Modelo de lentes interactivos para la visualización y comparación de taxonomías biológicas

Manuel Figueroa, *ITCR,* Nathalia Gonzalez, *ITCR,* Esteban Leandro, *ITCR*

MC-7205 Tema Selecto de Investigación

Instituto Tecnológico de Costa Rica

{*mfigueroacr, natgondu, elc790*}*@gmail.com*

***Resumen*—Se presenta un modelo de visualización alternativo para la comparación de taxonomías biológicas, que busca fortalecer el avance logrado en el sistema Diaforá [?]. Permitiendo a los taxónomos enfocarse en aspectos importantes de los árboles de clasificación y manteniendo al mismo tiempo un mapa de la totalidad de los árboles de taxonomía que están analizando. Dicha propuesta pretende ser evaluada por un panel de expertos en taxonomía, para verificar la eficacia de esta extensión al sistema Diaforá, de manera similar al análisis presentado en [?].**

***Index Terms*—LaTeX Visualización, Taxonomías biológicas, Diaforá, Enfoque y Contexto.**

I. Introducción

El problema descrito se deriva de la investigación realizada por los profesores Lilliana Sancho Chavarría y Erick Mata Montero del ITCR, como parte de su investigación en la comparación y visualización de taxonomías biológicas [**?**]. Las taxonomías biológicas son estructuras donde las especies son clasificadas de acuerdo a un sistema jerárquico propuesto por Linnaeus en el siglo 18 [**?**], y que incluye las categorías de dominio, reino, filo o división, clase, orden, familia, género y especie. La información de todos los seres vivientes conocidos se agrupa en árboles taxonómicos, que han sido creados y mantenidos por taxónomos a lo largo del mundo durante siglos. La reciente revolución digital ha permitido que gran parte de esa información pueda ser compartida y revisada por expertos. Debido a la naturaleza dinámica de estos datos es común que los taxónomos se enfrenten a distintas versiones de los datos que pueden ser corregidas y unificadas mediante la comparación de árboles taxonómicos. Las herramientas que ayuden a este grupo a analizar e identificar estas diferencias y facilitar el proceso de curación de las taxonomías permitiría un avance significativo en la calidad y fiabilidad de las clasificaciones biológicas de los seres vivos.

II. Trabajos relacionados

El problema de comparar grandes colecciones de datos es una necesidad común en el campo de la analítica visual [**?**], donde se identifica que a mayor escala se tienen como límite la capacidad cognitiva y perceptiva del usuario. Para solventar el problema de la escala se sugiere considerar como estrategías para el usuario:

- **Escanear secuencialmente:** el usuario puede examinar los objetos de manera secuencial.
- **Seleccionar un grupo:** el usuario analiza un grupo más pequeño de datos.
- **Resumir los datos:** presentar al usuario una abstracción que describa los datos.

Los trabajos relacionados en comparación de jerarquías de datos biológicos se centran en el estudio de árboles filogenéticos y taxonomías biológicas.

III. Análisis del uso de lentes interactivos en taxonomías biológicas

Uno de los sentidos más importantes de los seres humanos es la visión. Ésta es empleada para obtener la información visual del entorno, y en este caso específico la visualización se ha convertido en un medio para ayudar a las personas de diversos campos a obtener información relevante sobre los datos. Sin embargo, dado a que el tamaño de los datos aumenta constantemente, los enfoques de visualización tienen que resolver el problema de representaciones visuales exponenciales que dificultan la visualización de contenido relevante en una sola imagen de visualización [**?**]. Algunos investigadores como C. Tominski han tratado de abordar el desafío de la visualización con enfoque a través de exploraciones con grandes volúmenes de datos. Una de las técnicas para resolver los problemas de visualización son los lentes interactivos, una clase de métodos que permiten la exploración de datos con múltiples facetas. Se busca con el uso de lentes interactivos una vista alternativa de los datos presentes en una área específica de la pantalla, con el fin de enfatizar parte de esta información de una manera más clara para los usuarios [**?**]. Los datos estructurados en árboles son comunes en muchas disciplinas; este trabajo se enfocará específicamente en las clasificaciones biológicas para la detección de diferencias y detalles relevantes en una única pantalla, por ejemplo, los árboles filogenéticos que a diferencia de las categorizaciones taxonómicas estudian las relaciones de parentesco entre las especies. Se han estudiado diferentes técnicas de visualización que permiten enfatizar las similitudes y resaltar las diferencias existentes entre los árboles, como árboles de consenso y debido a que estos árboles cuentan en promedio con más de 50 nodos es necesario la utilización de estrategias para ordenar los árboles de manera automática entre

