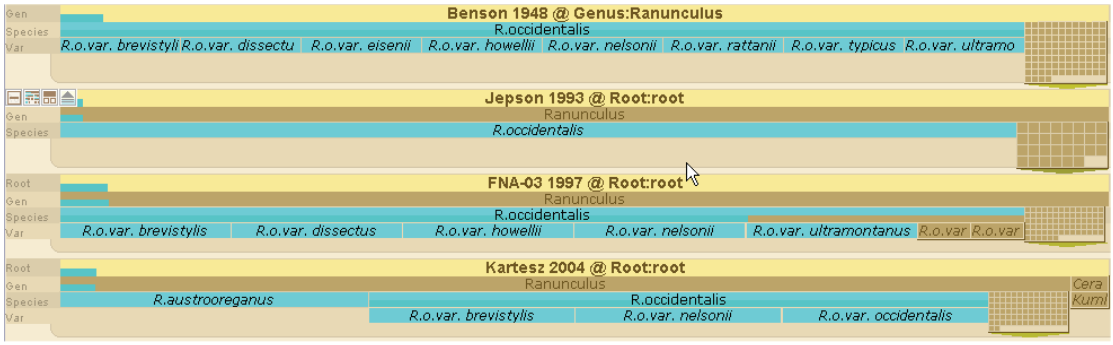


Figure 6. *Ranunculus occidentalis* has been selected in Benson's 1948 classification. Consequently, its children in that classification have been similarly highlighted. Taxa in the Jepson 1993 and FNA 1997 have been highlighted as they have the same names. In the Kartesz 2004 classification there are highlighted taxa such as *R. austooreganus* and *R. o. var occidentalis* that do not share the same names as Benson's concepts but are nevertheless highlighted, these are examples of highlighting through concept relationships.

Taxa that contain sub-taxa have a slightly more complex colouring system: they are coloured along the top half of their representations according to whether they themselves have been selected, and coloured along their bottom half according to the proportion of its sub-taxa that are also selected. The latter colouring metric allows a user to find selected taxa that are buried deep within a classification as an indication of their presence is communicated all the way up through its parent taxa to the classification root

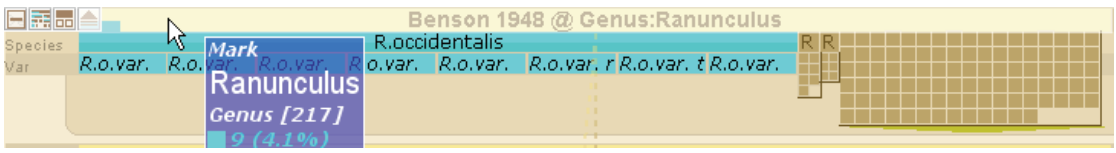


Figure 7. In the screenshot above it can be seen that *Ranunculus occidentalis* has been selected along with all its child taxa, so is completely coloured. *Ranunculus* itself has not been selected but the selection of *R. occidentalis*, totalling 9 out of 217 child taxa, results in its representation's bottom half being coloured in proportion to this.

An informative tooltip is displayed whenever the mouse pointer is above a taxon representation, displaying information such as name, rank, number of child nodes, and details of how many child nodes have been affected by user selections. Similarly

to the taxon detail tab shown in Figure 4, it displays information on relationships that involve this concept.

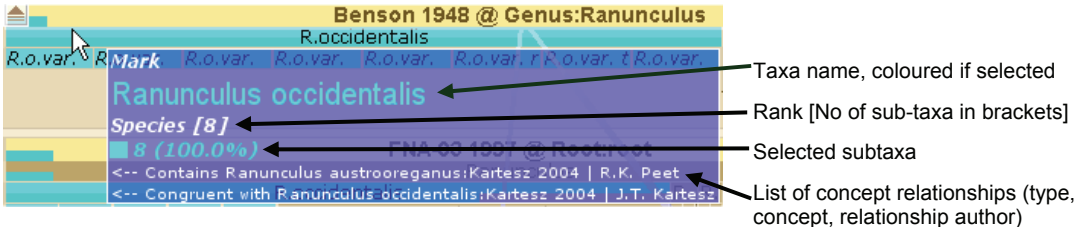
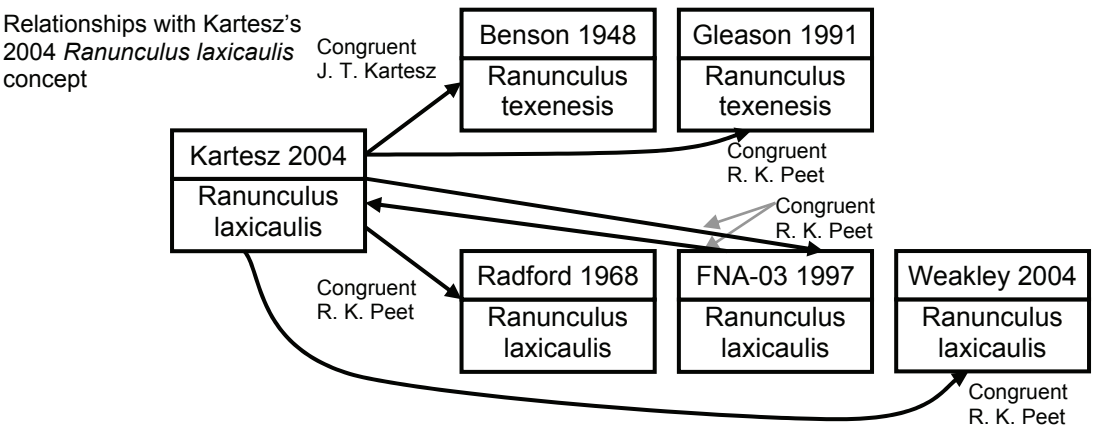


Figure 8. The tooltip displays information about the taxa concept currently under the mouse pointer along with its current selection state.

