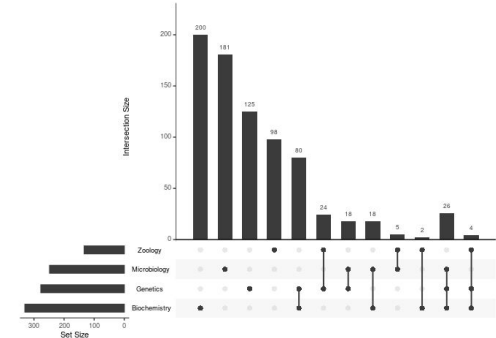
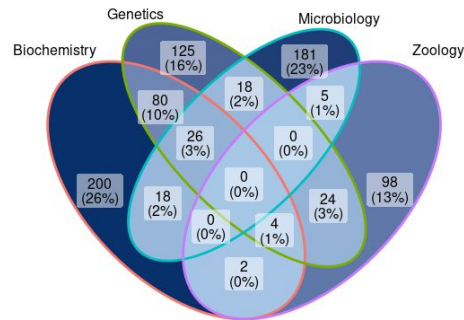
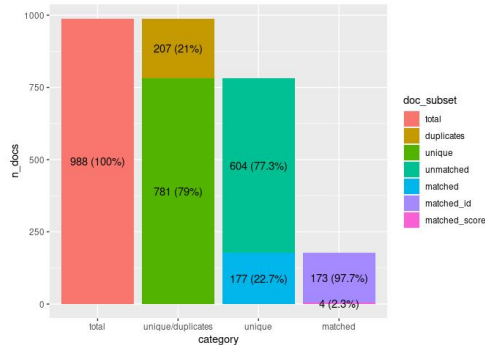


Workshop: Co-ocorrência de documentos em conjuntos de dados bibliográficos com o pacote R *biblioverlap*

Gabriel Alves Vieira
Orientadora: Prof^a Dr^a Jacqueline Leta



Introdução - Fontes de informação científica

- Primárias:

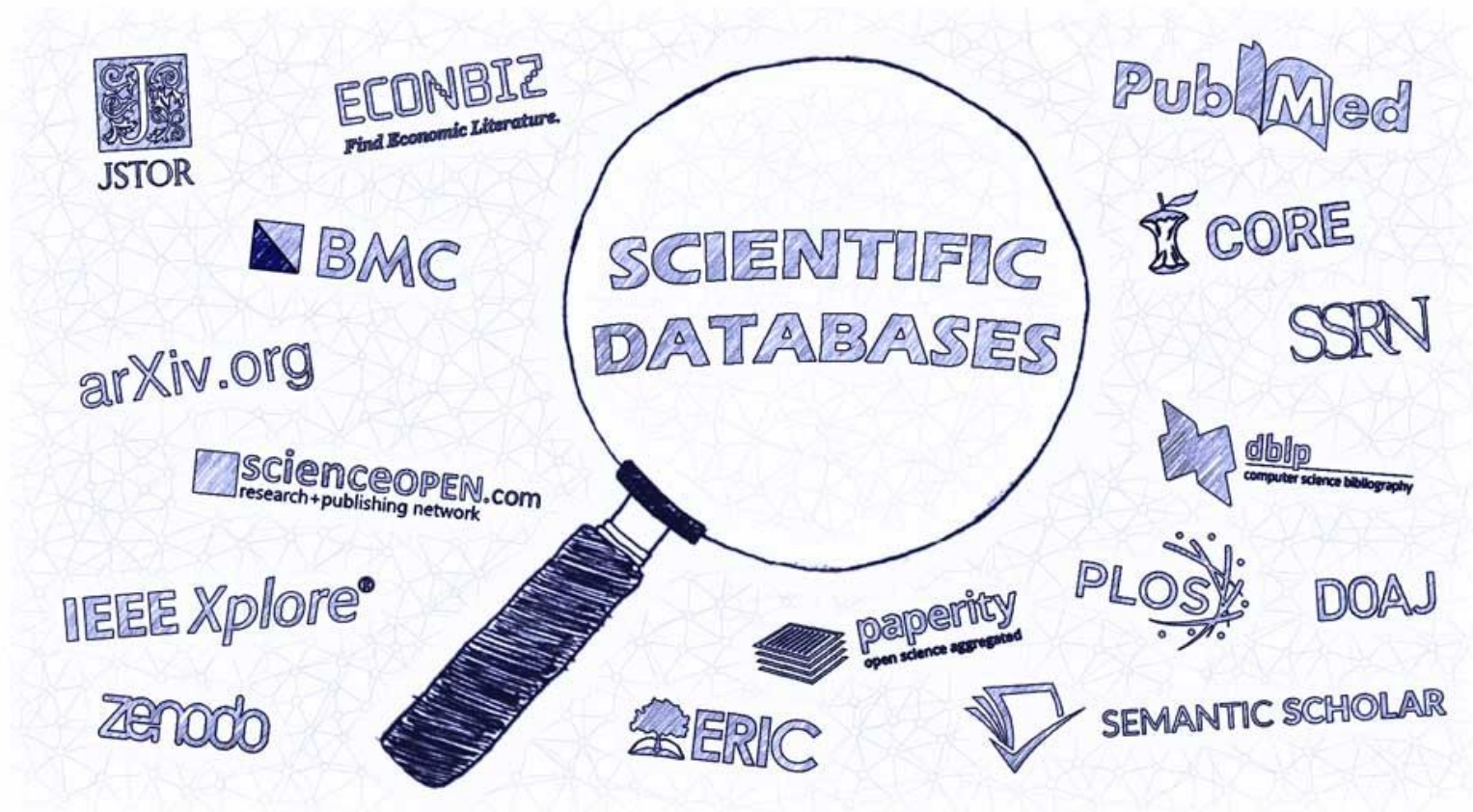
- Informação nova/original
- e.g. Publicações em periódicos

