An expert study on hierarchy comparison methods applied to biological taxonomies curation

Lilliana Sancho-Chavarria [Corresp., 1] , **Fabian Beck** [2] , **Erick Mata-Montero** [1]

[1] School of Computing, Costa Rica Institute of Technology, Cartago, Cartago, Costa Rica

[2] paluno, University of Duisburg-Essen, Duisburg, North Rhine-Westphalia, Germany

Corresponding Author: Lilliana Sancho-Chavarria
Email address: lsancho@tec.ac.cr

Comparison of hierarchies aims at identifying differences and similarities between two or more hierarchical structures. In the biological taxonomy domain, comparison is indispensable for the reconciliation of alternative versions of a taxonomic classification. Biological taxonomies are knowledge structures that may include large amounts of nodes (taxa), which are typically maintained manually. We present the results of a user study with taxonomy experts that evaluates four well-known methods for the comparison of two hierarchies, namely, *edge drawing*, *matrix representation*, *animation*, and *agglomeration*. Each of these methods is evaluated with respect to seven typical biological taxonomy curation tasks. To this end, we designed an interactive software environment through which expert taxonomists performed exercises representative of the considered tasks. We evaluated participants' effectiveness and level of satisfaction from both quantitative and qualitative perspectives. Overall quantitative results evidence that participants were less effective with *agglomeration* whereas they were more satisfied with *edge drawing*. Qualitative findings reveal a greater preference among participants for the *edge drawing* method. In addition, from the qualitative analysis, we obtained insights that contribute to explain the differences between the methods and provide directions for future research.

1 **An expert study on hierarchy comparison methods applied to**
2 **biological taxonomies curation**
3

5 Lilliana Sancho-Chavarría[1], Fabian Beck[2], Erick Mata-Montero[1]
6
7 [1] School of Computing, Costa Rica Institute of Technology, Cartago, Costa Rica
8 [2] paluno, University of Duisburg-Essen, Essen, Germany
9
10 Corresponding Author:
11 Lilliana Sancho-Chavarría[1]
12 ITCR, Cartago, Costa Rica
13 Email address: lsancho@tec.ac.cr

15 Abstract

16 Comparison of hierarchies aims at identifying differences and similarities between two or more
17 hierarchical structures. In the biological taxonomy domain, comparison is indispensable for the
18 reconciliation of alternative versions of a taxonomic classification. Biological taxonomies are
19 knowledge structures that may include large amounts of nodes (taxa), which are typically
20 maintained manually. We present the results of a user study with taxonomy experts that
21 evaluates four well-known methods for the comparison of two hierarchies, namely, *edge*
22 *drawing*, *matrix representation*, *animation*, and *agglomeration*. Each of these methods is
23 evaluated with respect to seven typical biological taxonomy curation tasks. To this end, we
24 designed an interactive software environment through which expert taxonomists performed
25 exercises representative of the considered tasks. We evaluated participants' effectiveness and
26 level of satisfaction from both quantitative and qualitative perspectives. Overall quantitative
27 results evidence that participants were less effective with *agglomeration* whereas they were more
28 satisfied with *edge drawing*. Qualitative findings reveal a greater preference among participants
29 for the *edge drawing* method. In addition, from the qualitative analysis, we obtained insights that
30 contribute to explain the differences between the methods and provide directions for future
31 research.

33
34 Introduction

35 Visual comparison of hierarchies has been prevalent in information visualization research
36 because it is relevant for a wide range of domains such as tracking changes in software projects,
37 comparing budgets, and describing dynamics of organizational structures, among others. In this
38 work, we study the comparison of hierarchies in the domain of biological taxonomies.
39 Taxonomic information has been scattered in publications for centuries. In spite of integration
40 efforts of global initiatives in the last decades, there are numerous taxonomic databases around
41 the globe (Ball-Damerow et al., 2019). A visualization tool for the identification of differences
42 and similarities between two versions of a taxonomy would contribute to such integration efforts.

43
44 Taxonomies are hierarchies created by experts to classify living organisms. Through
45 classification, mutually resembling organisms are placed together in categories known as
46 *taxonomic ranks,* which, in turn, make up the levels of the hierarchy. The main taxonomic ranks
47 include *domain*, *kingdom*, *phylum*, *class*, *order*, *family*, *genus*, and *species*. Each node within the
48 hierarchy is referred to as a taxon, that is, a name given to a group of organisms; for
49 instanceinstance, *Vertebrates* and *Mammals* are two taxa, the former is placed at the *phylum*
50 taxonomic rank and latter at the *class* taxonomic rank. Species and sub-species are placed at the
51 lower level of the hierarchy and their scientific names are expressed with a *binomial system of*
52 *nomenclature* that uses a Latin grammatical form. The first part is the genus and the second part
53 is the *specific epithet*. For instance, the domestic cat's scientific name is *Feliz catus* (formerly
54 known, for many years, as *Felix domesticus*), where the epithet name is *catus* and the genus is
55 *Feliz*. Upper levels of the cat's taxonomy are: family *Felidae*, order *Carnivora*, class *Mammalia*,
56 phylum *Chordata*, and kingdom *Animalia*. For a taxon to be recorded in a taxonomy, it must
57 have been described in a publication, either as a new group or as a review of an existing group of
58 organisms. Taxonomy records may include various data; but for comparison, they are required to
59 include at least taxon name, author's name, and the year of publication. This allows users to
60 determine under which judgment the classification was devised. After almost three centuries
61 since modern taxonomy was first established by Carl Linnaeus (Linné & Gmelin, 1767), one
62 might think that most organisms on Earth have been identified and classified, and that
63 taxonomies are rather static. However, on one hand, it is estimated that only about 1.5 million
64 from approximately 11 million species of macro organisms haven been identified and described
65 (Larsen et al., 2017). On the other hand, the dynamics of taxonomic work has lead experts
66 worldwide to end up with different versions of the classifications. Taxa names represent concepts
67 whose definition depends on the authors' criteria, which eventually gives rise to conflicting
68 versions of a taxonomy. These multiple versions will require corrections and re-classifications in
69 order to come to an integrated version that can more accurately document biodiversity. That is
70 how taxonomists often face the problem of reconciling different versions of a taxonomy. For
71 such reconciliation efforts, biological taxonomists require to perform a series of curation tasks.
72
73 Sancho-Chavarría et al. (2016, 2018) characterized curation tasks that involve taxonomic
74 changes when comparing two versions T_1 and T_2 of a taxonomy. Such characterization involved
75 interviews to six experts from three different countries and followed a two-stage analysis. During
76 the first stage, the authors reviewed literature and interviewed experts in order to obtain a list of
77 preliminary tasks. In the second stage, the tasks were shared and discussed with the experts, in
78 order to obtain a final list of ten tasks. Table 1 provides a description of those derived ten tasks
79 organized into three categories, namely, *pattern identification*, *query*, and *edition*. Some of these
80 tasks are domain-independent, and are frequently mentioned in information visualization
81 research (e.g., filtering, focus, retrieving details). Other are domain-specific to biological
82 taxonomy curation work (e.g., tasks in the pattern identification category such as identify splits
83 or identify merges).
84
85 • *Pattern identification* tasks include the identification of *congruent*, *merged*, *split*, *renamed*,
86   *moved*, and *added/excluded* taxa, as well as the *overview* of changes and the visualization of
87   a *summary* of the resulting comparison. Congruence refers to same taxonomic concepts
88   present in both versions of a taxonomy. A split occurs when taxonomists determine that a

89   group previously considered a unit actually consists of several groups of species that should
90   be described separately. Conversely, a merge happens when taxonomists decide that several
91   independent taxa should be combined into the same group. A change of name (i.e., rename)
92   is usually due to a typo that needs to be corrected. A taxon appears moved when it has been
93   re-classified and placed in a different location within the other version of the taxonomy. An
94   addition occurs when a new taxon is added to the taxonomy, either because it is a new
95   discovery or because it had not been previously recorded in the database. Exclusions refer to
96   taxa that are contained in version T_1 but that are missing in the alternative version T_2. It is
97   important to note that, from a taxonomic point of view, once a species is discovered, it is
98   kept in the taxonomy even if the species becomes extinct; however, in this work we consider
99   exclusions because it is important that taxonomists know when records are missing in the
100   database. The *overview changes* task refers to the possibility of globally overviewing all
101   differences between two versions of a taxonomy. The *summarize* task consists of obtaining
102   statistical information on changes.
103   •   Curation tasks in the *Query* category enable users to obtain detailed information on taxa.
104   The *retrieve details* task lets users obtain attributes of a taxon, for instance, the year of
105   publication and the authors' names. The *focus* task refers to the action that users perform
106   when focusing on a group of organisms. Through *filter*, users may find taxa that satisfy
107   some given conditions and through *find inconsistencies*, users may recognize differences due
108   to errors or missing information (e.g., typos or undefined names).
109   •   The *Edition* category comprises just one task that is rather ample, namely, the process
110   through which experts make changes to T_1 and/or T_2 after analyzing the results of a
111   comparison.
112
113   For this research, we considered tasks that are most relevant for the identification of changes
114   between two versions of a taxonomy, that is: 1. *Identify congruence*, 2. *Identify corrections*
115   (splits, merges, moves and renames), 3. *Identify additions/exclusions*, and 4. *Overview changes*.
116   Changes are more likely to occur at lower level taxonomic ranks, for instance, species level
117   (Vaidya, Lepage & Guralnick, 2018), therefore for this study the identification of changes will
118   be visualized only at species level.
119
120   Previous research has contributed with visual models and tools to support the comparison of
121   alternative versions of a taxonomy (Contian et al., 2016; Graham, Craig & Kennedy, 2008; Dang
122   et al., 2015). However, in practice, most taxonomists still rely on simple indented lists to carry
123   out the curation process with little computational assistance. Information visualization provides
124   visual representations that help people perform their tasks more efficiently (Munzner, 2015);
125   therefore, we believe that the identification of differences can be eased with the support of a
126   hierarchy comparison visualization system.
127
128   Graham and Kennedy (Graham & Kennedy, 2010) surveyed the comparison of hierarchies and
129   organized the visualization methods into five categories, namely, *edge drawing*, *coloring*,
130   *animation*, *matrix representation*, and *agglomeration*. The *edge drawing* method presents the
131   two hierarchies as separate structures where differences and similarities are represented by edges
132   from nodes in T_1 to the associated nodes in T_2. The *coloring* method represents similar nodes
133   with the same color. *Animation* shows changes as smooth transitions from one hierarchy to the
134   other. In a *matrix* representation, one hierarchy is placed along the vertical axis and the other one

