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

Lilliana Sancho-Chavarría, *ITCR,* Manuel Figueroa, *ITCR,* Nathalia Gonzalez, *ITCR,* Esteban Leandro, *ITCR*

MC-7205 Tema Selecto de Investigación

Instituto Tecnológico de Costa Rica

lsancho@tec.ac.cr, {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 [1], 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. A la vez, se realiza un análisis comparativo entre modelos de visualización con técnicas actuales de desarrollo de infraestructura informática para generar nuevos árboles o grafos con diferencias para una visualización de datos menos detallada que pueda perder al taxónomo de contexto, esto para una posible extensión del modelo presentado. 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 [2].**

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

I. Introducción

El problema descrito se deriva del trabajo realizado por Sancho-Chavarría et al. como parte de su investigación en la comparación y visualización de taxonomías biológicas [1]. Las taxonomías biológicas son estructuras donde las especies son clasificadas de acuerdo a un sistema jerárquico propuesto por Linnaeus en el siglo 18 [3], 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 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.

El manejo de información es cada vez más importante conforme las industrias, ciencias básicas y distintos mercados evolucionan. Los métodos de visualización brindan formas de comprender información voluminosa en poco tiempo. Buscamos brindar una herramienta y modelo útiles para que los taxónomos puedan comparar dos versiones de una taxonomía.

Cuando se busca una comparación de datos masiva como lo son las distintas taxonomías, se pueden encontrar múltiples formas de visualizar la información con distintas implementaciones gráficas, sin embargo no todas son útiles. Uno de los retos presentados fue investigar formas para visualizar las grandes taxonomías que permitieran mantener la coherencia de dichos datos y a la vez presentar al taxónomo una herramienta útil con la cual poder realizar comparaciones sin que le tome mucho tiempo.

A la vez, se realizaron distintas comparaciones con trabajos relacionados recientes para manejo de grandes cantidades de datos, dando propuestas durante un período inicial de investigación, visualizando los dos árboles taxonómicos de distintas formas, llegando a distintas conclusiones sobre la preferencia por utilidad del modelo propuesto en el presente trabajo. Este trabajo se enfocará específicamente en las clasificaciones biológicas para la detección de diferencias y detalles relevantes en una única pantalla.

Este artículo esta organizado de la siguiente manera. En la sección II se describen trabajos relacionados en el área de la visualización de datos voluminosos. La sección III presenta un análisis del uso de las técnicas de lentes interactivos en la visualización de taxonomías biológicas. Posteriormente, la sección IV describe el diseño de una extensión del sistema *Diaforá* con una visualización InterRing para la visualización de las taxonomías biológicas. La sección V describe los detalles del desarrollo de la extensión del sistema *Diaforá*. La sección VI describe la evaluación y validación recibida del modelo propuesto de InterRing por parte de un profesional en el área de biología. La sección VII muestra dos propuestas de trabajo futuro con distintas perspectivas de visualización. Finalmente las secciones VIII y IX presentan la discusión y las conclusiones del trabajo realizado.

II. Trabajos relacionados

El problema de comparar grandes colecciones de datos es una necesidad común en el campo de la analítica visual [4], 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 estrategias 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.

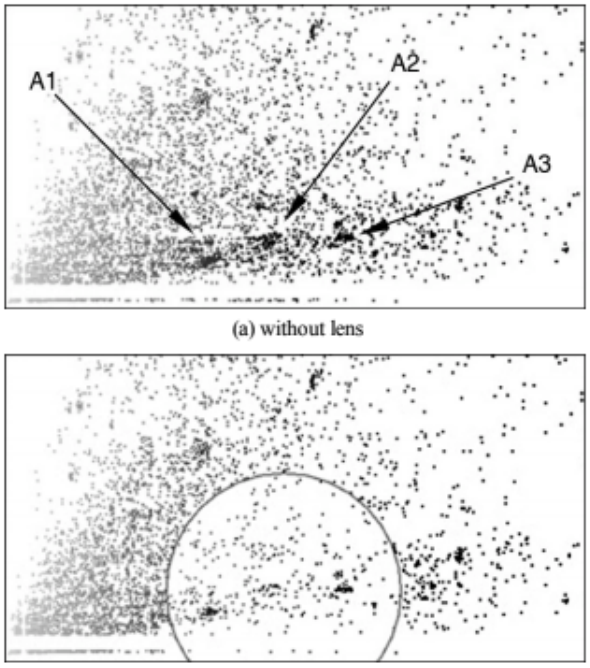
Los trabajos relacionados en comparación de jerarquías de datos biológicos se centran en el estudio de árboles filogenéticos y taxonomías biológicas.

