TEC | Tecnológico de Costa Rica

ESCUELA DE INGENIERÍA EN COMPUTACIÓN

Tecnológico de Costa Rica

Un enfoque semiautomático de extracción de conocimiento sobre biodiversidad a partir de descripciones textuales de especies botánicas

Reporte final

2017

Código y Título del proyecto

Código del Proyecto: 5402-1375-4301

Título del proyecto: Un enfoque semiautomático de extracción de conocimiento sobre biodiversidad a partir de descripciones textuales de especies botánicas

Autores y direcciones

José Enrique Araya Monge, Ph.D.
Escuela de Ingeniería en Computación
Tecnológico de Costa Rica
Coordinador

Erick Mata Montero, Ph.D.
Escuela de Ingeniería en Computación
Tecnológico de Costa Rica

Resumen

Este documento describe el estado final del proyecto. Primero se introduce la gran necesidad que se tiene de poder acceder a información textual sobre biodiversidad de una manera más estructurada y semánticamente más significativa. Luego se recapitulan los principales enfoques que han sido usados para enfrentar dicho problema. Se enfatizan los enfoques que se refieren a la estructuración de descripciones morfológicas y de distribuciones geográficas, por ser estas las áreas de interés principal del proyecto. A continuación se presenta en detalle la organización del proyecto y sus tres etapas principales: recolección y transformación de documentos fuentes, estructuración semántica de fragmentos de texto de interés, y finalmente, desarrollo de herramientas para aprovechar la información estructurada. Luego se presentan los resultados obtenidos por el proyecto: resultados y evaluaciones obtenidos en la estructuración semántica de descripciones morfológicas y distribuciones geográficas, así como el estado final de las herramientas desarrolladas para pre procesamiento de los documentos originales y para la consulta de fragmentos de texto estructurados semánticamente. Después de presentar los resultados se hace una comparación entre los diferentes objetivos planteados por el proyecto y los resultados obtenidos. Finalmente se hacen una serie de recomendaciones para que futuros proyectos aprovechen los estudios y herramientas producidos por este proyecto.

Palabras clave

Bioinformática para la computación
Extracción de entidades
Estructuración semántica

Tabla de contenido

Índice de Figuras

Índice de Tablas

1 Introducción

Para estudiar y conservar la riquísima biodiversidad del neo trópico se debe contar con información científicamente validada, accesible en formatos estándar, y abierta para el uso de la comunidad científica, tomadores de decisiones y el público en general. Si bien iniciativas mundiales como el GBIF (www.gbif.org) y nacionales como el INBio (www.inbio.ac.cr) en Costa Rica y la CONABIO (www.conabio.gob.mx) en México han articulado mecanismos para digitalizar e integrar información nueva que ha sido recopilada en bases de datos, hay grandes cantidades de información histórica que se ha generado a través de más de tres siglos y publicado en cientos de miles de artículos, libros y literatura gris. La extracción del conocimiento de esta literatura se hace típicamente de forma manual (investigación bibliográfica) y el conocimiento obtenido reside básicamente en la mente de los expertos y se convierte en más literatura accesible para unos pocos especialistas. El problema que este proyecto enfrentó es convertir este importantísimo acervo de conocimiento en estructuras digitales de conocimiento para apoyar la toma de decisiones que lleven a comprender mejor y a conservar la biodiversidad. Este proceso de conversión se hará de manera semiautomática por medio de técnicas de minería de datos como las que él investigador principal ya utilizó en el proyecto "Towards a New Generation of Naturalist Citizens: Generating and Delivering Multimedia Biodiversity Information and Knowledge to Empower Citizens" que desarrolló el INBio para la Fundación JRS de EE.UU.

La investigación propuesta consiste en extraer de repositorios de información biológica, tales como los manuales de flora, información acerca de la distribución geográfica de especies y las relaciones entre las especies; de modo que esto se pueda vincular con información sobre organismos, taxones y ecosistemas con el fin de complementar y enriquecer dicha información. Además, se aprovechará la experiencia de proyectos previos en la extracción de información morfológica para desarrollar algoritmos y herramientas para la identificación de especies, su distribución geográfica y relaciones ínter específicas a partir de información parcial disponible. El proyecto extraerá información de las descripciones de especies de árboles de Costa Rica contenidas en el Manual de la Flora de Costa Rica que ha estado desarrollando el Jardín Botánico de Missouri conjuntamente con el INBio.

Con el fin de hacer más efectivo el proceso de conversión, la investigación se dividió en dos aspectos: delimitación y clasificación de fragmentos de interés, y análisis y conversión profundos de los fragmentos para hacer explícita la semántica de la información allí contenida. El primer aspecto provee un contexto que es aprovechado por los algoritmos de estructuración semántica del texto.

El objetivo general del proyecto es facilitar la consulta de información sobre relaciones interespecíficas, distribución de especies e identificación de especies de plantas mediante el desarrollo de algoritmos y herramientas a partir de información extraída semi-

automáticamente por medio de algoritmos de análisis semántico y estadístico de literatura de flora de Costa Rica.

Los objetivos específicos del proyecto son los siguientes:

1. Establecer un modelo que indique los principales componentes de las descripciones biológicas

2. Desarrollar una herramienta de software para etiquetar manualmente grandes fragmentos de texto usando la lista de conceptos principales desarrollada anteriormente.

3. Aplicar la herramienta de etiquetado para crear colecciones de documentos con las principales secciones de interés explícitamente delimitadas.

4. Establecer listas de términos y abreviaturas relacionados con los principales componentes de las descripciones biológicas.

5. Extraer información semánticamente estructurada sobre distribución geográfica, relaciones interespecíficas y morfología de las especies a partir de descripciones biológicas. Desarrollar algoritmos y herramientas para estructurar dichas secciones. Evaluar el desempeño de esos algoritmos.

6. Desarrollar herramientas permitan consultar y operar con la información estructurada extraída.

El objetivo 5 es el punto principal de la investigación. Hubo que hacer un ajuste a dicho objetivo: debido a que las fuentes de información disponibles no incluían suficiente información sobre relaciones interespecíficas, se trabajó más bien con las claves dicotómicas de clasificación. Este punto sin embargo no se llegó a completar aunque aún está en desarrollo en una tesis del Programa de Maestría en Computación.

1.1 Marco Teórico

Debido a la urgente necesidad de aprovechar los repositorios de información biológica, ya ha habido proyectos tendientes a la extracción de dicha información. Del año 2007 al año 2010 INBio (Instituto Nacional de Biodiversidad) desarrolló un proyecto llamado *Towards a New Generation of Naturalist Citizens: Generating and Delivering Multimedia Biodiversity Information and Knowledge to Empower Citizens* para la JRS Foundation. Un componente de ese proyecto consistió en el desarrollo de un módulo de software para tomar las descripciones de los árboles maderables incluidos en el Manual de Flora de Costa Rica y enriquecerlas introduciendo etiquetas XML para hacer computacionalmente explícita la semántica de las diferentes partes de esas descripciones. Aunque los documentos originales eran muy ricos en estructura, la misma no era explícitamente mostrada. El módulo desarrollado incluye mecanismos de indexación y búsqueda que aprovecha la reestructuración del texto. Aunque no fue posible automatizar completamente la conversión, dicho proyecto ayudó a establecer pautas para hacer que la intervención manual fuera simple y eficaz.

Recientemente han sido desarrolladas varias herramientas muy relevantes para el manejo de descripciones biológicas. El proyecto GoldenGate es un editor XML que admite el marcado automático de texto con corrección manual. El proceso se basa en expresiones regulares y diccionarios pre compilados para etiquetar descripciones taxonómicas. El marcado automático detecta correctamente los nombres taxonómicos y tratamientos taxonómicos. Para lograrlo el sistema integra diferentes herramientas para el procesamiento del lenguaje natural como GATE [1].

Por otro lado, MARTT (MARkuper para tratamientos taxonómicos) es un sistema de marcado semántico automatizado basado en algoritmos de machine learning supervisados o no supervisado mejorados usando reglas de asociación. MARTT ha sido probado en diferentes publicaciones como Flora of China, Flora of North America, y Flora de North Central Texas. La aplicación ya ha etiquetado con éxito más de 15.000 descripciones. Los resultados experimentales muestran que el enfoque de aprendizaje automático con reglas de asociación tiene un recall y una precisión: 80% - 95% [2].

TaxonFinder es un proyecto uBio que proporciona un conjunto de servicios web para la identificación de especies y categorización automática de artículos basados en las especies mencionadas en ellos [3]. Estos servicios web son utilizados por Biodiversity Heritage Library (BHL) para extraer los nombres taxonómicos de la literatura disponible en dicho repositorio.

El proyecto TaxonGrab combina la exclusión basada listas con algunas reglas. El sistema busca todas las palabras que no están en el diccionario del lenguaje común (usando un diccionario del idioma inglés) y aplica acciones basadas en reglas para determinar si un término es el nombre de una especie o no. El rendimiento de TaxonGrab es 94% de recall y 96% de precisión aplicado a un solo volumen de 5.000 páginas sobre la taxonomía de las aves [4].

Por otro lado, LINNAEUS es un software de código abierto para reconocer nombres de especies. Dicho software utiliza un enfoque basado en diccionarios y un conjunto de heurísticas para resolver menciones ambiguas. El rendimiento de LINNAEUS es 98% de recall y 90% de precisión a nivel del documento [3].

GATE es una plataforma de código abierto para desarrollar y desplegar componentes de software que procesan el lenguaje humano. Ya tiene más de 15 años de desarrollo y está en uso para todo tipo de tareas computacionales que involucren el lenguaje humano (llamado con frecuencia procesamiento del lenguaje natural, análisis de texto, o la minería de texto). La estructura modular permite aprovechar módulos desarrollados previamente para tareas como división del texto en oraciones o reconocimiento de entidades (nombres), y extenderlos para ajustarlos a situaciones nuevas [5].

Como se mencionó anteriormente, el Global Biodiversity Outlook (GBIO) ha definido la extracción de conocimiento y codificación del mismo en forma de atributos de las especies o interacciones ínterespecíficas como un nuevo tema estratégico de la IB en los próximos diez años. Siguiendo esa recomendación, la Enciclopedia de la Vida implementó un primer esquema de estructuración (manual) de conocimiento sobre atributos de especies (trait bank) disponible en www.eol.org.

Este proyecto lleva estos esfuerzos un paso más allá al plantear un enfoque con componentes automáticos de procesamiento de textos lineales y componentes manuales de edición de éstos.

En la tesis de María Auxiliadora Mora, se señala que el análisis profundido de las descripciones biológicas es un problema del área de ***Extracción de Información*** (***Information Extraction***, **IE**). Dicha área desarrolla algoritmos para analizar el contenido de grandes volúmenes de texto no estructurado o semi-estructurado buscando documentar tipos predefinidos de eventos, entidades y relaciones. A partir del texto se identifica, recoge y normaliza información relevante para un usuario particular. La normalización consiste en usar una representación estructurada de la información, como por ejemplo una plantilla o un esquema XML. Para alcanzar este objetivo, se usan herramientas de lenguaje natural (NLP), inteligencia artificial y aprendizaje automático.

Thessen, Cui y Mozzherin [6] presentan una arquitectura de referencia para un sistema de extracción de información aplicado a descripciones morfológicas en inglés:

- se define un objetivo de extracción con su correspondiente plantilla
- se realiza un estudio del conocimiento disponible para complementar el proceso de extracción (ontologías, vocabularios controlados, etc.), de los estándares existentes y de los métodos y tecnología a utilizar

- las entradas pueden requerir el reconocimiento óptico de caracteres (OCR) con su corrección de errores

- además, las entradas pueden requerir la identificación de secciones de interés dentro de los documentos, para lo cual se usan métodos de extracción

- previo al proceso de extracción de información se usan técnicas de NLP para etiquetar componentes del texto y generar representaciones con más facilidad de procesamiento

- Finalmente, sobre esas representaciones se aplican los métodos de extracción, los cuales pueden ser supervisados, semi-supervisados, no supervisados o basados en reglas, con el fin de producir como salidas bases de datos, documentos XML o RDF.

"El NLP es un área de investigación que abarca un conjunto de técnicas para la generación, manipulación y análisis del lenguaje natural. Aunque la mayoría de las técnicas son heredadas de la Lingüística y la Inteligencia Artificial, también han sido influenciadas por áreas relativamente nuevas como el Aprendizaje Automático, la Estadística Computacional y la Ciencia Cognitiva". [7]

Con respecto a los métodos de extracción de información, Steven Abney en [8] presenta los cinco tipos más importantes de problemas en aprendizaje automático. Los cuatro primeros tienen que ver con la estimación de una función f(x) y están agrupados de acuerdo a si el algoritmo es supervisado o no supervisado y si la variable a predecir tiene un valor continuo o discreto. Consisten en algoritmos de clasificación, de clustering, de regresión y de estimación de densidad. El quinto tipo es el aprendizaje por refuerzo cuya entrada es continua y la supervisión es indirecta. Otros métodos de extracción de información son los basados en reglas [9].

Las métricas generalmente usadas en IE para evaluar los resultados son precisión y cobertura (precision and recall). Estas métricas miden el porcentaje de anotaciones correctas y lo completo del método de extracción, respectivamente. Además, se utiliza la medida F, que aplica la media armónica ponderada entre la precisión y la cobertura [10].

Las fórmulas que definen estás métrica se listan a continuación:

- Precisión = $\frac{Número\ de\ instancias\ correctamente\ identificadas}{Número\ total\ de\ instancias\ identificadas}$
- Cobertura = $\frac{Número\ de\ instancias\ correctamente\ identificadas}{Número\ total\ de\ instancias\ correctas}$
- F = $\frac{(b^2+1)\cdot(Precision*Cobertura)}{b^2\cdot(Precision+Cobertura)}$

  b representa la importancia relativa entre la precisión y la cobertura.
  Si b=1 ambas medidas tienen igual importancia.

Para la selección de la muestra en las evaluaciones se utilizó el algoritmo de la rueda de la ruleta [11] con el objetivo de darle más prioridad a las cláusulas con mayor cantidad de estructuras (indicador de complejidad).

Thessen y otros publicaron en [6] un panorama muy completo de las tecnologías de extracción de información de acuerdo con el énfasis de la extracción. Para el reconocimiento de nombres de entidades usan reglas Taxon Finder [12], Find All Taxonomic Names [13] (que además de reglas usa lógica difusa), TaxonTagger [13]. Por otro lado, entre los que usan modelos probabilísticos para el reconocimiento de nombres están NetiNeti [14]. La estructuración de textos completos consiste en anotar de manera semiautomática las secciones que los componen; en este caso se tiene GoldenGate [15] que es un editor XML con marcado automático, Curry y Connor [16] proponen un sistema para estructurar documentos usando heurísticas basadas en el estilo del texto organización y puntuación.

La extracción de características morfológicas es el área en que se ubica este proyecto. Trabajos en esta área incluyen Chinese Markuper for Taxonomic Treatments – Cmartt [17] para el chino; Charparser [18] que implementa un algoritmo de aprendizaje no supervisado utilizando Bootstrapping para anotar descripciones morfológicas a nivel de cláusulas; Phenex [19] utiliza Charparser y una ontología para anotar descripciones de fenotipos de organismos; o Markuper for Taxonomic Treatments – MARTT [20] basado en métodos de aprendizaje semi-automático inductivo y reforzado con reglas aprendidas durante el proceso; Multiflora [21] utiliza expresiones regulares, una ontología y la herramienta GATE; finalmente X-Tract [22] y Terminator [23] usan heurísticas y un diccionario de términos

Como señala Moisés Acuña en su tesis, el geo-parsing o análisis sintáctico de entidades geográficas consiste en la identificación de entidades geográficas de un texto. Se trata de un caso específico de NERC (Named Entity Recognition and Classification). Es una tarea que busca delimitar elementos de texto y clasificarlos en un rango de categorías predefinidas. Hay tres métodos principales para realizar geo-parsing [24]: búsqueda en gazetteers, aplicación de reglas, y aprendizaje de máquina.

En la búsqueda de gazetteers las palabras del texto son buscadas en una base de datos que contiene nombres de lugares junto con sus metadatos [25]. Para su implementación se usan tries, tablas de hash y hasta con tablas de bases de datos relacionales [26].

En la aplicación de reglas se usan expresiones regulares o gramáticas libres de contexto para decidir si una secuencia de caracteres es un topónimo o no. Las expresiones regulares permiten una búsqueda rápida con poca memoria, pero tienen una profundidad limitada [27] [28]. Por otro lado, las gramáticas libres de contexto se implementan de una manera menos eficientes, pero permiten trabajar con situaciones más complejas [29] [30].

En el aprendizaje de máquina un conjunto de caracteres o características es calculado haciendo un recorrido por el texto; se maneja una probabilidad de que una de estas características sea una instancia de un término. Este algoritmo se ejecuta sobre un corpus de entrenamiento que tiene

valores ya conocidos. Luego se utiliza inferencia estadística para decidir si un término es un topónimo o no. [31] [32] [33]

Cui et al [34] establecen un algoritmo, basado en aprendizaje de máquinas, para anotar texto no estructurado, con un alto nivel de precisión. Dicho algoritmo utiliza el mismo texto, sin entrenamiento. Esto reduce o elimina el trabajo manual, mejora la cobertura de la anotación, aprende conceptos nuevos y los reutiliza. Ventajas adicionales son su independencia del tamaño de la colección y una complejidad algorítmica lineal. Cui presenta CharaParser [18] es un parser que permite la anotación semántica de descripciones morfológicas de organismos biológicos.

Pouliquen et al [35] utilizan Gazetteers para identificar topónimos dentro de textos en los medios de comunicación. Como trabajan con al menos cinco idiomas su enfoque utiliza heurísticas estadísticas y no análisis lingüístico ni etiquetado PoS (Part-of-speech). Los autores indican que estas reglas mejoran los resultados.

Geo-parsing se define como el proceso en que se reconocen palabras como lugares, esto según Kimler [36]. En su tesis, utiliza varios de los heurísticos propuestos por Pouliquen et al [35] como gazetteers para los nombres en mayúscula. Se divide el proceso en dos pasadas, donde la primera le da un poco más de calidad a la segunda. La primera detecta los nombres importantes usando un umbral de importancia. En la segunda pasada se realiza una búsqueda más profunda que toma en cuenta los lugares con menor importancia que el umbral. Los nombres obtenidos en la primera pasada proveen un contexto geográfico que permite reducir la ambigüedad que se presenta en la segunda etapa.

Geo-Coding es el proceso que elimina la ambigüedad y asocia los topónimos con lugares reales. Kimler [36] utiliza los heurísticos que utiliza Pouliquen et al. [35] y agrega otros más implementados en Perl. Una de las heurísticas determina el geo-context como lugar de publicación, lugar de escritura o un parseo superficial del texto para determinar la región geográfica donde se lleva a cabo la mayoría de las acciones en un documento. Otra heurística consiste en el *shallow-deep parsing* que es filtrar a priori las referencias que son falsos positivos para un geo-context particular. Heurísticas adicionales son presentadas para manejar el caso de lugares que sean nombres de personas u otras palabras del lenguaje natural.

2 Metodología

2.1 Introducción general del proyecto

El proyecto busca extraer semi-automáticamente información biológica para producir una representación semántica altamente estructurada que facilite la consulta sobre aspectos como distribución de especies, identificación de plantas y claves de clasificación.

Para realizar lo anterior el proyecto se organizó en tres etapas:

1. Recolección y transformación de los documentos fuentes, así como la delimitación de los fragmentos de interés: descripción morfológica, distribución geográfica y claves dicotómicas.