135 along the horizontal axis; matrix cells indicate relationships between nodes of the compared
136 hierarchies. The *agglomeration* method visually merges both hierarchies into an integrated list.
137
138 The open question that we address in this work is how well these methods support the above
139 curation tasks between two versions T_1 and T_2 of a taxonomy. From the five methods, we leave
140 out coloring as an independent condition since color can be used across all methods. We
141 designed and conducted a user study where twelve expert taxonomists evaluate those four
142 methods. We wanted participants to interact with each of the four methods in a close-to-reality
143 scenario. We developed an interactive software environment that integrates the four methods and
144 allows users to easily navigate from one method to another while doing the assessment exercises
145 (see figures 1 through 4). We wanted to capture the essence of each method and avoided the
146 introduction of features that could potentially favor any particular method. We developed the
147 software taking into consideration the importance of reaching a balanced implementation; thus,
148 the software included same functionality and a common user interface design for all methods.
149 We also carefully selected the data. Datasets contain sufficient types of changes to carry out
150 exercises for all tasks and were also selected to avoid the introduction of any bias due to the
151 potential prior knowledge of the data by the experts. Participants performed the same exercises
152 with each method; however, the target taxa were not the same, also to avoid bias from a learning
153 effect. Immediately after performing the exercises related to a task, participants were asked to
154 answer a user satisfaction questionnaire in order to evaluate each method in relation to the
155 completion of that task. We registered the participants' answers along with their interactions and
156 thinking out-loud comments. We performed a quantitative analysis on the participants' responses
157 to the exercises (i.e., whether they answered correctly or not) and also on their user satisfaction
158 assessment. Additionally, we obtained qualitative findings based on the participants' feedback
159 throughout the session.
160
161 Our contribution with this work is twofold. On one hand, we assessed the effectiveness and level
162 of participants' satisfaction with each visualization method. Participants were likely effective
163 with three methods: *edge drawing*, *matrix*, and *animation*, and also preferred *edge drawing* over
164 the other methods. On the other hand, we obtained a set of themes that contribute to explain the
165 differences between the methods and provide valuable insights for future work on the design of
166 software tools for the comparison of biological taxonomies.
167
168 The rest of this paper is organized as follows. Section 2 presents related work. Section 3 details
169 the study design. It includes a description of the interactive software environment, the
170 participants' profile, the characteristics of the datasets, the user study protocol, and the
171 questionnaire. Section 4 describes the results of the study. In Section 5, we discuss the results
172 and present limitations and implications of the study. Finally, in Section 6, we present
173 conclusions and future work.
174
175

176 **Related Work**
177 Comparison—understood as the examination of two or more items to determine similarities and
178 differences—is a means that facilitates the process of interpreting information. Gleicher et al.
179 (2011) provide a comprehensive survey of visual comparison approaches focusing on comparing

180 complex objects. They analyze a number of publications, systems and designs, looking for
181 common themes for comparison, and propose a *general categorization of visual designs for*
182 *comparison* that consists of three general categories, namely, juxtaposition, superposition, and
183 explicit encoding. The juxtaposed designs place the objects to be compared separately, in time or
184 space. The superposed designs place the objects to be compared one over the other in an overlay
185 fashion. The explicit encoded designs show the relationships between objects explicitly and is
186 generally used when the relationships between objects are the subject of comparison. Hybrid
187 designs are also possible and usually combine two categories. More recently, a set of
188 considerations to understand comparison tasks and their challenges has been discussed (Gleicher,
189 2018), as well as comparison in the context of exploratory analysis (von Landesberger, 2018).
190 Guerra-Gómez et al. (2012) provide a classification of trees for comparison. The trees used in
191 this work are similar to Type 0 trees, which display only a label; however, nodes also contain
192 information about the author and year of publication of the species. Each node also contains the
193 list of synonyms. All these data are necessary in order to identify the types of changes.
194
195 As mentioned above, for the comparison of biological taxonomies we are considering the
196 methods surveyed by Graham and Kennedy (2010) for the comparison of two hierarchies,
197 namely, *edge drawing*, *matrix*, *animation* and *agglomeration*. Each of these methods can be
198 mapped to the mentioned *general categorization of visual designs for comparison*. So, the *edge*
199 *drawing* method comprises characteristics from both juxtaposition (hierarchies are placed
200 separately side by side) and explicit encoding (edges encode the relations between nodes); the
201 *matrix* layout corresponds to an explicit encoding design since the matrix cells can explicitly
202 indicate the relations among taxa; and *agglomeration* corresponds to a superposed design.
203
204 Previous works on hierarchy comparison match these categories. For instance, TreeJuxtaposer
205 (Munzner et al., 2003) compares phylogenetic trees by using a juxtaposed layout. It presents a
206 novel focus+context technique for guaranteed visibility and comparison is approached by
207 coloring. Holten and van Wijk (Holten & van Wijk, 2008) present a visualization method where
208 hierarchies are structured as icicle plots placed in juxtaposition. Relations are explicitly
209 represented by edges arranged through hierarchical edge bundles to reduce cluttering. The
210 Taxonomic Tree Tool (Contian et al., 2016) uses a juxtaposed layout to compare biological
211 taxonomies. It combines glyphs to explain the relations between taxa. ProvenanceMatrix (Dang
212 et al., 2015) compares two taxonomies using a matrix representation. Relations are explicitly
213 displayed through two mechanisms: glyphs and edges. Beck and Diehl (2010) compare two
214 software architectures that use a matrix. Hierarchies here are represented as icicle plots.
215
216 Examples that use animation for comparison are scarcer. Ghoniem and Fekete (2001) use
217 animation to visualize the transition between two alternative representations of the same tree laid
218 out as treemaps. Considering agglomeration-based designs, Beck et al. (2014) present a nested
219 icicle plot approach for comparing two hierarchies and Guerra-Gomez et al. (2012) contrast two
220 trees for the visualization of both node value changes as well as topological differences for the
221 comparison of budgets. Lutz et al. (2014) compared directory structures and conducted a
222 qualitative user study to identify usage strategies. Unlike the above-mentioned work, which did
223 not focus on biological taxonomies, Graham and Kennedy (2007) propose an agglomerated

224   visualization based on directed acyclic graphs for the comparison of multiple biological
225   taxonomies. They also analyzed set-based hierarchies and agglomerated graph-based
226   visualizations for the comparison of botanical taxonomies (Graham et al., 2000).
227
228   Our work differs from the above-mentioned previous studies in that, for the first time, four
229   visualization methods described in (Graham & Kennedy, 2010), are assessed for the comparison
230   of pairs of biological hierarchies with respect to typical curation tasks.
231

232

Study Design

234   The research question addressed is: "*How well does each method support carrying out biological*
235   *taxonomy curation tasks?*" This is assessed both quantitatively and qualitatively. On one hand,
236   we aim at obtaining a quantitative understanding of the participants' effectiveness and level of
237   satisfaction. On the other hand, we also aim at obtaining qualitative insights on the capacity of
238   each method to carry out tasks for the comparison of biological taxonomies. We therefore
239   explore how users interact with the visualizations and what their judgment of each method is. We
240   opted for a within-subject design study that involves four experimental conditions, one for each
241   method (*edge drawing*, *matrix*, *animation*, and *agglomeration*). In this way, each participant
242   could test and contrast all methods. The study fits into the category "Evaluating Visual Data
243   Analysis and Reasoning (VDAR)" (Lam et al., 2012). In this approach, the goal is to assess how
244   a visualization tool supports the analytic process for a particular domain. Accordingly, in our
245   study we want to evaluate how each implemented method supports the identification of
246   similarities and differences for the curation of biological taxonomies. In order to assess each
247   method for each task individually, participants were asked to solve a task with each method, then
248   giving feedback, before moving to the next task. We favored this design over one where
249   participants would first complete all tasks with one method (before moving to the next method)
250   because we wanted to gain task-specific insights on the differences of the tested methods.
251

The Software Environment

253   We developed an interactive web-based software environment that integrates in the same
254   environment the four methods for visual comparison of biological taxonomies described in the
255   Introduction section. The software environment was designed to investigate how these four
256   methods support the taxonomy curation tasks described in Table 1. We tried to conceive a
257   balanced design within the same environment, in order not to favor any method. We also decided
258   to develop a functional system in which participants get a realistic impression. We provided
259   remote web-based access to the software because many of the participants were located in
260   different parts of the world. Figure 1 illustrates the user interface for the agglomeration method,
261   Figure 2 presents the animation method, Figure 3 presents the edge drawing method, and Figure
262   4 presents the matrix method implementation. The two taxonomies to be compared, T_1 and T_2,
263   are displayed as indented lists. Each method implementation is accessible by easily clicking on a
264   tab. Users can inspect the data through the provided zooming features and by vertically scrolling
265   for all methods. Additionally, horizontal scrolling is provided for *matrix*. The visualization
266   layout is of course method-dependent. For *edge drawing* and *animation*, taxonomies are placed
267   juxtaposed. T_1 is placed on the left side of the screen and T_2 on the right side. For *matrix*,