estas se destacan la diferencia mínima de tripletas (MDT), y la semejanza máxima de ramas (MBS). Estos algoritmos buscan maximizar el alineamiento de las hojas de los árboles en una comparación cara a cara [**?**].

III-A. Lentes Interactivos

Según la definición encontrada en [**?**], un lente interactivo es una herramienta ligera, que intenta resolver un problema localizado de visualización, alterando temporalmente una parte seleccionada de la representación de los datos.

También siguiendo el trabajo de Tominski, se definen como propiedades importantes de los lentes interactivos:

- *Forma:* La forma del lente virtualmente no tiene restricción, sin embargo, es común que muchos sistemas intenten emular el modelo de un lente del mundo real, en su mayoría circulares, no obstante esta forma puede adaptarse según la naturaleza de los datos que se están explorando. La importancia radica en que el usuario pueda identificar el lente fácilmente y sobre cuales datos quiere que el lente realice su función.
- *Posición y tamaño:* Se consideran atributos parametrizables, y que el usuario pueda ubicar el lente y ajustar su tamaño sobre cualquier parte de los datos en el área de exploración.
- *Orientación:* Cuando se emplea el recurso de visualización en tres dimensiones, la orientación toma relevancia en la forma en la que se observan los datos, ya que dependiendo del ángulo de visión del punto de observación el modelo de datos presentado en pantalla puede variar.

III-B. Lentes Interactivos para Visualización

Las técnicas de lentes son herramientas que nos permiten enfocarnos temporalmente en un punto de interes, un lente es una selección de una visualización base donde se buscan localizar un punto especifico y una vez que se llega al punto de interes la visualización vuelve a su estado original. La selección captura lo que debe ser resaltado por un lente. Normalmente el usuario controla la selección a través de movimientos sobre la representación visual de los datos.

III-C. Lentes Interactivos para Visualización de Taxonomías

La taxonomía tiene su origen en un vocablo griego que significa "ordenación". Se trata de la ciencia de la clasificación que se aplica en la biología para la ordenación sistemática y jerarquizada de los grupos de animales o plantas. Es importante establecer además que la taxonomía está muy en relación con lo que se conoce por el nombre de sistemática. Esta puede definirse como la ciencia que se encarga de llevar a cabo el estudio de las relaciones de parentesco, también llamadas afinidades, que se producen entre las distintas especies. En este paper nos enfocaremos en la taxonomía biológica, la cual forma parte de la biología sistemática, dedicada al análisis de las relaciones de parentesco entre los organismos. Una vez que se resuelve el árbol filogenético del organismo en cuestión y se conocen sus ramas evolutivas, la taxonomía se encarga de estudiar las relaciones de parentesco. La visión más

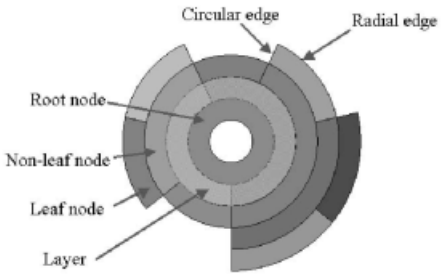

Figura 1. Ejemplo de visualización RSF. Tomado de [**?**]

extendida entiende a los taxones como clados (ramas del árbol filogenético, con especies emparentadas por un antepasado común) que ya fueron asignados a una categoría taxonómica.

El proceso de la taxonomía continúa con la asignación de nombres (de acuerdo a los principios de la nomenclatura), la elaboración de las claves dicotómicas de identificación y la creación de los sistemas de clasificación.

