



VIII WORKSHOP DE INFORMAÇÃO,  
DADOS E TECNOLOGIA

MARÍLIA • SP

# Extraíndo dados do OJS: introdução ao webscraping em linguagem R

**Prof. Rafael Gutierrez Castanha**

✉ [r.castanha@gmail.com](mailto:r.castanha@gmail.com) / [rafaelcastanha@unimar.br](mailto:rafaelcastanha@unimar.br)



# WEBCRAPING E OJS

- ❑ O webscraping coleta dados a partir de elementos (classes) presentes o código Hypertext Markup Language (HTML) de um site.
- ❑ É necessário verificar se o *web site* a ser raspado permite que processos automatizados de coleta de dados e/ou robôs sejam implementados sobre sua plataforma, e desta maneira, por mais versátil que o método se apresente, alguns portais eletrônicos não autorizam que a raspagem seja realizada.
- ❑ Assim sendo, o método é aplicável a *web sites* diversos, independentemente de seu conteúdo, desde que não apresentem restrições para este método de mineração de dados

# WEBSCRAPING E OJS

Para verificar as permissões do site geralmente utilizamos:

- url/**robots.txt**
- url/**terms**

# WEBSCRAPING E OJS

O processo que estamos interessados compreende:

- i) **leitura da *url*** do site pela linguagem de programação;
- ii) **identificação da classe a ser raspada;**
- iii) **extração** dos elementos identificados anteriormente;
- iv) **organização, estruturação e visualização dos dados.**

# WEBCRAPING E OJS

← → ↺ seer.ufrgs.br/index.php/EmQuestao/issue/view/5170 ☆

Elementos Console Fontes Rede Desempenho Memória >> 3 2 ⚙️ ⋮ ✕

## Artigo

Disponibilidade e reúso de dados governamentais sobre apreensão de drogas de abuso  
uma análise a partir da métrica DGABr

**h3.title** 302.06 × 83.56  
Color ■ #000000DE  
Font 14px "Noto Sans", -apple-system, BlinkM...  
ACCESSIBILITY  
Name Disponibilidade e reúso de dados gover...  
Role heading  
Keyboard-focusable

**plano de gestão de dados**  
o caso DataPB

Adriana Alves Rodrigues, Joana Ferreira de Araújo, Pedro Felipy, Vivianne Leal, Guilherme Ataíde, Alzira Karla Araújo, Wagner Junqueira

PDF/A Parecer A

**Domínios, campos e novas formas de produção do conhecimento**  
abordagem interdisciplinar a partir de uma socioantropologia da informação

```
...before
<h3 class="title">
  <a id="article-143901" href="https://seer.ufrgs.br/index.php/EmQuestao/article/view/143901"> == $0
    " Disponibilidade e reúso de dados governamentais sobre apreensão de
      drogas de abuso "
    <span class="subtitle"> uma análise a partir da métrica DGABr </span>
  </a>
</h3>
<div class="meta">...</div>
<ul class="galley_links">...</ul>
::after
</div>
::after
</li>
<li>...</li>
<li>...</li>
<li>...</li>
<li>...</li>
<li>...</li>
</li>
```

loc div.sections div.section ul.cmp\_article\_list.articles li div.obj\_article\_summary h3.title a#article-143901

# PORQUE O OJS?

OJS (Open Journal Systems) é um software de código aberto que serve para gerenciar e publicar revistas eletrônicas científicas online. É uma ferramenta flexível, desenvolvida pelo Public Knowledge Project (PKP), que oferece aos editores a possibilidade de configurar o processo editorial, desde a submissão de artigos até a sua publicação.

**Mas vc já percebeu que revistas que utilizam o OJS tem uma estrutura parecida?**

- ☐ **Essa estrutura visual também é refletida em seu código html!**
- ☐ **É fácil identificarmos as classes para raspar em revistas OJS**

# WEBCRAPING E OJS

## PROCESSOS DO MINICURSO

- Biblioteca a ser utilizada: **rvest**
- **Código base:**

```
extract <-read_html("meu site") %>%  
html_nodes("classe css") %>%  
html_text2 ()
```



**Retorna valor em formato textual**

# FERRAMENTAS

Ferramenta auxiliar – Extensão do chrome Selector Gadget

**Código disponível em:**

**[Github.com/rafaelcastanha/scraping\\_OJS](https://github.com/rafaelcastanha/scraping_OJS)**