Manual de uso de la herramienta de extracción de información en literatura

Esta herramienta está siendo diseñada por el equipo informático, en colaboración con los expertos en abejas nativas, plantas hospederas, e interacciones abejas nativas - plantas, también del proyecto.

Su objetivo principal es simplificar la extracción de la información provenientes de distintas fuentes bibliográficas (artículos científicos, documentos, tesis, etc.).

A continuación se describen los principales aspectos.

Identificación de taxones

Son paquetes de archivos generados automáticamente que corresponden a los nombres de taxones obtenidos de una referencia bibliográfica (e.g., artículo científico, tesis). Muestran ocurrencias individuales, globales y la señalización visual (*highlight*) dentro del documento PDF.

Se generan utilizando 2 fuentes de información:

1. El catálogo de especies de CONABIO. Para generar anotaciones acotadas a un subconjunto de especies.
2. El catálogo global de nombres de especies GNRD. Que entrega un conjunto complementario de nombres de especies no tan estricto y que <<puede>> ser de utilidad.

La descripción técnica de los procesos se encuentra en el siguiente enlace (aún en construcción).

¿Para que sirve?

Su propósito es reducir el tiempo que una persona invierte en analizar el contexto de un taxón mencionado en una referencia bibliográfica.

Sus objetivos son:

1. Identificar y señalar los taxones en el documento.
2. Mostrar si es posible hacer un mapeo hacia un catálogo.
3. Proveer documentos intermedios que reporten los nombres de taxones.
4. Proveer una interfaz sencilla para dar retroalimentación a la metodología diseñada.

¿Como se usa?

Hay 3 archivos resultantes que presentan los taxones obtenidos. A describir:

1. **ocurrences**

   Contiene las columnas que indican la posición dentro del documento (**page, ordinal)** de los matches **annotation-verbatim** y la fuente de la anotación (***ann_type***).

   Este archivo está destinado ayudar en la interpretación del documento anotado resultante (**ID_DOCUMENTO.pdf**)

Columnas e información que representan

annotation: Anotación resultante
verbatim: Texto original en el documento
page: Página en donde está la anotación (***en índice 0***)
ordinal: Párrafo de la página **`page`** en donde está la anotación
most_probable_parents: ***Si la anotación esta abreviada.*** Taxón superior más probable
all_parents: ***Si la anotación esta abreviada.*** Taxones superiores relacionados
es_util. Columna para indicar si la anotación obtenida es de utilidad.
parent_correcto: ***Si la anotación esta abreviada y la columna*** **most_probable_parents es incorrecta*.*** Columna para indicar el taxón superior relacionado.

Posición de las anotaciones.

(**page** y **ordinal**) indican la posición dentro del documento de una ocurrencia individual de una anotación (**annotation)** y el texto original **(verbatim)**, así como la fuente de la anotación (***ann_type***).

Para ocurrencias abreviadas, del tipo ***I. simulans***, además se muestran las columnas ***most_probable_parents***, que indican cual es el taxón superior más probable de acuerdo al análisis del texto y, ***all_parents***, que indican todos los taxones superiores que “podrían ser” los relacionados en el texto.

Ejemplo

Anotación: ***I. simulans***
Taxón más probable: ***leucophyllum simulans***
También podrían ser <<pero menos probables>>: ***lycieae simulans, lonicera simulans, litomegachile simulans, lantana simulans, …***

Tipo de las anotaciones

Pueden identificarse por la columna ***ann_type***, y corresponden a un color dentro de `**ID_DOCUMENTO.pdf**`.

0. Son anotaciones acotadas al catálogo. Son las más estrictas y, potencialmente , las más confiables. Señaladas en color **AZUL** en ID_DOCUMENTO.
2. Provienen de GNRD, se pueden considerar como un conjunto adicional de anotaciones de taxones que no están dentro del catálogo estricto. Sin embargo, son más fuzzies y la posición no es siempre la correcta (varían a lo máximo 1 o 2 ngramas de la posición actual). Señaladas en color **ROJO** en ID_DOCUMENTO.
3. Son las anotaciones menos confiables. Siempre son de 1-gram y también contienen palabras de vocabulario común. Véase <<sección>>. Señaladas en color **AMARILLO** en ID_DOCUMENTO