Un enfoque mencionado en [5] es utilizar navegadores de árboles hiperbólicos, donde podría tenerse en un árbol el detalle de un taxón y en cada conjunto de ramas de dicho árbol cada una de las subespecies, sin embargo no es muy útil a la hora de querer comparar una taxonomía completa de un año específico con la misma taxonomía de otro año, ya que se terminaría teniendo dos árboles con ramas y sub ramas bastante profundas y voluminosas, lo cual no brindaría ningún beneficio respecto al modelo actual de visualización, a la vez, el hecho de ubicar dos árboles tan grandes en una sola pantalla no sería conveniente, ya que se puede perder de vista las comparaciones específicas que se quieren realizar. Utilizando este método con dos árboles donde uno se sobrepone al otro (un phylum del 2010 por ejemplo y el mismo phylum en el 2012) podría dar una visualización de las taxonomías más apropiada, ya que de esta forma de entrada se podría visualizar algún cambio en las ramas de los árboles al notarse que no son iguales, o por el contrario notar que no han habido cambios pronunciados en otras ramas.

Otro modelo de visualización para grandes cantidades de información, es tener un lente sobre una cuadrícula con puntos distribuidos que representen concentraciones de datos [6], esta técnica representa un modelo muy útil cuando se trate de profundizar en una única taxonomía de un año específico, sin embargo a la hora de realizar una comparación con otro año, el lente podría ser confuso a la hora de sacar diferencias entre subespecies.

En [7] se puede observar un modelo de visualización con grafos de exploración, dicho modelo utiliza colores y distintas aristas para representar la información, a la vez cuenta con una serie de puntos de interconexión entre los puntos mostrados, este modelo tiene cierta similitud al modelo propuesto, sin embargo dependiendo de la cantidad de subespecies para una especie, podría ser confuso o utilizar tantos colores que lleguen a quitar el enfoque en las comparaciones taxonómicas. Dichos puntos al estar cerca unos de otros pueden llegar a distorsionar una rama. También sobreponer un grafo sobre otro eliminaría la comprensión de lo que se puede buscar, que es comparar sin invertir mucho tiempo, una especie con la misma en distinto período.

Se puede observar además un modelo de visualización tridimensional [8], donde se tienen distintas capas, una al lado de otra en secuencia, dichas capas podrían representar una especie o subespecie del phylum. También se muestran entre cada capa una serie de lineas con distintos colores, que podrían llegar a representar de alguna manera la información de cada subespecie, como pueden ser las cantidades de categorías o especies en específico que existen. Este modelo tridimensional



(a) without lens

Figura 1. Ejemplo de visualización usando técnica de lentes. Tomado de [6]

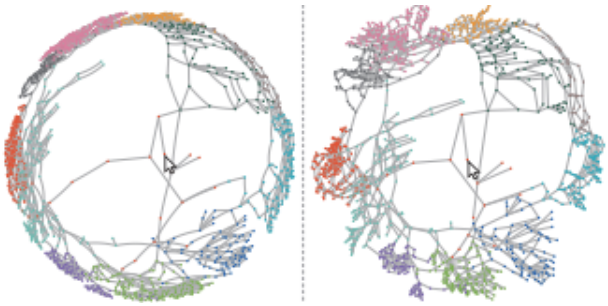


Figura 2. Ejemplo de visualización de gráficos hiperbólicos y la estructura propuesta. Tomado de [7]

es bastante útil si de visualizar información de un único año se tratara, ya que al modelo contener tantas lineas simétricas con distintos colores, sobreponer una sobre otra podría ser confuso, sin embargo se podría mantener dos esquemas uno encima del otro, mostrando por medio de las líneas la información comparativa entre los años del phylum. Una ventaja de este modelo es la capacidad de realizar giros tridimencionales, dando al taxónomo una perspectiva muy amplia desde más de un solo punto de vista, siendo útil talvez no específicamente para comparar un phylum entre distintos períodos, pero si para entrar en detalle del taxón en un único año y visualizar información del mismo.

En la Figura 1 se puede observar el modelo de lentes mencionado, donde puede realizarse una ampliación de los datos, pero el significado de cada uno de ellos a pequeña escala no se puede representar, no se sabría que representa cada punto al ser una visualización tan densa. En la Figura 2 se muestra un ejemplo de gráfico hiperbólico y su estructura propuesta.

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 de taxonomías se ha convertido en un medio para ayudar a las personas de diversos campos a obtener información relevante sobre los datos organizados jerárquicamente.

- **La taxonomía:** es la ciencia que estudia los principios, métodos y fines de la clasificación. Este término se utiliza especialmente en biología para referirse a una clasificación ordenada y jerarquizada de los seres vivos y en educación para ordenar y diseñar los objetivos del aprendizaje. [3]
- **La taxonomía en la biología:** clasifica de forma ordenada a los seres vivos [3]. La clasificación, niveles o categorías taxonómicas son importantes ya que ayudan a evitar la confusión entre las especies al regirse por un sistema universal y consensual. De esta manera, sirve para que la comunidad científica pueda definir sin errores al ser vivo que pretenden estudiar o nombrar.

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 [9]. Algunos investigadores como Tominski et al. 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 con información voluminosa 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 [9]. 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 [10] 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 [10].