- Secundárias:


- Informação fundamentada sobre documentos primários
- Informação sobre informação
- Facilitar o acesso à produção científica
- e.g. Revisões e **bases de dados bibliográficos**
- Bases -> Fomentam estudos



Grande diversidade de bases -> Estudos comparativos



The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis

Vivek Kumar Singh¹  · Prashasti Singh¹ · Mousumi Karmakar¹ · Jacqueline Leta² · Philipp Mayr³

Received: 26 September 2020 / Accepted: 9 March 2021
© Akadémiai Kiadó, Budapest, Hungary 2021

Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic

Martijn Visser¹  · Nees Jan van Eck¹  · and Ludo Waltman¹ 

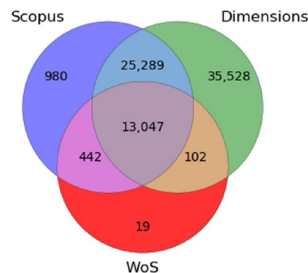
Centre for Science and Technology Studies, Leiden University, The Netherlands

RESEARCH ARTICLE

Comparison of bibliographic data sources: Implications for the robustness of university rankings

Chun-Kai (Karl) Huang¹  · Cameron Neylon²  · Chloe Brookes-Kenworthy³  · Richard Hosking⁴  · Lucy Montgomery⁵  · Katie Wilson⁶  · and Alkim Ozaygen⁷ 

Centre for Culture and Technology, Curtin University, Bentley 6102, Western Australia



ISSN

WoS was presented by Mongeon and Paul-Hus (2016). A similar comparison, including not only Scopus and WoS but also Dimensions, was carried out by Singh, Singh et al. (2020). However, both comparisons were performed at the level of journals rather than individual documents. Recently, Huang, Neylon et al. (2020) reported a document-level comparison of Scopus, WoS, and Microsoft Academic based on a fairly large amount of data (i.e., documents published by 15 universities). Their comparison has the limitation that documents in the different data sources are matched based only on Digital Object Identifiers (DOIs). Another recent document-

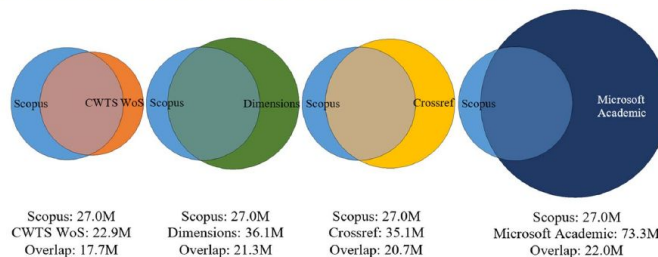
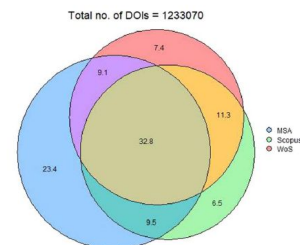


Figure 1. Overlap of documents between Scopus and the other data sources.