268 taxonomy T_1 is also placed on the left side but T_2 is at the top of the matrix. Finally, for
269 *agglomeration*, T_1 and T_2 are interleaved and centered horizontally.
270
271 The main menu is common to all methods, which adds uniformity to the user interface; however,
272 the implementations of the different visualizations were optimized to best reflect the intrinsic
273 characteristics of each method. The main menu is located at the top of the window and contains
274 eight toggle buttons that display the changes induced by each type of curation task that we are
275 considering, namely, *congruence*, *splits*, *merges*, *moves*, *renames*, *new*, *exclusions*, and an
276 additional *all* button. For example, when the *splits* switch is on, the visualization shows how
277 each taxon with a split in T_1 is divided into taxa contained in T_2. The system is flexible enough to
278 allow users to turn several buttons on at the same time, in case they want to have several types of
279 changes displayed simultaneously. For *animation* (see Fig. 2) additional controls to play and stop
280 animations were added.
281
282 The color-coding scheme of the toggle buttons (i.e., tasks) is also the same across all methods
283 and defines the types of changes to be visualized. We tried to use hues that were easily
284 distinguishable to the eye in order to avoid confusion: blue for congruent taxa, pink for splits,
285 orange for merges, light green for moves, light brownish purple for renames, red for exclusions,
286 and green for added taxa. Familiar codes were also used; that is, red for exclusions and green for
287 new taxa added. The representation of relations depends on the comparison method. For
288 agglomeration, relations have to be inferred since data is interleaved and no explicit additional
289 marks or lines can be included easily. Hence, for this method we decided to use an augmented
290 color code in order to have a cue that would make it easier for participants to recognize to which
291 taxonomy a node belongs and to highlight the types of changes between T_1 and T_2 (i.e., the
292 relations between nodes). For this, we use the same hues but with different intensity, so the light
293 nuanced nodes in the agglomerated structure indicate that they belong to the taxonomy of origin
294 T_1 while the darker nuanced nodes indicate that they belong to the taxonomy of destination T_2.
295 For example, a split of a taxon x in T_1 into taxa p, q, and r in T_2 would show x in light pink color
296 and p, q, and r in a darker hue of pink. A legend was added to explain this color-coding. In the
297 *agglomeration* method relations are permanent but not explicit.
298
299 In the *edge drawing* and *matrix* methods, relations are explicit and permanent. For instance, with
300 the *edge drawing* method, a split of a taxon x in T_1 into taxa p, q, and r in T_2 is shown as three
301 pink edges going horizontally from taxon x to p, q, and r in taxonomy T_2. In the *matrix* method,
302 the same split case is shown as marked colored cells $(x; p)$, $(x;q)$, and $(x; r)$ respectively.
303
304 For *animation*, we considered two design choices: "animation by movement" and "animation by
305 emergence". In the former, an animation consists of moving the target taxon from T_1 to its new
306 position in T_2. In the latter, the target taxon would fade out from T_1 and would gradually appear
307 in T_2. In either case, relations are explicit although temporary because they disappear when the
308 animation is finished. We chose the first option because the paths followed by each moving
309 taxon provides better traceability cues than the second one. Considering the split case described
310 above, the animation would show x moving towards taxonomy T_2. On its way, x splits and
311 disappears to let p, q and r appear and keep moving until each of them reaches its definitive
312 position in T_2. The *animation* method per se does not necessarily involve leaving an explicit trace
313 (as *edge drawing* does).

314
315   Two curation tasks do not involve relations between nodes in the alternative taxonomies,
316   namely, *identify new taxa added* and *identify excluded taxa*. Given that inclusions and exclusions
317   take place only in one of the taxonomies, the system visualizes these situations only in the
318   taxonomy in which they occurred. Thus, excluded taxa are visualized in red color in taxonomy
319   T_1 and included taxa in green color in taxonomy T_2. Without the use of color, asking users to
320   visually infer which taxa were excluded from T_1 and which ones were included into T_2 would
321   require too much mental effort, especially when taxonomies are large.
322
323   The software included interactions that we considered crucial, such as navigating, selecting, and
324   zooming. For zooming, we took advantage of the browser's functionality and it is offered via
325   three different alternatives: a) by using the zoom feature on the browser's menu, b) using the
326   browser controls, that is, by pressing the <ctrl> key and moving the mouse scroll wheel up or
327   down, or c) pressing the key *I* on the keyboard for *zoom in* and the key *O* for *zoom out*. Functions
328   such as undo/redo, edit, and statistics were left out of the prototype since they were not crucial
329   for the study. We developed the software incrementally through several iterations until we
330   reached balanced implementations of the four methods. At the end of each iteration, computer
331   science students tested the software. Tests were also conducted involving an experienced
332   taxonomist and a PhD student in computer science.
333

334   **Participants**
335   Twelve experts were recruited from our professional network. They received no compensation.
336   Table 2 summarizes the participants' profiles. Each participant was given an identification
337   number, ranging from E1 to E12. Eight of them are botanists (three of which are also Forestry
338   engineers), two biologists (one entomologist and one ichthyologist), one ecologist, and one
339   computing engineer (with 21 years of experience in biodiversity informatics). In addition, three
340   of them reported Biodiversity Informatics as a second area of expertise. One participant holds an
341   Engineering degree, five have a Master's degree, and six have a PhD degree. Their average
342   professional experience was 28 years and their average experience in the taxonomy field was 23
343   years; this includes taxonomic classification, taxonomy nomenclature, and curation of biological
344   taxonomies. Ten participants are male and two female. Three participants worked as full time
345   university professors and the rest worked full time at herbaria, museums, or biodiversity
346   conservation initiatives. Participants came from three different countries and their expertise was
347   with different taxonomic groups of organisms.
348

349   **Datasets**
350   We carefully selected and designed the datasets, taking into account the level of familiarity
351   participants might have with the data. Although taxonomists, in general, have extensive
352   knowledge on certain groups of species, in practice, a taxonomist is only expert on a limited
353   group of organisms; therefore, not all taxonomists need an overview of large subtrees. In
354   addition, because of the large number and complexity of groups of species, their expertise is also
355   geographically focused. Thus, despite of having ten botanists in our group of experts, all of them
356   specialize in different groups of plants. In order not to favor any participant and avoid the
357   eventual bias, we did not choose groups of species that were known by any of the experts.
358   Therefore, we chose an unfamiliar taxonomy. It should not be very large since we did not want
359   to burden participants by spending too much time performing the user study. However, at the

360   same time, the dataset should be large enough to contain representative cases of all types of
361   changes. We therefore used a small-size real taxonomy and derived artificial variants from it.
362
363   We downloaded a set of 66 species of amphibians from the Catalogue of Life website
364   (http://www.catalogueoflife.org/) with a total of 96 nodes. We called this the *seed taxonomy*
365   from which we derived variations on the datasets (derived taxonomies) to be used for each
366   method. It was important that the data sets were different, but at the same time similar enough to
367   be comparable. Therefore, we programmed an *artificial taxonomy generator* to which we input
368   as parameters the percentage of splits, merges, movements, renames, additions, and exclusions
369   that we wanted to add to the seed taxonomy T_1. The generator randomly selected the taxa to
370   introduce the changes and verified that only one change was introduced to each taxa to be
371   modified. In this way, we prevented data conflicts, since more than one change to a taxon could
372   generate inconsistent data. We also verified that, although questions were identical, the datasets
373   would produce different answers for each method. The number of nodes in the derived datasets
374   varied between 78 to 116 nodes. Table 3 describes the main characteristics of the four derived
375   datasets, that is, the number of nodes, of species, of splits cases, merges, moves, renames, new,
376   and excluded taxa. The goal of this setup was to ask experts to visualize changes in four pairs of
377   datasets: (T_1, T_2), (T_1, T_3), (T_1, T_4), and (T_1, T_5), with respect to *edge drawing*, *matrix*, *animation*,
378   and *agglomeration*, respectively. The derived datasets T_2, T_3, T_4 and T_5 are similar because they
379   are all obtained from the seed taxonomy and have roughly the same number of changes. We
380   avoided the use of the same pair of datasets across all visualization methods in order to neutralize
381   a potential bias introduced by a learning effect.
382

383   **User Study Protocol**
384   We planned the user study for a 2-hour session with each participant. During the session,
385   participants would work with the interactive software environment to perform some exercises
386   and to answer questions from a questionnaire. Seven out of the twelve experts lived overseas;
387   therefore, the session was conducted remotely via a video call for them. For the rest of
388   participants, sessions were face-to-face. For participants in remote sessions, at the beginning of
389   the session, we shared a link were the software and data were hosted. In case of the face-to-face
390   interviews, we supplied a laptop computer. In both settings, access to the software environment
391   was via web browser. We followed the same interview protocol for all participants.
392
393   A written guide and a 15-minute descriptive video of the software environment were available to
394   the participants at least two days before the session, so that they could get familiar with it.
395   Access to the software environment, datasets, and questionnaire was not provided before the
396   interview session.
397
398   A moderator was leading the session and assisted participants, while an observer was taking
399   notes. Participants did not have to write down the answers; both the moderator and the observer
400   would write the participants' answers on an answer sheet that they had previously prepared.
401   Audios of the interviews were recorded for later confirmation of answers and analysis. At the
402   beginning of the session, we checked to see if the participants had studied the guide and video
403   beforehand and if they had any questions. In case they had not done so or if they needed to
404   clarify any aspect, the moderator offered a demonstration of the software and resolved the
405   doubts. Exercises were not started until both the participant and moderator felt they were ready;

406   only then did the moderator provide the link for participants to access the interactive
407   environment. Working speed was not to be measured and participants were made aware of the
408   fact that they had no time limit to answer the questions and were able to express any inquiry,
409   doubt or suggestion at any time. Participants were also asked and reminded to think aloud while
410   solving the questions. Our goal was to get insights on how they carry out the data exploration
411   and the tasks.
412
413   We designed an instrument that consists of twelve task performing exercises, nine method
414   assessment questions and one open-ended comments section. The task performing exercises have
415   clearly correct answers and were intended to measure the participants' effectiveness. The method
416   assessment questions were intended to obtain participants' perception. The purpose of the open-
417   ended question was to obtain additional feedback on user satisfaction and suggestions for a
418   future design of an interactive visualization system. The study started with an exercise where
419   participants had to identify the most common type of change (overview task). Next, exercises
420   were targeted to identify splits, merges, renames, moves, added or excluded taxa, and ended with
421   an overview question again. Each task-performing exercise had to be answered with each
422   method. For instance, instructions such as "Use the Matrix method: Explore the visualization and
423   find into what taxa *Babina caldwelli* was split?" were followed by the same question for all
424   methods. However, the taxon to be used in each exercise (*Babina caldwelli*) was different for
425   each method.
426
427   We randomized the order in which participants used each method on each question. Participants
428   performed the exercises related to one task (for instance, identification of splits) and then were
429   asked to assess each method to perform such task. The nine method-assessment questions
430   consisted of five-level Likert scale items that assessed how good each method was to carry out
431   the task. In the course of the session, participants had access to a copy of the questions and
432   instructions, especially because taxa names were in Latin, and we wanted to avoid any
433   confusion.
434

