Extraindo dados do OJS: introdução ao webscraping em linguagem R

Prof.Rafael Gutierres Castanha

r.castanha@gmail.com / rafaelcastanha@unimar.br

WEBSCRAPING 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 site*s 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:

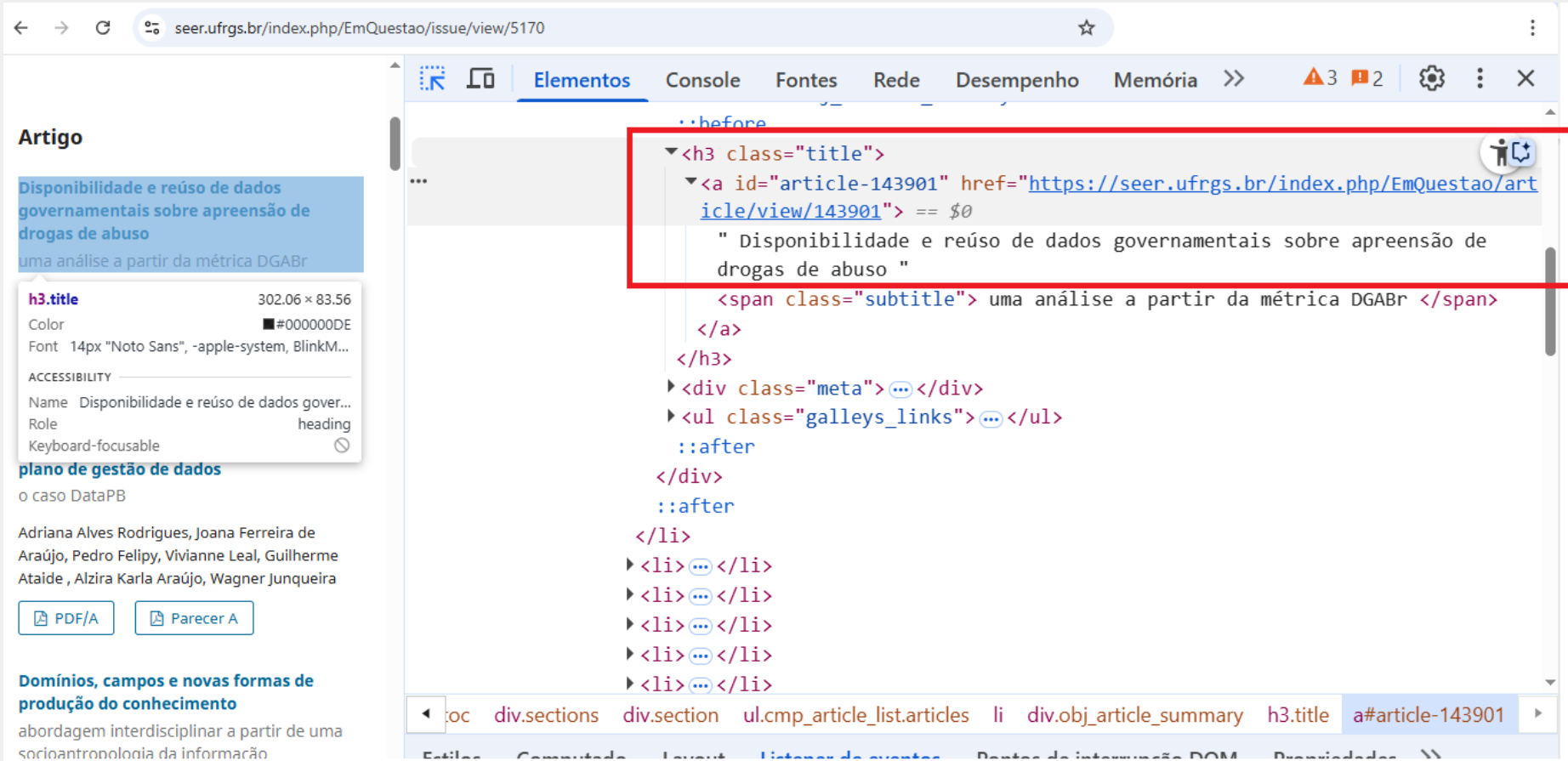
i) **leitura da *url*** do site pela linguagem de programação;

ii) **identificação da classe a ser raspada;**

iii) **extração** dos elementos identificados anteriormente;

iv) **organização, estruturação e visualização dos dados.**

WEBSCRAPING E OJS



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**

WEBSCRAPING 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 crhome Selector Gadget

Código disponível em:

Github.com/rafaelcastanha/scraping_OJS