Los taxones permiten clasificar a los seres vivos a partir de una jerarquía de inclusión (cada grupo abarca a otros menores mientras está subordinado a uno mayor). Las categorías fundamentales, desde la más abarcativa hasta la menor, son el dominio, el reino, el filo o división, la clase, el orden, la familia, el género y la especie.

IV. Diseño de visualización InterRing para el sistema taxonómico Diáfora

Las técnicas conocidad como *RSF* o *Radial, space-filling* por sus siglas en inglés, tienen ciertas ventajas para la visualización de jerarquías, utilizan el espacio en pantalla de manera eficiente mientras que proveen una vista intuitiva de la estructuras jerárquicas.

Haciendo uso de una visualización de tipo *InterRing* [**?**], que emplea el concepto de distorción circular extendemos la capacidad actual del sistema *Diaforá* para mantener el contexto de la estructura jerárquica que esta siendo desplegada en el árbol taxonómico.

V. Desarrollo del modelo de visualización en el sistema Diaforá

Según lo investigado en [**?**], el método *edge drawing* puede comunicar de manera clara las diferencias entre dos versiones de una taxonomía. El uso de colores y líneas permite de manera clara detectar los cambios, además de que la interacción del usuario con la visualización permite enfocarse en aquellos cambios que puedan llamar su atención.

La principal desventaja detectada sobre el sistema de visualización *edge drawing* consiste en la pérdida de contexto de la estructura jerárquica del árbol taxonómico debido a la gran cantidad de nodos presentes una taxonomía. El propósito de utilizar una visualización *InterRing* [**?**] secundaria pretende apoyar al usuario a conocer el ámbito local del árbol taxonómico presentando de manera gráfica parte de la jeraquía

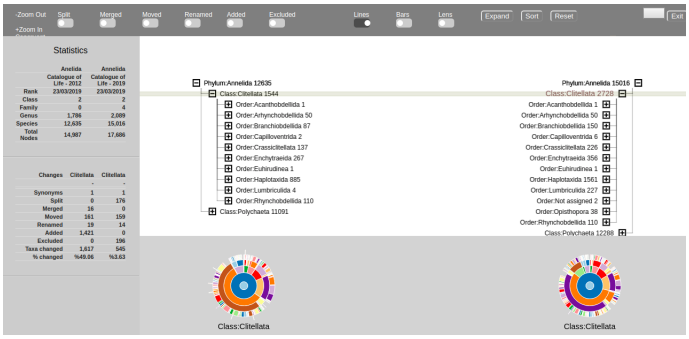

Figura 2. Sistema Diaforá en conjunto con visualización InterRing.

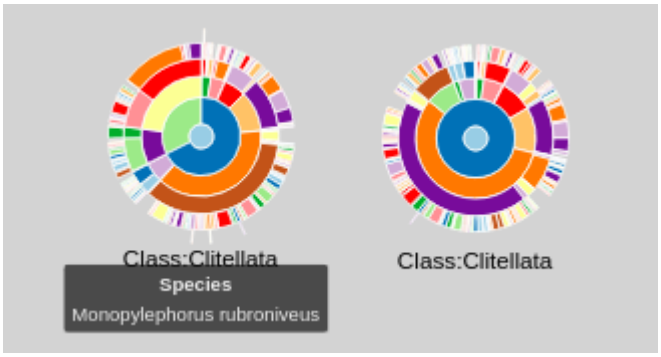

Figura 3. Comparación de dos clases usando la visualización InterRing

circundante que puede no ser visible debido a la cantidad de datos y que para poder visualizarlos requiere que el usuario haga *scroll* sobre la gráfica *edge drawing*.

El uso del método InterRing permite de manera compacta y simple renderizar una gran cantidad de nodos y permite al usuario navegar sobre las ramas del árbol taxonómico y explorar su composición, también al estar sincronizadas ambas gráficas se pueden realizar comparaciónes visuales sobre las imágenes resultantes de las visualizaciones InterRing que destacan las mayores diferencias entre distintas versiones de una taxonomía biológica.