DOI

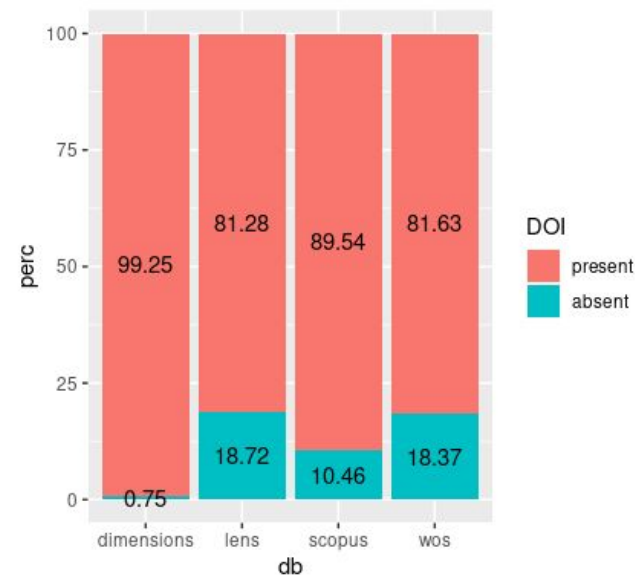
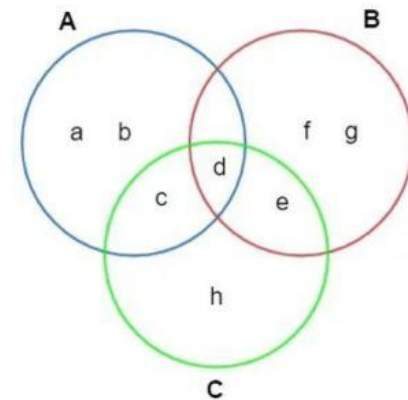
Score

$$S_{A,B} = 15m_{\text{DOI}} + 7m_{\text{first author}} + 14m_{\text{title}} + 5m_{\text{source}} + 14m_{\text{other}}.$$

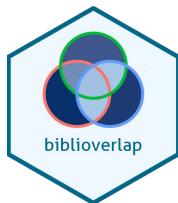
(Computacionalmente pesado)

Sobreposição (ou co-ocorrência) a nível de documento

- “Quantos/quais documentos de uma base são encontrados em outra(s) base(s)?”
- Teoria dos conjuntos - Diagrama de Venn
- Pacotes no R que fazem essa visualização
 - Recebem uma lista com os elementos de cada conjunto - retornam o diagrama
 - Identificadores únicos (e.g. DOI) podem ser utilizados nesse caso
- Muitas publicações sem DOI
 - Artigos mais velhos
 - Outros tipos de documentos (e.g. livros)
- Necessidade de estabelecer sobreposição com base em outros campos
- Nenhuma ferramenta específica para análise de co-ocorrência de docs em dados bibliográficos