Navigation and selection can also be made in the tabbed panels shown in Figure 4 using the same left-button press for navigation and right-button press for selection actions used in the main classification panel. These views are linked so that navigation or selection actions performed through one view are reflected in the other views.

Displaying Concept Relationships

Previously (Graham et al., 2005), we had indicated concept relationships with the same highlighting method as we had for name matching relationships. This was adequate for showing shared names across classifications as this relationship was strictly one-to-one - if you could see a highlighted or selected node in a classification that was the shared name node. However, concept relationships for a taxon to another classification can be one-to-many as a concept may overlap with or include several other concepts in an alternative taxonomy. This meant that there was always some uncertainty about how many related concepts were to be searched for in the visual representation. Lists of concept relationships in a tooltip and side panel were introduced, and could enumerate how many relationships should be present in each classification along with their details, but then visually searching for more than one target in the classification representations was an uncomfortable procedure. To overcome this it was logical to draw links between the queried concept representation and the related concepts if currently visible. For the same rationale that we do not show name overlaps unless requested, only the currently queried concept has its relationships displayed. To display all the relationships for all concepts leads to a dense, cluttered display of entangled and irresolvable links, a problem which plagues many visualisations of graph and network-oriented data.



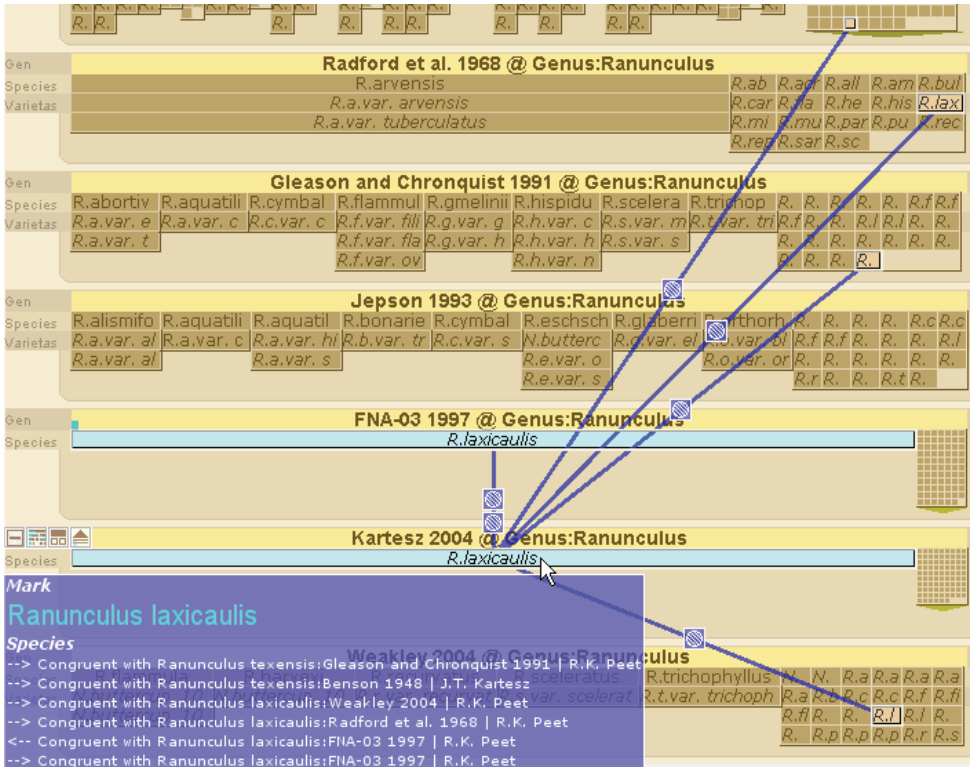


Figure 9. Concept relationships involving *Ranunculus laxicaulis* in Kartesz 2004 displayed firstly in diagrammatic form and secondly how they appear in the visualisation. The relationships are all congruent and all defined by the same author except for the relationship to Benson 1948. The equivalent reciprocal relationship has also been defined from FNA-03 1997 to Kartesz 2004.

Thus, distinct lines were drawn to indicate explicit relationships between concepts in different classifications as shown in Figure 9. The lines are given a simple colour-coding according to the following rules: blue colouring at one end of a line signifies the relationship involves everything in the concept at that end of the relationship; a white colouring at the end of a line means the relationship involves only part of the concept at the corresponding end. Thus a completely blue line means everything in both concepts matches, a congruent match; a completely white line means only some parts of both concepts match, an overlap. A line that is blue at one end, yet white at the other, indicates that all of one concept matches up to some of the other concept i.e. one concept is included within another. The direction of the line colouring indicates which way round the inclusion is – the concept at the blue end of the line is included within the concept present at the white-coloured end. Non-congruent links are also indicated with a dashed blue and white colouring to signify all we know is that the relationship is not a congruent one – technically all relationships that are not defined as congruent are 'not congruent' but in practice taxonomists only insert these into data sets for certain conditions, usually when the same name in different classifications is not equivalent, but the precise nature of the mismatch is unknown.