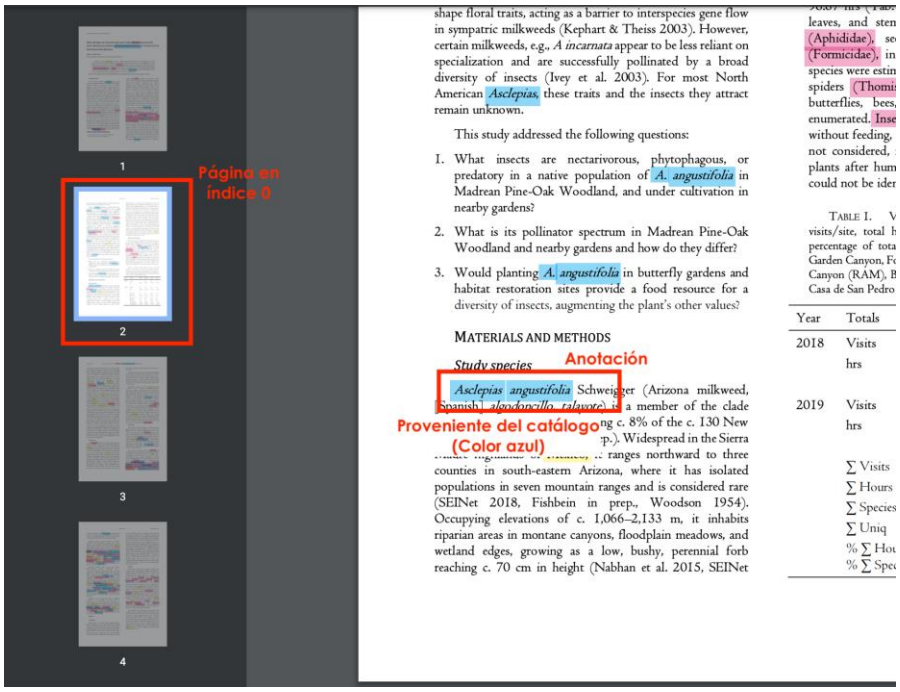
2. Documento

Es el documento anotado resultante, muestra las anotaciones y su tipo. Corresponde a los resultados en **taxa**.

Para identificar una anotación dentro del documento nos referimos a las columnas del documento de *ocurrences* ***annotation-verbatim, page, ordinal, ann_type,***

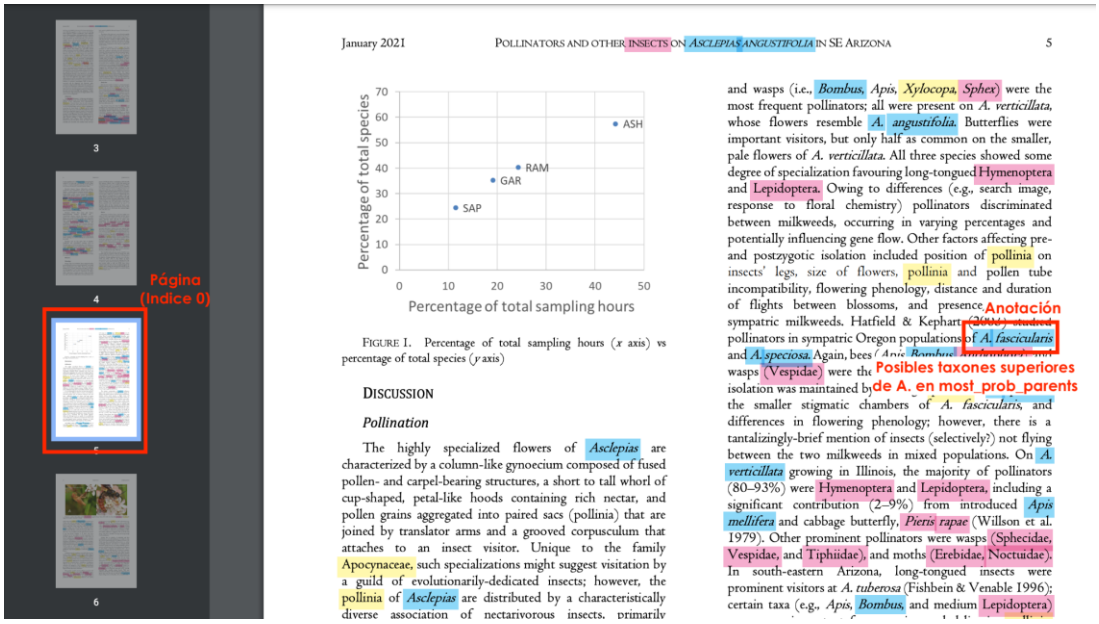
Ejemplo 1: Anotación proveniente del catálogo de CONABIO

annotation	asclepias angustifolia
verbatim	asclepias angustifolia
page	1
ordinal	9
ann_type	0
most_probable_parents	–
all_parents	–