Analysis

436   For the analysis of the results, we organized the participants' responses into a spreadsheet. We
437   gathered three types of data: a) the effectiveness data, i.e., whether the participants answered
438   each question correctly or incorrectly, b) the user satisfaction data, i.e., the Likert-scale ratings
439   that participants gave to each method after accomplishing each task, and c) the qualitative data,
440   i.e., the thinking-aloud comments and the suggestions that participants provided during the
441   session. Quantitative analysis was performed on data of types a) and b) by using a statistical
442   package. We used non-parametric statistics with alpha=0.05 and compared medians to determine
443   that differences are not due to chance. For the analysis of effectiveness, we used the Cochran's Q
444   test, which can be used when you have a group of people performing a series of tasks where the
445   outcome is dichotomic (e.g. success or failure). For the analysis of participants' satisfaction, we
446   used the Friedman test, which is appropriate for within-subjects designs that have three or more
447   conditions, and particularly it can be used for the analysis of ordinal data, such as the Likert-
448   scale responses (MacKenzie, 2013). When necessary, both tests were followed by pairwise
449   comparison using Dunn's test with Bonferroni correction.
450

451 For the qualitative analysis, we applied the following procedure. The responses were first placed
452 in the same order in which the questions were presented to the participants, and, then, they were
453 sorted by method. Two columns were designated for each participant, one to record their
454 comments and suggestions (e.g., E1) and another one to afterwards register the codes generated
455 during the qualitative analysis (e.g., E1-codes). Secondly, we listened to the audio recordings
456 checking for additional feedback from the participants, which we added to the spreadsheet.
457 Thirdly, we conducted a qualitative analysis: the first author made several coding passes using
458 *open coding* (Charmaz, 2006) to obtain a first coding version that was then shared with the other
459 authors. We coded participants' interactions and feedback. Repeated or related topics were
460 grouped together, revised and re-grouped through several refinement cycles until we reached an
461 agreement with twelve categories to finally conclude with four meaningful themes. During the
462 process, we also organized the positive and negative comments, as well as the participants'
463 suggestions for improving the methods.
464

Results

466
467 The study took 2:15 hours on average per participant. We first present quantitative results on
468 participants' effectiveness and satisfaction, and then findings from the qualitative analysis.
469

Effectiveness

471 The results of the participants' effectiveness on the task-performing exercises are summarized in
472 Table 4. Overall results indicate that participants obtained more correct answers with *matrix*
473 (94%), then with *edge drawing* (88%), followed by *animation* (87%) and then with
474 *agglomeration* (73%). We tested for statistical significance by using Cochran's Q test for N=12
475 and DF=3. We did not find significant differences on participants' responses between pairs of
476 methods (*matrix*, *edge*), (*matrix*, *animation*), and (*edge*, *animation*). However, we did find
477 differences (χ^2=40.480, p-value = 0.05) between agglomeration (73%) and the other methods,
478 meaning that participants were less effective with the agglomeration method.
479
480 We also did a quantitative analysis on responses to each exercise. We did not find significant
481 differences among participants' responses when identifying: a) into which taxa a taxon was split
482 (exercise 2), b) whether species were merged and how (exercises 4 and 5), c) whether species
483 were renamed (exercises 6 and 7), d) whether any species were added to a version of the
484 taxonomy (exercise 9), and e) whether any species were excluded (exercise 10). We found
485 significant differences in participants' responses in identifying: a) an overview of changes
486 (exercises 1 and 12), b) which species were most divided (exercise 3), c) moved taxa (exercise
487 8), and d) all changes on a taxon (exercise 11). For these cases, a post hoc pairwise comparison
488 was performed in order to determine where the differences occurred. Resulting less-effective
489 methods are highlighted in orange color in Table 4:
490 • Exercise 1. Overview of changes. We found differences (χ^2=25.500, p-value < 0.05) between
491   the following pairs of methods: (*animation*, *edge drawing*), (*matrix*, *edge drawing*) and
492   (*agglomeration*, *edge drawing*). This indicates that the effectiveness with *edge drawing* (8%)
493   was lower with respect to *agglomeration* (75%), *animation* (92%) and *matrix* (92%).
494 • Exercise 3. Identification of splits. We found differences (χ^2=17.571, p-value <0.05) between
495   pairs of methods (*edge drawing*, *agglomeration*) and (*matrix*, *agglomeration*). These results
496   indicate that the effectiveness with *agglomeration* was different with respect to the other

497   methods. Participants were less effective with *agglomeration* (33%) and more effective with
498   *edge drawing* (100%) and *matrix* (92%).
499 • Exercise 8. Identification of moved taxa. We found differences (x^2=21.000, p-value < 0.05)
500   between pairs of methods (*animation*, *agglomeration*), (*edge drawing*, *agglomeration*), and
501   (*matrix*, *agglomeration*). This indicates that participants were less effective with
502   *agglomeration* (42%) whereas they were more effective with the other methods (100%).
503 • Exercise 11. Focus on a taxon. We found differences (x^2=9.692, p-value < 0.05) between the
504   pair of methods (*agglomeration*, *matrix*). This indicates that participants were less effective
505   with *agglomeration* (50%) and more effective with *matrix* (100%).
506 • Exercise 12. Overview of changes. We found differences (x^2=10.714, p-value < 0.05)
507   between pairs of methods (*animation*, *agglomeration*), (*edge drawing*, *agglomeration*), and
508   (*matrix*, *agglomeration*). This indicates that participants were less effective with
509   *agglomeration* (17%) than with *animation* (58%), *edge drawing* (58%), and *matrix* (58%).
510
511 **Satisfaction Level**
512 Right after carrying out the task-performing exercises for each task, participants answered a
513 Likert-scale questionnaire to assess the methods. The questions had the following structure:
514 "How good do you think each method is in order to perform task t? For each method provide a
515 rating between 1 and 5, where 1 stands for 'poor', 2 for 'fair', 3 for 'good', 4 for 'very good',
516 and 5 for 'excellent'". We performed a statistical analysis on the participants' ratings using the
517 Friedman test. Table 5 summarizes the results for N=12 and DF=3. We did not find any
518 difference in participants' responses to accomplish the task for the identification of excluded
519 species (question 10e). Neither we found differences regarding the identification of added
520 species (question 9e) after running the post pairwise comparison. On the contrary, we found
521 differences in participants' responses to carry out tasks for the identification of the most common
522 type of change (1e), splits (3e), merges (5e), renaming (7e), moves (8e), changes to a taxon (12e)
523 and the general methods assessment question (13). The post hoc pairwise comparison gave the
524 following results (in Table 5, higher satisfaction level is shown in green and less satisfaction
525 level is highlighted in orange):
526
527 • Question 1e-overview (x^2=12.588, p-value =0.05). We found differences between pairs of
528   methods (*animation*, *edge drawing*). Participants gave a better rating to the *edge drawing*
529   method (median = 3.29 ~ 'good/very good') than *animation* (median = 1.79 ~ 'fair').
530 • Question 3e-splits (x^2=33.055, p-value = 0.05). We found differences between pairs of
531   methods (*agglomeration*, *matrix*) (*agglomeration*, *edge drawing*) (*animation*, *matrix*), and
532   (*animation*, *edge drawing*). There was no difference between *agglomeration* and *animation*,
533   and neither *matrix* and *edge drawing*. Participants ratings for *agglomeration* (median = 1.50
534   ~ 'poor/fair') and *animation* (median = 1.50 ~ 'poor/fair') were the lowest while for *matrix*
535   (median = 3.21 ~ 'good') and e*dge drawing* (median = 3.79 ~ 'good/very good') were the
536   highest ones.
537 • Question 5e-merges (x^2=20.050, p-value = 0.05). We found differences between pairs of
538   methods (*animation*, *edge drawing*) and (*agglomeration*, *edge drawing*). Participants
539   assessed the *edge drawing* method with the highest rating (median = 3.71 ~ 'good/very
540   good') compared to *agglomeration* (median = 2.0 ~ 'fair') and *animation* (median = 1.71 ~
541   'poor/fair').

542 • Question 7e-renames (χ^2=21.559, p-value < 0.05). We found differences between pairs of
543   methods (*animation*, *edge drawing*) and (*agglomeration*, *edge drawing*). Participants
544   assessed the *edge drawing* method with the highest rating (median = 3.67 ~ 'good/very
545   good') than *agglomeration* (median = 2.13 ~ 'fair') and *animation* (median = 1.50 ~
546   'poor/fair').
547 • Question 8e-moves (χ^2=24.295, p-value = 0.05). We found differences between pairs of
548   methods (*agglomeration*, *edge drawing*) and (*animation*, *edge drawing*). Participants
549   assessed the *edge drawing* method with the highest rating (median = 3.83 ~ 'very good') than
550   *agglomeration* (median = 1.54 ~ 'poor/fair') and *animation* (median = 1.88 ~ 'poor/fair').
551 • Question 12e-focus (χ^2=16.057, p-value = 0.05). We found differences between pairs of
552   methods (*animation*, *edge drawing*) and (*agglomeration*, *edge drawing*). Participants
553   assessed the *edge drawing* method with the highest rating (median = 3.46 ~ 'good/very
554   good') than *agglomeration* (median = 1.88 ~ 'fair') and *animation* (median = 1.79 ~ 'fair').
555 • Question 13-general assessment of all methods. (χ^2=16.057, p-value = 0.05). We found
556   differences between pairs of methods (*animation*, *edge drawing*) and (*agglomeration*, *edge*
557   *drawing*). Participants assessed the *edge drawing* method with the highest rating (median =
558   4.97 ~ 'excellent') than *agglomeration* (median = 2.97 ~ 'good') and *animation* (median =
559   2.84 ~ 'good').
560
561 It is important to notice that response for question 13 summarizes the participants' level of
562 satisfaction.
563
564 **Findings from Qualitative Analysis**
565 Regarding *edge drawing*, all participants referred to this method throughout all exercises with
566 expressions such as: "easy", "very direct", "I can easily relate taxa", "it is very fast", "I do not
567 have to think too much", and "you can see ... at a glance". One participant (E6) said that it was
568 the best because "you can clearly see the origin and the destination". Another participant (E7)
569 considered that *edge drawing* "is familiar, it is similar to an 'associate' type of exercise".
570
571 Participants' feedback on *matrix* highlighted this method as good for the visualization of general
572 overviews and the identification of patterns, eight participants mentioned it (E2, E3, E7, E8, E9,
573 E10, E11, and E12). Five participants (E6, E7, E10, E11 and E12) mentioned that *matrix* was the
574 fastest one. Another participant (E2) recognized that "it is easy to see in a row the changes to a
575 taxon". Three participants (E10, E11 E12) considered that the required vertical and horizontal
576 scrolling add complexity and two other participants (E2 and E10) mentioned that scaling could
577 be a problem. One participant (E1) complained that he had to use his fingers on the screen to
578 follow the relations in the two dimensions. Eleven participants complained about the vertical
579 name implementation in the top hierarchy (all except E2), and two participants (E1 and E10)
580 found navigation difficult because parts of the hierarchies were off the screen.
581
582 Regarding *animation*, two participants (E1 and E8) rated it positively indicating that it was
583 "dynamic", and therefore "fun". However, eight out of the twelve participants described this
584 method in negative terms such as "difficult", "ineffective", "hard to follow", "complicated", "not
585 intuitive", and "waiting until the end of the animation is a waste of time". Five participants
586 considered that the *animation* was not necessary. Participants emphasized that changes between
587 the two taxonomies were very difficult to follow because they could very soon forget what