Results

We now explore detailed examples of how the visualisation can guide users to explore concept data and discover relevant information, using a data set that covers nine full or partial classifications of *Ranunculus* and has been populated with the appropriate concept relationships by an expert (Robert Peet at North Carolina State University) in *Ranunculus*.

Table 1 lists the classifications used, all of which cover North America. Concepts have been defined from two classifications – Kartesz 2004 and FNA 1997 – to the other classifications. Some of the classifications are only partial, Kartesz with 2 taxa and Tutin with 7 being the obvious standouts, as only taxa involved with the concept relationships have been recorded, along with their parent taxa to family level.

Ranunculus classification	No. of taxa	Relationships from *Kartesz 2004*						Relationships from *FNA 1997*					
To		=	⊃	⊂	∩	≠	∑	=	⊃	⊂	∩	≠	∑
Kartesz 2004	145							100	41	13	1		155
Weakley 2004	36	30	3				33						
Kartesz 1999	2	2					2						
FNA 1997	130	99	21	22	1	5	148						
Tutin et al 1993	7	5					5						
Jepson 1993	46	36	4	6			46	39	3	5			47
Gleason & Chronquist 1991	51	34	12	3			49						
Radford et al 1968	22	18	3				21						
Benson 1948	218	115	63	15	4	12	209						
Totals	657	339	106	46	5	17	513	139	44	18	1		202

Table 1. Concept relationships between Ranunculus data sets. Key: (=, congruent), (⊂, is contained in), (⊃, contains), (∩, overlaps), (≠, is not congruent).

Scenario 1

As a first example scenario we consider how an ecologist could integrate the results of a recent survey, classified using the Flora of North America's (FNA) 1997 taxonomy, with a historical survey classified using Benson's earlier taxonomy of *Ranunculus* from 1948.

A simple example begins with *Ranunculus laxicaulis* in the FNA's 1997 classification. As shown in Figure 10, the taxon name list on the right-hand side of the interface reveals that the *laxicaulis* name does not exist in Benson's 1948 classification, the left hand column for that name remains greyed out – and even if it did, it might not be a correct match for the 1997 version. Pressing the mouse's right-button when over the active concept in the list for the FNA classification's *laxicaulis* colours the concept and any explicitly related concepts. As previously stated, the different views in the visualisation are linked such that this selection is mirrored in the main classification display panel, where the appropriate concept representations expand to take more space.

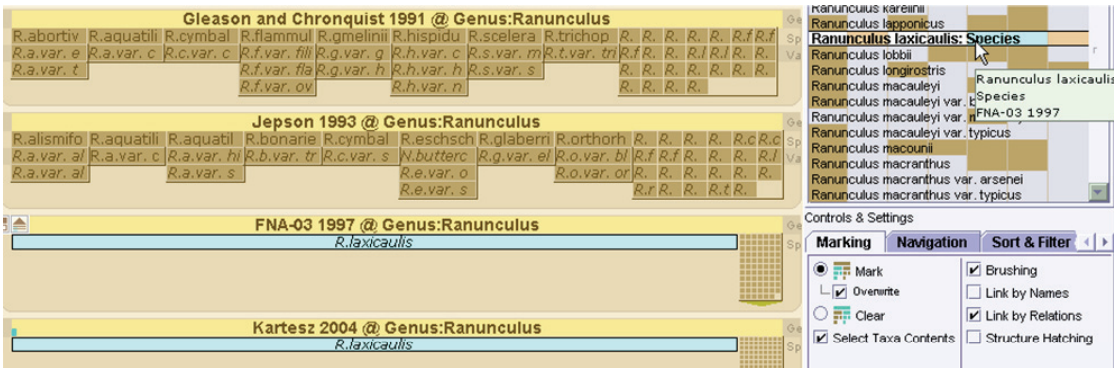


Figure 10. Finding *Ranunculus laxicaulis* in the list and then selecting it expands these concepts in the main window.

Moving the mouse over the representation in the FNA classification, as shown in Figure 11, reveals it has one congruent relationship to *Ranunculus laxicaulis* in Kartesz's 2004 classification. Selecting the right mouse button here allows a user to choose to fix the line representation of this relationship in place.

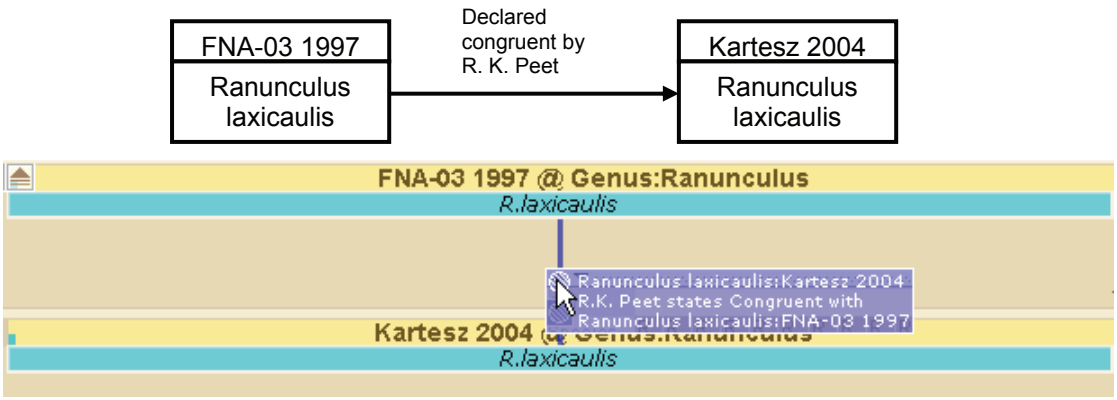


Figure 11. A simple relationship between the two *Ranunculus laxicaulis* concepts shown diagramatically and as displayed in the visualisation.