III-A. Lentes Interactivos

Según la definición encontrada en [9], 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 interés 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 relacionada 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 artículo 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 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.

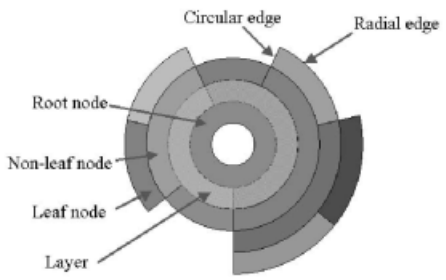


Figura 3. Ejemplo de visualización RSF. Tomado de [11]

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.

Es por esta razón que hemos optado por implementar un diseño de árbol radial [2] para representar los datos jerárquicos y debido a que la comparación de árboles es una tarea común realizada en árboles filogenéticos se podrá ver como todos los nodos de hoja o superiores se extienden hasta la parte inferior del gráfico, los nodos se colorean de acuerdo a su nivel facilitando que se pueda reorientar y reposicionar libremente.

- **Árbol radial:** En un diagrama, donde las ideas son expuestas de una manera ordenada y sistemática permitiendo mostrar las relaciones entre ellas. [12] El objetivo es inducir a construir estructuras mentales identificando ideas principales e ideas subordinadas según el orden lógico.

Haciendo uso de una visualización de tipo *InterRing* [11], que emplea el concepto de distorción circular extendemos la capacidad actual del sistema *Diaforá* para mantener el contexto de la estructura jerárquica que está siendo desplegada en el árbol taxonómico. De esta manera se proporciona la capacidad de mover elementos de la interfaz libremente, permitiendo a los usuarios trabajar con su propia organización en los elementos taxonómicos, ayudando a los colaboradores a crear y mantener modelos mentales de un conjunto de datos que contiene variaciones a través de los años. Al mover libremente el árbol taxonómico, los usuarios del sistema pueden diseñar sus propias categorizaciones, evitando que se dé una recuperación tardía de los elementos o datos específicos. Por ejemplo, relacionar una versión del 2010 con otro árbol del 2012 podría requerir una búsqueda exhaustiva para localizar el cambio respectivo en la otra representación.

Implementando InterRing logramos resaltar datos en representaciones taxonómicas, ayudando a las personas cuando cambian entre datos grupales como especie, género o familias.

IV-A. Interacción:

El sistema Diáfora proporcionará un acceso amigable al historial de cambio de datos. Las técnicas de interacción con el sistema Diaforá es a través de elementos comunes conocidas desde el escritorio, como doble clic, clic izquierdo o derecho. Por otro lado, las tareas de análisis de la información requieren de poca manipulación de widgets y diálogos de interfaz, facilitando la comprensión de los datos. La nueva visualización de trabajo en el sistema Diáfora es individual, lo cual facilita que múltiples usuarios puedan tener acceso a distintas versiones de un árbol y trabajar de manera local. Logrando evitar que, si un miembro del área desea eliminar o modificar familias, los datos del resto de usuarios no se vea afectados.

IV-B. Rendimiento:

Para casos prácticos, solo se usan dos lentes simultáneamente durante la exploración de datos. Sin embargo, para generar lentes para nodos superiores, se puede requerir una gran cantidad de lentes. Nuestra implementación se aplica a la visualización, y no al tamaño del conjunto de datos o la cantidad de capas de atributos que se representan.

Para nuestra propuesta, generamos un conjunto de lentes distribuidos de manera uniforme sobre la superficie del árbol radial en el cual el tiempo de respuesta no se ve afectado por la prueba de profundidad.

IV-C. Tamaño de la visualización:

Normalmente para una correcta visualización taxonomica será necesario utilizar el programa Diaforá en modo Maximizado, es decir, que la ventana ocupe todo el espacio de pantalla.

IV-D. Resolución:

La resolución es un problema tanto para la salida (display) y para la entrada. La resolución de la pantalla tiene gran influencia en la legibilidad de las visualizaciones de información por lo que se recomienda una pantalla grande.

V. Desarrollo del modelo de visualización en el sistema Diaforá

Según lo investigado en [1], 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.

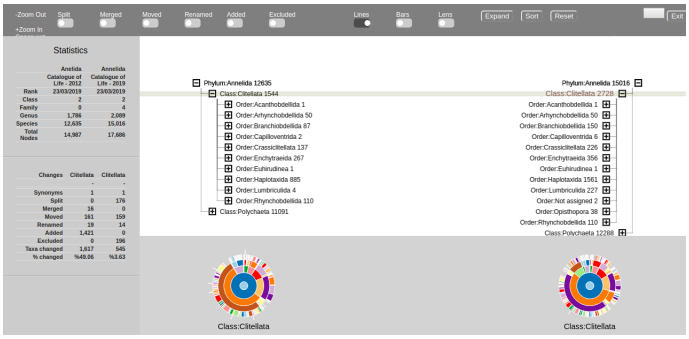


Figura 4. Sistema Diaforá en conjunto con visualización InterRing.