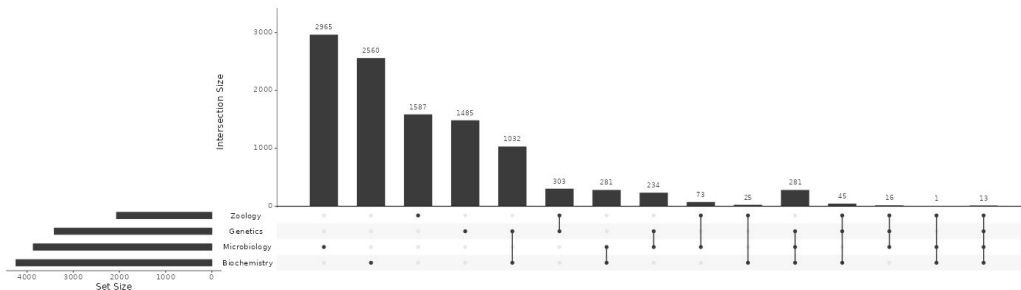
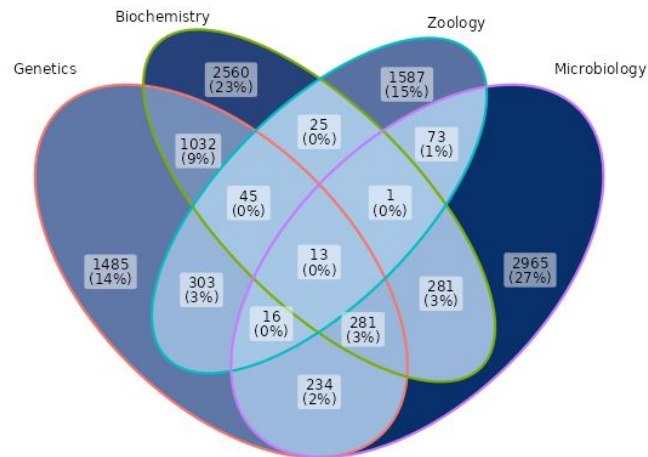
Bibliooverlap



- Pipeline de sobreposição -> Pacote do R
- Comparações:
 - DOI (ou outro identificador único), TI, SO, AU e PY
 - Entre bases distintas
 - Dentro da mesma base
 - *Proxy* - sobreposição entre diferentes áreas
 - Computação paralela - ↑ velocidade
 - Projeto do doutorado (375609 docs)
 - 16 threads - 40 minutos
 - RAM: Pico de 22 GB

- *Output:*

- Conjuntos de dados + UUID
- Sumário (tabela)
- Gráficos (sumário, venn e upset)



Interface gráfica (Shiny)

Bibliolap Input Data Plots Merge Files

Input data options

Column names
DI TI SO AU PY
DOI / unique identifier
DOI

Score matching
Change parameters

Number of threads
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Number of bibliographical datasets
1 2 3 4 5 6 7

Set1:
Name File Sep
Upload file:
Browse... math.csv
Upload complete

Set2:
Name File Sep
Upload file:
Browse... physics.csv
Upload complete

Submit

Data Summary

Download Data

Column visibility Show 10 entries

SET_NAME	UUID	Lens ID	Title	Date Published	Publication Year	Publication Type
Physics.21	Physics	57a18780-d924-4ee0-a488-d5a168...	070-309-925-113-539	Measurement of the Prompt D ⁰ ...	2023-09-06	2023 journal article
Math.607	Math	0341cc59-c7cb-4199-9f83-e309ec...	134-791-720-390-987	Searches for exclusive Higgs a...	2023-09-05	2023 journal article
Math.650	Math	20d62cac-1262-4842-94b8-6f67ab...	157-698-106-525-835	Investigation of role of anti...	2023-09-05	2023 journal article
Math.348	Math	1f3105a-9904-4fde-85ef-11ca14...	061-722-761-892-902	Lapse risk modeling in insuran...	2023-09-01	2023 journal article
Math.480	Math	fed1d9de-610a-4a8e-a8e5-ecb3d9...	091-728-517-693-712	Search for the rare decays ...	2023-09-01	2023 journal article
Math.487	Math	018a7b4e-6d6a-40a4-931d-217508...	093-754-127-030-835	Measurements of differential c...	2023-09-01	2023 journal article
Math.575	Math	7ee72a7e-d3b0-4fa7-bbda-c402ce...	124-013-855-882-977	Search for a light charged Hig...	2023-09-01	2023 journal article
Physics.8	Physics	0da7cd49-d8f9-4c02-b7d8-0218b9...	022-372-975-714-333	Measurement of the Branching F...	2023-08-29	2023 journal article
Physics.30	Physics	d85b461b-453e-4868-bb28-d55abd...	191-832-634-587-519	Measurement of the Time-Integr...	2023-08-29	2023 journal article
Math.659	Math	18802e26-bffd-4546-af4e-addb0c...	161-550-348-752-76X	Observation of the B ⁺ → J/ψ K ⁺ ...	2023-08-25	2023 journal article

Showing 1 to 10 of 730 entries

Previous 1 2 3 4 5 ... 73 Next

Bibliolap Input Data Plots Merge Files

General plot options

Summary Venn Diagram UpSet Plot

Modify plot

Label type Both Label color Black Label size 4.5 Label alpha 0.5 Set size 4.5

Math Physics

700 (96%) 0 (0%) 30 (4%)

count

400 200 0

Merging files

Bibliolap accepts a single csv file for each dataset. However, there are cases when a query has to be split between multiple files.