In turn, moving the mouse to Kartesz's representation of *laxicaulis* we find multiple relationships to other classifications, as demonstrated in Figure 9. In this case they are all congruent relationships to other instances of *lexicaulis* and *texensis*, with at the most only one relationship per classification, excepting the bi-directional reciprocal relationship between FNA and Kartesz's concepts of *laxicaulis*. Amongst these is a relationship to Benson's 1948 classification, stating Kartesz's *laxicaulis* concept is congruent to *Ranunculus texensis* in Benson. Thus we find a historical equivalence, via Kartesz's classification, between the concept *Ranunculus texensis* in FNA 1997 and *Ranunculus laxicaulis* in Benson 1948. Figure 12 shows a small screenshot of the tooltip exploring Benson's *Ranunculus texensis* concept to confirm this.

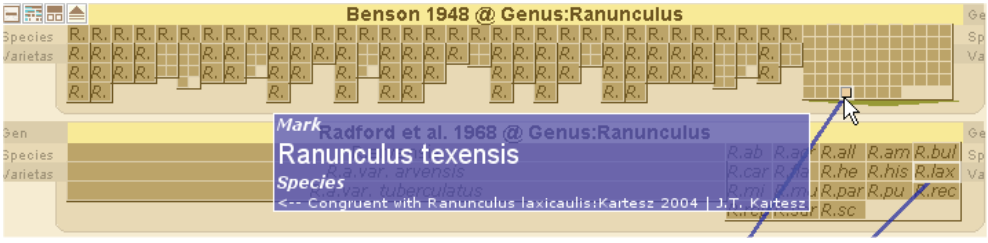


Figure 12. Finding further relationships from Kartesz's *laxicaulis* concept. The bottom part of the figure reveals the relationship to *texensis* in Benson's 1948 classification.

This example also reveals that not all classifications have to be directly inter-related, often it suffices for one classification to be matched to the others under consideration and through this classification transitive relationships can be deduced – though this can lead to pitfalls which are discussed later.

Scenario 2

A more complex example supposes a collector has, according to the FNA 1997 classification, an example of *Ranunculus aquatilus var. diffusus.* A quick search in the list finds the name, but also reveals via the shading behind the name that it occurs only in the FNA classification in this data set. There are other named *aquitilis/aquitilus* varieties clustered about that name but it is not obvious which, if

any, match to the *diffusus* variety. As before, selecting the concept colours it and any related concepts in the main visualisation panel.

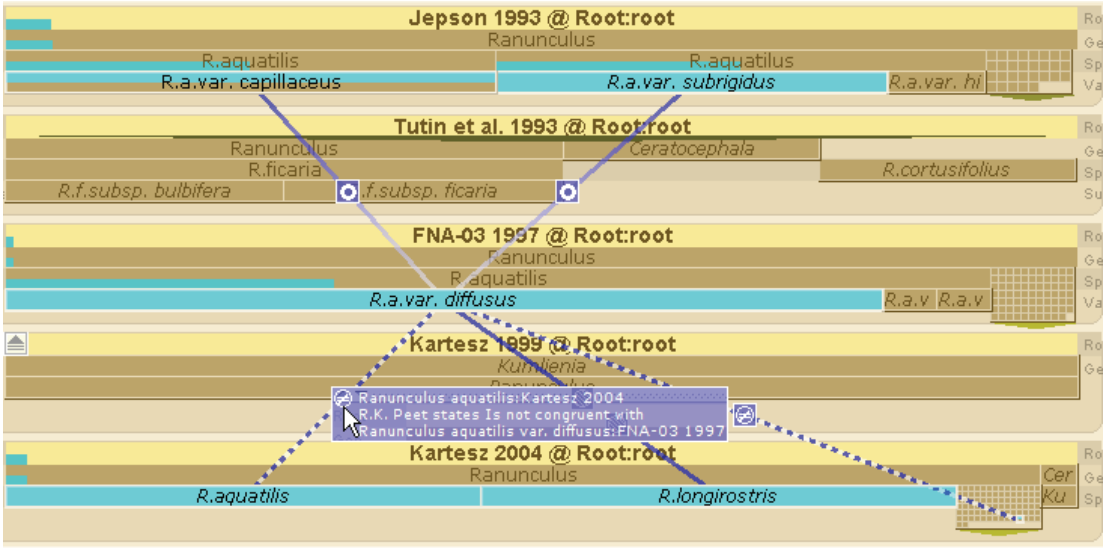


Figure 13. Concept relationships of *Ranunculus aquitilis var. diffusus* in FNA-03 to taxon concepts in other classifications. Mouse pointer is currently over the 'not congruent' relationship to *Ranunculus aquitilis* in Kartesz 2004, marked by a dashed line.

Moving the mouse over to the FNA classification's *diffusus* representation reveals that several concept relationships have been declared with it, the most interesting being a congruency to *Ranunculus longirostris* in Kartesz's 2004 classification, and to a combination of *Aquatilus subrigidus* and *Aquatilis capillaceus* varieties in Jepson 1993 (note the variations in spelling between *Aquatilis* and *Aquitilus*.) Again, right clicking and marking the concepts makes these relationships a fixture in the visualisation.