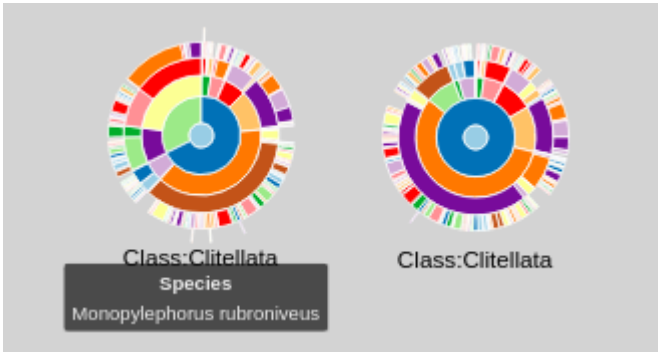


Figura 5. Comparación de dos clases usando la visualización InterRing

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* [11] secundaria pretende apoyar al usuario a conocer el ámbito local del árbol taxonómico presentando de manera gráfica parte de la jeraquía 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 5, 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 dos 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* [13]. En la figura 6 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 [14].

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.

Si el usuario sigue explorando el órden *Haplotaxida* y compara una de las familias existentes en este órden, como por ejemplo la familia *Sparganophilidae* se puede apreciar en la gráfica *InterRing* (figura 7)que existen bastantes diferencias para este grupo taxonómico en las versiones de la taxonomía *Annelida* del año 2012 y 2019 respectivamente.

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* [1]. El sistema original esta desarrollado como una aplicación web, haciendo uso de la librería Processing [15]. 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 [16].

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*.

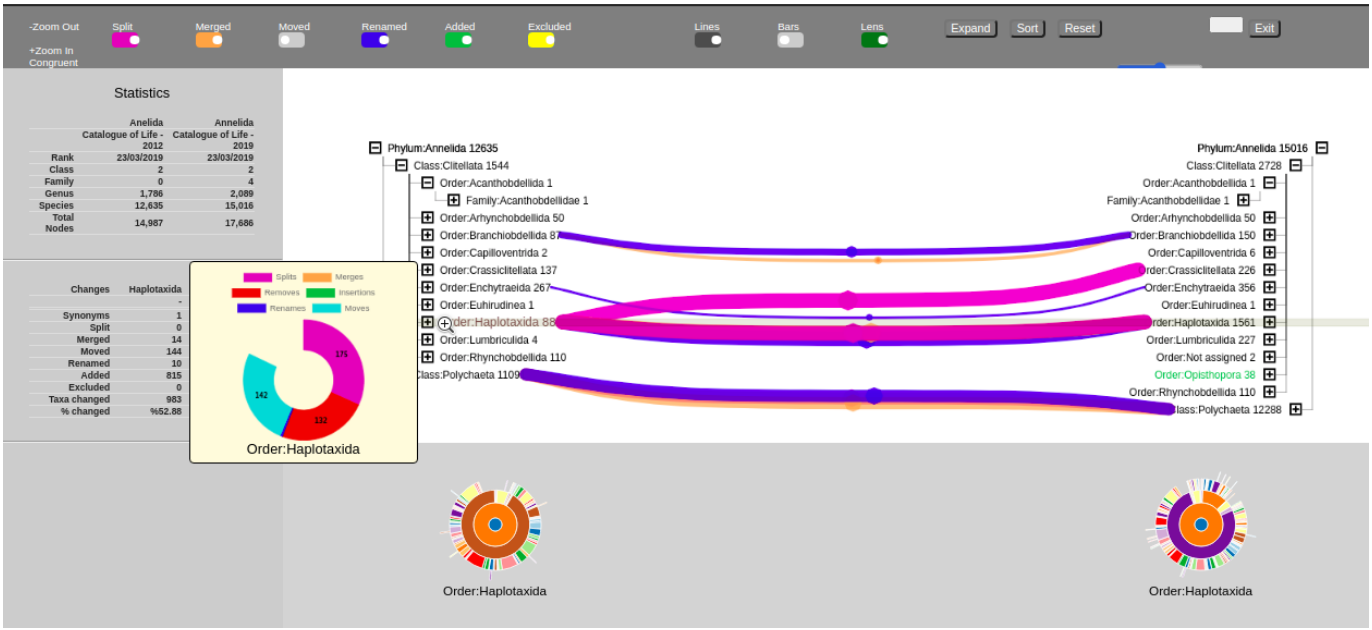


Figura 6. Caso de uso: Orden Haplotaxida

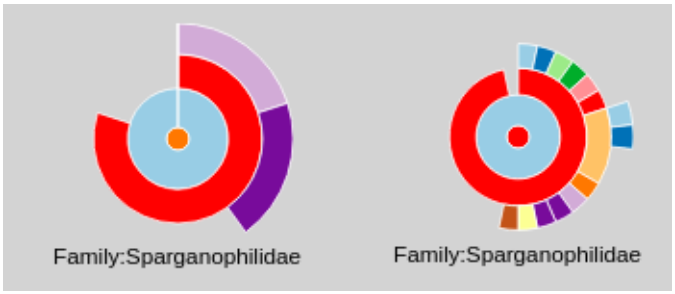


Figura 7. Comparativa familia *Sparganophilidae*. Izquierda versión 2012, derecha versión 2019.