2. Estructuración semántica de fragmentos de interés para marcar explícitamente el contenido de esos fragmentos y así generar una representación estructurada de la información contenida.

3. Desarrollo de herramientas para aprovechar la información estructurada mostrando su uso en consultas mucho más específicas que una simple búsqueda textual.

La Figura 2.1 muestra en detalle el flujo de información que orientó el desarrollo de este proyecto de investigación. Para la etapa de recolección y transformación se desarrollaron dos herramientas: un filtro para limpiar el texto de información irrelevante como encabezados de página, y un etiquetador que permite seleccionar fragmentos de los documentos, seleccionar una categoría que describe su contenido y registrar esa información en una base de datos. Dicho etiquetador tiene la posibilidad de extraer fragmentos de una misma categoría para así poder formar las tres colecciones que serán usadas en la etapa siguiente: fragmentos con descripciones morfológicas, fragmentos con distribuciones geográficas y fragmentos con claves dicotómicas.

En la siguiente etapa se hace un análisis profundo de los fragmentos con el fin de hacer explícita la semántica de la información. Esta etapa produjo dos tesis de Maestría en Computación. La tesis de María Auxiliadora Mora hizo el análisis de las descripciones morfológicas y la tesis de Moisés Acuña realizó el análisis de las distribuciones geográficas. El análisis de las claves dicotómicas y su uso para generar claves matriciales tipo Lucid [referencia] es tema de una tercera tesis que no está terminada al concluir este proyecto. Con el fin de proveer insumo de claves para la última etapa del proyecto, se hizo un análisis más superficial de las claves. Los tres análisis anteriores produjeron colecciones estructuradas de información morfológica, distribución geográfica y claves de clasificación.

La etapa final consistió en desarrollar tres herramientas que pudieran consultar las colecciones estructuradas producidas por la etapa anterior y aprovechando la estructura provista por el análisis realizado. Las herramientas desarrolladas están al nivel de prueba de concepto y

forman la base para un futuro estudio y desarrollo más elaborado sobre el uso de la información estructurada.

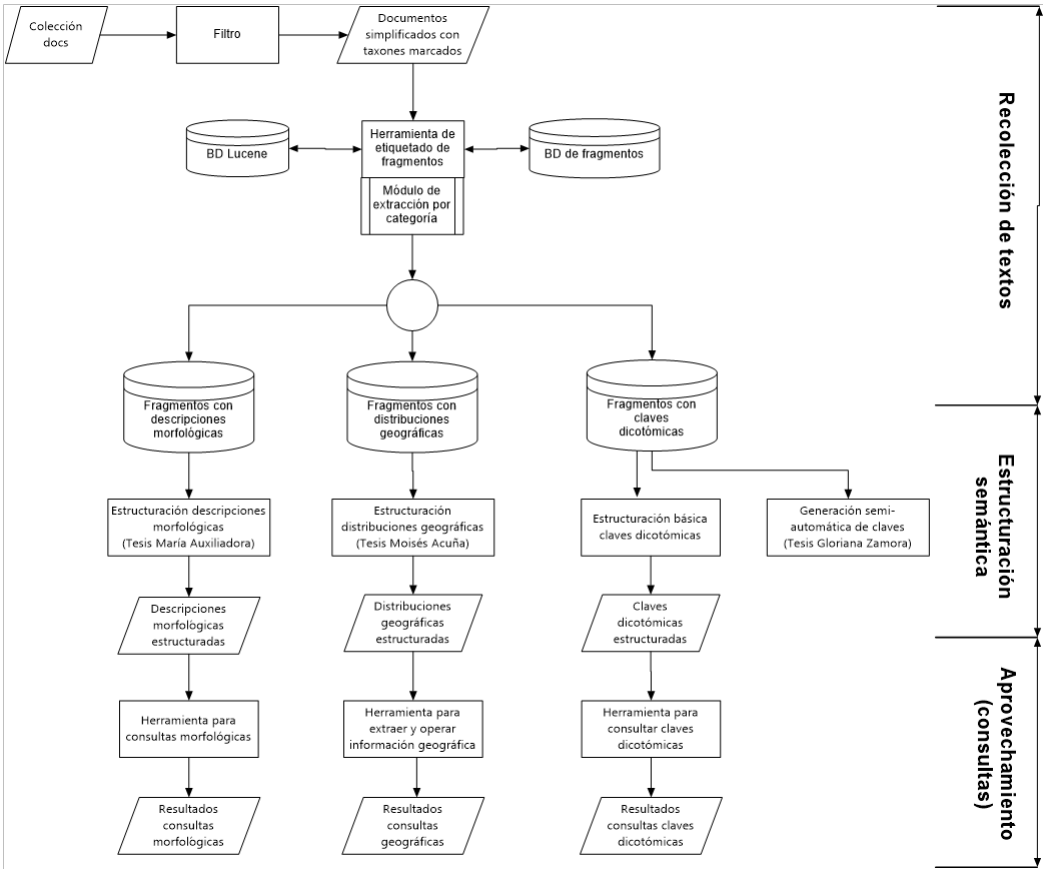


Figura 2.1 Plan general del proyecto

A continuación, se describen con más detalles las diferentes herramientas y procesos realizados por el proyecto.

2.2 Etapa 1: Recolección y transformación de los documentos fuentes, así como la delimitación de los fragmentos de interés

Estructurar automáticamente texto abierto que puede referirse a diferentes tópicos resulta una tarea extremadamente dura porque se debe primero determinar a qué tópico se refiere cada fragmento del texto. Como el proyecto busca estructurar texto en tres áreas:

descripciones morfológicas, distribuciones geográficas y claves dicotómicas, un primer paso consistió en desarrollar herramientas para delimitar semi-automáticamente los fragmentos de texto de esas áreas y así recolectar colecciones especializadas de fragmentos de texto. Las herramientas que se describen a continuación fueron desarrolladas para estandarizar la entrada, delimitar el fragmento de texto en que se presenta cada taxón y etiquetar fragmentos de texto con alguna categoría de interés.

2.2.1 Filtro de pre-procesamiento

Cada fuente de información puede tener un formato distinto, por lo que se requiere de una herramienta de filtrado que estandarice el texto. El Filtro de Flora de Costa Rica fue desarrollado para procesar archivos de texto extraídos del Manual de Plantas de Costa Rica. El programa se encarga de leer los archivos, filtrar la información no relevante como los pies de página, números de página, encabezados, pies de imágenes. También analiza la estructura del documento buscando jerárquicamente los taxones que en él se describen. La lista de taxones obtenida es presentada al usuario para su revisión y corrección.

La Figura 2.2 muestra el texto de entrada con su información irrelevante como el número de página. La Figura 2.3 muestra el resultado de eliminar la información redundante y establecer marcas con el taxón al cual se refiere el texto.

```
...

Matudaea

3 spp., S Méx.–CR, Col.

Matudaea trinervia Lundell, Lloydia 3: 210. 1940. Guayabillo, Murta.

Árbol, 10–25 m, hermafrodita. Hojas con el pecíolo 0.4–1.7 cm; lámina 7–17  2.5–8
cm, elíptica a lanceolada,
ocasionalmente asimétrica en la base, entera, muy esparcidamente estrellado-
pubescente en ambas
caras (pero glabrescente con la edad), triplinervada. Infls. ca. 1.3 cm,
glomeruladas. Fls. bisexuales, sésiles;


3



MPCRv6.001_025 7/19/07 9:36 AM Page 4

4 Manual de Plantas de Costa Rica

sépalos valvados, 2–4 mm; estambres 20–24; anteras dehiscentes
longitudinalmente, el conectivo generalmente prolongado
por hasta ca. 1.2 mm; ovario súpero. Frs. 0.8–1.5 cm, oblongo-
ovoides, densamente estrellado-lepidotos; semillas negras, lustrosas,
ca. 8 mm.

...
```

Figura 2.2 Texto original

```
...

GENERO::Matudaea
3 especies, S Méx.-CR, Colombia.

ESPECIE::Matudaea trinervia

Matudaea trinervia Lundell, Lloydia 3: 210. 1940. Guayabillo, Murta.

Árbol, 10-25 metros, hermafrodita. Hojas con el pecíolo 0.4-1.7 cm; lámina 7-17 x 2.5-8 cm,
elíptica a lanceolada, ocasionalmente asimétrica en la base, entera, muy esparcidamente
estrellado-pubescente en ambas caras (pero glabrescente con la edad), triplinervada.
Inflorescencias aproximadamente 1.3 cm, glomeruladas. Flores bisexuales, sésiles; sépalos
valvados, 2-4 mm; estambres 20-24; anteras dehiscentes longitudinalmente, el conectivo
generalmente prolongado por hasta aproximadamente 1.2 mm; ovario súpero. Frutos 0.8-1.5 cm,
oblongo-ovoides, densamente estrellado-lepidotos; semillas negras, lustrosas,
aproximadamente 8 milímetros.

...
```

Figura 2.3 Texto filtrado

La Figura 2.4 muestra la herramienta. En el Apéndice 8.3 Herramientas desarrolladas se describe con más detalle esta herramienta.

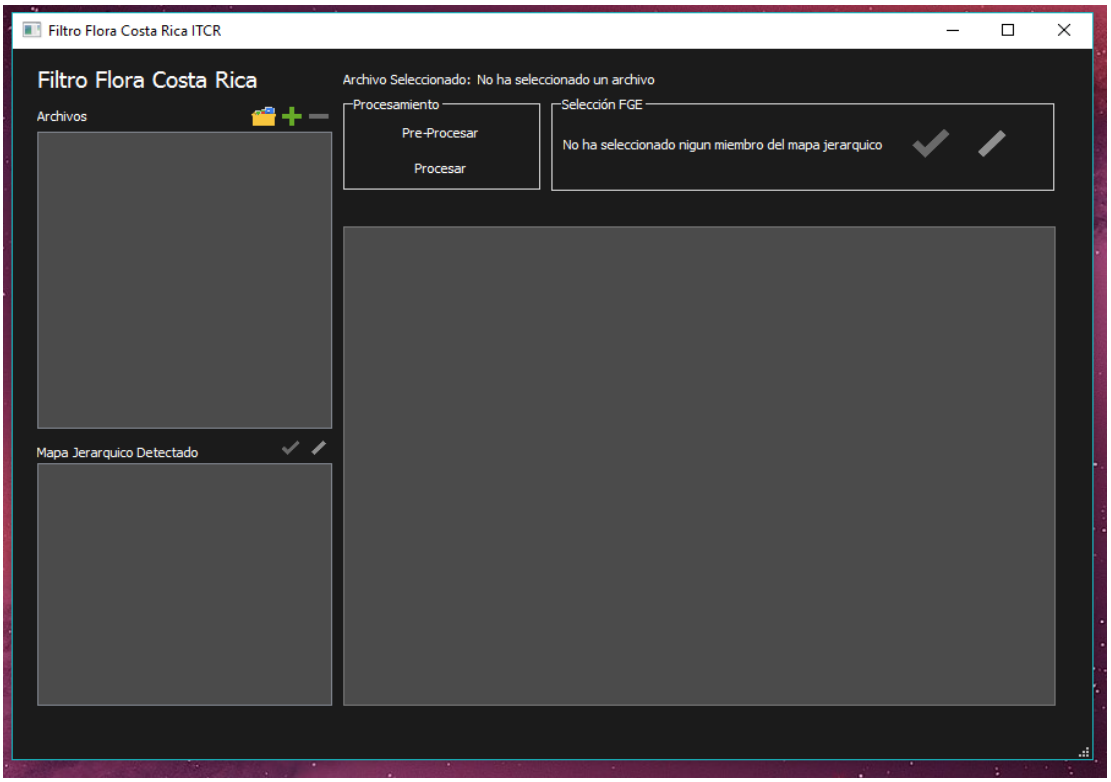


Figura 2.4 Interfaz Filtro Flora de Costa Rica

2.2.2 Herramienta de clasificación manual de texto

Determinar automáticamente si un fragmento de un documento biológico es una descripción morfológica o es algún otro tipo de información, no era un objetivo de este proyecto. Por esa razón, se desarrolló una herramienta manual que permita a un usuario conocedor señalar distintos fragmentos de un texto y etiquetarlos indicando el tipo de información que contienen.

La herramienta no modifica los archivos sino que registra el rango de posiciones de cada fragmento y almacena en una base de datos Mongodb la siguiente información: posiciones de inicio y fin, copia del fragmento escogido, categoría asignada. Esto permite extraer luego los fragmentos de una colección para que herramientas de más alto nivel estructuren su contenido.

La Figura 2.5 muestra un ejemplo en el que se han marcado fragmentos de texto correspondientes a la descripción morfológica y a la distribución geográfica de la especie Matudea trinervia.

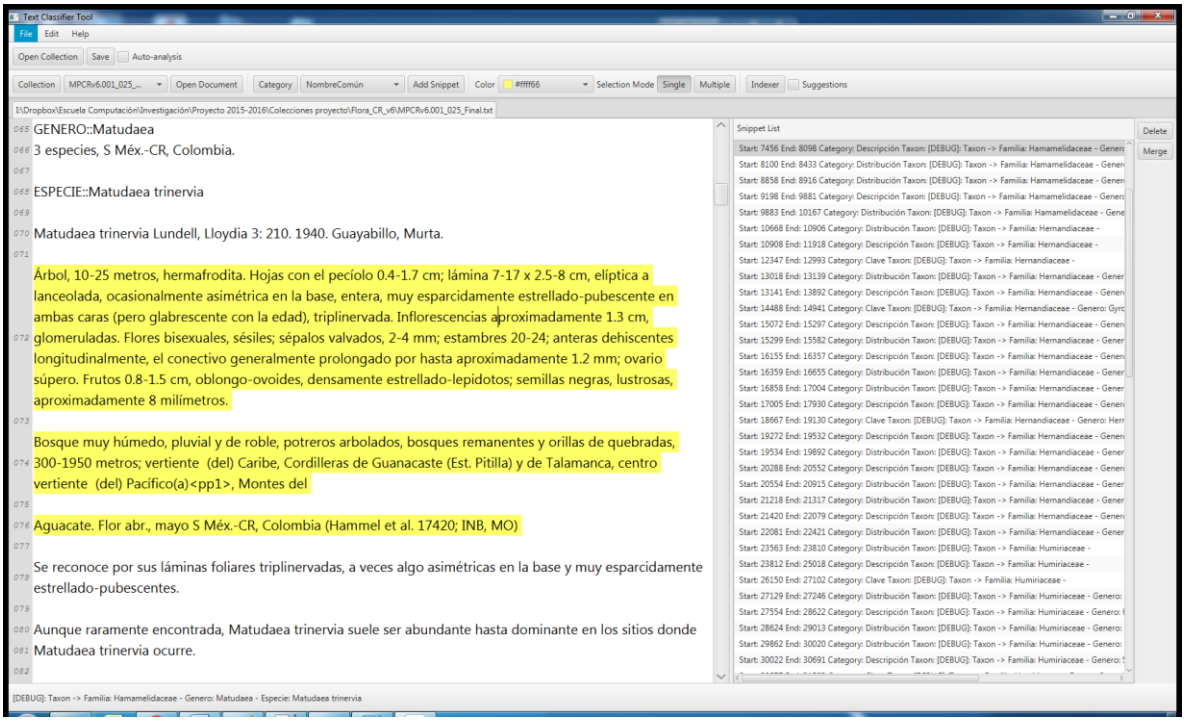


Figura 2.5 Herramienta de marcado y clasificación

La herramienta permite que una vez marcados los fragmentos de interés sea posible extraerlos en archivos XML tal como los muestra la Figura 2.6.

En el Apéndice 8.3 Herramientas desarrolladas Herramientas desarrolladas se describe con más detalle esta herramienta.

```
?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<snippets>
    <snippet>
        <category>
            <id/>
            <name>Descripción</name>
        </category>
        <taxon>
            <id/>
            <family>Hamamelidaceae</family>
            <genre>Matudaea</genre>
            <species>Matudaea trinervia</species>
        </taxon>
        <text>Árbol, 10-25 metros, hermafrodita. Hojas con el pecíolo 0.4-1.7
              cm; lámina 7-17 x 2.5-8 cm, elíptica a lanceolada,
              ocasionalmente asimétrica en la base, entera, muy esparcidamente
              estrellado-pubescente en ambas caras (pero glabrescente con la
              edad), triplinervada. Inflorescencias aproximadamente 1.3 cm,
              glomeruladas. Frutos 0.8-1.5 cm, oblongo-ovoides, densamente
              estrellado-lepidotos; semillas negras, lustrosas,
              aproximadamente 8 milímetros. </text>
    </snippet>...
</snippets>
```

Figura 2.6 Fragmentos de interés extraídos

Primera versión de un clasificador automático

Se desarrolló una opción básica de clasificación automática. Usando una colección de fragmentos previamente categorizados (Árboles de Costa Rica), se montó una base de datos usando Lucene de Apache. Usando dicha base de datos se procedió a aplicar un algoritmo de vecinos más próximos (Nearest Neighbors) cada vez que el usuario selecciona un fragmento de texto en un nuevo documento. De ese modo se obtienen los vecinos más próximos ya categorizados y se escogen las categorías que aparecen más veces y se le muestran al usuario como sugerencias. Aunque este desarrollo no es parte del proyecto, se realizó para explorar rápidamente posibles temas de investigación futura.

2.2.3 Colección de descripciones biológicas con principales componentes etiquetados explícitamente

Se usó la herramienta de marcado y clasificación para etiquetar los componentes de las descripciones morfológicas, las distribuciones geográficas y las claves dicotómicas para dos documentos fuentes:

- Manual de Flora de Costa Rica, volúmenes 5 y 6 [37].
- Árboles de Costa Rica, libro 3 [38].

Las versiones más actualizadas de esas colecciones etiquetadas se encuentran almacenadas en una base de datos Mongodb. De la cual se han extraído documentos XML específicos para cada grupo:

- Descripción.XML
- Distribución.XML
- Clave.XML

Dichos archivos fueron usados como fuentes de entrada para el desarrollo de algoritmos de extracción de información estructurada.

2.3 Desarrollo de algoritmos para la extracción de información estructurada a partir de las descripciones morfológicas.

En esta etapa las diferentes categorías de fragmentos son analizados en profundidad para hacer explícita la semántica de la información. Se realizaron dos tesis de Maestría en Computación, cada una enfocada en un tipo de información particular. En la tesis de María Auxiliadora Mora se trabajó con las descripciones morfológicas; mientras que en la tesis de Moisés Acuña se trabajó con las distribuciones geográficas. Adicionalmente se hizo un análisis más superficial de las claves dicotómicas, preparando el camino a una tercera tesis.

2.3.1 Algoritmos y herramientas de análisis semántico para enriquecer las descripciones biológicas usando marcadores altamente estructurados.

María Auxiliadora Mora presentó en su tesis en junio 2016, la cual fue aprobada con la máxima distinción. En ella implementó y evaluó un algoritmo para etiquetar las descripciones morfológicas y producir un conjunto de datos altamente estructurados y relacionados con ontologías estándar de la botánica.

Para realizar la estructuración del texto de las descripciones morfológicas presentes en dos fuentes bibliográficas (Manual de Plantas de Costa Rica: MPCR [37], Árboles de Costa Rica: ACR [38]), se usaron técnicas de procesamiento de lenguaje natural, reglas morfo-sintácticas y ontologías. Se usó tecnología existente como la biblioteca de NLP Freeling, el Organizador de Términos de Ontología (OTO), la Ontología de Plantas (PO) y el Glosario inglés-español, español-inglés para la Flora Mesoamericana.