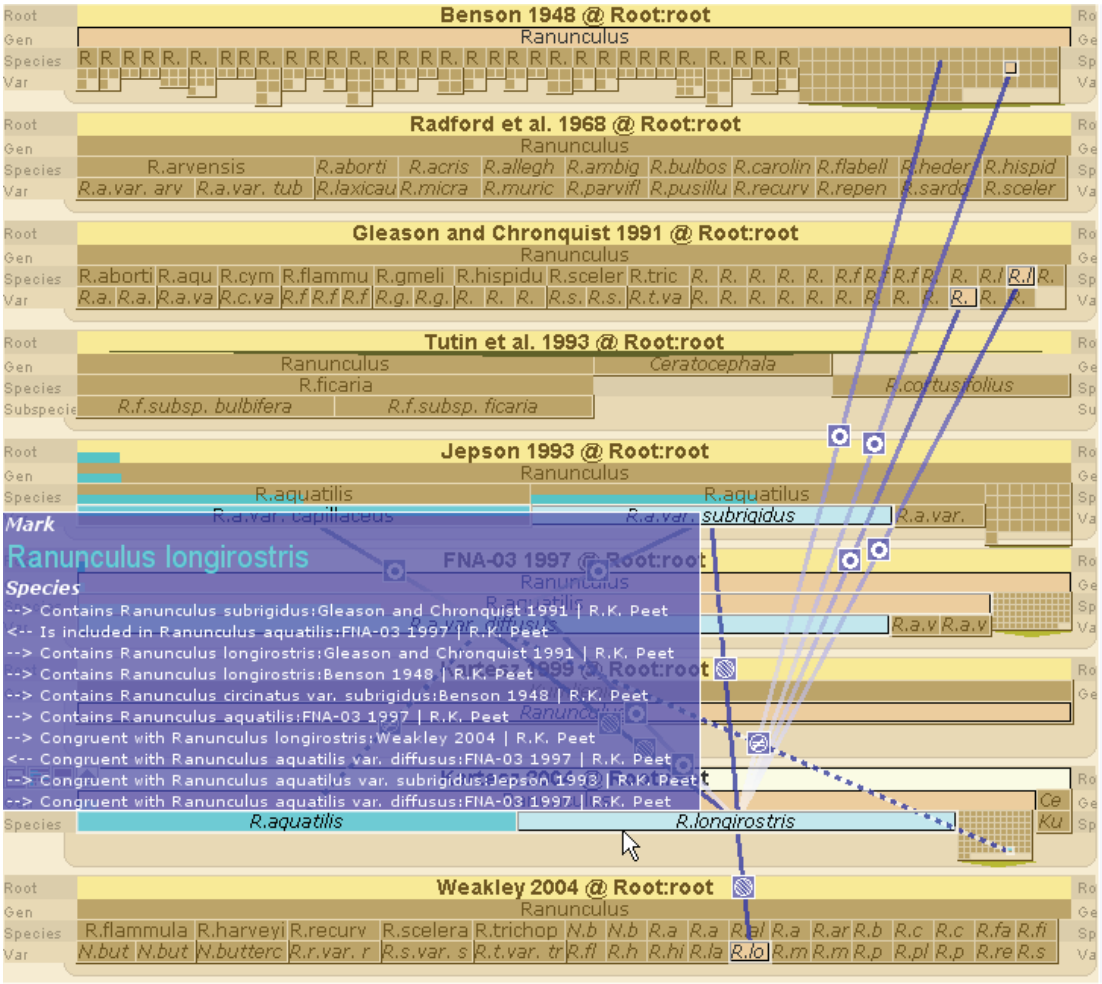


Figure 14. Concept relationships involving *Ranunculus longirostris* go back as far as Benson's 1948 classification

This only gives us matching concepts as far back as Jepson 1993. However, by moving the mouse pointer over to *Ranunculus longirostris* in Kartesz more relationships are revealed as shown in Figure 14, this time back to Gleason and Chronquist's 1991 classification and Benson's 1948 classification. Here the concept in Kartesz is split between two concepts in each of these classifications, in the case of the older Benson classification being formed from their *longirostris* and *circinatus var subrigidus* concepts. In this manner, the original concept, *Ranunculus aquatilis var. diffusus* in the FNA classification has been traced back to be equivalent to a combination of two taxa in the earlier 1948 classification. Though it would have been preferable to arrive at only one concept, it is the more accurate relationship according to the expert who has analysed these classifications.

Thus we have examples that reveal that though the same name in two classifications is often a congruent match, as demonstrated in the first example, there are also situations exemplified by the two scenarios where certain names do not re-occur or names are not exactly congruent between classifications. The latter condition if matching solely by names would lead to imperfect decisions being made, whilst the former condition could result in an ecologist being simply unable to integrate data sets named according to different classifications. Concept matching alleviates these difficulties and allows data collected and then named under different classifications to be reconciled accurately.

User Feedback

Demonstrations have been given at four sites to assess requirements in terms of the operations the visualisation currently supports and those tasks that representative users would like to see it support in the future. These demonstrations were given remotely at Kansas Natural History Museum using Skype, GoToMyPC, and a lot of local help, and in person at the National Evolutionary Synthesis Center (NESCent), Raleigh, North Carolina, followed by The Natural History Museum, London, the Global Biodiversity Information Facility (GBIF) headquarters in Copenhagen and the National Centre for Ecological Analysis and Synthesis (NCEAS) in Santa Barbara.