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

Los archivos que contienen las taxonomías siguen el formato original de la primera versión del sistema [1].

VI. Validación del modelo propuesto

En esta sección se muestra una evaluación que realizó un biólogo para validar que el modelo de visualización propuesto de InterRing cumple con los requisitos para que los taxónomos tengan una visualización más adecuada de las taxonomías y comparación de las mismas respecto al modelo ya existente de árboles con todas las subespecies.

- *"La clasificación en relaciones filogenéticas es la principal función de un taxónomo, para realizar el trabajo de una manera más rápida y eficiente, es necesario contar con herramientas para un mejor rendimiento de los datos y obtener formas de simplificar los mismos de una forma visual y no tan transaccional. Cuando se llega a tener una gran cantidad de datos para un análisis filogenético, es posible que se dificulten determinar las relaciones sin una visualización gráfica, al contar cada phylum con una gran cantidad de subespecies, por lo tanto, la utilización de estas visualizaciones, en específico, la herramienta "InterRing", tiene mucha utilidad en estos casos. El modelo actual donde el taxónomo se ve obligado a realizar "scroll"de forma tan expandida, definitivamente no es método eficaz y funcional para un taxónomo, ya que toma mucho tiempo por analizar las especies, así como se la complejidad que nace a la hora de comprender las relaciones entre los elementos de interés.*
  Es por esto, que la herramienta "InterRing" es una opción mucho más viable, ya que no requiere realizar "scroll", más bien, la herramienta permite al taxónomo visualmente entender con mucha más precisión si existieron cambios con el paso del tiempo. Ésta opción le permite al usuario explorar en otras ramas del árbol taxonómico".

  Biólogo BSc. André Leandro C.

VII. Trabajo Futuro

El modelo actual junto con su implementación, permite a los taxónomos tener una herramienta para visualizar árboles taxonómicos y comparar unos con otros, así como ver ciertos datos estadísticos. Por limitantes en el tiempo para el desarrollo de la presente investigación, se llevaron a cabo una serie

de propuestas respecto a posibles extensiones de la presente implementación del sistema de visualización de taxonomías, mezclando con técnicas actuales para definir infraestructura computacional, una tarea llevada a cabo durante varias décadas por departamentos técnicos. Se muestran más a detalle dichas propuestas en esta sección.

Para un departamento de IT el concepto de aprovisionar de forma manual o automatizada una infraestructura para un sistema de cómputo [17], 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 [18] 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) [19] son muy utilizados en aplicaciones en la nube [20], 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 [21], [22], 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 [23] 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 [24], [25] 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.

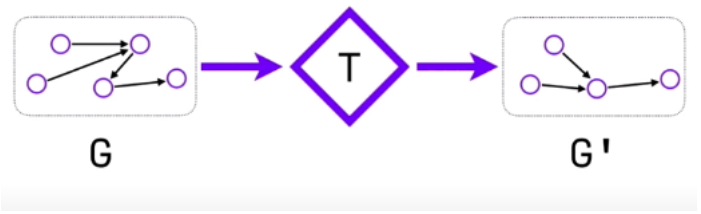


Figura 8. Ejemplo de transformación de los grafos propuestos, mostrando una entra y una salida, la cual podría representarse como un árbol inicial y la salida como el resultado de la diferencia propuesta, siendo un árbol o grafo más pequeño, más sencillo y rápido de analizar. Tomado de google images. Paul Hinze, Hashicorp.

VII-A. I Propuesta

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.

En la Figura 8 se observa un ejemplo conceptual de una transformación entre dos grafos, lo cual podría servir como base para realizar dicha diferencia de árboles y mostrar al usuario un árbol solo con sus diferencias, obviando los datos que se mantuvieron iguales.

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.

VII-B. II Propuesta

Comparación estadística de múltiples períodos de un mismo Phylum.