In this page, the user can upload multiple csv files (from the same bibliographical database) and merge all records into a single file. Duplicates are automatically removed.

Files Sep

Upload files

Browse... 2 files

Upload complete

Download Merged File

Column visibility Show 2 entries

Authors	Author full names	Author(s) ID	Title	Year	Source title	Volume	Issue	Art. No.	Pa	stz
1	Neltzke-Montinelli V.; Calòba ...	Neltzke-Montinelli, Vanessa (5...)	55352183800; 57225014686; 5720...	Differentiation of Memory CD8 ...	2022	Frontiers in Immunology	13		840203	
2	Campana E.H.; Kraycheto G.B.; ...	Campana, Eloiza H. (1654910030...)	16549100300; 56582488000; 5606...	Description of a new non-Tn440...	2022	Journal of Global Antimicrobia...	29			207

Showing 1 to 2 of 245 entries

Previous 1 2 3 4 5 ... 123 Next

- Limitações:
 - Máximo de 7 datasets
 - Aceita apenas .csv

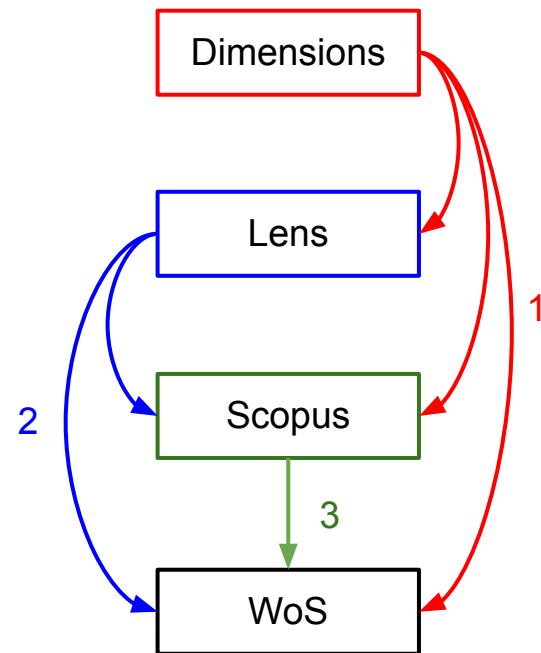
Sobreposição a nível de documento

- Tratamento dos dados

- Campos alfanuméricos: Conversão para maiúsculas e padrão ASCII
- Remoção de 'whitespace' de todos os campos
- Conversão de valores inválidos para NA
- Extração do primeiro autor de cada documento
- Adição de identificador único (**UUID**) - cada linha de cada base

- Comparações par-a-par entre bases

- 4 bases - Combinação simples: 6 possibilidades
- Cada comparação: db1 e db2
 - *Match*: documento de db1 encontrado em db2
 - Registro em db2 <- UUID do match em db1 - Direcionalidade
 - Registos de db2 modificados - Excluídos prox. comparações
- 3 “rodadas” de comparações - fixa db1, modifica db2
- Cada comparação é, na verdade, duas:
 - Registros com DOI - *Match*: DOI idêntico
 - Registros sem DOI - *Match*: Score maior/igual a 1
 - Campos: TI, PY, AU, Source (abreviado)
 - Distância levenshtein - campos alfanumericos



Scopus: “Frontiers in Antibiotics”
WoS: “Frontiers Antibiotics”

$$\begin{aligned} \text{Score} = & 0.6 - \text{lev}(\text{TI}) * 0.1 + \\ & 0.3 * \text{PY} + \\ & 0.3 - \text{lev}(\text{AU}) * 0.1 + \\ & 0.3 - \text{lev}(\text{Source}) * 0.1 \end{aligned}$$