Como se puede observar en la figura 3, es posible detectar diferencias entre las figuras que representan una misma clase (*Clitellata*) mediante las variaciones existentes en ambas figuras, de esta manera se espera poder contribuir con el trabajo de refinamiento de las taxonomías al resaltar y hacer más evidentes las diferencias entre versiones de una taxonomía.

Como detalles del modelo propuesto en la extensión del sistema Diaforá podemos listar:

- **Gráficas InterRing:** Se agregan un par de gráficas circulares para representar los árboles taxonómicos de las dos versiones de la taxonomía que se está comparando.
- **Soporte interactivo:** Las gráficas además de representar visualmente los árboles taxonómicos permiten la navegación interactiva por parte del usuario. Permitiendo escoger algún nivel específico en el árbol, lo que de manera automática se ve reflejado en la gráfica *edge drawing*.
- **Etiquetas interactivas:** Utilizando el control *Lens* se incluyen etiquetas interactivas que permiten saber cuantas diferencias y de que tipo existen en cada uno de los niveles del árbol taxonómico.

V-A. Caso de Uso: Orden: Haplotaxida

Utilizando el sistema *Diaforá* para analizar el órden taxonómico *Haplotaxida* correspondiente al filo *Annelida* [**?**]. En la figura 4 podemos apreciar la comparación entre la versión de la taxonomía del año 2012 y la versión de la taxonomía de 2019 del Catalogue of Life [**?**].

Como es posible apreciar en el resumen de las etiquetas interactivas el órden *Haplotaxida* tiene al menos 175 *splits* o divisiones de los taxones del grupo. Y eso se refleja adecuadamente en las diferencias de la gráfica *InterRing* en la parte inferior del área de comparación.

Es importante destacar, la sincronía existente entre las gráficas *InterRing* y el árbol taxonómico con *edge drawing*, por lo que cuando el usuario selecciona un nodo en alguna de las dos visualizaciones se refleja en la otra para tener en todo momento el contexto del sub-árbol que esta siendo sujeto de comparación.

V-B. Posibles extensiones a futuro del sistema Diaforá

Se recomienda como posibles temas de extensión al sistema, la posibilidad de incoporar la edición de las taxonomías biológicas en el sistema, así como la incorporación de un módulo de análisis de diferencias que incluya el resumen de los cambios y un conjunto alternativo de visualizaciones incluyendo una visualización matricial de los datos.

V-C. Detalles del desarrollo de la herramienta

El desarrollo al que hace referencia este documento es una extensión del sistema *Diaforá* [**?**]. El sistema original esta desarrollado como una aplicación web, haciendo uso de la librería Processing [**?**]. Los componentes adicionales que corresponden a la visualización de la gráfica *InterRing* y las etiquetas interactivas están desarrolladas haciendo uso de la librería Data Driven Documents [**?**].

Al ser una aplicación web, se permite un fácil acceso y disponibilidad para el uso de los taxónomos y se admite el mismo formato de árbol taxonómico que en la versión previa del sistema *Diaforá*.

La aplicación se encuentra publicada en el url: https://diafora2.herokuapp.com/ donde se puede acceder y probar las extensiones mencionadas al sistema [á].

Los archivos que contienen las taxonomías siguen el formato original de la primera versión del sistema [**?**].

VI. Validación del modelo propuesto

En esta sección se muestran una serie de evaluaciones que realizaron distintas personas para validar que el modelo de visualización propuesto cumple con los requisitos para que los taxónomos tengan una visualización más adecuada de las taxonomías a la ya existente.