La meta era identificar estructuras, subestructuras, estado de los caracteres, y restricciones. Así como relacionar los caracteres con la estructura/subestructura que corresponde y determinar relaciones entre estructuras.

La Figura 2.7 muestra un ejemplo de estructuración de una cláusula del libro ACRv4. En este ejemplo, la frase “hojas simples” genera un objeto de tipo biological_entity para la estructura “hojas“ y un objeto de tipo character para el carácter “architecture” con estado “simples“. El nombre del carácter “architecture“ se toma de la PO debido a que comúnmente estos componentes no aparecen explícitamente en las descripciones. A veces se presenta ambigüedad para asociar un estado con un carácter; en este caso el sistema agrega una nota de “Carácter repetido“ para que en una etapa posterior algún experto determine a cual carácter corresponde el estado.

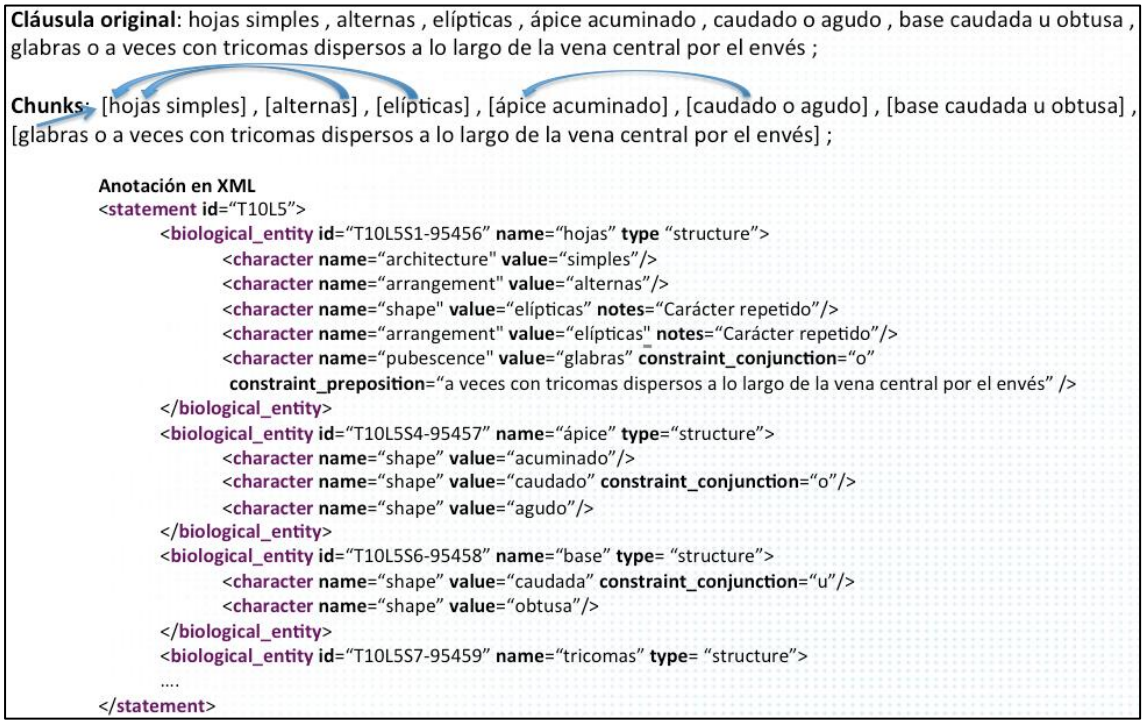


Figura 2.7 Ejemplo de estructuración

El algoritmo asocia cada carácter con la estructura o subestructura a la que corresponde utilizando el orden de aparición de los tokens y su concordancia en género y número. La Figura 2.8 muestra para la misma descripción, las entidades biológicas reconocidas (en rojo), los

Figura 2.8 Ejemplo asociación caracteres con estructuras

estados de los caracteres (en verde) y la relación entre entidades biológicas y caracteres (flecha).

El esquema de datos utilizado para la estructuración fue propuesto por Cui [18]. Además de secciones asociadas a metadatos generales, ese esquema define conceptos que permiten estructurar la taxonomía superior y conceptos asociados a las descripciones morfológicas como biological_entity, character y relation.

El algoritmo propuesto fue implementado utilizando Java porque facilitaba la integración de la tecnología seleccionada. El sistema almacena los resultados de la estructuración en una base de datos PostgreSQL.

La Figura 2.9 presenta el diagrama de flujo del algoritmo. La entrada inicial del sistema está compuesta por documentos en formato tabular que contienen las descripciones morfológicas y los nombres científicos de las especies a procesar.

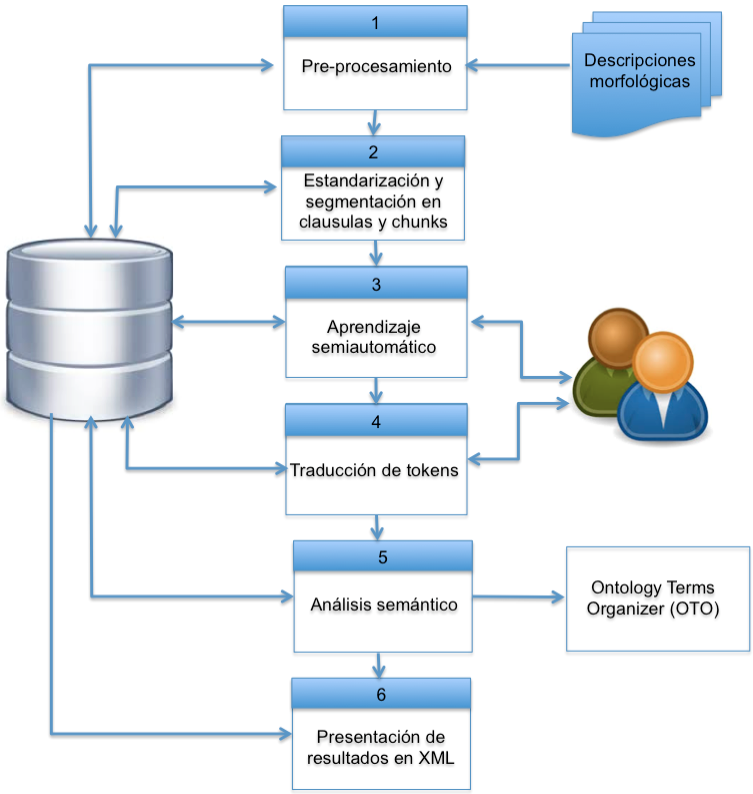


Figura 2.9 Diagrama de flujo algoritmo de estructuración de descripciones morfológicas

El algoritmo consta de las siguientes etapas:

Etapa- 1. Pre-procesamiento de los textos de las descripciones morfológicas. Eliminar comillas y agregar puntos finales.

Etapa- 2. Estandarización y segmentación de las descripciones en cláusulas y chunks. Separar usando puntos, dos puntos y punto y coma; estandarizar convirtiendo a minúsculas y separando por un solo espacio; separar en chunks usando comas.

Etapa- 3. Aprendizaje semiautomático de nombres de estructuras y estado de los caracteres.

Etapa- 4. Traducción de tokens al inglés para que coincidan con entradas en la PO.

Etapa- 5. Anotación semántica de las descripciones.

Etapa- 6. Generación de resultados en XML siguiendo el esquema propuesto por Cui.

Para evaluar el algoritmo se usaron las métricas de precisión, cobertura y la medida F. Para evaluar la correctitud de la estructuración de las cláusulas se usó el concepto de "razonable" propuesto por Cui:

- una estructura es identificada de forma razonable si se determina bien su nombre
- una estructura es identificada de forma estricta si además de razonable, se identifican correctamente el modificador y las restricciones
- los caracteres y estados se identifican de forma razonable si el nombre aparece en la ontología y se identifican bien el valor, la unidad de medida y los rangos
- los caracteres y estados se identifican de forma estricta si además de razonable se identifican correctamente las restricciones

Para la evaluación se tomó una muestra de 5% de las cláusulas por libro. La selección de cláusulas fue realizada utilizando el método de selección por la rueda de la ruleta. Esto permitió escoger aleatoriamente los chunks de texto para obtener casos tanto de cláusulas simples como de cláusulas complejas.

2.3.2 Algoritmos y herramientas de análisis semántico para enriquecer las distribuciones geográficas usando marcadores altamente estructurados.

La tesis de Moisés Acuña implementa un algoritmo que desglosa los párrafos en los que se tratan las distribuciones de las especies e identifica las entidades geográficas mencionadas y las codificadas siguiendo alguna autoridad estandarizadora.

La Figura 2.10, tomada de dicha tesis, presenta la arquitectura general del software que implementa los algoritmos de geo-parsing y geo-coding.

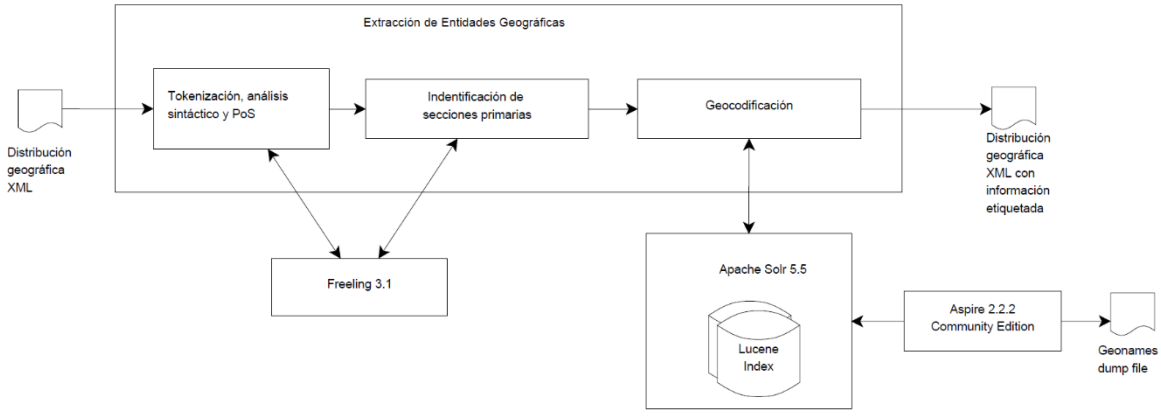


Figura 2.10 Extracción de elementos geográficos - Arquitectura general

Para el procesamiento de la información sobre distribución geográfica se procedió de la siguiente manera.

1. Selección del gazetteer

    - Se analizaron dos gazetteers disponibles para localizar nombres geográficos para Costa Rica y el mundo: NGA GEOnet Names Server (GNS) [4], y GeoNames [3].

    - El gazetteer GeoNames fue seleccionado por contener en general mayor contenido, en particular contiene nombres para Estados Unidos que resultan importantes al procesar la distribución mundial de las especies.

2. Arquitectura general

    - Se construyó un módulo de programación para ingresar datos del gazetteer en el motor de búsqueda Apache Solr: se alimentó el archivo del gazetteer a la plataforma Aspire Community [2], la cual procesó el archivo separado por comas de GeoNames y lo indexó usando Apache Solr.

    - Se utilizó un parser gramatical para hacer el geo-parsing y extraer información adicional como zonas de vida, elevación y meses de floración.

    - Hay un submódulo que realiza el análisis de los párrafos: tokenización, análisis sintáctico y etiquetado POS. Luego un submódulo de geo-parsing divide los párrafos y encuentra los posibles puntos geográficos.

    - El módulo siguiente realiza la geocodificación para asociar los puntos geográficos encontrados con entradas en el gazetteer; para ello se usa el motor de búsqueda Solr.

3. Diseño experimental

    - La evaluación de la precisión de los algoritmos de geo-coding y geo-se realizó de la siguiente manera: se escogió un subconjunto aleatorio de los topónimos extraídos; para cada topónimo escogido, se determinó si fue extraídos correctamente y si fue localizado correctamente en el gazetteer.

    - Se trabajó con tres subconjuntos de topónimos seleccionados al azar. Esos conjuntos contienen 1%; 2,5% y 5% del conjunto total de los topónimos obtenidos para la distribución en Costa Rica, en el mundo.

    - El volumen VI del MPCR fue usado para el desarrollo y ajuste de los algoritmos, por lo que se usó el volumen V del MPCR para la evaluación.

    - Los subconjuntos de prueba fueron evaluados manualmente y a cada caso se le asignó un nivel de BUENO, MALO y DESCONOCIDO.

2.4 Herramientas para consultar y operar la información extraída

El Objetivo 6 del proyecto indica que se deben desarrollar herramientas permitan consultar y operar con la información extraída. Como resultado, la última etapa del proyecto se concentró en el desarrollo de dichas herramientas.

Herramienta para consultas morfológicas

Esta herramienta de consulta permite la identificación de especies mediante un proceso de consulta incremental que en base a los datos conocidos y a las descripciones estructuradas de los árboles que cumplan con esos datos, se presente al usuario taxones alternativos que cumplan con esas condiciones.

Herramienta para consultas geográficas

Esta herramienta de consulta permite extraer y operar la información geográfica extraída. Se requiere usar catálogos de localizaciones geográficas para aproximar las descripciones textuales a mapas y operar con dichos mapas por medio de operaciones de intersección y unión.

Herramienta para consultas a claves dicotómicas

Aunque actualmente se dispone de herramientas muy flexibles para ayudar a clasificar un espécimen (por ejemplo, LUCID [39]), no deja de ser útil poder aprovechar la información incluida en claves dicotómicas de fuentes ya establecidas. Hay mucho conocimiento incluido por expertos en esas claves y no debe ser descartado. Una herramienta que permita explorar dichas claves facilitaría a expertos acceder a información muy valiosa mientras elaboran claves matriciales más modernas.

Esta herramienta debe permitir consultar claves dicotómicas mediante patrones sin tener que seguir la secuencia lineal impuesta por dichas claves.

3 Resultados

Los resultados obtenidos serán presentados siguiendo el orden de desarrollo del proyecto descrito en la sección 2.1 Introducción general del proyecto.

3.1 Recolección y transformación de los documentos

Esta etapa produjo como resultado el desarrollo exitoso de dos herramientas:

- Filtro de pre-procesamiento
- Herramienta de clasificación de texto

Ambas herramientas se encuentran documentadas en documentos anexos. Su código fuente se encuentra almacenado en el repositorio bitbucket.org. Usando esas dos herramientas se produjeron varias colecciones de fragmentos de las secciones de descripción morfológica, distribución geográfica y claves dicotómicas de los volúmenes 2 y 4 del MPCR. La Tabla 1 Describe el tamaño y el contenido de las colecciones de fragmentos desarrolladas. Dichas colecciones fueron usadas para los algoritmos de análisis más profundo realizados en la etapa siguiente.

Tabla 1 Colecciones de fragmentos

Colección	Tamaño	Contenido
2016-01-28 Clave-Flora CR	100	claves dicotómicas
2016-01-28 Distribución-Flora CR	796	distribuciones geográficas
2016-01-28 Descripción-Flora CR	734	descripciones morfológicas
2015-11-29 Descripción-Árboles de CR	237	descripciones morfológicas
2016-08-09 Distribución-Flora CR v2	1330	distribuciones geográficas

3.2 Resultados tesis de María Auxiliadora Mora

Luego de crear una colección de fragmentos de texto con las descripciones morfológicas disponibles en ACRv3 y ACRv4 así como una selección tomada del MPCR, se realizó el procesamiento de estructuración descrito en la metodología. Dicho trabajo fue realizado como tesis de Maestría en Computación de María Auxiliadora Mora. Los resultados obtenidos se evaluaron tomando una muestra del 5% del total de cláusulas generadas. La selección fue realizada usando el algoritmo de la ruleta con el fin de tomar en cuenta la complejidad de las cláusulas. La Tabla 2 muestra la cantidad promedio de estructuras y caracteres para cada libro.

Tabla 2 Cantidad promedio de estructuras y caracteres en las cláusulas evaluadas

Libro	Cantidad de descripciones	Cláusulas	Cláusulas (muestra)	Promedio (en la muestra)	
				Estructuras	Caracteres
ACRv3	233	1.738	87 (5%)	2,85	3,62
MPCR	237	2.230	106 (5%)	3,42	3,69

La complejidad de las cláusulas en la muestra tomada del libro ACRv3 estuvo bien distribuida. Un 52% de las cláusulas eran simples y un 48% complejas. Se estimó que una cláusula era simple si tiene dos o menos estructuras, y era complejas si tiene más de dos estructuras. La complejidad de la muestra del MPCR estuvo también muy bien distribuida. Un 53% de las cláusulas eran simples y un 47% complejas

Se calculó la precisión, cobertura y medida F para la muestra de cada libro de forma individual. La Tabla 3 muestra el resultado obtenido para la medida F, la cual combina precisión y cobertura.

Tabla 3 Rendimiento (F) del algoritmo al ser aplicado a la muestra de los libros ACRv3 y MPCR

Libro	Identificación de estructuras (F)		Identificación de caracteres (F)		Asociación caracteres a estructuras (F)	Asociación conjunc. (F)	Promedio (valores razonable)
	Razonable	Estricto	Razonable	Estricto			
ACRv3	98,7	97,9	99,1	98,5	98,7	96,4	98,2
MPCR	99,7	98,1	98,1	93,3	86,4	92,4	94,1

Los resultados (precisión y cobertura) obtenidos al evaluar el algoritmo de aprendizaje no supervisado (bootstrapping) con datos de los libros ACRv4, MPCR y ACRv3 se muestran en la Tabla 4. El algoritmo clasifica nombres, verbos y adjetivos en estados de carácter (A), estructuras (E) y verbos (V) que son almacenados en la base de conocimiento.

Tabla 4 Resultados al evaluar el algoritmo no supervisado (bootstrapping)

Libro	Tipo de token	Precisión	Cobertura	Rendimiento (F)
ACRv4	Valor de carácter (A)	99,6% (486/488)	94,2% (486/516)	96,8
ACRv4	Estructura (E)	83,1% (138/166)	98,6% (138/140)	90,1
ACRv4	Verbo (V)	89,5% (17/19)	100% (17/17)	94,4
MPCR	Valor de carácter (A)	99,8% (628/629)	96,3% (628/652)	98
MPCR	Estructura (E)	86% (130/151)	98,5% (130/132)	91,8
MPCR	Verbo (V)	83,3% (35/42)	100% (35/35)	90,9
ACRv3	Valor de carácter (A)	100% (183/183)	87,1% (183/210)	93,1
ACRv3	Estructura (E)	57,8% (37/64)	100% (37/37)	73,2
ACRv3	Verbo (V)	100%	100%	100%

3.3 Resultados tesis de Moisés Acuña

Las pruebas realizadas para evaluar la extracción de información geográfica, usaron fragmentos de texto extraídos del libro Manual de Plantas de Costa Rica. VI: Dicotiledoneas (Haloragaceae-Phytolaccaceae) [16]. Dichos fragmentos fueron obtenidos usando las herramientas de filtrado y marcado de texto desarrolladas por este proyecto. Los fragmentos y sus atributos se almacenan en elementos XML llamados snippets. A partir de los fragmentos se extrajeron cláusulas que especifican entidades geográficas identificadas.

La evaluación fue realizada manualmente sobre tres muestras aleatorias (1.25%, 2.5%, 5%), tomadas sobre las cláusulas generadas para las distribuciones en Costa Rica y las distribuciones en el mundo. La Tabla 5 detalla el número total de cláusulas en las distribuciones y el número de cláusulas tomadas para la muestra del 5%.