Location	Users	Date	Data Sets
Kansas NHM (remote)	Collection Managers (5)	Jan 23rd, 2006	MANIS
NESCent, North Carolina	Ecologists (2) Taxonomists (1) Research Scientist (1)	Feb 1st, 2006	Moss/Ranunculus
Natural History Museum, London	Taxonomists (1) Research Scientists (2) IT Professionals (2) Other – Unspecified (1)	Feb 27th, 2006	Moss/ Ranunculus/ Fish
GBIF, Copenhagen	Database Administrators & Taxonomists (5)	Sept 11th, 2006	ITIS/ MANIS/ Ranunculus/ Fish
NCEAS, Santa Barbara	Ecologists (4)	Dec 7th, 2006	Ranunculus / Fish / ITIS

Table 2. Details of visualisation demonstrations

In each case we initially collected biographical data such as job descriptions and qualifications from the observers, followed by demonstrations of the visualisation using various TCS-based data sets and finally gathered their opinions via a post-demo questionnaire. Each group were initially shown an appropriate data set, e.g. museum collection data for museum curators, concept-based data such as *Ranunculus* for taxonomists and ecologists, but additional data sets such as large-scale annual ITIS revisions were shown where time permitted.

The demonstrations were controlled by one of the visualisation developers for two reasons: firstly, the complexity of the controls was such that it was felt that asking the observers themselves to use it would have inevitably deflected attention away from what the visualisation could do towards the details of how to use the visualisation interface. Our aim was to discover whether the tasks supported by the visualisation were of use, finding interface bugs could come later. Secondly, the number of participants at each demonstration numbered 4 or 5 on average, thus a one-to-many demonstration was much more feasible than multiple one-to-one sessions given time and logistical constraints. In practice this gave each of the demonstrations the air of a focus group approach (Nielsen, 1993), the aim being to elicit feedback on whether the interface was performing the right tasks and gather further functional and usability requirements. As such our data gathering consisted of post-demonstration subjective questionnaire responses rather than empirical task performance measurements. These questionnaires captured opinions on the utility of the tasks demonstrated, such as showing the different types of concepts, finding the occurrence of a given name in a set of classifications and more GUI-oriented aspects such as the utility of animation when changing navigation viewpoints.

The course of the demonstrations themselves were driven by the goal of exhibiting the relevant functionality (for museum curators, comparing whole or sub-parts of museum collections, for ecologists and taxonomists demonstrating the concept facets of the visualisation), but allowed sufficient leeway for observers to comment

and question on current actions or other aspects of the application that had attracted their interest. Small example scenarios similar to those described in the previous section of this paper were used to show the utility of concepts as opposed to name matching and the ability to follow modest transitive relationships.

Question	Avg (1-5)	No.
Find those concepts that are congruent	4.62	13
Find those concepts that are included in other concepts	4.46	13
Compare taxonomic concepts through explicit relationships.	4.42	12
Find those concepts that include other concepts	4.30	13
Determine which taxonomies to involve in a comparison.	4.23	13
Find those concepts that overlap	4.23	13
Find the occurrence of a particular taxa across my collection or a set of collections. Find the occurrences of a particular taxonomic name across a set of taxonomies.	4.11	18
Display the overlaps between my collection and other collections. Determine the similarity of one particular classification to a set of other classifications. Compare classification name coverage against other classifications	3.94	18
Based on this demonstration, rate the overall usefulness of this visualization tool.	3.88	17
Show how my collection is organized differently compared to other collections. Discover structural differences between taxonomies (based on names).	3.82	17
Find the first occurrence of a particular taxonomic name across a set of taxonomies.	3.5	12
Find taxonomic names that occur only in one particular classification in a set of such classifications Display what is unique to my collection as compared to other collections.	3.44	16

Table 3. Results from questionnaire where questions were answered by 12 people or more. Multiple questions reflect changes in emphasis when asking questions to museum curators or ecologists/taxonomists.

Table 3 displays those abilities of the visualisation ordered by perceived usefulness as judged by the observers (the most poorly regarded feature in terms of usefulness was the ability to find names unique to one particular classification or collection.) These results indicated the ability to display concept-based data was of the utmost importance, and a comment from one observer in the North Carolina session summarised this need by stating they wished *"to relate different community classifications to one another in a standard way."* The ability of the visualisation to trace the history of a name or related concepts by ordering the classification chronologically was favourably received as this allowed users to find the first use of a concept in a particular data set, a requirement pointed out as a need for ecologists by Gotelli (2004). Other notable comments included *"Can it find and display linked relations?"* from an observer at the Natural History Museum demonstration, which has relevance to the discussion on transitive relationships found in the next section.

The observers also revealed that they wished to control the types of relationship that were displayed, as some did not want to see dissimilar or non-congruent relationships. To this end, the types of relationship to be displayed can be filtered

using a drop-down list of checkboxes found in one of the control panel tabs. Those types of relationship that have their checkboxes selected are those that are then displayed in the main panel.