Otra propuesta para un posible trabajo a futuro o un añadido al presente modelo, es implementar otra pantalla en el software

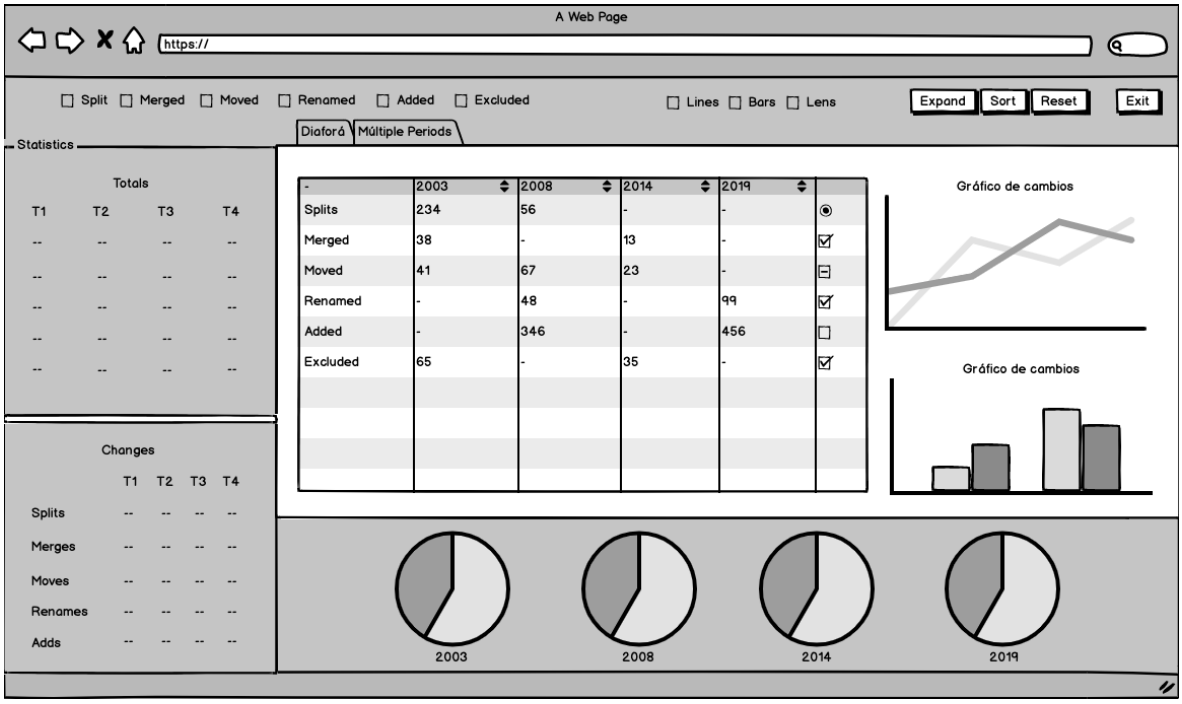


Figura 9. Modelo conceptual para una propuesta de comparaciones estadísticas utilizando múltiples períodos de un mismo phylum.

taxonómicos donde el taxónomo pueda realizar comparaciones a nivel del mismo phylum pero en más de dos años, por ejemplo introducir los dataset del phylum para el año 2010, 2012, 2015 y 2019 para posteriormente en una nueva visualización detectar cuáles han sido los cambios más pronunciados. Dicha visualización no sería representada a nivel de árbol como la presente, sino que sería más una representación estadística. Se podría introducir las taxonomías deseadas y dicho modelo podría iterar sobre cada árbol y presentar por ejemplo cuál es el año en donde se presentaron más eliminaciones de especies, cuál fue el o los años en donde se agregaron más especies a una subespecie del phylum, también cuáles fueron los años en donde menos cambios a nivel de agregación y eliminación hubieron.

Si bien es muy útil visualizar más información que la mencionada en esta segunda propuesta a trabajo a futuro, también es importante que el taxónomo no tenga una única forma de visualizar los datos, entre más perspectivas tenga, puede ser más útil y la toma de decisiones puede tomar menos tiempo. Esta idea surge como una propuesta donde el taxónomo quiera tener una especie de resumen que vaya más allá que sólo dos períodos, sino que pueda ir más allá.

En la Figura 9, se presenta un diseño o estilo conceptual respecto a la segunda propuesta para trabajo a futuro, donde se puede tener un espacio de visualización estadístico sin incluir todas las subespecies o subcategorías del phylum. Dicha visualización podría dar una perspectiva más amplia de los cambios que se han realizado no solo entre un período, sino entre distintos períodos, se podría observar cuales años han sufrido más cambios. El diseño de dicha pantalla podría incluirse a modo de pestaña en el UI general de la aplicación, permitiendo al taxónomo mantener la pestaña inicial con la comparación entre los dos árboles (modelo actual junto con los gráficos de InterRing) y añadir una segunda o tercer pestaña en donde se pueda hacer el importe de los distintos datasets para cada periodo del mismo phylum y visualizar a nivel estadístico cual o cuales son los años en donde se da ciertas condiciones. Los menús con botones y demás opciones, podrían des habilitarse cuando el usuario se encuentre en la presente pantalla, ya que son opciones más relacionadas a los árboles del Diaforá.

Cabe mencionar que dicho modelo debe contar con una validación rigurosa a la hora de importar los datos, ya que el taxónomo por error, en el importe de los datos puede cometer un error y agregar a la lista de archivos un phylum que no corresponde al que se quiere visualizar, dando espacio a que los resultados no sean del todo coherentes.