Tabla 5 Cantidad de cláusulas generadas y seleccionadas para una muestra del 5%

Tipo de distribución	Número total de cláusulas	Número de cláusulas en la muestra
Distribución en Costa Rica	6638	331
Distribución en el mundo	3052	152

La evaluación clasificó las cláusulas en una de las siguientes categorías:

BUENO: término geográfico correctamente identificado, encontrado en gazetteer y concuerda con el contexto

MALO: término encontrado no es un punto geográfico, o el término fue encontrado en el gazetteer pero no concuerda con el contexto geográfico, o el término no fue encontrado en el gazetteer pero sí lo contenía

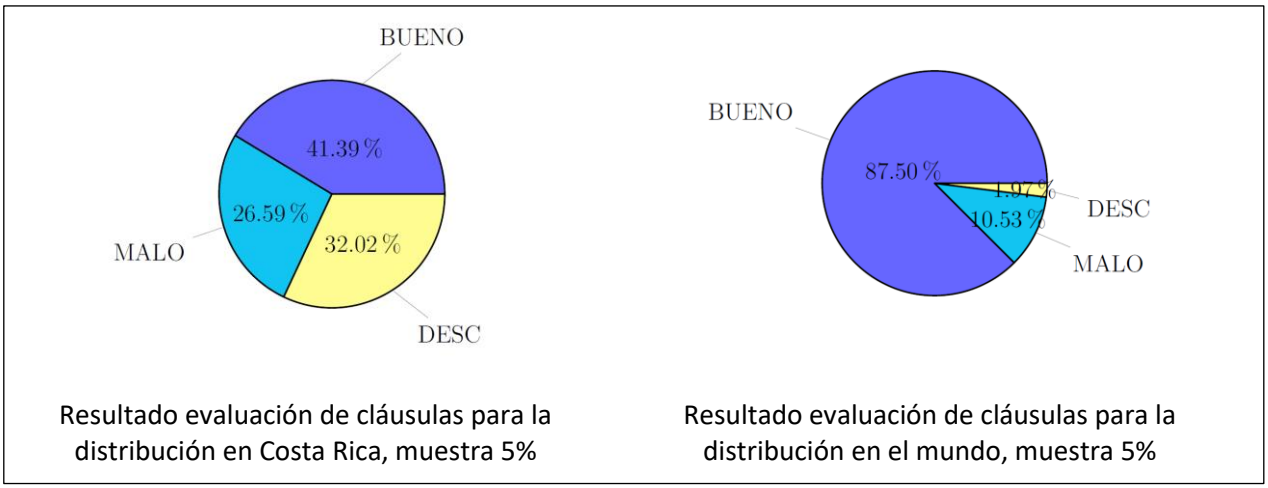
DESCONOCIDO:
término encontrado si es un punto geográfico, pero no está contenido en el gazetteer.

La Tabla 6 contiene el resultado de la evaluación realizada.

Tabla 6 Resultados de evaluación de geo-parsing y geo-codificación al 5%

Tipo de Cláusula	5% (frecuencia / %)		
	BUENO	**MALO**	**DESCONOCIDO**
Distribución en Costa Rica	137 41,39%	88 26,59%	106 32,02%
Distribución en el mundo	133 87,50%	16 10,53%	3 1,97%

La Figura 3.1 muestra en forma gráfica esos mismos resultados.



Resultado evaluación de cláusulas para la distribución en Costa Rica, muestra 5%

Resultado evaluación de cláusulas para la distribución en el mundo, muestra 5%

Figura 3.1 Resultado de evaluación para distribución en Costa Rica y en el mundo para muestra al 5%

Para la muestra más grande (5%), se analizó más profundamente el tipo de error (MALO) y se desglosó en dos subcategorías:

GEO-PARSING: el término encontrado no corresponde a un posible punto geográfico

GEO-CODING: el término es encontrado en el gazetteer, pero no concuerda con el punto geográfico al cual se refiere el contexto; también se considera el caso en que el término no fue encontrado en el gazetteer, pero sí estaba contenido en el mismo

La Tabla 7 desglosa el tipo de error según las dos subcategorías presentadas arriba.

Tabla 7 Tipo de error de las cláusulas que fueron clasificadas con MALO, muestra 5%

	Distribución en Costa Rica (frecuencia /%)	**Distribución en el mundo (frecuencia /%)**
GEO-PARSING	66 75,0%	6 37,5%
GEO-CODING	22 25,0%	10 62,5%
Total	88	16

La Figura 3.2 muestra el tipo de error de las cláusulas clasificadas como MALO para la muestra del

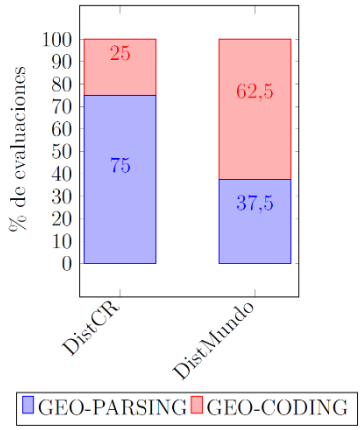


Figura 3.2 Tipo de error de las cláusulas clasificadas con MALO

5%

3.4 Herramientas de consulta

La última etapa del proyecto consistió en desarrollar una serie de herramientas de consulta con el fin de explorar el aprovechamiento de la estructuración obtenida para las secciones de interés. A continuación, se describe el estado final de dichas herramientas.

Herramienta para consultas morfológicas

Fue desarrollada exitosamente. Permite la identificación de especies mediante un proceso de consulta incremental que en base a los datos conocidos y a las descripciones estructuradas de los árboles que cumplan con esos datos, se presente al usuario alternativas para refinar la identificación. La interfaz de esta herramienta se muestra en la Figura 3.3.

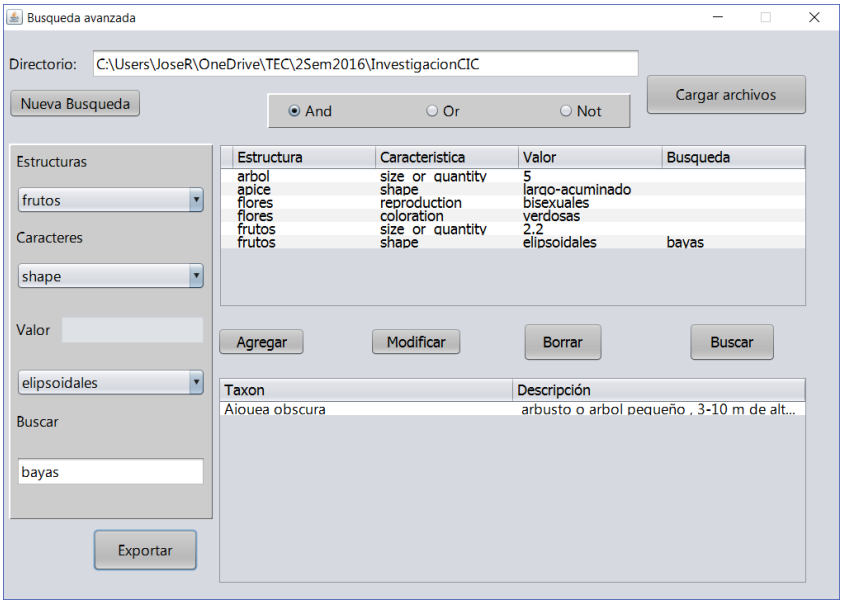


Figura 3.3 Interfaz herramienta para consultas morfológicas

Herramienta para consultas geográficas

Fue desarrollada exitosamente. Permite seleccionar taxones y mostrar gráficamente la información contenida en frases que se refieren a la distribución mundial de la especie. También muestra un ítem de información adicional, meses de floración, que estaba disponible. La herramienta además permite realizar operaciones con las distribuciones geográficas de varios taxones (unión, intersección, diferencia). La interfaz de esta herramienta se muestra en la Figura 3.4.

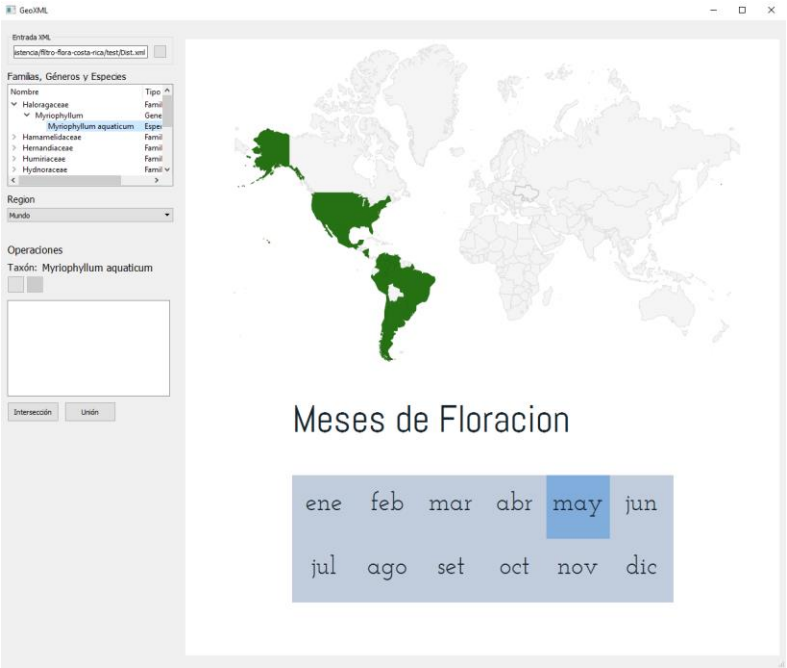


Figura 3.4 Interfaz de la herramienta para consultas geográficas

Herramienta para consultas a claves dicotómicas

Fue desarrollada exitosamente. Permite consultar claves dicotómicas, extraídas de los documentos fuentes, mediante patrones sin tener que seguir la secuencia impuesta por dichas claves. La interfaz de esta herramienta se muestra en la Figura 3.5.

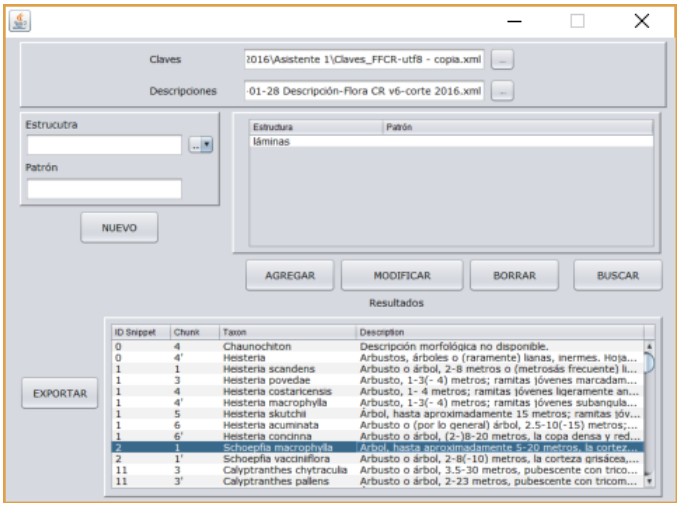


Figura 3.5 Interfaz herramienta para consulta de claves dicotómicas

4 Discusión y conclusiones

Objetivo propuesto	Relación con los resultados obtenidos
1. Establecer un modelo que indique los principales componentes de las descripciones biológicas	Se obtuvo la lista de conceptos estructurada jerárquicamente que se muestran en el Apéndice 8.1. Se planeaba usar para el desglose automático de las descripciones biológicas, pero no fue requerida dada la flexibilidad de las herramientas de procesamiento de lenguaje usada y a las ontologías disponibles.
2. Desarrollar una herramienta de software para etiquetar manualmente grandes fragmentos de texto usando la lista de conceptos principales desarrollada anteriormente.	La herramienta fue desarrollada exitosamente. Permite marcar fragmentos de texto y asociarlos a categorías de información. Permite además aprovechar si fuera del caso información más explícita del texto para categorizarlo automáticamente. También permite aprovechar la información previamente categorizada para sugerir una categoría para nuevos fragmentos.
3. Aplicar herramienta genérica de etiquetado para obtener colección de descripciones biológicas con sus principales componentes marcados.	Se obtuvieron varias colecciones de fragmentos de las secciones de descripción morfológica, distribución geográfica y claves dicotómicas para los volúmenes 2 y 4 del Manual de Plantas de Costa Rica.
4. Establecer listas de términos y abreviaturas relacionados con los principales componentes de las descripciones biológicas.	El análisis que permitió identificar los términos más significativos en las descripciones biológicas, resulta útil para un análisis textual, pero resultó innecesario en un análisis más profundo como el realizado por María Auxiliadora Mora en su tesis[MARI MORA]. Para dicha tesis, resultó más útil usar las facilidades de ontologías como OTO, además de que el análisis gramatical permite establecer claramente los diferentes componentes de las descripciones biológicas. De todas formas en el Apéndice 8.2 se muestran las listas de términos obtenidas.
5. Extraer información semánticamente estructurada sobre distribución geográfica, relaciones interespecíficas y morfología de las especies a partir de descripciones biológicas.	El algoritmo desarrollado para la estructuración de las descripciones morfológicas de plantas alcanzó resultados muy competitivos logrando un rendimiento promedio del 94,1%. Aunque el algoritmo implementado se basa en el lenguaje telegráfico utilizado por la comunidad de expertos botánicos, este puede generalizarse a otros grupos biológico pre-procesando los textos de las descripciones para omitir algunas palabras funcionales. Hay mucha información disponible en las formas preposicionales de las descripciones morfológicas, las cuales debido a su gran variedad no fue posible de procesar en

	esta ocasión. El algoritmo desarrollado para la estructuración de las distribuciones geográficas mostró una gran diferencia entre la precisión obtenida para distribuciones en el mundo (87,5%) y distribuciones en Costa Rica (41,39%). Lo cual es provocado por dos factores: • Las distribuciones geográficas para Costa Rica contienen oraciones gramaticalmente más complejas que las oraciones de las distribuciones en el mundo. • El gazetteer usado no incluye muchos puntos geográficos de Costa Rica que son requeridos para procesar adecuadamente las distribuciones geográficas para Costa Rica. El procesamiento de claves dicotómicas no se realizó a profundidad por falta de tiempo. Sin embargo, está en desarrollo una tesis para completar este punto. Finalmente se elaboró un script de tipo sintáctico para realizar una primera estructuración de las claves dicotómicas.
6. Desarrollo de herramientas permitan consultar y operar con la información extraída.	Las herramientas de consulta desarrolladas muestran que la información estructurada puede ser aprovechada por los usuarios por medio de interfaces sencillas y precisas Tanto la herramienta búsqueda morfológica como la herramienta para operaciones geográficas y la herramienta búsqueda claves dicotómicas alcanzaron satisfactoriamente el nivel de prueba de concepto.

5 Recomendaciones

Tanto para la estructuración de las descripciones morfológicas como para la estructuración de las distribuciones geográficas se determinó la necesidad de estructurar una herramienta de traducción de tokens del español al inglés con el fin de aprovechar los recursos más avanzados que hay disponibles. Dicha herramienta permitiría aprovechar plenamente las ontologías y los gazetteers existentes.

En la estructuración de descripciones morfológicas hubo información que no fue extraída (sintagmas preposicional o verbal), debido a la complejidad del análisis. Un futuro proyecto debe refinar debe enfocarse en extraer la información contenida en estos sintagmas. Además, se deben explorar heurísticas tanto para resolver la ambigüedad que a veces se presenta en la asociación entre caracteres y estructuras, como para mejorar la heurística simple de género y número usada en este proyecto.

Para la estructuración de la distribución geográfica se debe profundizar el algoritmo de geo-parsing para que pueda reconocer mejor los términos dentro de oraciones enumerativas y con preposiciones espaciales, composición, y posesivas. También se debe estudiar cómo resolver el problema de cómo localizar los nombres correctos en inglés en el gazetteer, cuando el término en el texto fuente está en español.

En cuanto a las herramientas desarrolladas, se requiere integrarlas en un solo ambiente: filtro, marcador, buscadores. Incluir también dentro de esta integración el acceso por medio de web services a las herramientas avanzadas de estructuración desarrolladas por las tesis. Además, como ya se dispone de una colección de fragmentos ya clasificados, se pueden aprovechar para desarrollar un mecanismo de categorización automática de texto.

Finalmente, se debe retomar el procesamiento de relaciones entre especies; fue omitido por dificultades para obtener un conjunto de datos que proveyera suficiente información, pero se trata de una temática de muy alto interés.

6 Agradecimientos (opcional)

Se agradece profundamente la colaboración brindada y la entrega entusiasta de los asistentes de este proyecto: Andrés Aguilar, Alejandro Rojas, José Rodolfo Garita y Kennet Quirós.

Es un gusto también reconocer el privilegio de haber trabajado con profesionales de tan alto nivel como María Auxiliadora Mora y Moisés Acuña, quienes desarrollaron sus trabajos de tesis con un alto sentido de la calidad académica y de la búsqueda de la excelencia.

7 Referencias

[1] A. D. B. K. &. K. C. Sautter G, «Creating digital resources from legacy documents – an experience report from the biosystematics domain,» de *Proceedings of ESWC*, Heraklion, Greece 2009, 2009.

[2] H. Cui, «Converting Taxonomic Descriptions to New Digital Formats,» *Biodiversity Informatics,* pp. 20-40, 2008.

[3] M. N. G. &. B. C. Gerner, «LINNAEUS: A species name identification system for biomedical literature,» *BMC Bioinformatics,* p. 11:85, 2010.

[4] S. I. &. M. T. Koning D, «TaxonGrab: Extracting taxonomic names from text,» *Biodiversity Informatics,* pp. 79-82, 2006.

[5] M. D. B. K. Cunningham H, Text Processing with GATE, Gateway Press CA, 2011.

[6] H. C. D. M. Thessen A. E., «Applications of natural language processing in biodiversity science,» *Adv. Bioinformatics,* 2012.

[7] N. Madnani, «Getting started on natural language processing with Python,» *Crossroads,* p. 1–16, 2007.

[8] A. Steven, Semisupervised Learning for Computational Linguisticts, Chapman & Hall/CRC, 2007.

[9] E. C. a. D. McDermott, Artificial Intelligence, Addison-Wesley Publishing Company, 1985.

[10] J. H. a. E. Riloff, «Information Extraction,» de *Handbook of Natural Language Procesing*, Second ed., F. D. I. Nitin, Ed., CRC Press, 2010, p. 511–526.

[11] G. Algorithms, Introduction to Genetic Algorithms, 2008.

[12] P. Leary, «TaxonFinder org,» [En línea].

[13] K. B. a. D. A. G. Sautter, «A combining approach to Find All taxon names (FAT) in legacy biosystematics literature,» *Biodivers. Informatics,* vol. 3, p. 46–58, 2006.

[14] L. M. Akella, C. N. Norton y H. Miller, «NetiNeti: discovery of scientific names from text using

machine learning methods,» *BMC Bioinformatics,* vol. 13, p. 211, 2012.

[15] C. Klingenberg, G. Sautter, D. Agosti y T. Catapano, «GoldenGATE XML Markup Editor Introduction and Manual for the Generation of TaxonX-based Legacy Literature Documents using the GoldenGATE Editor,» p. 34.

[16] G. B. Curry y R. J. Connor, «Automated Extraction of Biodiversity Data from Taxonomic Descriptions,» *Biodiversity Databases Techniques, Politics, and Applications,* 2007.

[17] H. Duan, Y. Hei y Z. Cui, «Heuristics based semantics annotation of biodiversity documents in Chinese,» *Chinese J. Libr. Inf. Sci.,* 2013.

