WIDaT 2018

II WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

WEB SCRAPING DO RESEARCHERID: proposta de sistema para o monitoramento de Índice H de pesquisadores no Brasil

Alexandre Ribas Semeler
Bibliotecário (IGEO/UFRGS)
Contato: alexandre.semeler@ufrgs.br

Adilson Luiz Pinto
Professor PGCIN/UFSC
Contato: adilson.pinto@ufsc.br

Arthur Oliveira
Bolsista de IC IGEO/UFRGS
Contato: arthur.holiver@gmail.com

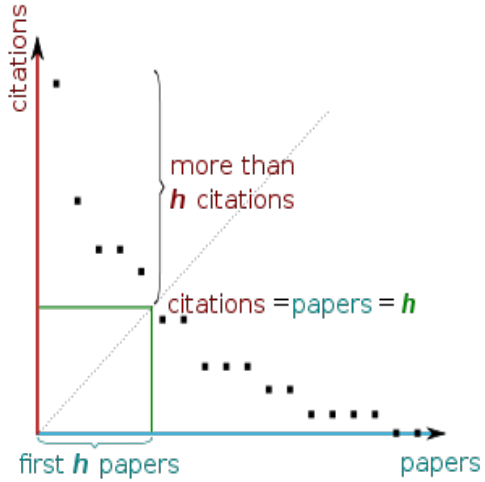
INTRODUÇÃO

RESEARCHERID

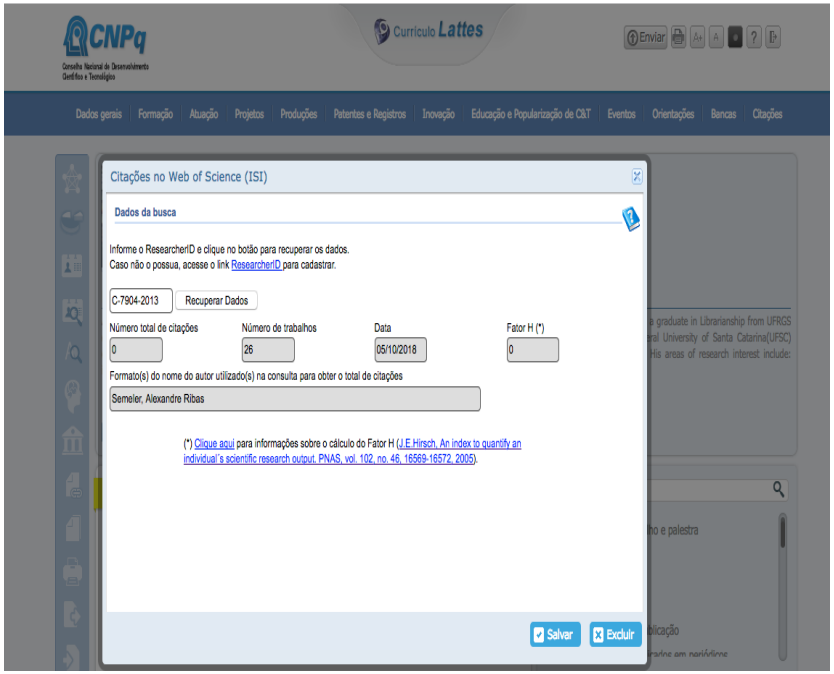
(J-9183-2016)

RANK DE USUÁRIOS/PAÍS

122.963 RUS
117.200 USA
108.893 BR
73.194 ESP
101.523 CH
46.468 IND



Indicador de quantificação para a produtividade e visualização do impacto de cientistas baseando-se nos seus artigos mais citados.



O Índice H é um indicador usado pelo CNPq para pontuar a distribuição de Bolsas de Pesquisa e fomentar projetos de Pesquisa.

Ex: um pesquisador com h = 5 tem 5 artigos que receberam 5 ou mais citações.

Desenvolver scripts em Python para monitorar o Índice H de pesquisadores brasileiros cadastrados no ResearcherID.

Metas específicas:

a) ***automatizar a coleta de dados (Índice H)*** *de pesquisadores registrados no ResearcherID;*

b) identificar o índice H *dos pesquisadores brasileiros cadastrados no ResearcherID.*