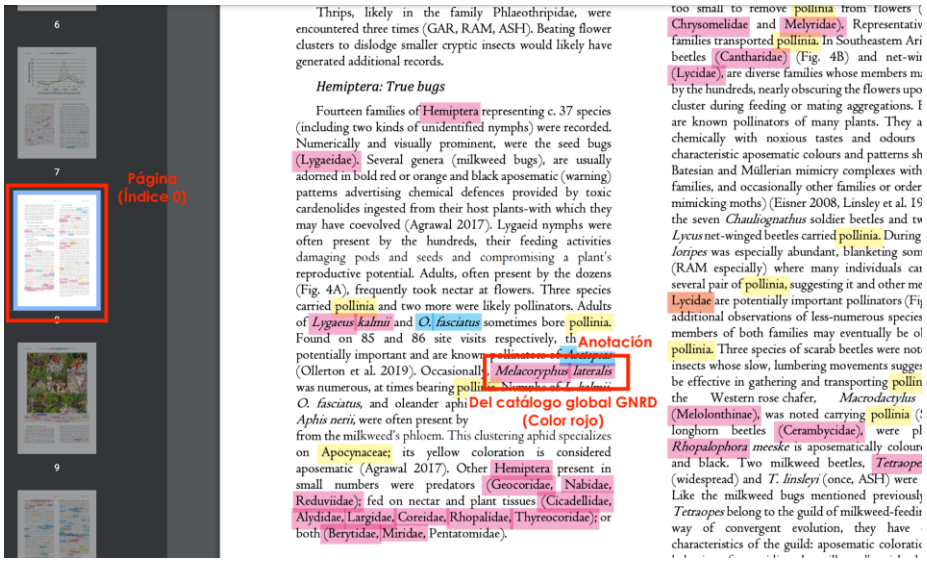
Ejemplo 2: Anotación abreviada del catálogo de CONABIO

annotation	a. fascicularis
verbatim	a. fascicularis
page	4
ordinal	24
ann_type	0
most_probable_parents	asclepias
all_parents	['augusta', 'apidae', 'anothea', 'augochlorini', 'apocynaceae', . . .]



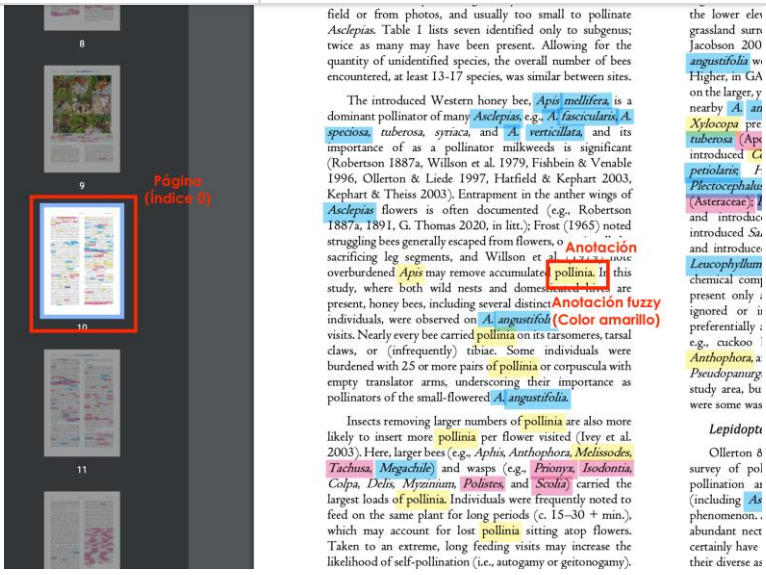
Ejemplo 3: Anotación del catálogo global GNRD

annotation	Melacoryphus lateralis
verbatim	Melacoryphus lateralis
page	7
ordinal	9
ann_type	2
most_probable_parents	–
all_parents	–



Ejemplo 4: Anotación fuzzy de ambos catálogos.

annotation	pollinia
verbatim	pollinia.
page	9
ordinal	2
ann_type	3
most_probable_parents	–
all_parents	–



3. **template**

   Este es un documento intermedio que será complementado por medio de las anotaciones del usuario y contiene las ocurrencias globales del documento. Tiene como fin construir un template de reporte como el mostrado.

   ¿Por que es un reporte intermedio? Si existiera la premisa de que todas las anotaciones entregadas son útiles y todos los parents son correctos, se pudiera implementar la construcción automática de este template. En la práctica este escenario todavía no es factible.

Premisas

Los dos procesos principales, marcar una especia y mapearla a un catálogo, parten de lo siguiente.

1. Para las anotaciones del catálogo de CONABIO. Las anotaciones pueden ser asociadas a un ID de categoría, es decir, para generarse previamente debieron existir y mapearse con el texto de la referencia.

¿Como se puede dar la retroalimentación?

Normalmente un proceso de aprendizaje automático necesita retroalimentación del usuario. Esto se obtiene con las columnas ***es_util*** y ***parent_correcto*** del documento ***ocurrences***.

- Si la anotación es de utilidad la columna ***es_util***, para ese renglón, se marca con 1.
- En caso de que la anotación provenga de un especie abreviada. Se sugerirá un taxon superior que sea lo más probable, de acuerdo a la metodología, en **most_probable_parents**. En caso de que esta sugerencia no sea correcta, se agregará el nombre del taxón que lo sea utilizando **all_parents.**