[18] H. Cui, «CharaParser for Fine-Grained Semantic Annotation of Organism Morphological Descriptions,» *J. Am. Soc. Inf. Sci. Technol.,* vol. 63, nº 4, pp. 738–754,, 2012.

[19] J. P. Balhoff, W. M. Dahdul, C. R. Kothari, H. Lapp, J. G. Lundberg, P. Mabee, P. E. Midford, M. Westerfield y T. J. Vision, «Phenex: Ontological annotation of phenotypic diversity,» *PLoS One,* vol. 5, nº 5, p. 1–10, 2010.

[20] H. Cui y M. Studies, «MARTT: A General Approach to Automatic Markup of Taxonomic Descriptions with XML,» *CAIS,* p. 1–11, 2005.

[21] M. M. Wood, L. S.J., V. Tablan, D. Maynard y H. Cunningham, «Populating a Database from Parallel Texts using Ontology-based Information Extraction,» 2004.

[22] A. Rocio y S. Afredo, «X-tract: Structure extraction from botanical textual descriptions,» de *Proc. string Process. Inf. Retr. Symp. Int. Work. Groupw.*, 1999.

[23] J. Diederich, R. Fortuner y J. Milton, «Computer-assisted data extraction from the taxonomical literature,» Department of Mathematics, University of California, Davis, [En línea]. Available: https://www.math.ucdavis.edu/~milton/genisys/terminator.html. [Último acceso: 19 3 2015].

[24] J. L. Leidner y M. D. Lieberman, «Detecting geographical references in the form of place names and associated spatial natural language,» *SIGSPATIAL Special,* vol. 3, nº 2, p. 5–11, 2011.

[25] L. L. Hill, Georeferencing: The Geographic Associations of Information (Digital Libraries and Electronic Publishing), The MIT Press, 2006.

[26] J. L. Leidner, G. Sinclair y B. Webber, «Grounding spatial named entities for information extraction and question answering,» de *Proceedings of the HLTNAACL 2003 Workshop on Analysis of Geographic References*, Stroudsburg, PA, 2003.

[27] J. E. Hopcroft, R. Motwani y . J. E. Ullman, Introduction to Automata Theory, Languages, and Computation, 3 ed., Reading, MA: Addison-Wesley, 2006.

[28] K. R. Beesley y L. Karttunen, Finite State Morphology, Center for the Study of Language and Inf, 2003.

[29] F. Bilhaut, T. Charnois, P. Enjalbert y Y. Math, «Geographic reference analysis for geographic document querying,» de *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Stroudsburg, PA, 2003.

[30] F. Schilder, Y. Versley y C. Habel, Extracting spatial information : grounding, classifying and linking spatial expressions, 2008.

[31] J. R. Finkel, T. Grenager y C. Manning, «Incorporating non-local information into information extraction systems by gibbs sampling,» de *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2005.

[32] J. R. Curran, S. Clark y J. Bos, «Linguistically motivated large-scale nlp with c&c and boxer,» de *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Prague, Czech Republic, 2007.

[33] C. D. Manning y H. Schütze, Foundations of Statistical Natural Language Processing, C. D. Manning y H. Schütze, Edits., The MIT Press, 1999.

[34] H. Cui, D. Boufford y P. Selden, «Semantic annotation of biosystematics literature,» *Journal of the American Society for Information Science and Technology,* vol. 61, nº 3, pp. :522–542,, 2010.

[35] B. Pouliquen, R. Steinberger, C. Ignat y T. De Groeve, «Geographical information recognition and visualization in texts written in various languages,» de *Proceedings of the 2004 ACM Symposium on Applied Computing*, New York, NY, USA, 2004.

[36] M. Kimler y R. Göbel, Geo-coding: Recognition of geographical references in unstructured text, and their visualisation, 2004.

[37] B. E. Hammel, Manual de plantas de Costa Rica, Manual de plantas de Costa Rica, 2004.

[38] N. Zamora, Q. Jiménez y L. Poveda, Árboles de Costa Rica vol. III (Spanish Edition), Editorial INBio, 2004.

[39] Lucidcentral.org, 13 6 2017. [En línea]. Available: URL http://www.lucidcentral.com/. [Último acceso: 13 6 2017].

[40] A. Hippisley, “Lexical Analysis,” in Handbook of Natural Language Procesing, Second ed., 2010, p. 31–58.

[41] C. G. a. A. C. Schalley, “Semantic Analysis,” in Handbook of Natural Language Procesing, Second ed., I. N. a. F. Damerau, Ed., CRC Press, 2010, p. 93–120.

[42] M. W. Ljunglof P., “Syntactic Parsing,” in Handbook of Natural Language Procesing, Second ed., 2010.

8 Apéndices

8.1 Conceptos identificados para el desglose de las descripciones biológicas

Lista de conceptos estructurada jerárquicamente (solo uno o dos niveles) que serán usados como pistas para el desglose automático de las descripciones biológicas.

Descripciones morfológicas

Presenta las características morfológicas de un taxón. Se identifican los siguientes dos niveles de conceptos en las descripciones morfológicas:

Parte: Partes principales del organismo vegetal. Separadas unas de otras por punto y seguido. Puede repetirse.
Ejemplos: árbol, arbusto, hierba, lianas, plantas, hoja, inflorescencia, flor, fruto.

SubParte: Subdivisiones en las descripciones de las partes. Separadas unas de otras por punto y seguido. Puede repetirse.
Ejemplos: tallos, estípulas, lámina, miembros del perianto, sépalos, pétalos, estambres, anteras, estaminodios, pistilo, ovario, óvulo, placentación, estilo, estigma, semilla.

Distribuciones

Presenta las características morfológicas de un taxón. Se identifican los siguientes dos niveles de conceptos en las distribuciones:

Para familias y géneros:

Número de géneros de esa familia en el mundo
Número de especies de esa familia o ese género en el mundo
Regiones geográficas del mundo en que está distribuida esa familia o género
Número de géneros de esa familia en Costa Rica
Número de especies de esa familia o ese género en Costa Rica

Para especies:

Zonas de vida de la especie
Rango de elevación en la que se encuentra la especie
Regiones en Costa Rica en que está distribuida esa familia o género
Meses de floración de la especie
Regiones geográficas del mundo en que está distribuida esa especie

Claves dicotómicas

Una clave dicotómica es un herramienta que permite a los biólogos identificar casos específicos dentro de un grupo de organismos. A partir de características morfológicas se van elaborando cláusulas booleanas, las cuales al contestarse falso o verdadero llevan a determinar un organismo específico o a evaluar cláusulas dicotómicas adicionales más refinadas. Las cláusulas se organizan en niveles. Cada nivel tiene un par de cláusulas, las cuales ofrecen dos opciones, una de las cuales debe ser escogida. En caso de que una cláusula determine un taxón específico, se asocia dicho taxón con la cláusula. En caso de que una cláusula no permita determinar un taxón específico, se desglosa el caso usando un nuevo nivel (par de cláusulas) y se continua con el proceso.

Ejemplo:

```
1  Hojas 3-5-lobuladas; frutos alados  . . . . . Gyrocarpus
1' Hojas nunca lobuladas; frutos sin alas.
   2  Plantas arbóreas; frutos lisos . . . . . Hernandia
   2' Plantas escandentes; frutos acostillados . Sparattanthelium
```

Para las claves dicotómicas se identificaron los siguientes dos niveles de conceptos:

Nivel de las cláusulas

Modificador de cláusula que indica si es la primera o la segunda del par de ese nivel

Descripción morfológica con las características que deben cumplirse

Característicasmorfológica que debe cumplirse. Pueden ser varias.

Taxón asociado a una cláusula

8.2 Listas de términos para descripción morfológica, distribución y claves dicotómicas

Se analizó la estructura seguida la información presentada en los manuales de flora estándar, en particular el Manual de Flora de Costa Rica, y se identificaron los siguientes componentes en las tres áreas de interés.

1. Descripciones
   Las descripciones se caracterizan por seguir una estructura jerárquica de acuerdo con las diferentes partes del organismo.

   En particular se identificaron los siguientes componentes, los cuales fueron separados de acuerdo con la frecuencia con que aparecen.

Alto	Medio		Bajo	
flor	semilla	pistilo	bractea	tubo
hoja	peciolo	planta	tepalo	copa
fruto	apice	rama	foliolo	bracteola
base	enves	endocarpo	estaminodio	pedicelo
lamina	ramita	endosperma	liana	valvula
petalo	nervio	verticilo	tronco	borde
estambre	filamento	hierba	axila	mericarpo
ovario	tricoma	estigma	lobulo	anillo
sepalo	antera	corteza	caliz	arbolito
arbol	placentacion	margen	espiga	pericarpio
arbusto	perianto		tallo	

   La importancia de los diferentes términos se puede visualizar mejor usando una nube de palabras:

2. Distribuciones

   Las distribuciones no solo incluyen referencias a áreas geográficas tanto políticas (países) como geológicas (vertientes) sino que incluyen conceptos como zonas de vida y altitudes. Adicionalmente incluyen información que sería conveniente capturar como cantidad de especies y géneros así como períodos de floración.

   Los siguientes conceptos han sido identificados:

Zonas de vida
Altitud
Áreas geográficas en el mundo
Áreas geográficas en CR
Cantidad de especies
Cantidad de géneros
Meses floración

En este caso los anteriores son conceptos generales. La siguiente nube de palabras muestra los términos particulares con su importancia relativa.

3. Claves

Las claves dicotómicas encontradas en los manuales de flora son una fuente de información muy valiosa. Siguen una estructura similar a las descripciones, aunque no son exhaustivas y solo mencionan las características de los organismos requeridas para poder clasificarlos.

Para este tipo de información, se identificaron los siguientes componentes, los cuales fueron separados de acuerdo con la frecuencia con que aparecen.

Alto	**Bajo**		
lamina	enves	endocarpo	estipula
flor	ovario	tricoma	arbol
fruto	peciolo	axila	liana
inflorescencia	bractea	apice	corola
hoja	estilo	lobulo	epidermis
planta	nervio	ramita	nervadura
estambre	arbusto	sepalo	perianto
petalo	margen	costilla	espina
	base	hierba	pistilodio

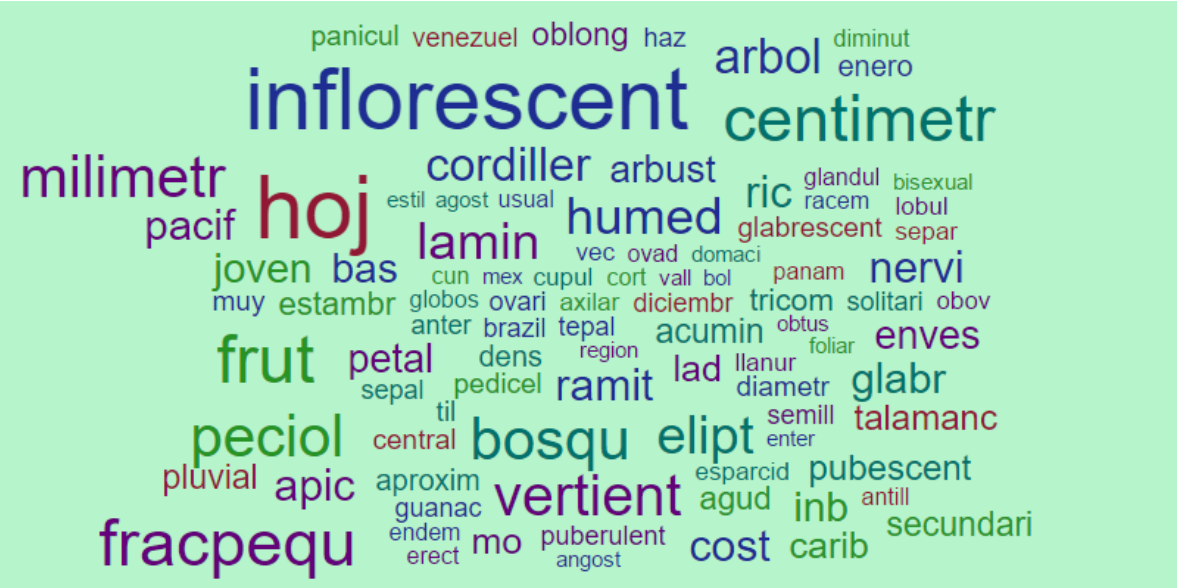
La importancia de los diferentes términos se puede visualizar mejor usando una nube de palabras:

LISTA ADICIONAL

ANÁLISIS DE LOS TÉRMINOS MÁS DISCRIMINANTES PARA CLASIFICAR LOS COMPONENTES DE INFORMACIÓN BIOLÓGICA

Tomando como base el conjunto de descripciones, distribuciones y claves extraídas de Flora de Costa Rica, volumen 6, realizó un análisis usando la métrica de Ganancia de Información [citar libro Baeza] con el fin de determinar cuáles son los términos más relevantes a la hora de distinguir entre esos tres grupos. La siguiente tabla muestra en orden los mejores 100 términos que se pueden usar para distinguir entre descripciones morfológicas, distribuciones y claves dicotómicas. También se incluye una nube de palabras para hacer más explícita la importancia relativa de cada uno de los términos.

hoj	glabr	oblong	venezuel	cort
inflorescent	pacif	estambr	guanac	globos
frut	inb	agud	tepal	cun
centimetr	bas	tricom	ovari	antill
#fraccionpequeña	joven	dens	haz	erect
bosqu	enves	enero	solitari	racem
peciol	#10-99	s	brazil	usual
milimetr	petal	aproxim	o	endem
vertient	arbust	til	vec	angost
elipt	mo	diametr	esparcid	domaci
lamin	lad	pedicel	cupul	obtus
humed	carib	muy	separ	foliar
arbol	talamanc	glabrescent	llanur	region
#1000-9999	#1-9	panicul	axilar	bisexual
cordiller	pubescent	central	obov	mex
apic	acumin	semill	lobul	vall
cost	#100-999	diciembr	glandul	agost
ramit	secundari	sepal	ovad	estil
ric	n	puberulent	diminut	bol
nervi	pluvial	anter	panam	enter


panicul venezuel oblong haz arbol diminut enero
inflorescent centimetr
milimetr cordiller arbust ric glandul racem bisexual lobul
hoj estil agost usual humed glabrescent separ
pacif lamin vec ovad domaci
joven bas cun mex cupul cort vall bol panam nervi
muy estambr globos ovari axilar diciembr tricom solitari obov
anter brazil tepal acumin obtus foliar enves
frut petal dens region lad llanur diametr glabr
sepal pedicel ramit
til semill talamanc
peciol central bosqu elipt enter
pluvial apic aproxim esparcid pubescent
guanac vertient agud inb antill
fracpequ endem mo puberulent cost carib secundari
erect angost

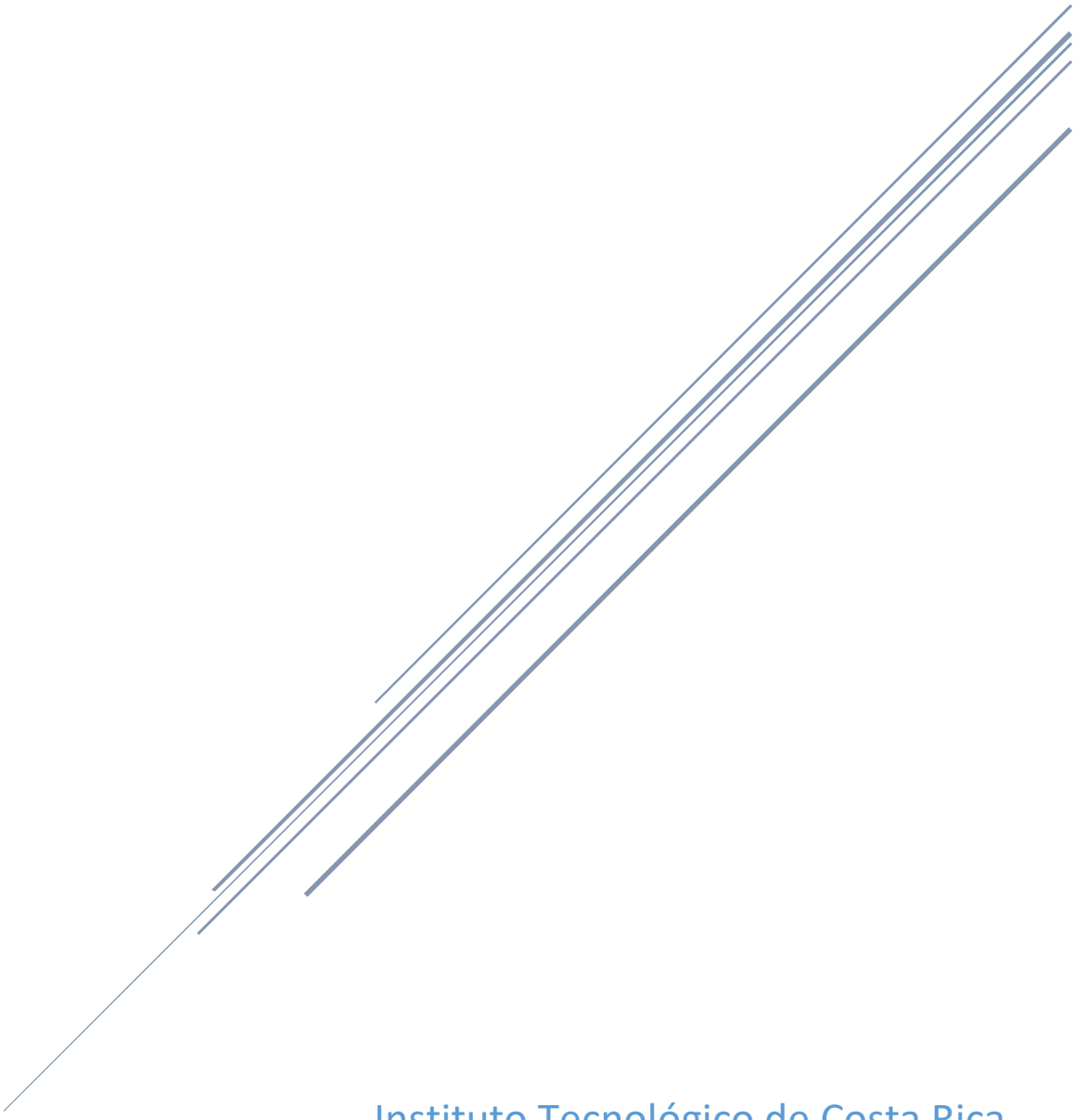
8.3 Herramientas desarrolladas

En este apéndice se describen con más detalle dos de las herramientas desarrolladas:

- ***Text Markup Tool***, herramienta de selección y clasificación de fragmentos de texto.
- ***Procesador de texto del Manual de Plantas***, herramiento de pre-procesamiento.