588 happened, especially if taxonomies were large. Five participants indicated that, while taxa were
589 moving, it was easy to lose track of the relation between origin and destination because the taxa
590 were moving. Most participants speeded up the animation, giving the impression that they
591 wanted it to get to an end quickly, but some had to execute it several times before being able to
592 solve the exercise.
593
594 Participants indicated that it was very difficult to carry out the tasks with the *agglomeration*
595 method, except for the identification of excluded taxa. Eight participants (E2, E5, E7, E8, E9;
596 E10, E11, E12) referred to *agglomeration* as very good when looking for specific taxa or to
597 focusing on a small part of the taxonomy. However, all participants also described it in negative
598 terms, such as "difficult", "very complicated", "requires too much effort", "not evident",
599 "confusing", and "very difficult to know origin and destination". Participants complained that
600 this method involved many variables that were difficult to remember (that is, many color hues)
601 and that it required considerable effort to recognize differences. However, two participants
602 thought that the *agglomeration* view could be complementary to *edge drawing*, and that it could
603 work well for small taxonomic groups.
604
605 We coded and organized the participants' feedback until we reached themes that we considered
606 meaningful. Our observations show four specific issues that are relevant when performing tasks
607 for the curation of biological taxonomies:
608
609 • **Explicit representation of changes**. Changes are visualized through relations among taxa. Nine
610   out of twelve participants clearly indicated that being able to identify the origin and
611   destination of relations was very important to recognize changes when comparing biological
612   taxonomies. Participants' suggestions such as; "add edges to animation", "add edges to
613   matrix", or "add numbers to each change in the agglomeration method" are indications that
614   they would prefer to see relations explicitly and, therefore, prefer methods that explicitly
615   represent the changes.
616 • **Efficiency**. Participants often commented about speed and time needed to solve the exercises.
617   Across exercises, they referred to the importance of understanding what is going on at a
618   glance. They expressed feeling frustrated when having to wait for the animation to end. They
619   speeded up the animation when they felt that solving the exercise was taking too much time.
620   Participants considered that having to scroll horizontally and relate rows and columns of the
621   matrix or having to interpret different colors as in agglomeration were steps that consumed
622   time.
623 • **Multiple views**. Several participants commented that the methods could be complementary;
624   for instance, that the *edge drawing* and the *matrix* methods could be used to visualize all
625   cases at once whereas the *animation* and the *agglomeration* methods could be useful when
626   analyzing smaller groups of species. They explained that, by combining several methods, the
627   advantages of one method could overcome the disadvantages of another one. On the other
628   hand, the experts also emphasized the convenience of having both overview and detailed
629   views; the first one to obtain a general understanding of changes and the second one to obtain
630   detailed information on a focused part of the taxonomies.
631 • **Visual and numerical summaries**. When asked for number of taxa that match a certain
632   condition, participants expressed their frustration because they had to count manually and
633   suggested to add statistics to the software environment. Although obtaining statistical

634 information is one of the tasks for the curation of biological taxonomies, we decided to leave
635 it out of this study on purpose in order to force participants focus on the visualizations. The
636 intention of quantity-related questions was to see if participants were able to visually identify
637 magnitude of changes (for instance, matrix resulted good). We obtained confirmation on the
638 importance of providing numerical understanding of changes.
639
640 **Suggested Improvements**
641 The methods that received most suggestions for improvement were the ones that had the lowest
642 participants' effectiveness and preference. Suggestions for agglomeration focused on
643 mechanisms that would make the relations explicit somehow and allow them to be recognized
644 quickly. For instance: a) add numbers instead of different color hues to indicate taxa of the origin
645 and taxa of the destination, b) use a different color hue for each change (not only for each type of
646 change), c) instead of using several color hues, consider glyphs or some other visual cues, d)
647 color the background of the text instead of the text, e) visually separate the taxonomies on user's
648 demand, f) separate the legend so that hues associated to origin are placed on the left side and the
649 hues associated to the destination are placed on the right side of the screen. For animation
650 participants provided suggestions such as: a) add a time slider and a rewind button, b) identify
651 each change with a number, c) add traces as in edge drawing, d) identify each specific change
652 with a different color, e) maybe consider the use of animation for comparing a small part of the
653 taxonomies. For matrix, participants also made suggestions to improve the visualization of
654 relations; for instance, a) add a feature to freeze rows or columns in place and ease the
655 visualization of relations when vertical or horizontal scrolling is needed, b) add horizontal and
656 vertical guiding lines to ease following the relations, c) add a colored rectangle around the
657 excluded or added taxa in order to highlight these changes, d) add edges in order to make the
658 relations more explicit, e) consider the matrix method as a way to feed a database with the
659 relations between the two taxonomies. Suggestions for edge drawing included: a) use more
660 intense color hues and b) provide features that ease the comparison between taxa of higher
661 taxonomic ranks (for instance, at the family level). Three participants (E3, E5, E6) mentioned on
662 several occasions that the identification of relations was easier when the involved taxa were
663 closer together, within the same view. They expressed this thought as they were solving matrix
664 and agglomeration exercises. On the other hand, four participants (E3, E6, E10, E12) mentioned
665 that vertical scrolling was fine in edge drawing, since "it is very familiar". Feedback obtained
666 from the open-ended comments section of the questionnaire also included suggestions for
667 enhancements to the implementation of the methods. Regarding the representation of hierarchies,
668 participants' suggestions included the elimination of lines that indicate hierarchical structure and
669 to use only indentation as a visual cue to recognize hierarchy (some participants believed that the
670 visualization could look cleaner). They also mentioned that the visual clutter caused by long
671 names might be overcome by using abbreviations when possible (for instance, for the genus part
672 of the species names).
673

Discussion

675

676 The research question we investigated was how well each of the four methods of hierarchy
677 comparison supports the tasks of contrasting two versions of a taxonomy. The quantitative and
678 qualitative results revealed differences among the methods. The difference in effectiveness
679 occurred only with respect to agglomeration, as the participants were the least effective with this

680 method. One likely explanation is that in all methods, the changes between the two taxonomies
681 were consistently represented throughout predetermined colors for each type of change, but in
682 agglomeration each change was represented by two tonalities of the predetermined color, one to
683 indicate the taxon of origin and the other to indicate the taxon of destination. This might have
684 added complexity that affected the participants' effectiveness with this method. Overall,
685 participants were very effective with the other three methods, which might be because the
686 participants could take as much time as they considered necessary to solve the exercises. The
687 number of correct answers with animation and with edge drawing was quite similar (125 and 126
688 respectively). In spite of this similarity in the participants' accuracy, and that both methods used
689 juxtaposed layouts, the user satisfaction results indicate greater participants' preference for edge
690 drawing. Comparing the amount of correct answers, participants showed similar performance in
691 many exercises with edge drawing and matrix. The difference between these two methods comes
692 mainly from the responses to the overview exercise where they had to identify "1. Which is the
693 most common type of change?" where only one participant answered correctly with edge
694 drawing and 11 participants with matrix. This might indicate that matrix works well to get a
695 general overview of changes. This is reflected also in the participants' feedback when they
696 highlighted that this method was good for pattern recognition.
697
698 The effectiveness on recognizing new taxa or excluded taxa was similar with all methods. This is
699 explained by the fact that both new and excluded species are visualized only in one of the
700 taxonomies, and require no relations between the involved taxonomies.
701
702 In another respect, we noticed that participants were more effective at identifying changes at the
703 lowest level of the taxonomy (i.e., species level) than when trying to recognize changes at upper
704 levels (such as at the genus level, exercise 12). This might suggest the convenience of having
705 summary overviews on changes at higher-level taxa.
706
707 Both the quantitative and the qualitative analysis coincide that the agglomeration method ranked
708 last. In spite that the results on effectiveness did not show clear differences among animation,
709 edge drawing and matrix, both the participants' feedback on satisfaction and the qualitative
710 findings suggest edge drawing in the first place.
711
712 **Threats to Validity**
713 Various factors can limit the validity of the results of visualization user studies, including the
714 number of participants, the choice of the datasets, the design of the study, and the data analysis.
715
716 The study included twelve expert users. A small number of participants can affect the quality of
717 results, especially with respect to quantitative results. However, with respect to qualitative
718 findings, the number was sufficient: After completing a first set of nine interviews, we recorded
719 the data on a spreadsheet and did a preliminary processing. Afterwards, when we finished all
720 twelve interviews, we noticed that the qualitative results repeated (that is, with 33% more
721 participants than in the preliminary processing). Such consistency is an indication of
722 dependability regarding the qualitative results.
723
724 Despite the task randomization and the careful selection of data, a learning bias might have been
725 introduced. Learning effects are a risk to all small-scale studies as randomization only partly

726   counter balances for small numbers and individual opinions of participants might have been
727   influenced by the sequence of tasks.

