



# Bibliometría usando Biblioturtle

Aplicación a un tema de salud

por

Claudio Bustos

Financiado por Agencia Nacional de Investigación y Desarrollo / FONDECYT Regular N° 1201681



Biblioturtle

Adherencia al tratamiento y depresión

# Biblioturtle



Biblioturtle es el software de análisis bibliométrico creado en Python, que hemos desarrollado dentro del contexto de Fondecyt Regular N°1201681. Se creó en función de los siguientes principios:

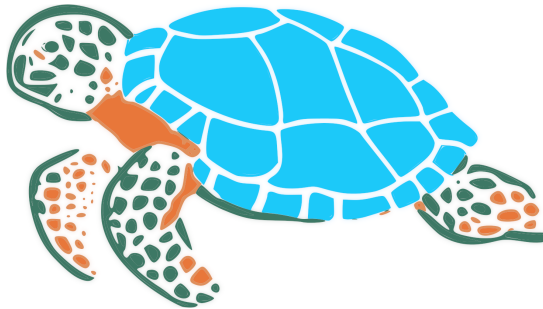
- ❖ Modularidad. Las salidas de cada parte del proceso pueden usarse en distintos módulos internos, en otros programas o incluso entre análisis.
- ❖ Integrar distintas bases de datos. Específicamente, ahora soporta Web of Science, Scopus y Scielo.
- ❖ Control estricto de inconsistencias y duplicaciones, tanto para autores/as, instituciones y artículos.
- ❖ Apoyo en el proceso de deduplicación, usando aprendizaje semi-supervisado. Estamos trabajando con miles de datos, no decenas o cientos como otros estudios.



El proceso de análisis bibliométrico usando Biblioturtle tiene 4 etapas bien definidas:

1. Recolección de información desde las bases de datos
2. Análisis por base: Primera estandarización de nombres de autores/as y filiaciones.
3. Deduplicación de artículos. Considera una segunda estandarización, cruzando información
4. Análisis de información conjunta

En todo momento se puede volver a una etapa previa para resolver inconsistencias



Logo provisional Biblioturtle

# Adherencia al tratamiento y depresión



Como ejemplo, se propone realizar un estudio bibliométrico de **depresión y adherencia a tratamiento**, usando WoS y Scopus con artículos entre los años 2018 y 2023.

La consulta en Scopus se descarga como **BibTex**:

```
TITLE ( treatment AND adherence AND depression ) AND PUBYEAR > 2017
```

La consulta en WOS se descarga como archivo plano de texto:

```
TI=( treatment AND adherence AND depression ) AND (FPY=2017 OR FPY=2018 OR FPY=2019 OR FPY=2020 OR FPY=2021 OR FPY=2022 OR FPY=2023)
```



## Etapa 2.1: estandarización de filiaciones



1. Se procesan los archivos de Wos y Scopus a un formato estándar interno.
2. Se detectan las afiliaciones que no se encuentran en el tesoro. Se sigue el siguiente proceso:
  - 2.1 Incorporo el nombre de la filiación en bruto dentro de la base de datos respectiva.
  - 2.2 Asigno el país usando un algoritmo basado en SVM, que tiene cerca de un 97 % de precisión.
  - 2.3 Estandarizo el nombre de la afiliación usando un algoritmo Naive Bayes. Opera con seguridad, pero no en todos los casos nuevos
  - 2.4 Se revisa manualmente las nuevas afiliaciones con un editor ad-hoc
  - 2.5 Se recrea el tesoro



1. Incorporo criterios de inclusión y exclusión de los artículos. En este caso, sólo se dejaron aquellos que están en las categorías de *artículos* y *revisiones*.
2. Se ejecuta un algoritmo estándar de detección de inconsistencias entre autores y de información de afiliaciones:
  - ❖ Número de autores en bruto vs número de autores estandarizados.
  - ❖ Se espera que todas las filiaciones tengan asignado un país, y estén incluidas en el tesoro.
  - ❖ Todos los autores deben tener género asignado.

## Etapa 3: unión de bases



1. Utilizando un algoritmo propio de deduplicación de artículos, se identifican duplicados dentro y entre las bases. Se almacena la información a comparar (citaciones)
2. Se verifica iterativamente que los autores y afiliaciones sean iguales entre artículos duplicados entre bases. En el caso de inconsistencias, el sistema propone posibles alternativas, las que se incorporan en los tesauros respectivos.
3. Si se detectan diferencias en el número de autores de un paper entre bases (error grave), se busca el artículo y se corrige a mano la información en la base que corresponda. Esto permite reutilizar los artículos entre análisis.



El sistema calcula descriptivos básicos por entidad (artículo, autor, afiliación y revista), y aplica el algoritmo de conglomerados de Liu et al (2014).

El sistema exporta diferentes archivos, que cuentan con la información necesaria para obtener los indicadores clásicos y realizar cruces de información:

1. Excel con información estándar: artículos, autores, afiliaciones, revistas
2. Excel con información de conglomerados: conglomerados, autores, artículos.

Liu, R., Feng, S., Shi, R., & Guo, W. (2014). Weighted Graph Clustering for Community Detection of Large Social Networks. *Procedia Computer Science*, 31, 85-94. <https://doi.org/10.1016/j.procs.2014.05.248>