VIII. Discusión

El análisis taxonómico requiere de la visualización de grandes cantidades de datos correspondientes a las clasificaciones biológicas de los seres vivos, organizadas en el sistema propuesto por *Carl aeus* durante el siglo *XVIII* [3]. Este análisis de los seres vivos ha sido desarrollado por diferentes entidades y biólogos a lo largo del mundo, muchas veces recopilando información sobre las mismas especies y clasificándolas según su criterio, por lo que se ha encontrado con el problema de inconsistencias en las taxonomías biológicas que han sido creadas por diferentes científicos a lo largo de la historia [26]. Como parte de la investigación previa a este trabajo se determinó que una parte importante de la labor de los taxónomos actuales es la curación y refinamiento de las taxonomías existentes definiendo y corrigiendo los cambios haciendo divisiones, unificaciones, eliminando o agregando

taxones a las clasificaciones ya existentes. El sistema Diaforá [1] permite a los taxónomos contar con una herramienta de visualización de datos que les permite comparar distintas versiones taxonómicas que han sido creadas y modificadas en el tiempo.

Uno de las principales problemáticas de la solución presentada en el sistema Diaforá, consistía en que debido a la magnitud de los datos se pierde el contexto del área taxonómica que se está comparando, por lo que es necesario encontrar una manera de apoyar la visualización con una herramienta que permita explorar el entorno del árbol taxonómico.

El sistema propuesto *InterRing* [11], permite condensar una gran cantidad de información en un área visualmente pequeña por lo que permite proveer al usuario de la información que corresponde a la sub-área taxonómica en la que está trabajando y poder navegar más fácilmente.

Adicionalmente, se determina que la incorporación de las gráficas *InterRing* en el sistema permiten una comparación visual inmediata de diferencias marcadas en secciones del árbol. Por lo que es sencillo determinar si se han incorporado nuevos nodos a alguna de las taxonomías.

El sistema Diaforá como herramienta de apoyo visual en el refinamiento de taxonomías biológicas ayuda a mejorar el análisis de diferencias en grandes colecciones de datos, uno de los grandes retos de la labor de los taxónomos.

IX. 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.

A pesar de no contar con una implementación del software completa en totalidad y pruebas exhaustivas de la misma, creemos que el modelo propuesto junto con el trabajo a futuro mencionado, pueden brindar al taxónomo múltiples puntos de vista, útiles para tomar decisiones u observaciones.

Según se puede observar en modelos de grafos [7] y de lentes [6] cuando se trata de comparar información, el modelo de InterRing es más apto, ya que permite en menos tiempo poder darse cuenta si hay cambios drásticos, normalmente el taxónomo no va a comparar todo el árbol completo múltiples veces al día, sin embargo el modelo InterRing le permite comparar la subespecie en distinto período y si necesita más información abrir cada una de las subramas.

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.

IX-A. *Autoevaluación del trabajo realizado*

Según los objetivos presentados en la propuesta original de este trabajo:

Objetivo general:

- Creación de un modelo de visualización que permita la comparación de árboles taxonómicos manteniendo el contexto de la totalidad de la clasificación taxonómica.

Objetivos específicos:

1. Incorporar el uso de lentes interactivos para resaltar los detalles de las diferencias entre versiones de clasificaciones taxonómicas.
2. Diseñar una visualización adicional para el sistema Diaforá que permita observar la totalidad del árbol taxonómico en una única pantalla.
3. Implementar el modelo de visualización propuesto como una extensión del sistema Diaforá.
4. Validar la efectividad de la visualización propuesta con usuarios del sistema Diaforá.

Podemos determinar un nivel de completitud aceptable para cada uno de los mismos lo que detallamos en la siguiente tabla de autoevaluación:

Objetivo	Porcentaje	Justificación
1	10	Se cumplió con el objetivo al presentar un mecanismo de visualización para resaltar las diferencias entre versiones de taxonomías.
2	10	El sistema de visualización InterRing [11] permite condensar la totalidad del árbol taxonómico en una área reducida.
3	10	Se implementó una extensión al sistema utilizando el mecanismo de visualización propuesto. El mismo se puede acceder mediante el url https://diafora2.herokuapp.com/.
4	9	Se logró validar el modelo propuesto con un biólogo real y con uno de los autores del sistema Diaforá [1], sin embargo creemos que para dar como completo este objetivo es requerido una evaluación más extensiva del modelo y de sus beneficios.

Como se observa la mayoría de los objetivos propuestos se cumplió a cabalidad por lo que podríamos calificar el trabajo al que hace mención este artículo con un 9 de 10.

REFERENCIAS

[1] L. Sancho-Chavarría, C. Gómez-Soza, F. Beck, and E. Mata-Montero, "Diaforá: A visualization tool for the comparison of biological taxonomies," *Communications in Computer and Information Science High Performance Computing*, p. 423–437, 2019.