PROCEDIMENTOS METODOLÓGICOS

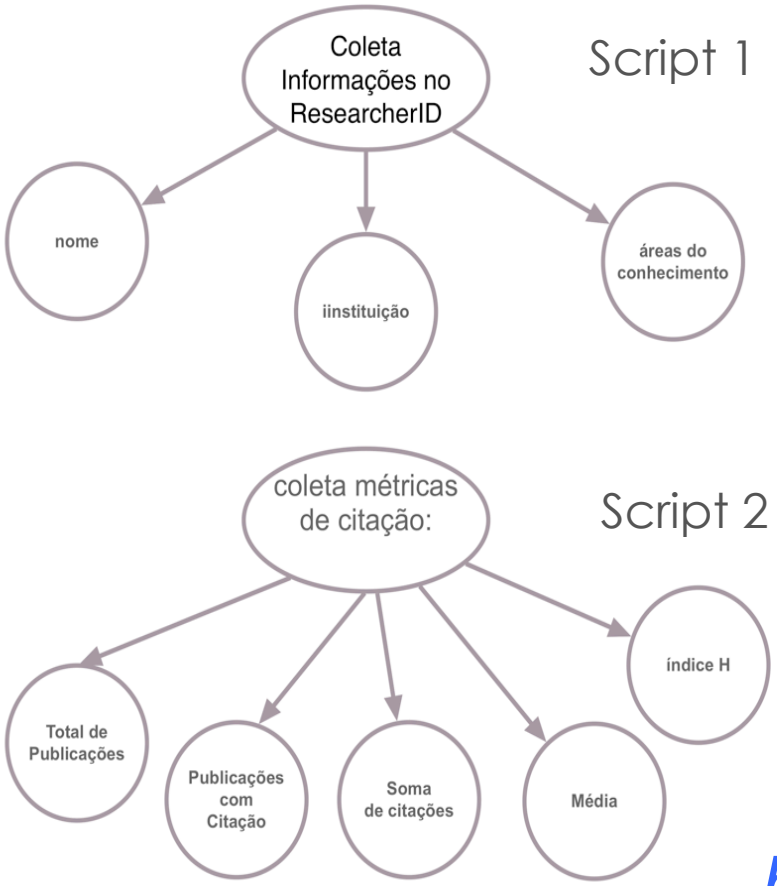
WIDaT2018
II WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

TIPO DE PESQUISA	Exploratória e Descritiva
ESTRATÉGIA	Uso de linguagens de programação para coletar o esquema de métricas do ResearcherID
NATUREZA	Quantitativa
PANORAMA GERAL DA INVESTIGAÇÃO	Índice H de pesquisadores brasileiros
CORPUS TOTAL	108.000 - IDS em 17 set. 2018
INSTRUMENTOS DE COLETA	*Web scraping, Automação de Navegação Web*
SOFTWARES	*IDE = Pycharm* *Python=2.7* *Módulos* = Codecs, BeautifulSoup, Selenium, Multiprocessing
FONTE DE COLETA DE DADOS	ResearcherID

Web Scraping

Automação de navegação web

RESULTADOS



Obs: ambos scripts possuem um dispositivo de **tratamento de falhas**, que garante que quando a extração é interrompida, em caso de *timeout* na página, a extração é reiniciada do último ponto válido e ao reiniciar verifica-se existem diferenças nas listas nos valores H.

Os scripts são **paralelizados em 8 processos iguais,** cada processo coleta informações de (1/8) das listas e métricas de citação do ReseracherID.

Código fonte Disponível em:

https://github.com/AlexSemeler/widat2018-H-index

RESULTADOS (Frente de Pesquisa)

WIDaT2018
II WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

A média do Índice H nacional é (7,97)

Rank/ ID		Total in Pub. List	With Citation Data	Sum of Times Cited	Average Citations/ Article	H-index	Área
1	B-2946-2012	438	438	31.845	72.87	96	Física (UNICAMP)
2	L-6239-2016	408	406	29.667	73.07	92	Física (USP)
3	D-3532-2012	833	820	39.622	48.50	86	Física (UNESP)
4	L-1621-2016	866	745	39.544	53.22	85	Física (UERJ)
5	C-4007-2013	619	619	37.172	60.15	84	Física (UFBA)
6	D-4476-2013	520	520	27.898	53.65	84	Medicina (UFPEL)
7	C-7679-2016	611	298	34.566	115.99	83	Medicina (UFRJ)
8	G-9573-2012	755	755	25.892	34.29	80	Psicologia (UFRGS)
9	L-4142-2016	501	488	28.821	59.30	77	Física (UERJ)
10	E-8874-2010	459	321	18.207	56.72	75	Física (USP)

Bolsistas de Produtividade em Pesquisa do CNPq

As 10 maiores frequências de índice a H estão entre 1 e 10: (**1=4006, 2=3454,3=3244, 4=2752, 5=2572, 6=2141, 7=1823, 8=1651, 9=1410, 10=1231**)

CONSIDERAÇÕES FINAIS

WIDaT2018
II WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

- A automatização da coleta de dados no ReseracherID é relevante para os estudos métricos que visem monitorar o output da produção científica nacional de impacto internacional.
- Os scripts elaborados neste trabalho podem ser utilizados para monitorar comunidades menores (PPGS, Grupos de Pesquisa e Departamentos Acadêmicos).
- ***Próxima etapa:***
  - Desenvolver scripts para validar os IDs do ResearcherID junto ao Lattes;
  - Interface gráfica;
  - Comparação com o índice de outros paises.

REFERÊNCIAS

HEIRICH, J. An index to quantify an individual's scientific research output. **PNAS, v.** 102, n. (46), 2005. Disponível em: https://doi.org/10.1073/pnas.0507655102>. Acesso em: set. 2018.

GLEZ-PEÑA, D. et al.. Web scraping technologies in an API world. **Briefings in Bioinformatics**, v. 15, n. 5, p. 788-797, 2013. Disponível em: <http://bib.oxfordjournals.org/content/15/5/788>. Acesso em: set. 2018.

ResearcherID. Disponível em: < http://www.researcherid.com > Acesso em: set. 2018.

RICE, R.; SOUTHALL, S. **The data librarian's handbook**. London: Facet Publishing, 2016.

WEBSTER, S. **What Is Scraping**? The Basics For Everyone. 2015. Disponível em: https://myhelpster.com/what-is-scraping-the-basics-for-everyone/. Acesso em: set. 2018.

PYTHON. Disponível em: <https://www.python.org/>. Acesso em set. 2018.