728

729   A restriction of the study is the small size of the datasets, which contained between 78 and 116
730   nodes. As the data were unknown to all participants, the datasets had to be restricted to a small
731   size in order to carry out the study in reasonable time. Still, the datasets contained all types of
732   required changes to perform the tasks. In the case of the dataset for matrix, it turned out a little
733   smaller than the other ones, but large enough to allow users to experience both horizontal and
734   vertical scrolling, so we considered it was fine for the study; however, the difference in size
735   compared to the other datasets might have added some bias to the study. For visualizing larger
736   excerpts or whole taxonomies, other visualizations might be necessary, and our results may not
737   generalize to this. Also, the binding of datasets to methods could have introduced a slight
738   confounding factor but seemed an acceptable trade-off, which we took to simplify the evaluation
739   logistics.

740

741   Although the transferability of results is limited by a) the specific domain application, b) the
742   tasks studied, and c) the data sets, the comparison of hierarchical structures is independent of the
743   application domain; thus, some features of the study might contribute to other contexts where
744   users need to identify divided, joined, moved or renamed nodes between two hierarchies.
745   The within-subject design is applied in studies with a small number of participants. All
746   participants interacted with every method. In this way, we expected to reduce biases associated
747   with individual differences. However, the within-subject design may bias participants because of
748   the carry over effect; that is, once participants perform a task with one method, they may expect
749   certain conditions to happen in the next method to evaluate. We tried to counterbalance bias from
750   the learning and tiring effects by asking participants to interact with the methods in a randomly
751   established order. In addition, although the questions were identical, the data sets would produce
752   different answers for each method in order to prevent the learning effect.

753

754   A bias may be introduced by the design of the study or the design of the interactive software
755   environment. We tried to implement the essence of the methods as well as to keep a standard
756   user interface for all visualizations (same main menu and color codes throughout all methods),
757   but limitations may come from the design choices and implementation decisions. For instance,
758   zooming and scrolling features were limited and not designed for very large amounts of data;
759   also, participants were unsatisfied regarding the vertical text orientation of the matrix
760   implementation. Applying coloring was more difficult in the agglomeration approach because, in
761   this method, color is also required to indicate the different versions of the taxonomy. We decided
762   for a mixed brightness and color encoding as a tradeoff, which seemed introducing a smaller
763   bias, instead of avoiding the use of color for encoding types of changes or introducing an
764   inconsistent use of color across the methods.

765

766   We aimed to objectively examine the collected data. For the quantitative part of the study, we
767   used statistical tests to analyze if the differences between the medians were significant. For the
768   qualitative part of the study we carefully organized the data and coded the participants' feedback
769   and interactions through several refinement cycles. By counting and grouping similar feedback
770   from participants, we were able to define the codes and themes. Nonetheless, the interpretation
771   of the data may be subject to the perspective of the researchers.

772

Implications

773

774   The participants performed well with edge drawing and consider it, in general, the best method;
775   despite they did not have the best performance with it. For overview tasks, participants showed
776   similar effectiveness with matrix and animation, however they preferred matrix.
777

778   Results indicate that identifying explicitly the origin and destination of taxa is very relevant for a
779   more efficient identification of changes; edge drawing and matrix methods seem to have
780   facilitated it. The participants' need to determine origin and destination may explain that edge
781   drawing outperformed animation in both effectiveness and user satisfaction. Both methods
782   present the taxonomies in a juxtaposed layout, however, relations are not explicit in animation.
783   Some participants considered that animation could be useful to focus on changes in small
784   taxonomic groups, which reaffirms the scope of animations, as indicated by Graham and
785   Kennedy [11].
786

787   During the sessions, we noticed that sometimes taxonomists wanted to see the big picture and
788   then focus on a smaller group of organisms of their interest. Also, sometimes they wanted to go
789   directly to the group they want to inspect. Thus, future research should consider easily toggling
790   between overview and detailed views as well as search and filtering functions. Text (that is, taxa
791   names) is crucial when comparing taxonomies. Visual cues such as color, size, shapes or glyphs
792   are not enough to recognize the differences and similarities. Text must be legible. The users
793   would have to read names, which would need to be accommodated efficiently avoiding
794   cluttering. Unlike other studies, we were not assessing how participants use the tool [18], neither
795   we were measuring the prototype efficiency for comparing hierarchies [12]. Instead, our
796   contribution lays in the assessment of the four visualization methods for the comparison of pairs
797   of biological hierarchies with respect to curation tasks.
798

799   A final remark about methods and tools for the comparison of biological taxonomies is that the
800   methods we evaluate in this work are those that have been mainly discussed in the scientific
801   literature; however, their implementation in tools is limited. The Taxonomic Tree Tool (Lin et
802   al.) is, as far as we know, one of the few tools available online that allows us to explore
803   relationships (such as ancestor-descendant relationships) when comparing two biological
804   taxonomies. In a post-study contact with some participants, we asked about the software they use
805   for taxonomy comparison. We obtained five responses and only one indicated to occasionally
806   use the Taxonomic Name Resolution Service (http://tnrs.iplantcollaborative.org) to compare a
807   list of plant names with an authoritative database of published names. We do not know the
808   precise reasons why the methods are not in widespread use, but some possibilities could be that
809   the developments only reached prototype versions and are not available online, that the
810   visualizations could be too complex, or the difficulty that comes with the lack of standardization
811   of data in different communities.
812

813

814   **Conclusions and Future Work**

815

816   This study contributes insights on the capacity of four visualization methods for hierarchy
817   comparison in typical biological taxonomy curation tasks. Twelve expert taxonomists took part
818   in a study and provided feedback. We performed quantitative as well as qualitative analysis. The
819   results clearly show differences among the methods, on both users' effectiveness and
820   satisfaction: the edge drawing method was preferred over other methods.
821
822   In this study, the data sets were selected to avoid bias, all participants used the same datasets, and
823   participants were able answer the questions in reasonable time. However, it will be interesting to
824   design a similar study with larger datasets. Another approach would be to design a study in
825   which the data would be specific and familiar to each participant.
826
827   Enhancements such as providing multiple views, adding visual cues at inner taxonomic rank
828   levels, and avoiding overloading caused by long names and hierarchical structure lines, are
829   insights for future research. Functions for searching, statistics and queries to retrieve the
830   information of a taxon will be considered in a future design of a visualization environment. We
831   also plan to further research on visual summary views to facilitate the comparison at different
832   taxonomic rank levels. Finally, it will be important to make biological taxonomy comparison
833   visualization tools available for use. This implies overcoming challenges such as promoting the
834   standardization of data to facilitate data sharing and comparison.
835
836

Acknowledgements

838   The authors would like to thank all expert taxonomists who collaborated with this research,
839   dedicated the time to participate and provided very valuable feedback. We would also like to
840   thank Eros Hernández-Romero and Ronald Andrés Bolaños-Rodríguez for their collaboration
841   during the software development and the interview sessions, and Shivam Agarwal for his
842   collaboration as test participant.

843
844

References

846
847   Ball-Damerow, J.E, Brenskelle, L, Barve, N, Soltis P.S., Sierwald P., Bieler R., LaFrance R.,
848   Ariño A.H., Guralnick R.P. Research applications of primary biodiversity databases in the
849   digital age. PLOS ONE 2019;14. Ed. by Silva, DdP:e0215794.
850   Beck F, Diehl S. 2010. Visual comparison of software architectures. In: *Proceedings of the 5th*
851   *international symposium on Software visualization - SOFTVIS '10*. New York, New York,
852   USA: ACM Press, 183. DOI: 10.1145/1879211.1879238.
853   Beck F, Wiszniewsky F-J, Burch M, Diehl S, Weiskopf D. 2014. Asymmetric visual hierarchy
854   comparison with nested icicle plots. In: *Joint Proceedings of the Fourth International*
855   *Workshop on Euler Diagrams and the First International Workshop on Graph Visualization*
856   *in Practice*. GraphVIP. 53–62.
857   Charmaz K. 2006. *Constructing grounded theory : a practical guide through qualitative*
858   *analysis*. London; Thousand Oaks, Calif.: Sage Publications.

859 Contian L, Huijie Q, Jiangning W, Liqiang J.Taxonomic Tree Tool -. *Available at*
860   *http://ttt.biodinfo.org/TF/* (accessed January 1, 2016).
861 Dang T, Franz N, Ludäscher B, Forbes A. 2015. ProvenanceMatrix. *International Workshop on*
862   *Visualizations and User Interfaces for Ontologies and Linked Data, VOILA 2015 - co-*
863   *located with 14th International Semantic Web Conference, ISWC 2015.*
864 Ghoniem M, Fekete J-D. 2001. Animating treemaps. In: *Proceedings of 18th HCIL Symposium-*
865   *Workshop on Treemap Implementations and Applications.*
866 Gleicher M. 2018. Considerations for Visualizing Comparison. *IEEE Transactions on*
867   *Visualization and Computer Graphics* 24:413–423. DOI: 10.1109/TVCG.2017.2744199.
868 Gleicher M, Albers D, Walker R, Jusufi I, Hansen CD, Roberts JC. 2011. Visual comparison for
869   information visualization. *Information Visualization* 10:289–309. DOI:
870   10.1177/1473871611416549.
871 Graham M, Craig P, Kennedy J. 2008. Visualisation to Aid Biodiversity Studies through
872   Accurate Taxonomic Reconciliation. In: *Sharing Data, Information and Knowledge*. Berlin,
873   Heidelberg: Springer Berlin Heidelberg, 280–291. DOI: 10.1007/978-3-540-70504-8_29.
874 Graham M, Kennedy J. 2007. Exploring Multiple Trees through DAG Representations. *IEEE*
875   *Transactions on Visualization and Computer Graphics* 13:1294–1301. DOI:
876   10.1109/TVCG.2007.70556.
877 Graham M, Kennedy J. 2010. A survey of multiple tree visualisation. *Information Visualization*
878   9:235–252.
879 Graham M, Kennedy JB, Hand C. 2000. A comparison of set-based and graph-based
880   visualisations of overlapping classification hierarchies. In: *Proceedings of the working*
881   *conference on Advanced visual interfaces - AVI '00*. New York, New York, USA: ACM
882   Press, 41–50. DOI: 10.1145/345513.345243.
883 Guerra-Gómez JA, Buck-Coleman A, Plaisant C, Shneiderman B. 2012. TreeVersity: Interactive
884   Visualizations for Comparing Two Trees with Structure and Node Value Changes. In: *Proc.*
885   *Conference of the Design Research Society-DRS2012*. 10.
886 Holten D, van Wijk JJ. 2008. Visual Comparison of Hierarchically Organized Data. *Computer*
887   *Graphics Forum* 27:759–766. DOI: 10.1111/j.1467-8659.2008.01205.x.
888 Lam H, Bertini E, Isenberg P, Plaisant C, Carpendale S. 2012. Empirical Studies in Information
889   Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer*
890   *Graphics* 18:1520–1536. DOI: 10.1109/TVCG.2011.279.
891 von Landesberger T. 2018. Insights by Visual Comparison: The State and Challenges. *IEEE*
892   *Computer Graphics and Applications* 38:140–148. DOI: 10.1109/MCG.2018.032421661.
893 Larsen BB, Miller E, Rhodes MK, Wiens JJ 2017. Inordinate Fondness Multiplied and
894   Redistributed: the Number of Species on Earth and the New Pie of Life. *Quarterly Review*
895   *of Biology* 92.
896 Linné Carl von, Gmelin Johann Friedrich. 1767. *Systema naturae per regna tria naturae :*
897   *secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis,*
898   *locis*. Lipsiae :impensis Georg. Emanuel. Beer,.
899 Lutz R, Rausch D, Beck F, Diehl S. 2014. Get Your Directories Right: From Hierarchy
900   Visualization to Hierarchy Manipulation. In: *Proceedings of the 2014 IEEE Symposium on*
901   *Visual Languages and Human-Centric Computing*. VL/HCC. IEEE, 25–32. DOI:
902   10.1109/VLHCC.2014.6883017.
903 MacKenzie IS. 2013. *Human-Computer Interaction: An Empirical Research Perspective*. San
904   Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