[2] L. Sancho-Chavarria, F. Beck2, and E. Mata-Montero1, "An expert study on hierarchy comparison methods applied to biological taxonomies curation," Aug 2019. [Online]. Available: https://peerj.com/preprints/27903/

[3] C. v. Linne, *Systema naturae, per regna tria naturae : secundum classes, ordines, genera, species cum characteribus, differentiis, synonymis, locis.* Vindobonae [Vienna] :Typis Ioannis Thomae ...,, 1767, vol. v. 1, pt. 1, https://www.biodiversitylibrary.org/bibliography/559 — .Ad editionem duodecimam reformatam Holmiensem.– Bound in the first part of v. 3 is Mantissa plantarum generum editionis VI et specierum editionis II. — Contents : v. 1. Regnum animale (2 pts.) – v. 2. Regnum vegetabile – v. 3. Regnum lapideum. — Soulsby — 116 — Stafleu (2nd) — 4832. [Online]. Available: https://www.biodiversitylibrary.org/item/10325

[4] M. Gleicher, "Considerations for visualizing comparison," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, p. 413–423, 2018.

[5] P. Pirolli, S. K. Card, and M. M. V. D. Wege, "Visual information foraging in a focus + context visualization," *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01*, 2001.

[6] G. Ellis, E. Bertini, and A. Dix, "The sampling lens," *CHI '05 extended abstracts on Human factors in computing systems - CHI '05*, 2005.

[7] "Structure-aware fisheye views for efficient large graph exploration," *IEEE Trans Vis Comput Graph*.

[8] F. S. Tim Dwyer, "Optimal leaf ordering for two and a half dimensional phylogenetic tree visualisation," 2006.

[9] C. Tominski, S. Gladisch, U. Kister, R. Dachselt, and H. Schumann, "Interactive lenses for visualization: An extended survey," *Computer Graphics Forum*, vol. 36, no. 6, p. 173–200, 2016.

[10] W. Zainon and P. Calder, "Visualising phylogenetic trees," *Conferences in Research and Practice in Information Technology Series*, vol. 50, pp. 145–152, 01 2006.

[11] J. Yang, M. Ward, and E. Rundensteiner, "Interring: an interactive tool for visually navigating and manipulating hierarchical structures," *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*

[12] G. Book and N. Keshary, "Radial tree graph drawing algorithm for representing large hierarchies," *University of Connecticut*, 2001.

[13] T. T. Nguyen, A. D. Nguyen, B. T. Tran, and R. J. Blakemore, "A comprehensive checklist of earthworm species and subspecies from vietnam (annelida: Clitellata: Oligochaeta: Almidae, eudrilidae, glossoscolecidae, lumbricidae, megascolecidae, moniligastridae, ocnerodrilidae, octochaetidae)," *Zootaxa*, vol. 4140, no. 1, p. 1, 2016.

[14] O. T. N. D. B. N. K. P. B. T. D. R. D. W. N. E. v. Z. J. P. L. e. Roskov Y., Ower G., "Species 2000 and itis catalogue of life," *2019 Annual Checklist*, 2019.

[15] T. P. Foundation. (2008) p5js.org. [Online]. Available: https://p5js.org/

[16] M. Bostock. (2011) Data driven documents. [Online]. Available: https://d3js.org/

[17] "End-to-end automation in cloud infrastructure provisioning," *26th International Conference on Information Systems Development (ISD 2017)*.

[18] A. M. Ariel Powell, "Sistemas de virtualización," 2019.

[19] e. a. M. Artac, T. Borovssak, "Devops: introducing infrastructure-as-code," 2017.

[20] K. Morris, "Infrastructure as code: Managing servers in the cloud. oreilly & associates incorporated," 2016.

[21] M. Guerriero, M. Garriga, D. A. Tamburri, and F. Palomba, "Adoption, support, and challenges of infrastructure-as-code: Insights from industry," pp. 580–589, 2019.

[22] M. Hüttermann, "Infrastructure as code: Managing servers in the cloud. oreilly & associates incorporated," p. pp. 135–156, 2012.

[23] P. A. Networks, "Automating security deployments with terraform and aansible - white paper," 2017.

[24] R. J. W. Joan M. Aldous, *Systema naturae, per regna tria naturae : secundum classes, ordines, genera, species cum characteribus, differentiis, synonymis, locis.* Springer-Verlag London, 2004, https://www.biodiversitylibrary.org/bibliography/559 — .Ad editionem duodecimam reformatam Holmiensem.– Bound in the first part of v. 3 is Mantissa plantarum generum editionis VI et specierum editionis II. — Contents : v. 1. Regnum animale (2 pts.) – v. 2. Regnum vegetabile – v. 3. Regnum lapideum. — Soulsby — 116 — Stafleu (2nd) — 4832. [Online]. Available: https://www.biodiversitylibrary.org/item/10325

[25] "An algorithm for drawing general undirected graphs," *Information Processing Letters*.

[26] J. C. Avise and J.-X. Liu, "On the temporal inconsistencies of linnean taxonomic ranks," *Biological Journal of the Linnean Society*, vol. 102, no. 4, p. 707–714, 2011.