Probably most interesting though was the feedback we gleaned on what further features the users would want to see in such a tool. For instance, geographical data was stated to be of importance to analyse distributions of taxa. We had previously incorporated a geography hierarchy into the museum collection data version of the visualisation (Graham et al., 2006), based on geography data for specimen collections from the MANIS portal. These specimens could then be arranged underneath a geo-hierarchy of county, state, country/ocean etc in a similar manner to which they were classified by species, genus and family, such that a specimen group was placed under one, and one only, category in this hierarchy. However, species themselves will cover a range rather than be mapped to one exact location as is the case with individual specimens, and such a feature if it is to be developed will need to cope with a one-to-many relationship of species to geographical entities, even if the information is readily available. Fortuitously, the concept relationship model that allows one concept to have multiple relationships to another classification does not have to be limited to taxonomic classifications. We could thus extend the model to enable congruent etc relationships between species and multiple geographic entities.

Discussion

The visualisation supplies a novel method of discovering relationships and equivalences between names in different classifications. Discovering transitive relations between two indirectly related nodes, a need also identified by our observer groups, is thus supported in our application by gradually building up a display of links by selecting relationships that emanate from individual concepts.

However, it could be asked why we do not simply output a name for a classification if supplied with a name in another classification. The answer is that the examples shown so far are simple, involving only two or three links in the chain of relationships, and still in some of these cases an exact match cannot be made. Extra information on who constructed the relationships is needed if a judgement on whether any implied relationships are meaningful is to be made, and our visual interface presents this information in the context of the taxonomies involved in the relationship path.

Calculating the type of overall relationship implied between two concepts based on intermediate relationships is also more problematic than at a first glance. A chain of two or three congruent relationships may well be interpreted with some degree of confidence as forming a congruent relationship between the concepts at the end of the chain. Similarly, a chain of unbroken includes or is_included relationships form an overall relationship with the same meaning as the link relationships. However, the old adage that a chain is only as strong as its weakest link, or in this case, weakest relationship, also applies. A chain involving just one overlap relationship would reduce the entire relationship between the two end concepts to an overlapping relation. A two-step chain of an includes relationship followed by an is_included_in type would in itself translate to an overlapping relationship. The vague ‘not congruent’ relationship would also negate any more specific relationships in a chain. Specifically, with the information present in concepts at the moment, adding together a chain of different relation types to produce an overall type or strength of relationship would quickly dissolve into a fuzzy sense of overlap, much like mixing together different colours of paint swiftly forms a muddy shade of brown.

To complicate matters further, there is also the question of whether transitive relations are only meaningful if the involved component relations are defined by the same person. Different people have different ideas of which concepts are related,

joining those together without consideration may give multiple and conflicting paths between two concepts i.e Peet says A is congruent to B which is congruent to C, however Kartesz may directly state that A only overlaps C, and neither is wrong, they are simply different opinions.

Another common situation that occurs is detailed in Figure 15. A species in one classification (say X in A) is declared congruent to a species in a second classification (Y in B), and a variety of Y (Yv in B) is declared congruent to a variety in a third classification (Zv in C). Given that the parent-child relationship of Y to Yv is essentially an 'includes' relationship internal to a classification, can we then deduce an 'includes' relationship from X through to Zv, if the inter-classification relationships were put in place by one expert, and the internal parent-child relationship being the judgement of the taxonomist who constructed classification B? Is the composer of the inter-classification relationships relying on the transitive path being deduced to avoid the placing of redundant relationships i.e. X includes Zv? It would seem to depend on context, since this expert has placed no relationships indicating he thinks Y and Yv are not inclusive, a user could logically follow the relationship transitively.

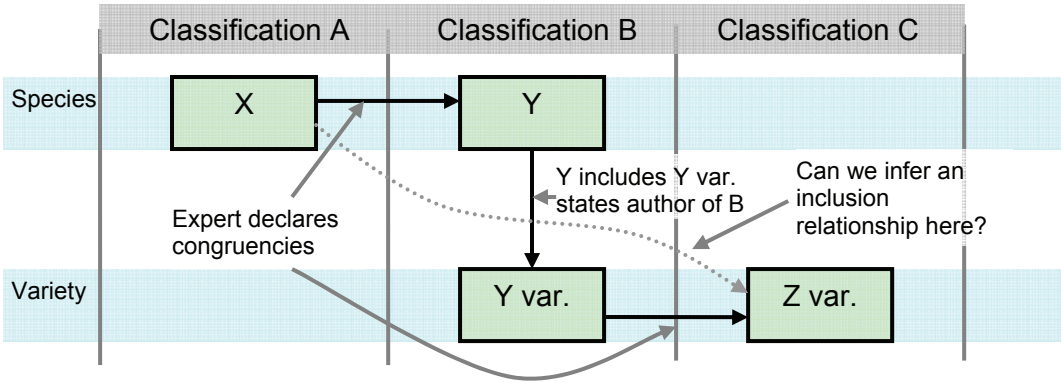


Figure 15. Can we infer a relationship from a chain of two or more authors?

This has some parallels with the ideas of inference in the semantic web in that further findings could be deduced from combining atomic statements. However, the problem here isn't that information stored under different ontologies needs to be mapped and combined, the information here is all defined in the same schema (TCS), rather the problem is whether the thinking behind two or more taxonomist's reasoning is compatible or contradictory and this isn't captured formally in TCS. Comparison of relationships by different authors can reveal obvious contradictions, perhaps revealing an indication of their general compatibility, but cannot be relied upon for a specific situation such as that given in Figure 15. The related idea of trust is similarly hazy, there is no reliable method of deciding whether a taxonomist's findings are trustworthy, as all taxonomies are matters of informed opinion rather than concrete facts.