- *"La clasificación en relaciones filogenéticas es la principal función de un taxónomo, para realizar el trabajo*

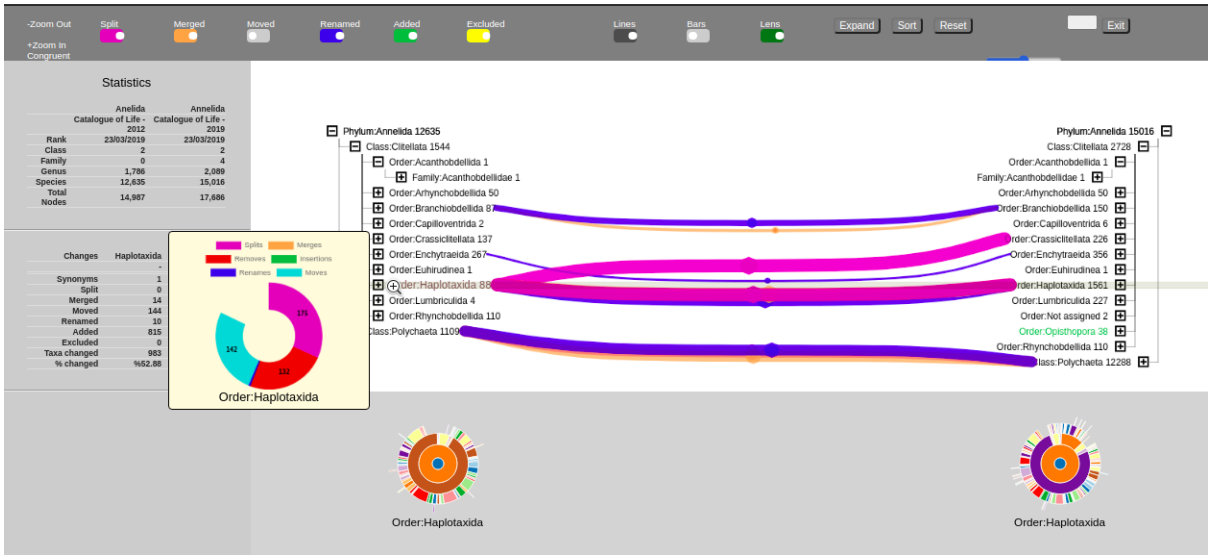

Figura 4. Caso de uso: Orden Haplotaxida

de una manera más rápida y eficiente, es necesario herramientas visuales para un mejor rendimiento de los datos y obtener formas de simplificar los datos de una forma visual. Cuando se llegan a tener una gran cantidad de datos para un análisis filogenético, es posible que se dificulten determinar las relaciones sin una visualización gráfica, por lo tanto, la utilización de estas visualizaciones, en específico, la herramienta "InterRing", tiene mucha utilidad en estos casos.
Esta herramienta permite al taxónomo tener una claridad mayor en los datos seleccionados, ya que no requiere que haga "scroll", y pueda explorar en otras ramas del árbol taxonómico".
Biólogo BSc. André Leandro C.

- Evaluación 2.

VII. Trabajo Futuro

Para un departamento de IT el concepto de aprovisionar de forma manual o automatizada una infraestructura para un sistema de cómputo [**?**], termina siendo el conjunto de tareas que tratan en preparar un conjunto de servidores, así como software adicional, configuración de redes, seguridad a la misma y demás, con el fin de ejecutar dentro de ellos aplicaciones de software que realizan distintas tareas, cumpliendo con los múltiples requerimientos dentro de una lógica de negocio.

Tendencias e implementaciones en el campo del aprovisionamiento de infraestructura para sistemas de software ha tomado un giro importante, donde cada vez se le brinda al usuario un mayor aprovisionamiento y menor responsabilidad de la configuración del hardware, el uso de máquinas virtuales [**?**] es cosa diaria respecto al lugar donde se ejecutan las aplicaciones, dando paso a procesos de automatización de infraestructura, que quitan aún más la responsabilidad al usuario de preocuparse por temas de hardware.

Modelos recientes sobre infraestructura como código (IaC) [**?**] son muy utilizados en aplicaciones en la nube [**?**], donde por medio mayormente de archivos con extensión .yaml, se describen un conjunto de árboles de configuración, las cuales después por medio de herramientas de automatización, se ejecutan todas las tareas deseadas. Desarrollar infraestructura como código [**?**], [**?**], permite la idempotencia en una arquitectura e infraestructura, lo cuál es una enorme ventaja respecto a configuraciones manuales.