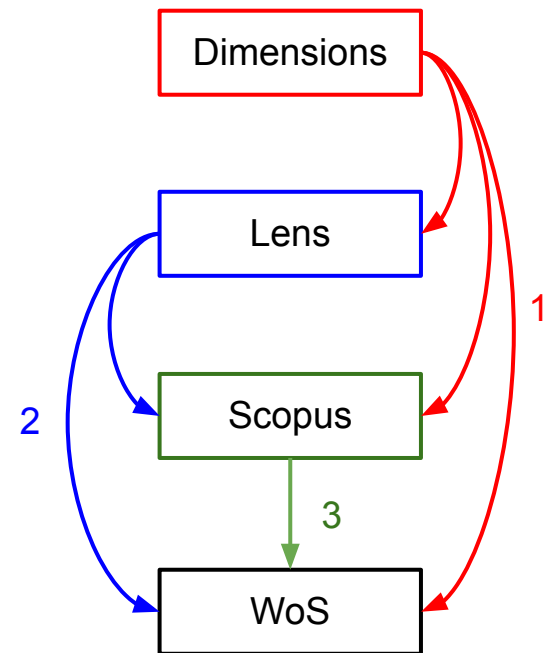
Sobreposição a nível de documento

- Tratamento dos dados

- Campos alfanuméricos: Conversão para maiúsculas e padrão ASCII
- Remoção de 'whitespace' de todos os campos
- Conversão de valores inválidos para NA
- Extração do primeiro autor de cada documento
- Adição de identificador único (**UUID**) - cada linha de cada base