In summary, when dealing with transitive relationships, it is perhaps best to simply display the data that is present, and allow a user to use their experience to draw conclusions. To this end, the visualisation allows users to mark relationships emanating from individual concepts and to chain these together to produce paths between concepts of interest. It will however not attempt to deduce relationships, leaving this to the judgement of the qualified user.

Conclusion

We have presented a visual application that allows the tracing and comparison of names and concepts across multiple classifications based on the Taxonomic Concept Schema (TCS) data standard. For ecologists the main advantages of this

approach are that it offers the opportunity to accurately discover related concepts for named specimens in alternative classifications in a manner that openly reveals how such a relation is arrived at. Whilst these advantages are inherent to concepts, this is to our knowledge the first visualisation that combines the display of concept relationships with the classifications the concepts reside in. With concept annotation, specimens from surveys or collections that were initially categorised under different classifications can be compared in the context of one of the classifications, or they can be all translated to a different classification altogether if the appropriate concept relationships exist. In the visualisations, related taxon concepts in the classifications can also be noted and explored if desired.

Acknowledgements

We are grateful for the feedback provided on this prototype from interested users at the University of Kansas Natural History Museum; NESCent, Raleigh, NC, USA; Natural History Museum, London; and GBIF, Copenhagen. Particular thanks to Jim Beach, Laura Downey, Bob Peet, Malcolm Scoble, Donald Hobern and Jim Regetz in facilitating these feedback sessions. Further thanks to Bob Peet for adding concept relationship data over the Ranunculus classifications and to Xianhua Liu for subsequently converting this data to TCS format. This work was funded through the SEEK project (Science Environment for Ecological Knowledge) - NSF Grant award 0225676. We would also thank the reviewers of the original draft of this paper for their constructive and helpful feedback.

References

Card, S. K., Mackinlay, J. D. and Shneiderman, B., (Eds.), 1999. Readings in Information Visualization: Using Vision to Think. The Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann, San Francisco, 686 pp.

Godfray, H. C. J., 2002. Challenges for taxonomy. Nature, 417 (6884): 17-19.

Gotelli, N. J., 2004. A taxonomic wish-list for community ecology. Philosophical Transactions of the Royal Society B: Biological Science, 359 (1444): 585-597.

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

Graham, M., Kennedy, J. and Downey, L., 2006. Visual Comparison and Exploration of Natural History Collections. Advanced Visual Interfaces (AVI), Venice, Italy, May 23-26, 2006, p.310-313: ACM Press.

Herman, P. M. J. and Heip, C., 1988. On the Use of Meiofauna in Ecological Monitoring: Who Needs Taxonomy? Marine Pollution Bulletin, 19 (12): 665-668.

Kennedy, J. B., Hyam, R., Kukla, R. and Paterson, T., 2006. Standard Data Model Representation for Taxonomic Information. OMICS: A Journal of Integrative Biology, 10 (2): 220-230.

Kennedy, J. B., Kukla, R. and Paterson, T., 2005. Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. 2nd International Workshop on Data Integration in the Life Sciences San Diego, California, USA, July 20-22, 2005, p.80-95: Springer.

Koperski, M., Sauer, M., Braun, W. and Gradstein, S. R., 2000. Referenzliste der Moose Deutschlands. LV Druck im Landwirtschaftsverlag GmbH, Münster-Hiltrup, 519 pp.

Krell, F.-T., 2004. Parataxonomy vs. taxonomy in biodiversity studies – pitfalls and applicability of ‘morphospecies’ sorting. Biodiversity and Conservation, 13: 795-812.

Lee, B., Parr, C. S., Plaisant, C., Bederson, B. B., Veksler, V. D., Gray, W. D. and Kotfila, C., 2006. TreePlus: Interactive Exploration of Networks with Enhanced Tree Layouts. IEEE Transactions on Visualization and Computer Graphics, 12 (6): 1414-1426.

Levkowitz, H. and Herman, G. T., 1992. Color Scales for Image Data. IEEE Computer Graphics & Applications, 12 (1): 72-80.

MANIS, 2001. Mammal Networked Information System, 9 January, 2006, http://manisnet.org/.

Michener, W., Beach, J., Bowers, S., Downey, L., Jones, M., Ludäscher, B., Pennington, D., Rajasekar, A., Romanello, S., Schildhauer, M., Vieglais, D. and Zhang, J., 2005. Data Integration and Workflow Solutions for Ecology. Data Integration in the Life Sciences (DILS), San Diego, California, USA, July 20-22, 2005, p.321-324: Springer.

Munzner, T., Guimbretière, F., Tasiran, S., Zhang, L. and Zhou, Y., 2003. TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility. ACM Transactions on Graphics, 22 (3): 453-462.

Nielsen, J., 1993. Usability Engineering. Academic Press Professional, Boston, 362 pp.

Sifer, M., 2006. Filter co-ordinations for exploring multi-dimensional data. Journal of Visual Languages and Computing, 17 (2): 107-125.

TCS, 2006. Taxonomic Names/Concepts Sub-group, Taxonomic Concept Schema, October 2005, http://tdwg.napier.ac.uk.

Yoon, I., Yoon, S., Martinez, N., Williams, R. and Dunne, J., 2005. Interactive 3D visualization of highly connected ecological networks on the WWW. ACM Symposium on Applied Computing Santa Fe, New Mexico, USA, March 13-17, 2005, p.1207-1212: ACM Press.