Herramientas para Iac como Terraform o Ansible [**?**] realizan tareas automatizadas para aprovisionar infraestructura y desplegar aplicaciones en dicha infraestructura. Como resultado del aprovisionamiento, Terraform en el fondo utiliza la Teoría de Grafos [**?**], [**?**] para definir y mantener dicha especificación en código de la infraestructura en tiempo real.

Existe una configuración en archivos .yaml que define un estado $deseado$ de la infraestructura, Terraform monitorea en tiempo real cual es el estado $actual$ de y realiza por medio de reducción transitiva una diferencia de grafos, con el fin de comparar y saber si existe una mínima diferencia entre dichos estados $(deseado, actual)$ y en caso de existir, realizar a cabo una serie de procesos automatizados que ponen de nuevo el estado actual a como se desea que esté la infraestructura en todo momento.

Un ejemplo de esta diferencia de grafos en Terraform junto con el proceso de recuperación, podría ser tener una definición inicial deseada con 5 servidores corriendo en todo momento, si en algún momento un servidor se cae, Terraform va a realizar la comparación entre el estado $deseado$ de la infraestructura (5 servidores) y el estado $actual$ (4 servidores al estar 1 servidor caído) y va a levantar una instancia para volver a tener el estado actual igual al estado deseado.

Una propuesta para continuar con esta investigación sobre visualizaciones para taxonomías biológicas y una posible implementación de dicho modelo, más allá de un prototipo, va muy de la mano con todas estas tendencias de IaC, aplicando un enfoque similar al de la teoría de grafos para comparación

de árboles taxonómicos.

Se propone diseñar otra vista para el taxónomo, donde no va a visualizar todo el despliegue de cada árbol en un año distinto, sino que se podría almacenar en un árbol el contenido del $phylum$ del año inicial de la comparación y en otro árbol distinto el contenido del $phylum$ para el siguiente año que se esté comparando. Posterior a eso se puede llevar a cabo una diferencia de árboles y guardar dicha diferencia en un tercer árbol, el cual va a servir para visualizar únicamente los cambios que han surgido en el $phylum$ en los años que esté comparando.

Dicho árbol posiblemente contenga considerablemente menos información que cualquiera de los árboles previos, al menos que se estén comparando las mismas especies con una diferencia muy marcada de años. De esta manera, se puede contar con la pantalla de visualización actual junto con los InterRing y a la vez con una segunda pantalla donde se muestre la diferencia o los cambios que han habido entre los años, sin mostrar lo que no haya cambiado.

Dicha pantalla podría constar de datos que representen los cambios que hubieron, puede representarse como un único árbol y a la vez se pueden agregar gráficos de InterRing para visualizar la información de una manera distinta.

VIII. Conclusiones

Después de analizar distintos modelos para visualizar grandes cantidad de información, es recomendable tener más de una forma de visualizar las taxonomías, contar únicamente con dos árboles donde cada uno representa un $phylum$ de la misma especie en distinto año y cada uno de estos árboles se despliega hasta el fondo del mismo, no es la manera más sencilla de comparar una especie de un año a otro, sin embargo es útil para que el taxónomo no pierda el contexto de lo que está viendo.

Un enfoque que a manera de validación del modelo propuesto nos ha dado valor, es el hecho de utilizar más de un tipo de visualización, donde se tiene la visualización actual con el desglose total del $phylum$ en distintos períodos, una visualización de InterRing que permite tener de entrada una comparación más abstracta donde más rápidamente podemos ver si han habido cambios, a esto uniéndole la diferencia de grafos de árboles, el taxónomo no depende de una única pantalla para realizar comparaciones, sino que puede optar con distintas pestañas que pueden reducir las comparaciones.

Además, de distintos modelos estudiados durante la investigación de modelos de visualización, cuando se trata de grandes cantidades de datos los que queremos ver en una pantalla, el modelo InterRing es de los más utilizados, dando de entrada una visualización donde se le permite al taxónomo ver si existen diferencias sin entrar en detalle.