MANUAL DE USUARIO

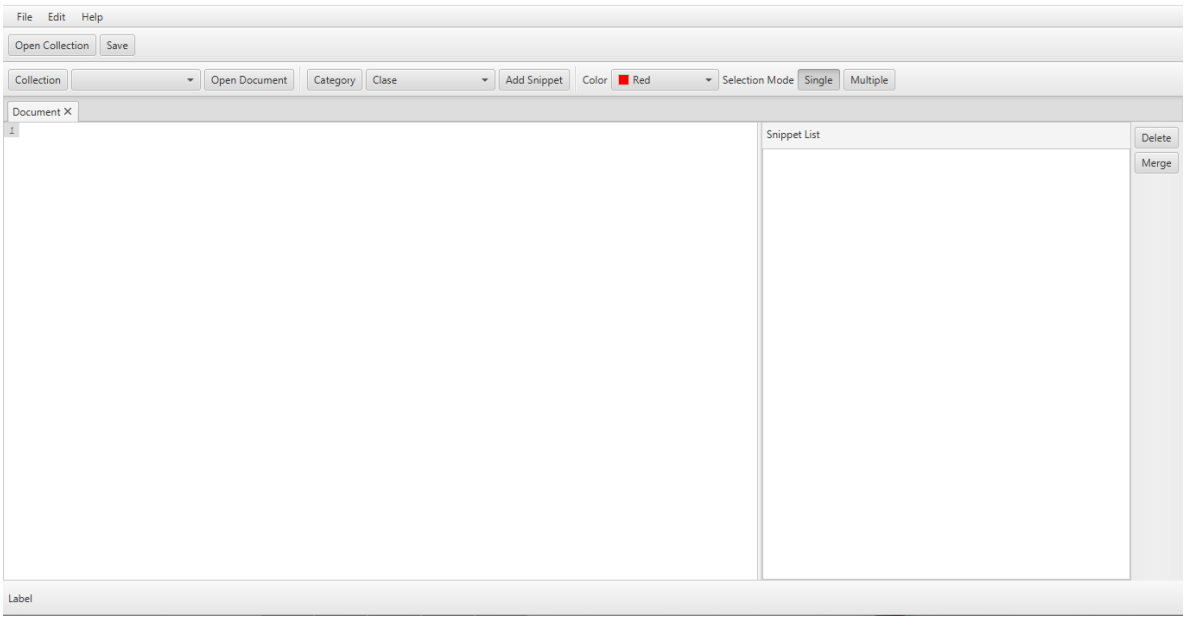
Text Markup Tool



Instituto Tecnológico de Costa Rica
Centro de Investigaciones en Computación

Introducción

Aplicación



La aplicación principal consta de dos barras de barras de herramientas:

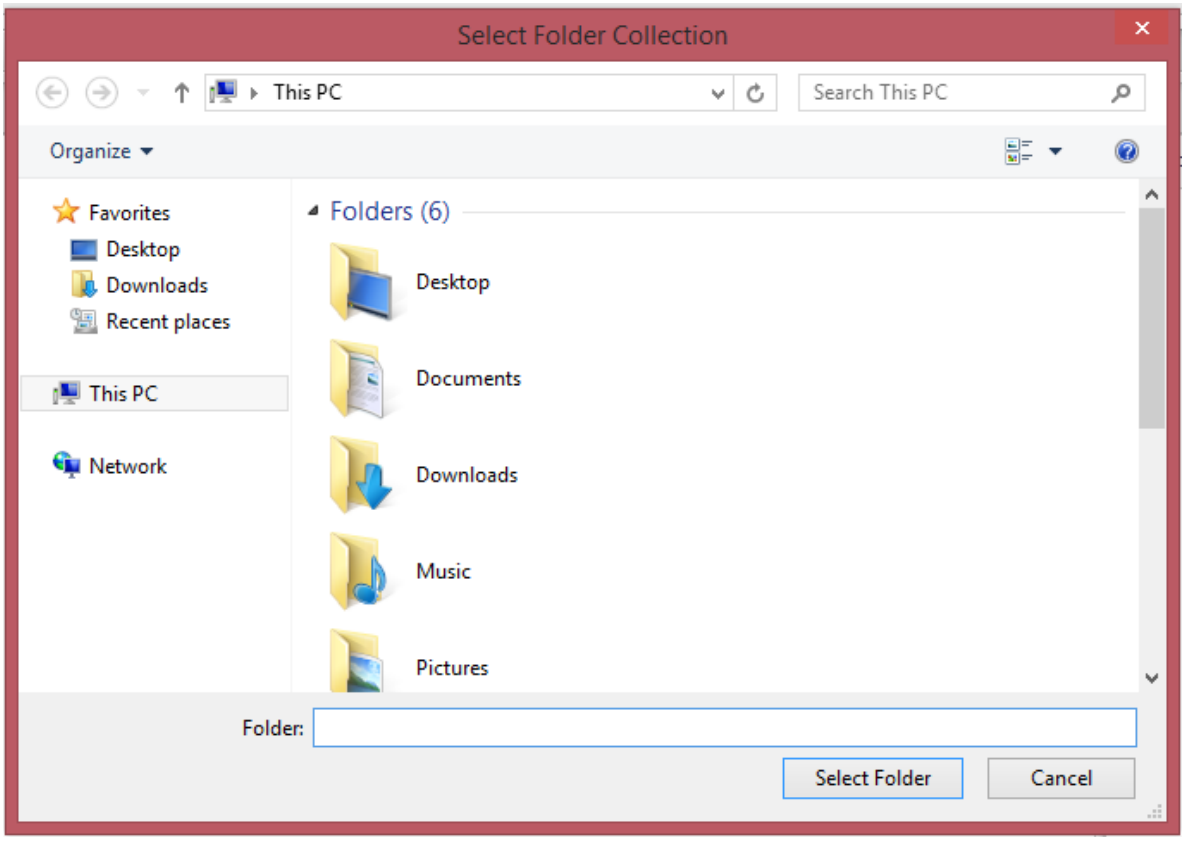
1. La primera barra ofrece las funcionalidades de abrir una colección nueva y de salvar el documento.
2. La segunda barra tiene las funcionalidades de abrir un documento, mantenimiento de categorías, agregar trozos de texto, seleccionar el color de selección de texto y seleccionar el modo de selección.

Además la aplicación tiene un menú principal el cual ofrece las funciones de exportar los trozos guardados en la base de datos en un XML, cambiar la configuración de la aplicación y cerrar la aplicación.

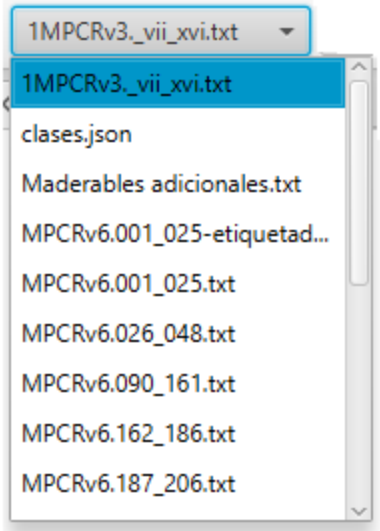
Abrir una colección nueva

Para abrir una colección nueva es necesario seleccionar el botón 

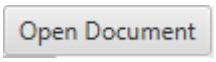
el cual abrirá una ventana para seleccionar el directorio deseado.



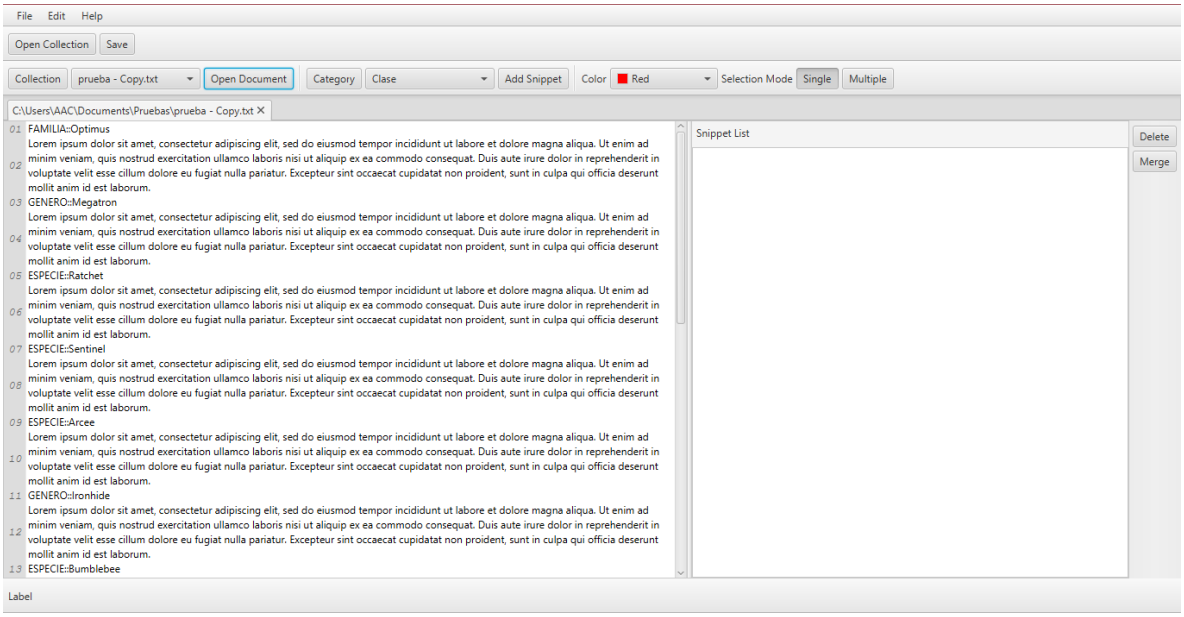
Luego de seleccionar el directorio, se cargaran todos los documentos en el combo box de documentos como se muestra en la siguiente imagen:



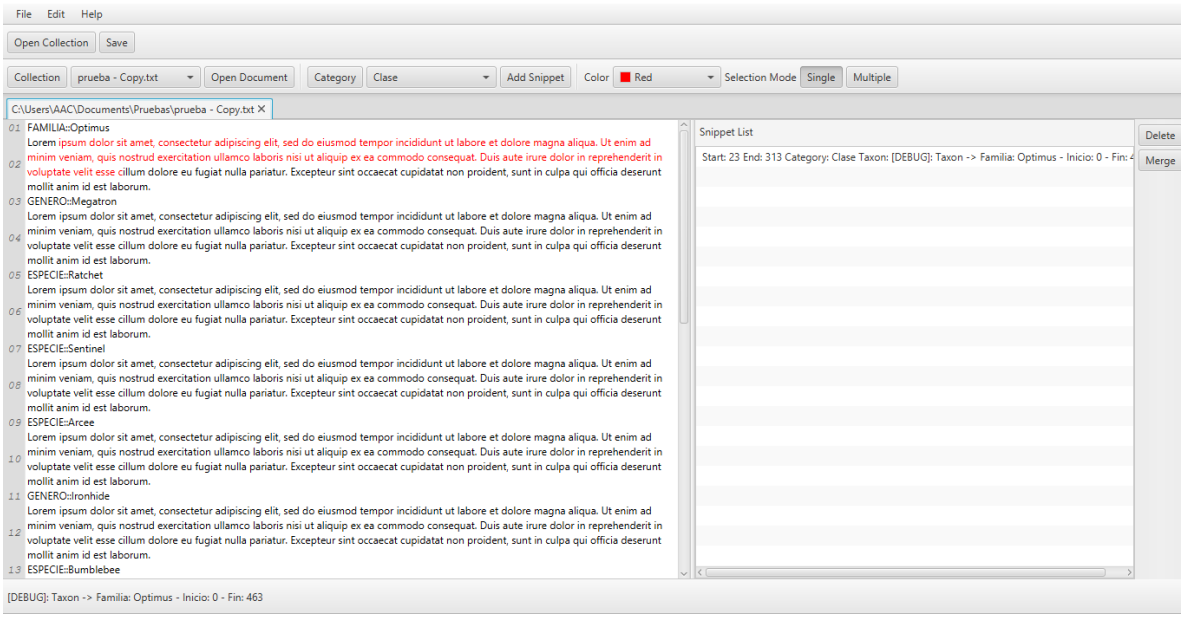
Abrir documento

Para abrir un documento, solo se necesita haber seleccionado algún documento del combo box, luego seleccionar el botón 

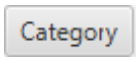
el cual abrir el documento y mostrará el contenido del documento en la aplicación:

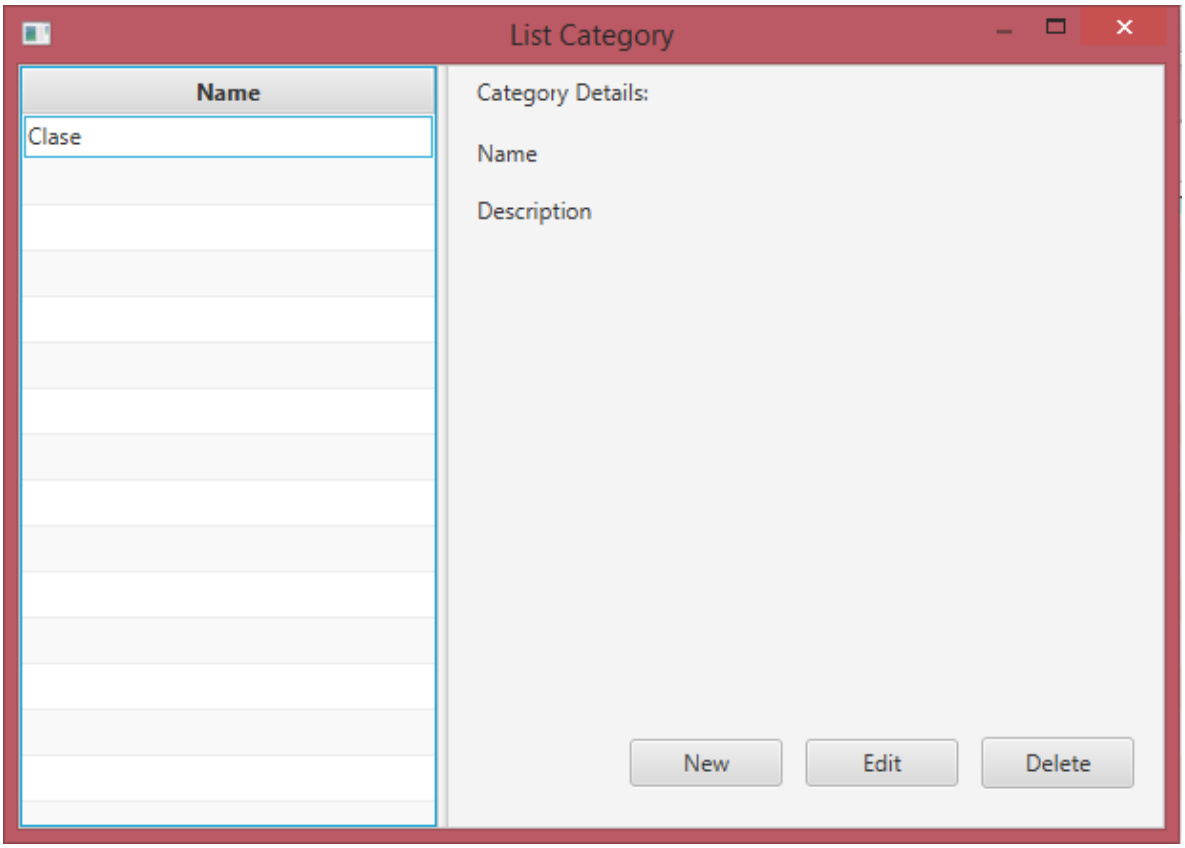


Nota: Si el documento ya tiene trozos de texto ya guardados en la base de datos, se cargarán en la lista y también saldrán marcados en el texto:



Mantenimiento de categorías

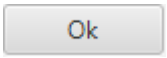
Para ejecutar esta funcionalidad primero hay que seleccionar el botón  y se abrirá el mantenimiento de categorías:



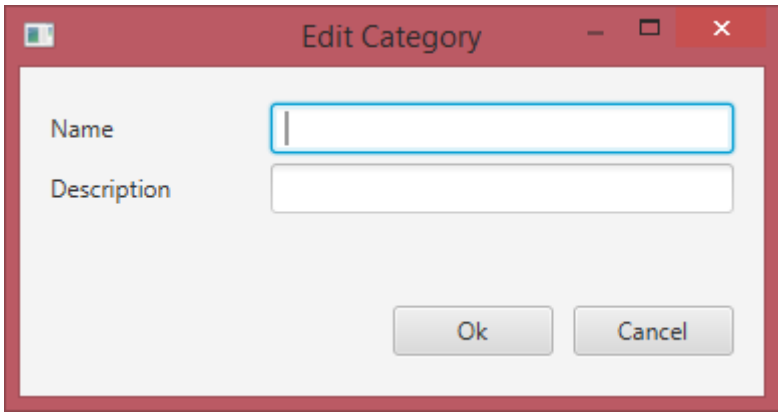
Donde saldrán las opciones para crear, editar y borrar una categoría.

Crear nueva categoría

Para crear una nueva categoría solo hay que seleccionar el botón 

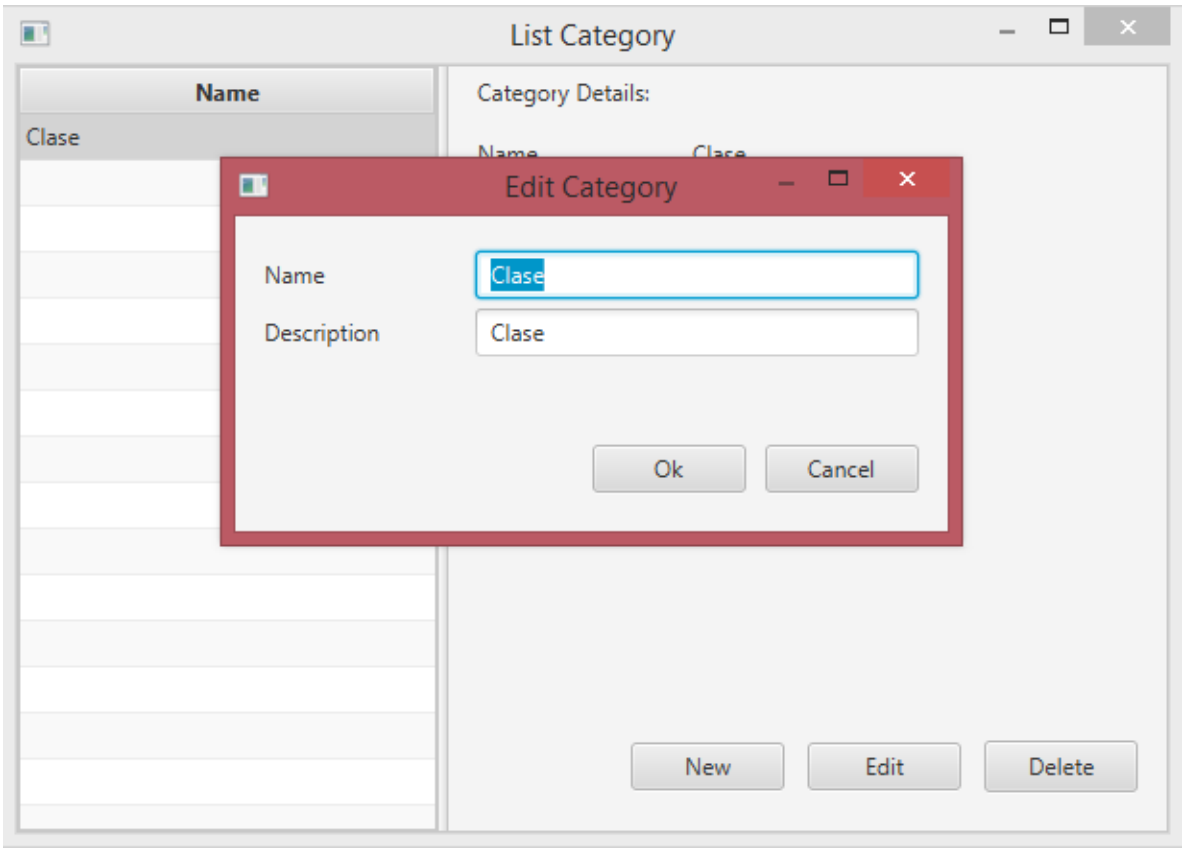
y llenar el pequeño formulario, luego seleccionar el botón 

y la categoría se almacena en la base de datos:



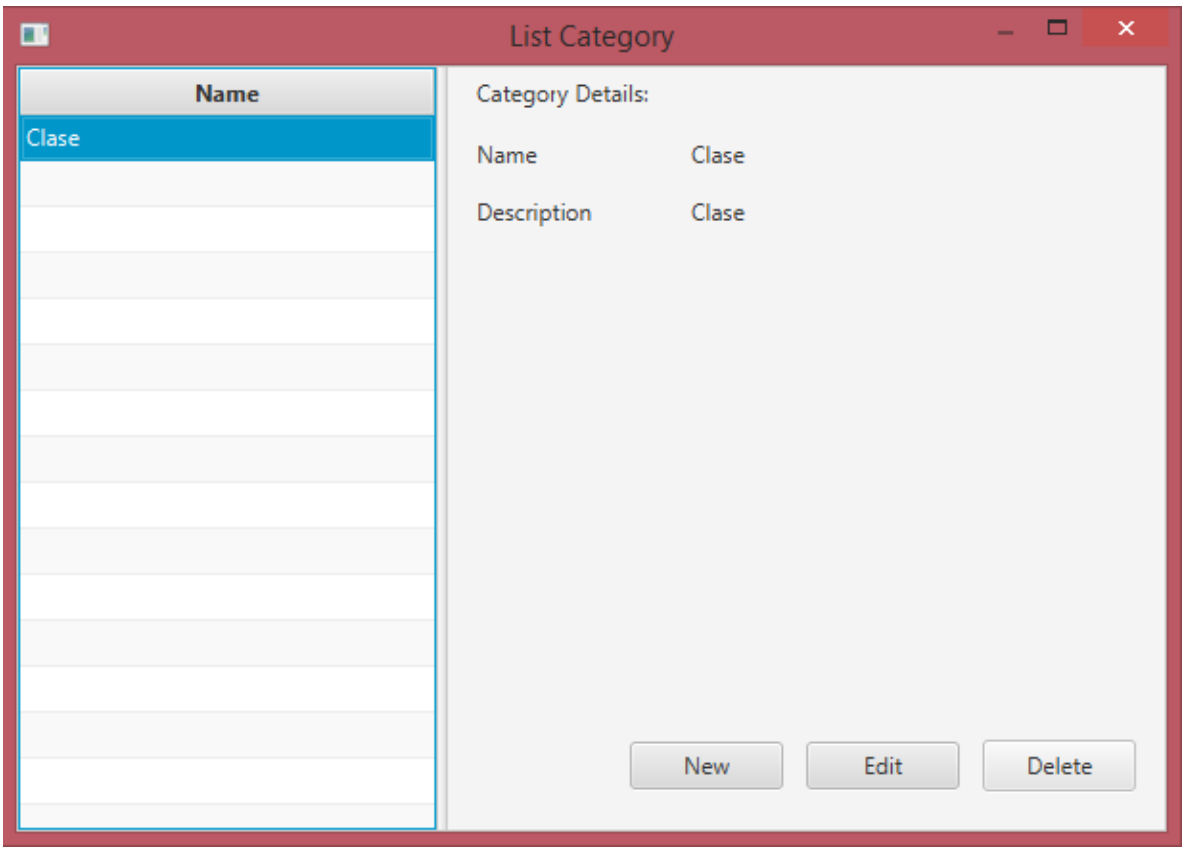
Editar categoría

Para editar una categoría solo es necesario seleccionar una categoría de la lista que aparece en al lado derecho y seleccionar el botón Edit. Luego aparecerá el formulario con los datos de la categoría a editar:



Eliminar categoría

Para eliminar una categoría solo tiene que seleccionar una categoría de la lista derecha y seleccionar el botón Delete y la categoría se eliminará de la base de datos:

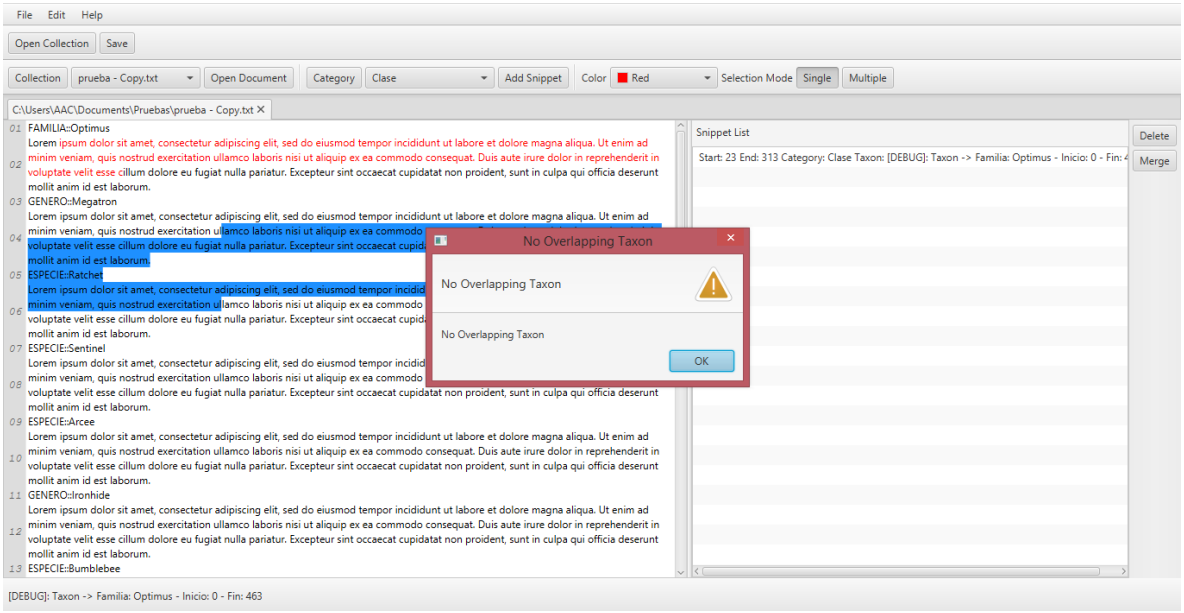


Agregar un snippet (trozo de texto seleccionado)

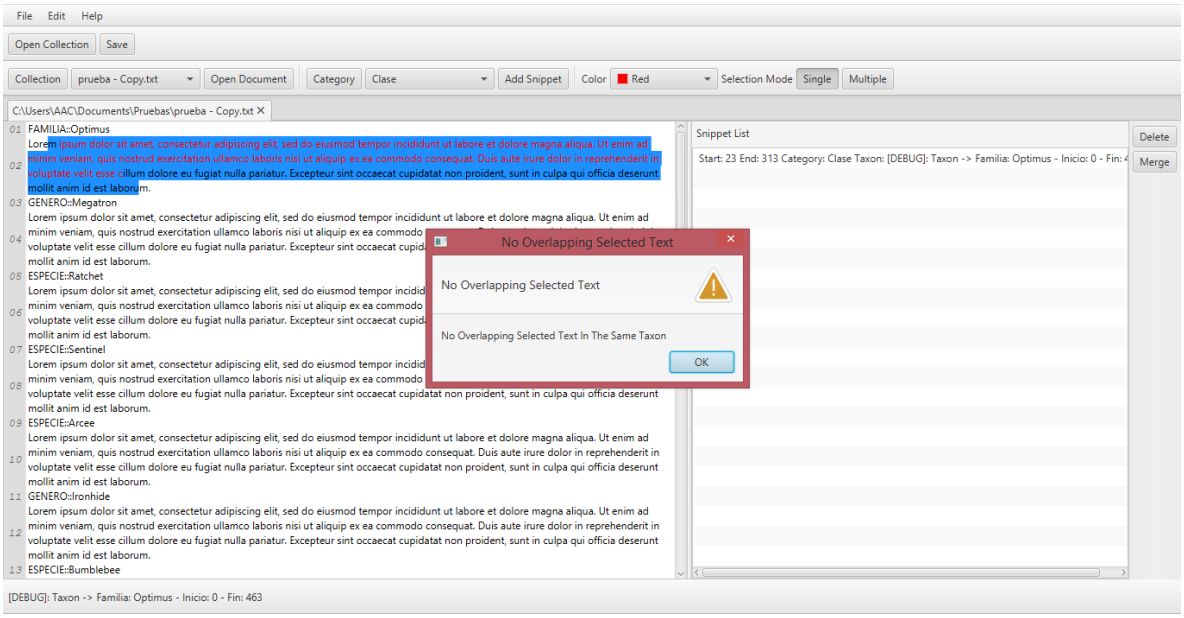
Para agregar un snippet solo hay que marcar el texto seleccionado que deseas y seleccionar una categoría, luego seleccionar el botón Add Snippet.

En el momento de marcar el texto existen dos restricciones:

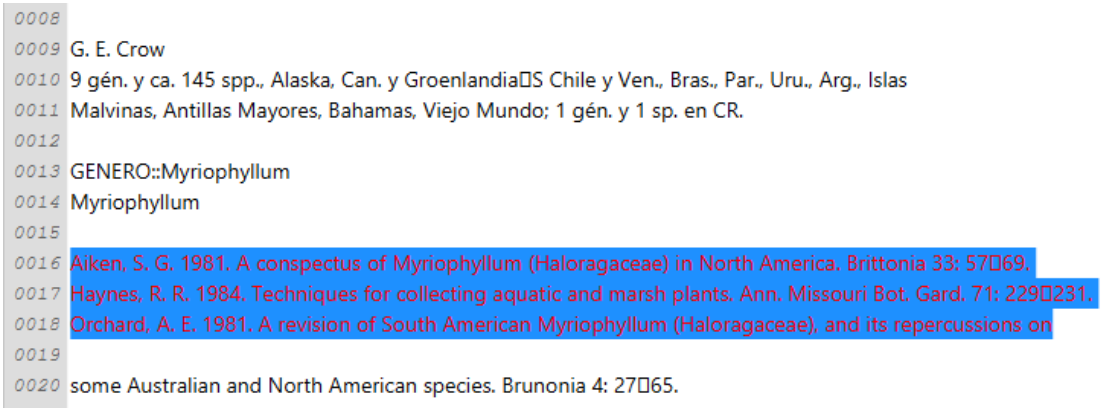
1. No se puede seleccionar texto de otro taxón.



2. No se puede traslapar texto seleccionado de los snippets ya agregados.



Nota: Hay documentos donde los párrafos están separados por saltos de línea. Para facilitar la selección de estos párrafos se implementó una funcionalidad que permite seleccionar estos párrafos mediante la combinación de la tecla Ctrl + Clic Principal (clic izquierdo). La selección del párrafo tiene como límite un salto de línea en blanco como se muestra en la siguiente imagen:



Eliminar un snippet

Para eliminar un snippet solo se ocupa seleccionar un snippet de la lista y luego seleccionar el botón Delete. Cuando un snippet se elimina, se quita de la lista de snippets y también se eliminar el color en el texto marcado. Se pueden seleccionar varios snippets de la lista con la combinación de Ctrl o Shift + Clic y se borran todos los snippets seleccionados.

Combinar dos snippets

Para combinar dos snippets, solo se necesitan seleccionar dos snippets de la lista, y luego seleccionar el botón Merge. Para combinar dos snippets existen algunas restricciones:

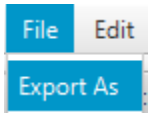
1. Los snippets tienen que estar continuos, es decir no debe existir otro snippet en medio de los que quiere combinar.
2. Ambos snippets tienen que tener el mismo taxón y la misma categoría.

Salvar documento

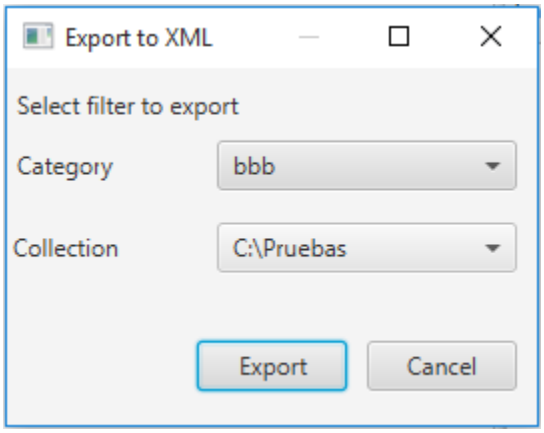
Luego de agregar todos los snippets deseados, puede guardar el documento en la base de datos. Para hacer estos solo es necesario seleccionar el botón Save y aceptar el cuadro de confirmación. La siguiente vez que abra el mismo documento se cargarán en la lista los snippets guardados.

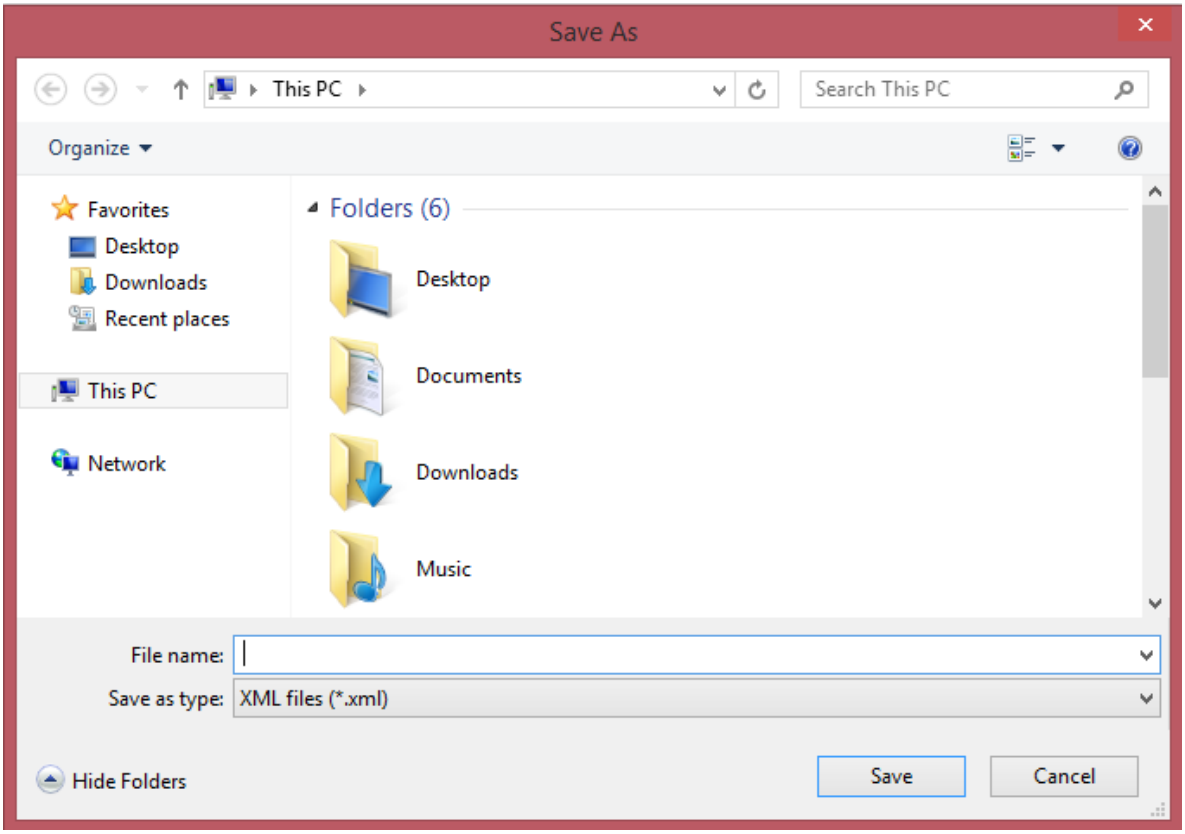
Exportar snippets

Para exportar todos los snippets guardados en la base de datos en un XML solo se ocupa ir File > Export as como lo muestra la siguiente imagen:



Luego saldrá una ventana donde se pueden seleccionar dos filtros: el primero es la categoría y el segundo filtro es la colección. Luego se selecciona el botón  y se selecciona la ruta donde se generará el XML y escribir el nombre del archivo. El XML se generará con todos los snippets que pertenezcan a la categoría y colección seleccionados:



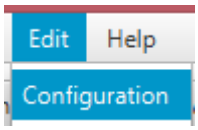


La estructura del XML generado lo muestra la siguiente imagen:

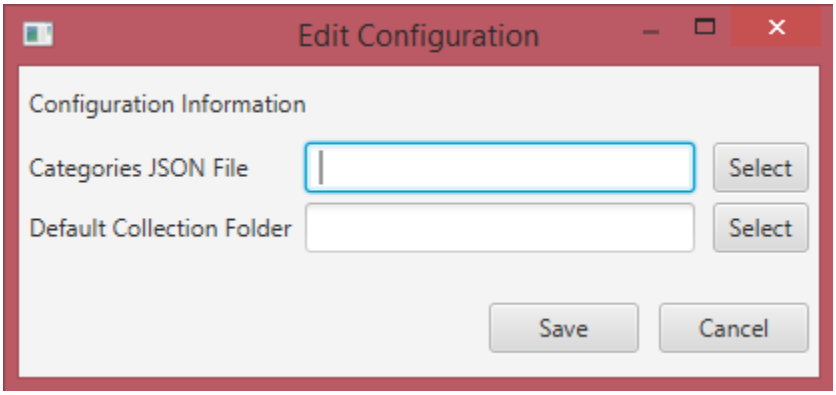
```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<snippets>
    <snippet>
        <category>
            <name>Clase</name>
        </category>
        <taxon>
            <family>Haloragaceae</family>
            <genre>Myriophyllum</genre>
            <species></species>
        </taxon>
        <text>Aiken, S. G. 1981. A conspectus of Myriophyllum (Haloragaceae) in North America. Brittonia 33: 5769.
Haynes, R. R. 1984. Techniques for collecting aquatic and marsh plants. Ann. Missouri Bot. Gard. 71: 229231.
Orchard, A. E. 1981. A revision of South American Myriophyllum (Haloragaceae), and its repercussions on</text>
    </snippet>
</snippets>
```

Editar configuración

Para editar la configuración de la aplicación solo hay que ir a Edit > Configuration



Luego aparecerá una pantalla con las configuraciones que desea editar:



Procesador de texto del Manual de Plantas de Costa Rica

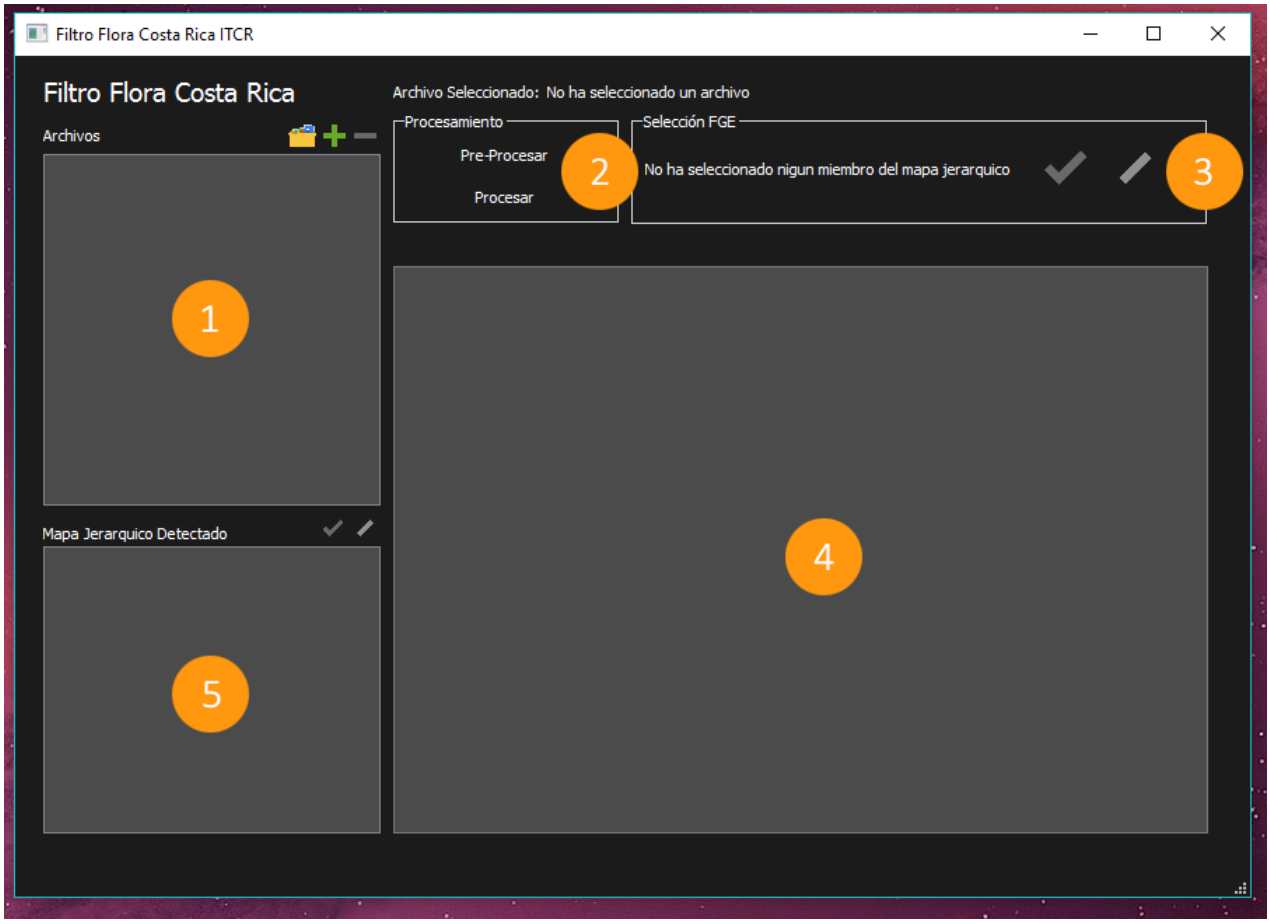
Alejandro Rojas

Diciembre 2016

1 Introducción

En este documento se explica el funcionamiento del programa Filtro de Flora de Costa Rica desarrollado para procesar archivos de texto generados del Manual de Plantas de Costa Rica. El programa se encarga de leer los archivos, filtrar la información no relevante como los pies de pagina, números de página, encabezados, pies de imágenes. También analiza la estructuración del documento buscando jerarquicamente los taxones que en el se describen, generando una posible lista de taxones para ser revisados posteriormente por el usuario. El programa también analiza los claves dicótomas que pueden encontrarse desordenadas en el texto.

2 Interfaz del programa



- **1. Lista de archivos a procesar:** Aquí se ubican los archivos a procesar. El icono de la carpeta que ubica arriba permite agregar todos los archivos de texto que se ubican en una misma carpeta. El símbolo de "+" permite agregar un solo archivo a la lista de archivos.

- **2. Operaciones principales:** El procesamiento se da en 2 etapas, pre-procesamiento y procesamiento. En la primera, lo que se hace es ubicar las páginas, borrar encabezados, y detectar los posibles taxones. Se les denomina posibles porque el usuario debe verificar que sean correctos. En la segunda etapa se asume que los taxones marcados como posibles son correctos. Por lo que se genera el archivo final en esta

etapa. En esta etapa también se borran los contenidos de los pies e imágenes.

- **3. Selección FGE:** Una vez procesado, el usuario puede escoger el taxón para ubicarlo en el texto y determinar así si fue correctamente marcado o no. Aquí se mostrará el taxón escogido. El check y la barra inclinada permiten decirle al programa que el taxón fue correctamente determinado, en caso contrario se marca la barra la barra inclinada. Note que los taxones automáticamente se consideran como ciertos, por lo q solo debe marcar los que están incorrectos en primera instancia.

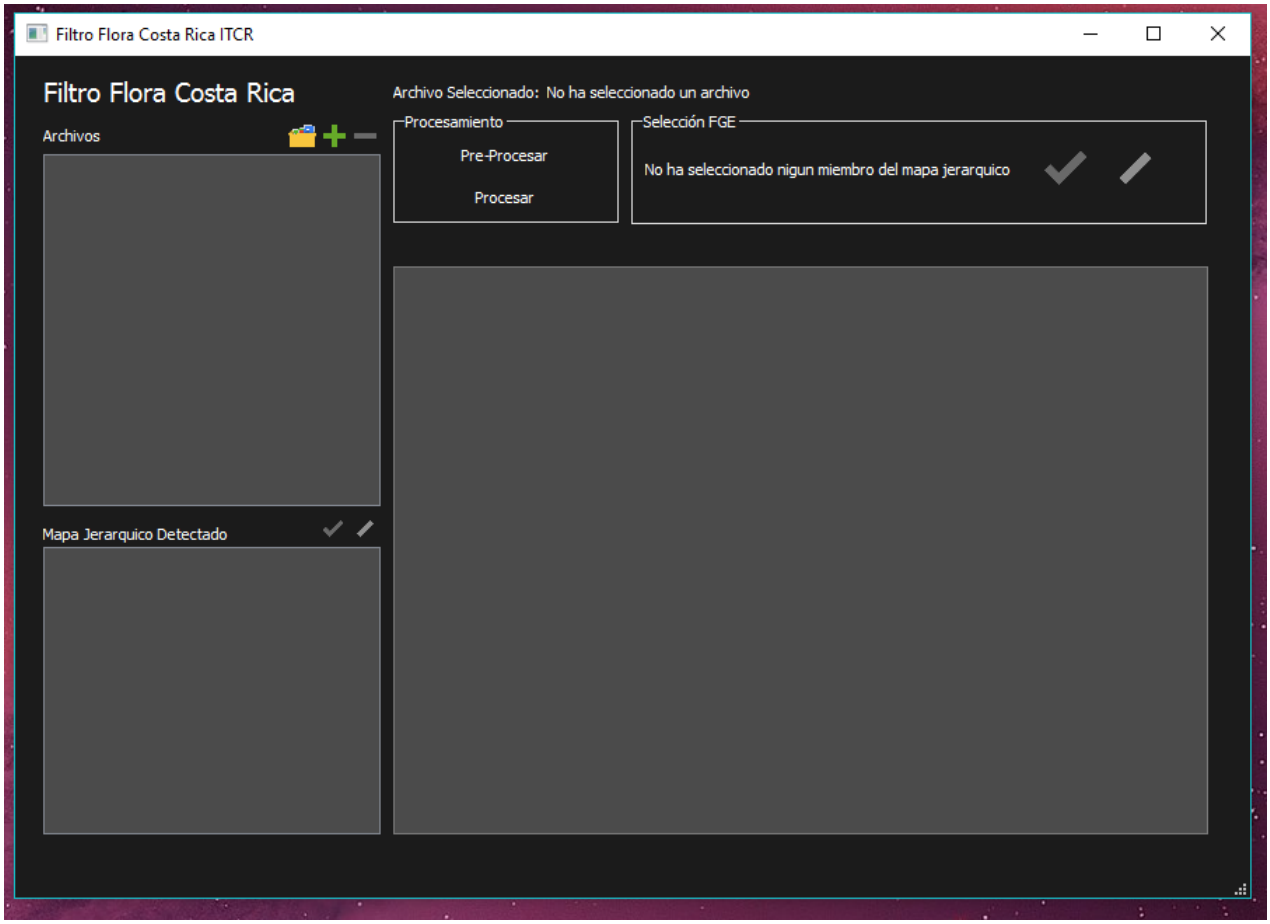
- **4. Visor de Texto:** Muestra el texto correspondiente a cada de etapa de análisis de taxones o archivo final.

- **5. Mapa Jerárquico detectado:** Aquí se muestran todos los taxones que se lograron recabar del texto, al hacer doble clic, el visor de texto ubicará en el texto, el taxón seleccionado. Así se facilita el análisis del taxón.

3 Funcionamiento del Programa

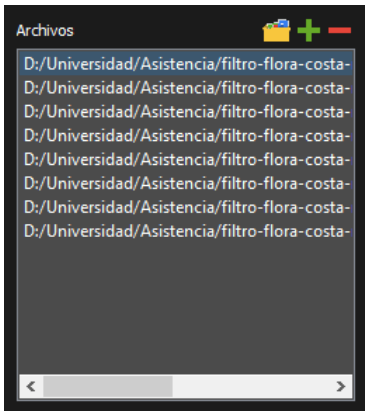
3.1 Cargar Archivos

Para comenzar a usar el programa, debe abrirse el ejecutable Filtro Flora de Costa Rica.exe.



Se empieza ubicando el archivo .txt que se desea procesar. Se asume que el archivo de texto (.txt) se generó con algún visor de PDF, preferiblemente Adobe Reader (Ver para el procedimiento). El programa permite procesar varios archivos a la vez, ya sea que se agreguen uno por uno o todos los archivos de texto de una determinada carpeta.

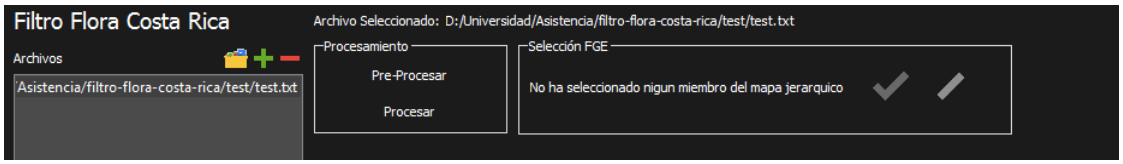
 Haciendo clic en el icono de la carpeta, se puede ubicar la carpeta deseada para obtener los archivos de texto. De otro modo, en el símbolo de "+" se puede agregar archivos individualmente. Una ves incluidos los archivos que se desean procesar, el programa los procesará uno a uno, empezando de arriba hacia abajo. se observará como en la imagen de de abajo.



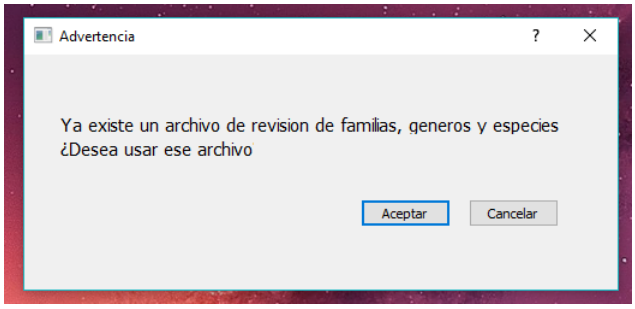
Para eliminar un archivo, basta dar doble clic sobre el archivo que no se desea procesar, veremos como se habilita el botón de guión en rojo 

3.2 Preprocesamiento

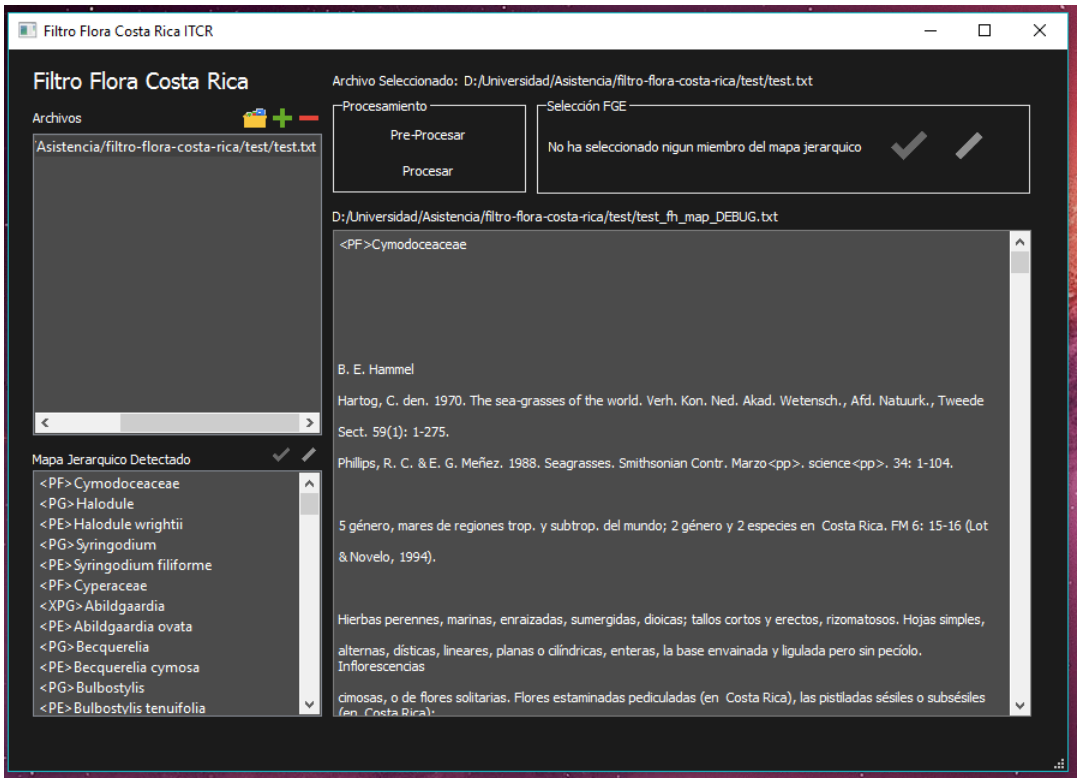
Una ves seleccionado el archivo que queremos procesar (Seleccionándolo en la lista de archivos) veremos que ahora, en la parte superior del programa muestra el nombre del archivo seleccionado.



Habiendo seleccionado el archivo, se habilita el botón "Pre-procesar", que nos permite realizar dicha acción sobre el archivo. Una ventana como la que se observa abajo puede emerger. En síntesis, esta ventana es para cuando ya tenemos un archivo de revisión de taxones de algún procesamiento anterior y no se desea volver a revisar desde el principio otra vez.

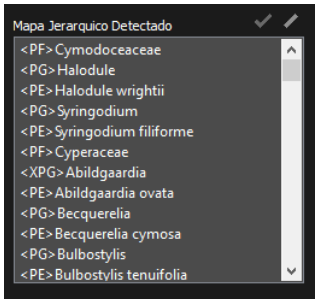


Si queremos cargar el archivo que la ventana nos indica, se da clic en *aceptar*, de lo contrario *cancelar* creará un archivo nuevo y sobrescribirá el existente. Una vez procesado el archivo, veremos que el programa llena la lista de taxones y el Texto leído en el recuadro de la derecha. Este texto se utiliza para ubicar el taxón escogido dentro del texto.



La parte más importante del pre procesamiento es la identificación de los

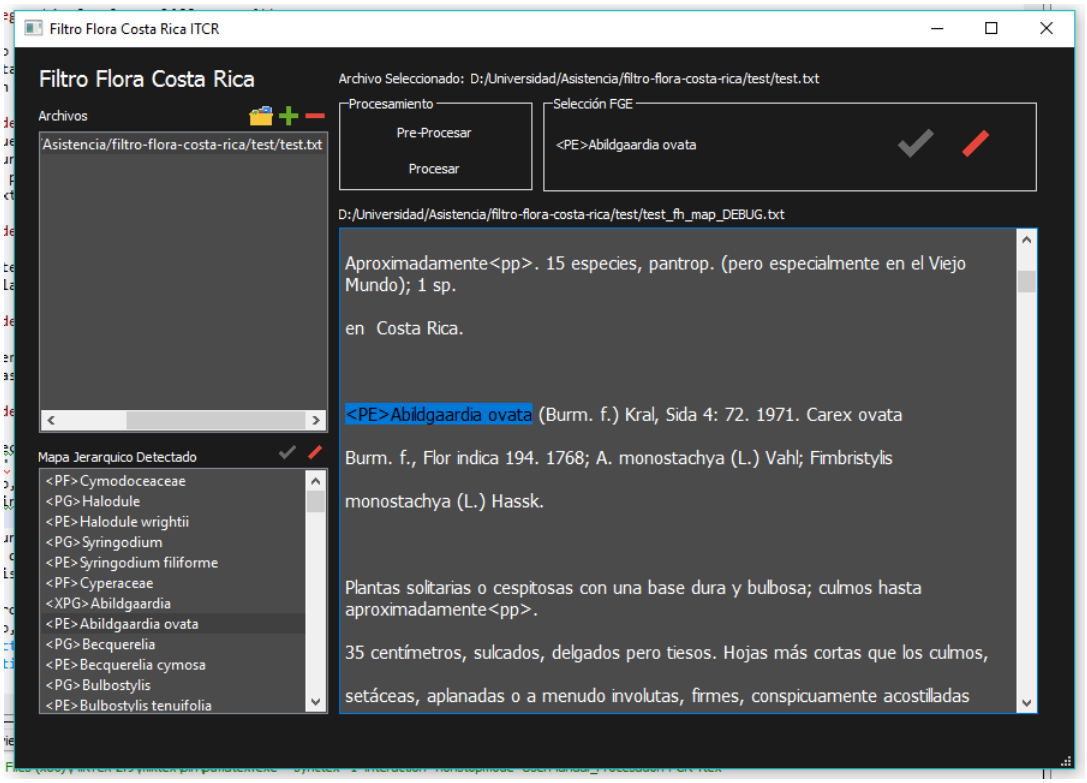
taxones erróneos. Para eso podemos hacer doble clic sobre la lista de taxones, en un taxón especifico.



Se observa que ocurren dos cosas, la primera es que se hablita el "Slash" que aparece justo arriba del recuadro de taxones, y otro mas arriba, en el recuadro de "Selección FGE".

Este recuadro de arriba indica el taxón que tenemos seleccionado, y el "slash" de color rojo, hace la misma acción que el "slash" que se mencionó anteriormente. Que es básicamente descartar ese taxón como posible, anteriormente se mencionó que por defecto, los taxones se reconocen como correctos, por lo que solo se puede marcar como "incorrectos". Habiéndolos encasillado como "incorrectos", podemos notar que el "check" se habilita, esto ocurre para volverlos al estado anterior de "correcto".

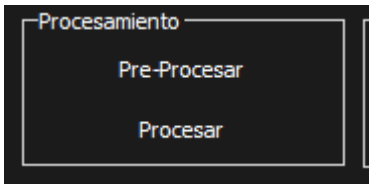
La segunda actividad que ocurre, es que el recuadro de la derecha, el que nos muestra el texto procesado, se posiciona en la región del texto donde se logró identificar ese taxón. Esto para ayudar a la identificación de taxones erróneos mediante el análisis del contexto.



Este proceso debe hacerse con todos los taxones incorrectos, que se encuentran en la lista de taxones. Una vez realizado este proceso, se procede al procesamiento de texto.

3.3 Procesamiento

Esta parte del procesamiento se da una ves que se esta seguro que el archivo se revisó correctamente. Es mas sencillo que el pre procesamiento. Basta hacer clic en el botón de "procesar" que se encuentra en la parte superior del programa.



Seguido de esto nos mostrará el texto final. Nótese que ya no esta el archivo que se proceso en la lista de archivos, ahora solo se debe repetir este proceso en los otros archivos para completar el procesamiento de los archivos deseados.