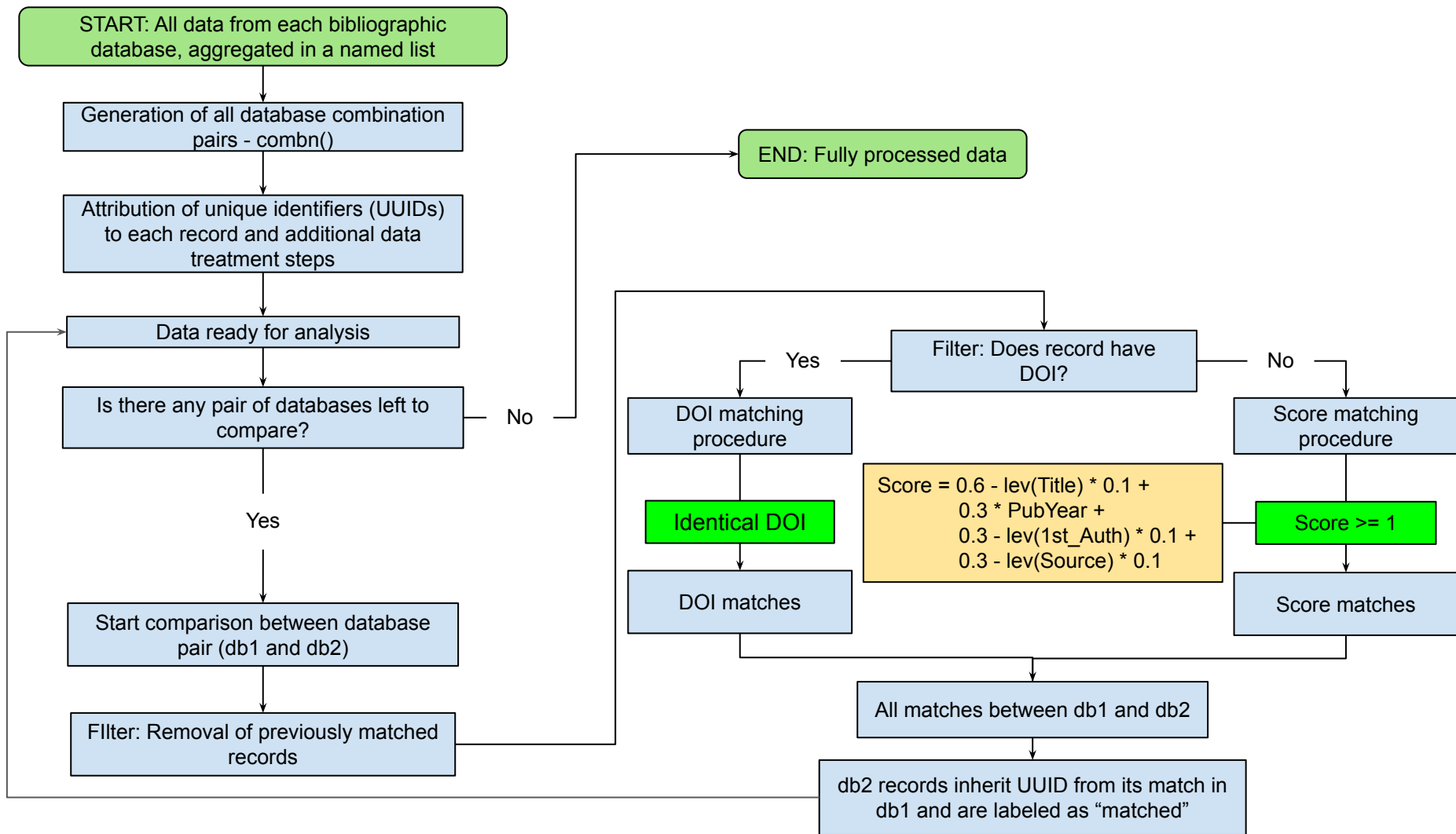
- Comparações par-a-par entre bases

- 4 bases - Combinação simples: 6 possibilidades
- Cada comparação: db1 e db2
 - *Match*: documento de db1 encontrado em db2
 - Registro em db2 <- UUID do match em db1 - Direcionalidade
 - Registos de db2 modificados - Excluídos prox. comparações
- 3 “rodadas” de comparações - fixa db1, modifica db2
- Cada comparação é, na verdade, duas:
 - Registros com DOI - *Match*: DOI idêntico
 - Registros sem DOI - *Match*: Score maior/igual a 1
 - Campos: TI, PY, AU, Source (abreviado)
 - Distância levenshtein - campos alfanumericos



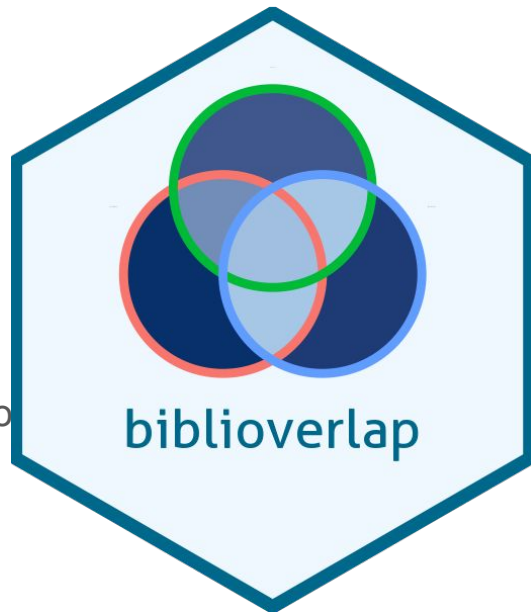
Scopus: “Frontiers in Antibiotics”
WoS: “Frontiers in Antibiotics”
Levenshtein: 3

$$\text{Score} = 0.6 - \text{lev}(\text{TI}) * 0.1 + \\ 0.3 * \text{PY} + \\ 0.3 - \text{lev}(\text{AU}) * 0.1 + \\ 0.3 - \text{lev}(\text{Source}) * 0.1$$



Em suma:

- O que o pacote espera (*input*):
 - Pelo menos dois *datasets* bibliográficos nomeados
 - Nome das colunas usadas na análise
 - Todos os conjuntos de dados devem ter o mesmo nome para essas colunas
- O que ele retorna (*output*):
 - Conjunto de dados originais + UUID
 - Possibilidade de análises adicionais
 - Sumário do processo de sobreposição
 - Shiny - Gráficos
- Disponível:
 - [CRAN](#) - Versão “estável”
 - [GitHub](#) - Versão “desenvolvimento”



Installation [↗](#)

You can install the [stable version](#) from CRAN with:

```
install.packages("biblioverlap")
```

It's also possible to install the [development version](#) from GitHub:

```
# install.packages("devtools")
devtools::install_github("gavieira/biblioverlap")
```

Dúvidas?



Contato:
gabriel.vieira@bioqmed.ufri.br