905   Munzner T. 2015. *Visualization Analysis and Design*. CRC Press.
906   Munzner T, Guimbretière F, Tasiran S, Zhang L, Zhou Y. 2003. TreeJuxtaposer: Scalable Tree
907   Comparison Using Focus+Context with Guaranteed Visibility. *ACM Trans. Graph.* 22:453–
908   462. DOI: 10.1145/882262.882291.
909   Sancho-Chavarria L, Beck F, Mata-Montero E, Weiskopf D. 2016. Visual Comparison of
910   Biological Taxonomies: A Task Characterization. In: *Poster aceptado para ser presentado*
911   *en EuroVis*.
912   Sancho-Chavarria L, Beck F, Weiskopf D, Mata-Montero E. 2018. Task-based assessment of
913   visualization tools for the comparison of biological taxonomies. *Research Ideas and*
914   *Outcomes* 4:e25742. DOI: 10.3897/rio.4.e25742.
915   Vaidya G, Lepage D, Guralnick R. 2018. The tempo and mode of the taxonomic correction
916   process: How taxonomists have corrected and recorrected North American bird species over
917   the last 127 years. *PLoS ONE* 13.
918

Category	Task
Pattern Identification	*1. Identify congruence*: Identify same taxonomic concepts.
	2. Identify corrections: Identify splits, merges, moves, renames.
	3. Identify additions/exclusions: Identify new or missing taxa.
	4. Overview changes: Obtain a global view of changes.
	5. Summarize: Obtain numerical understanding of change.
Query	*6. Find inconsistencies*: Recognize violation of rules (e.g. repeated names)
	7. Filter: Find cases that satisfy certain conditions.
	8. Retrieve details: Retrieve the attributes of a particular concept.
	9. Focus: Navigate and see the information in detail.
Edit	*10. Edit*: Make changes to the taxonomies.

Id	Area	Degree	Professional Experience (years)
E1	Botany, Bioinformatics	PhD	32
E2	Botany, Forestry	MSc	10
E3	Informatics, Bioinformatics	Engineer	21
E4	Botany, Forestry	PhD	28
E5	Botany	MSc	15
E6	Botany, Bioinformatics	PhD	31
E7	Botany, Forestry	Master	21
E8	Botany	MSc	21
E9	Botany	PhD	23
E10	Biology	MSc	30
E11	Ecology	PhD	12
E12	Ichthyology	PhD	32

	T_2	T_3	T_4	T_5
Nodes	116	78	105	114
Species	86	55	75	84
Splits	6	6	7	9
Merges	7	3	5	3
Moves	6	4	4	6
Renames	7	4	6	6
New	6	4	3	6
Excluded	4	3	4	8

Each row corresponds to a task-performing exercise. Each cell of the row contains the results of the participants' effectiveness with each method for the evaluated exercise, as well as the chi-square and p-value resulting from the Cochran's Q test. Resulting less-effective methods for each exercise are shaded in orange.

Exercise	Frequency (%)				χ^2	p-value
	Agg	**Ani**	**Edg**	**Mat**		
1. Which is the most common type of change?	9 (75)	11 (92)	1 (8)	11 (92)	25.500	0.000
2. Into what taxa was taxon "t" split?	11 (92)	10 (83)	11 (92)	11 (92)	3.000	0.392
3. Which species was split most?	4 (33)	7 (58)	12 (100)	11 (92)	17.571	0.001
4. Was species "s" merged with any other species?	11 (92)	11 (92)	12 (100)	12 (100)	3.000	0.392
5. With which other species was taxon "t" merged?	12 (100)	11 (92)	12 (100)	12 (100)	3.000	0.392
6. Which is the new name of taxon "t"?	12 (100)	12 (100)	12 (100)	12 (100)	–	–
7. Which was the previous name of "t"?	11 (92)	12 (100)	12 (100)	12 (100)	3.000	0.392
8. Which species were moved to genus "g"?	5 (42)	12 (100)	12 (100)	12 (100)	21.000	0.000
9. Which family has the most species added?	11 (92)	11 (92)	12 (100)	12 (100)	2.000	0.572
10. Genus to which more than one species were excluded?	11 (92)	12 (100)	12 (100)	12 (100)	3.000	0.392
11. What types of changes occurred to taxon "t"?	6 (50)	9 (75)	11 (92)	12 (100)	9.692	0.021
12. Identify which genus shows most changes.	2 (17)	7 (58)	7 (58)	7 (58)	10.714	0.013
Effectiveness (overall)	**105 (73)**	**125 (87)**	**126 (88)**	**136 (94)**	**40.480**	**0.000**

Each row corresponds to a task-performing exercise. Each cell of the row contains the results of the participants' satisfaction level with each method for the evaluated exercise, as well as the chi-square and p-value resulting from the Friedman test. Higher satisfaction level is shaded in green and less satisfaction level in orange.

Question How good do you think is each method in order to …	Median				χ^2	p-value
	Agg	**Ani**	**Edg**	**Mat**		
1e- … identify the most common type of change?	2.08	1.79	3.29	2.83	12.588	0.006
3e- … identify splits?	1.50	1.50	3.79	3.21	33.055	0.000
5e- … identify merges?	2.00	1.71	3.71	2.58	20.050	0.000
7e- … identify renaming of taxa?	2.13	1.50	3.67	2.71	21.559	0.000
8e- … identify moves?	1.54	1.88	3.83	2.75	24.295	0.000
9e- … identify new species added?	2.75	2.63	2.96	1.67	9.539	0.023
10e- … identify excluded taxa?	2.67	1.88	2.79	2.67	7.062	0.070
12-e- … identify changes to a taxon?	1.88	1.79	3.46	2.88	16.057	0.001
13- … visualize differences and similarities between two taxonomies? (Overall)	2.97	2.84	4.97	4.09	25.064	0.000

Figure 1

Taxonomies are agglomerated into one hierarchical structure. Changes are indicated by relationships between the two taxonomies. In order to distinguish to which taxonomy the taxon belongs to, a dual use of color was introduced. For each taxon's specific change, light nuanced nodes indicate that the taxa belongs to the taxonomy of origin and the darker nuanced nodes indicate that they belong to the taxonomy of destination. Through the colored toggle buttons on the menu, users can visualize either one or several types of changes at once. Users can select a specific taxon in order to visualize its changes.
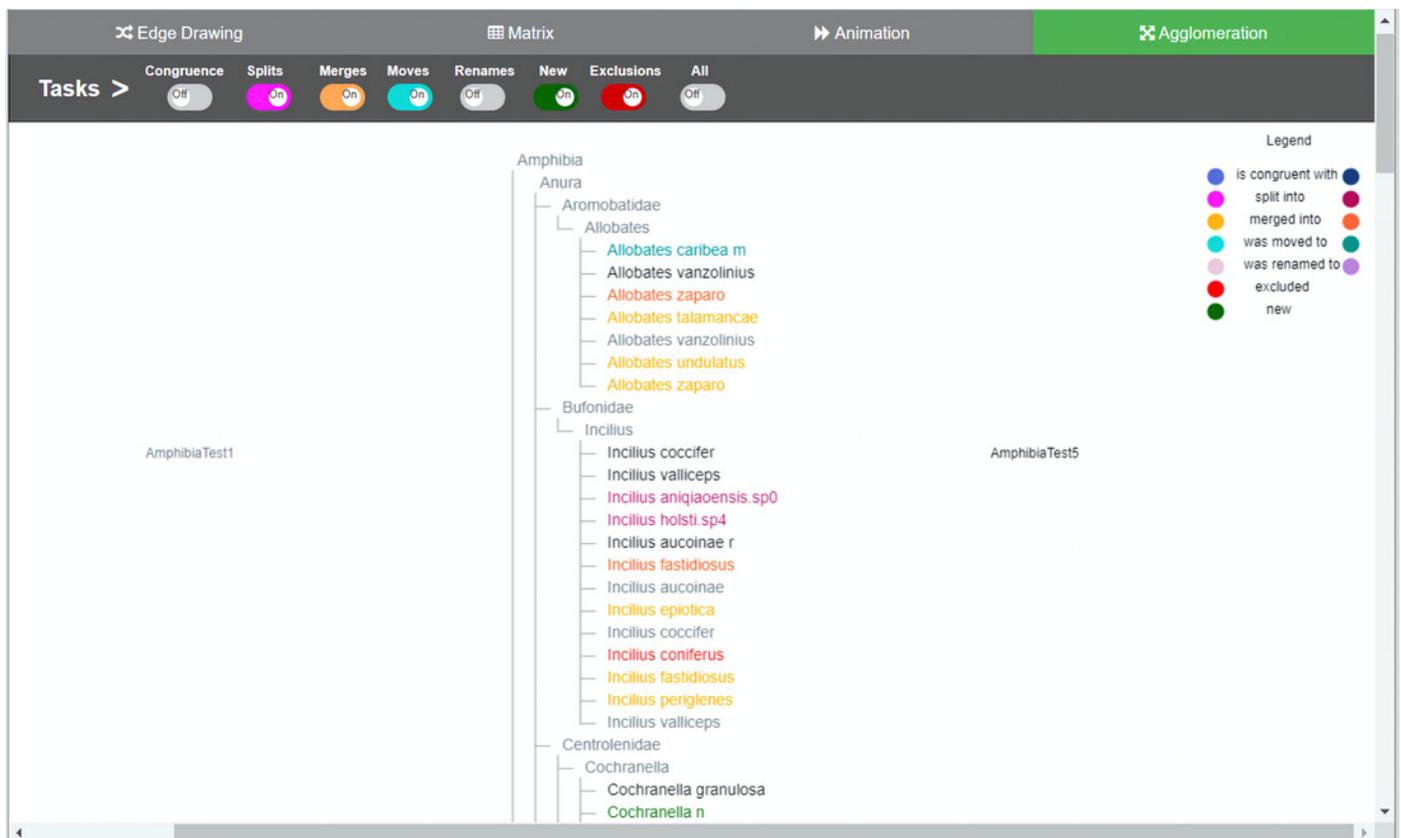
Figure 2

Changes are presented through animation. Besides the common toggle buttons on the main menu for all methods, the animation method includes a play/stop button, a speed slider, and a button that controls whether the animation will present the changes either one by oner or simultaneously. During animation, taxa moves from the origin taxonomy on the left side to the destination taxonomy on the right, displaying the transformation. Through the colored toggle buttons on the menu, users can animate either one or several types of changes at once. Users can select a specific taxon in order to visualize its changes.
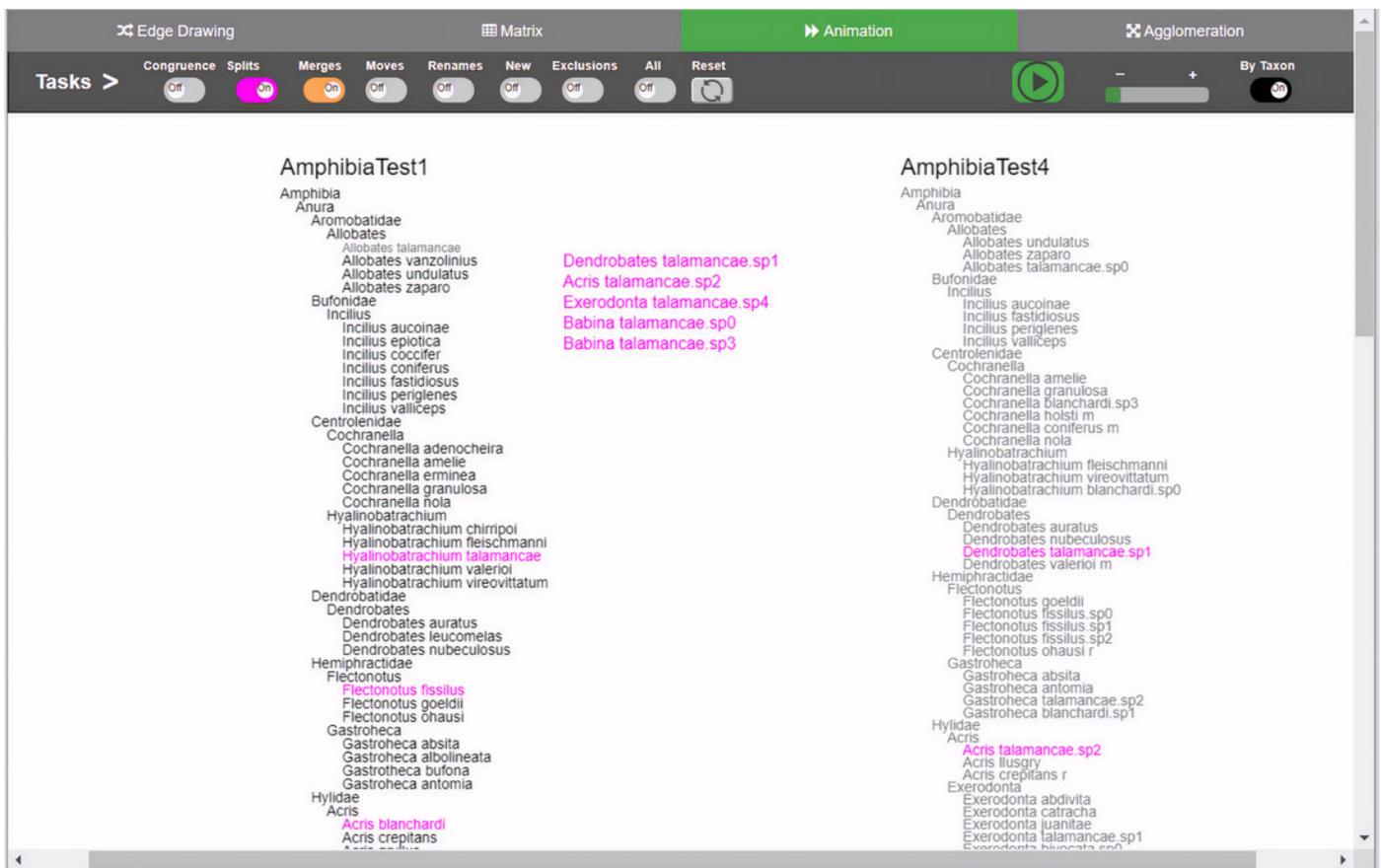
Figure 3

Changes are indicated by relationships between the two taxonomies. In the edge drawing implementation, relationships are visualized by color-coded edges that go from one taxonomy to the other. Through the colored toggle buttons on the menu, users can visualize either one or several types of changes at once. Users can select a specific taxon in order to visualize its changes.
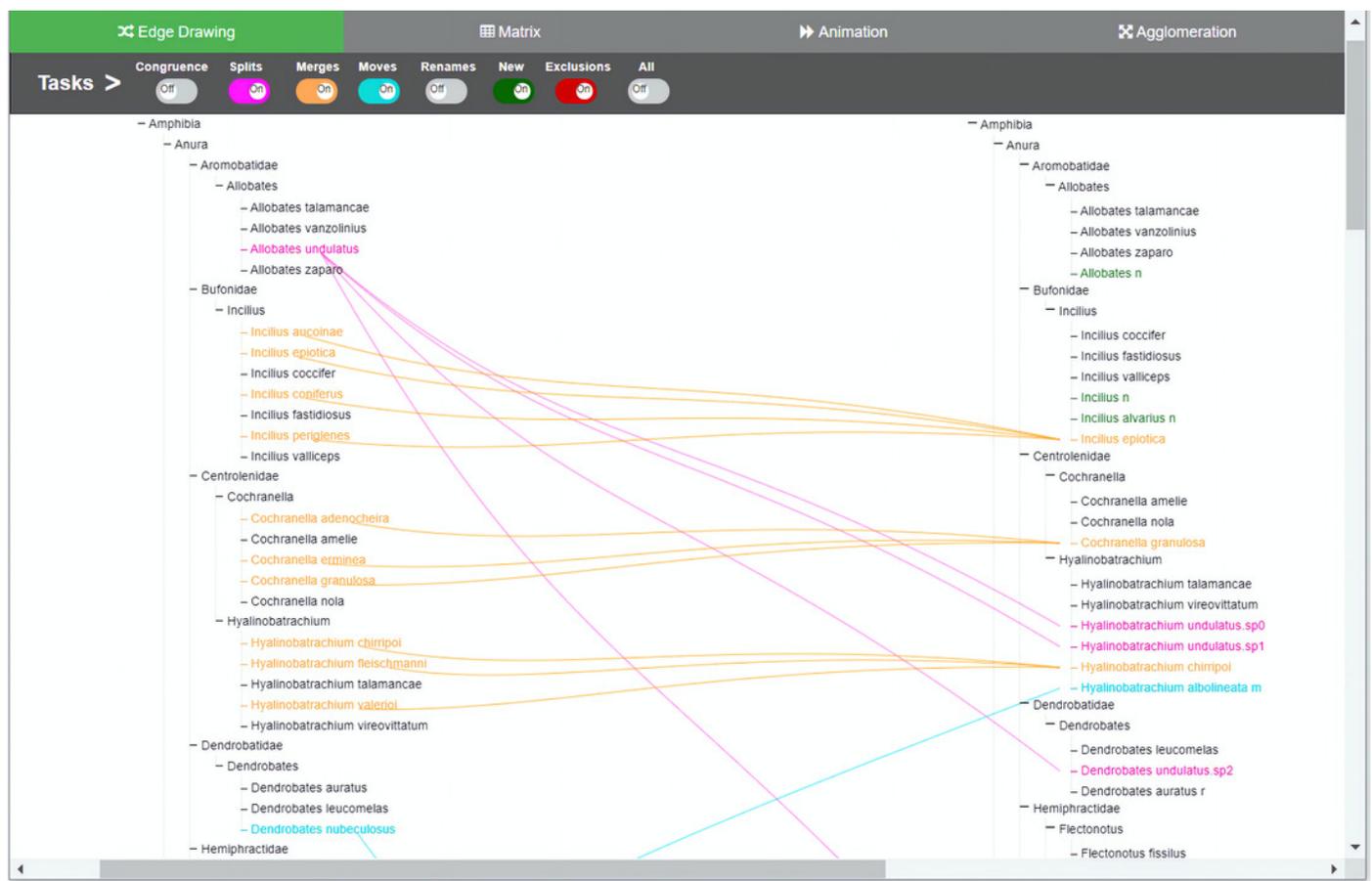
Figure 4

The two taxonomies are arranged along the horizontal and vertical axes of the matrix.
Changes between the two comparable taxonomies are highlighted thorugh the color-coded
cells. Through the colored toggle buttons on the menu, users can visualize either one or
several types of changes at once. Users can select a specific taxon in order to visualize its
changes.