*Scan*COVID-19: Monitoramento da produção de conhecimento sobre COVID-19

Destaques

Trata-se do desenvolvimento, em caráter piloto, de um conjunto de técnicas e métodos automatizados denominado ***scan*COVID-19**, orientado ao monitoramento do esforço de pesquisa em Covid-19, com o objetivo de traçar um cenário atualizado do conhecimento produzido sobre o tema. São monitorados repositórios temáticos com **produção *preprint*** (medrxiv.org; biorxiv.org; arxiv.org, ssrn.com e Scielo Prepint), **produção revisada por pares** (PubMed, Scopus e Scielo), **Ensaios Clínicos** (*Clinical Trials*, ICTRP e REBEC), o **portfólio de vacinas em desenvolvimento** (WHO e London School of Hygiene & Tropical Medicine) e os protocolos de pesquisa que envolvem seres humanos submetidos para análise ética na **CONEP**. Adicionalmente, são monitoradas as **retratações** feitas aos artigos publicados. O monitoramento das fontes cobre o período a partir de janeiro de 2020, e é atualizado diariamente, sempre que disponível.

Sobre o conceito de monitoramento: Trata-se do exame de fontes de informação para detectar os primeiros sinais de desenvolvimentos importantes, apontando para tendências, desafios e oportunidades mais amplos. Ele fornece um aviso prévio de "sinais", em lugar de um estudo abrangente ou detalhado de seu impacto.

Objetivo principal: Traçar um cenário da produção de conhecimento sobre Covid-19 a partir do monitoramento de um conjunto de fontes de informação, qualificadas, com vistas a traçar um cenário atualizado do conhecimento produzido sobre o tema.

Meta: Dispor de um ponto de convergência, um espaço organizado de registros de conhecimento a partir de onde se possa ter acesso a um quadro síntese sobre

as publicações mais recentes relacionadas ao Covid-19. Sempre que possível, e em consonância com à política de acesso aberto das fontes de informação monitoradas, são coletados arquivos com texto completo.

Públicos: Pesquisadores e/ou gestores de saúde que busquem se atualizar sobre novos temas e/ou abordagens de pesquisas e ensaios clínicos sobre Covid-19.

O ***scan*COVID-19** é um projeto desenvolvido no âmbito do Programa de Pós-Graduação em Comunicação e Informação em Saúde (PPGICS), em parceria com o Laboratório de Informação Científica e Tecnológica em Saúde (LICTS), ambos do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Icict/Fiocruz. Por ser um projeto, está em constante desenvolvimento, agregando novas fontes de registros de conhecimento sobre COVID-19, e buscando novas e aprimoradas formas de acesso aos dados recuperados.

Equipe responsável

Rosane Abdala Lins (LICTS/Icict) – Coordenação Geral
http://lattes.cnpq.br/7822248387163786
Maria da Conceição Rodrigues de Carvalho (LICTS/Icict) – Coordenação Adjunta
http://lattes.cnpq.br/4654682564649838
Gustavo Barbosa (PPGICS/Icict)
http://lattes.cnpq.br/4632224013999422
Matheus Andrade Monteiro (LICTS/Icict)
http://lattes.cnpq.br/1947781197060334
Cícera Henrique da Silva (PPGICS/Icict)
http://lattes.cnpq.br/5879940619015415

Rosangela Cordeiro de Souza Assef (LICTS/Icict)

http://lattes.cnpq.br/5543134006644694

Eduardo Henrique Olímpio de Gusmão (LICTS/Icict)

http://lattes.cnpq.br/0311047260859477

Maria Cristina Soares Guimarães (PPGICS/Icict)

http://lattes.cnpq.br/8852127703130337

Introdução

A dinâmica acelerada de produção de conhecimento em tempos de pandemia amplia o desafio de acompanhar, e se atualizar, sobre as tendências e novas perspectivas de pesquisa sobre o Covid-19. O monitoramento da informação, enquanto o exame de fontes de informação para detectar os primeiros sinais de desenvolvimentos importantes, se apresenta como estratégia valiosa em meio ao volume crescente de registros de informação cientifica divulgados diariamente. Ao monitoramento cabe apontar tendências, desafios e oportunidades mais amplos. Ele fornece um aviso prévio de "sinais", em lugar de um estudo abrangente e detalhado de seu impacto.

O ***scan*COVID-19** tem como principal objetivo traçar um cenário da produção de conhecimento sobre Covid-19 a partir do monitoramento de um conjunto de fontes de informação. São monitorados repositórios temáticos com **produção *preprint*** (medrxiv.org; biorxiv.org; arxiv.org, ssrn.com e Scielo *Preprint*s), **produção revisada por pares** (PubMed, Scopus e Scielo), **Ensaios Clínicos** (*Clinical Trials*, ICTRP, REBEC), o **Portfólio de vacinas em desenvolvimento** (WHO e London School of Hygiene & Tropical Medicine) e os protocolos de pesquisa nacionais que envolvem seres humanos submetidos para análise ética na **CONEP**. Adicionalmente, são monitoradas as **retratações** feitas aos artigos publicados

A meta é dispor de um ponto de convergência, um espaço organizado de registros de conhecimento a partir de onde se possa ter acesso a um quadro síntese sobre as

publicações mais recentes relacionadas ao Covid-19. Sempre que possível, e em consonância com a política de acesso aberto das fontes de informação monitoradas, são coletados arquivos com texto completo.

Na perspectiva tecnológica, o ***scan*COVID-19** é um sistema de leitura e coleta de dados automatizado, em linguagem Python, com banco de dados Mysql, que usa três bibliotecas de *web crawler* (Beatiful Soap, Requests e Pymysql), com código versionado em *GIT*. A coleta dos registros é realizada diretamente nos portais de informação que usam protocolo *http ou https*, e também por meio de sistemas de serviço web, tipo API e/ou REST. O código encontra-se versionado em repositório em nuvem – GITLAB (shorturl.at/awAHL) e o sistema em servidor em nuvem com Sistema Operacional Linux / Ubuntu com configuração (2GB de RAM / 100 GB de ROM).

Além da classificação e agrupamento dos dados de fontes diferentes, o ***scan*COVID-19** permite que sejam realizadas buscas no conjunto de registros que foram recuperados, em cada repositório gerado. São elegidos duas maneiras de busca. Por padrão, a busca pode ser realizada nos títulos dos registros, e sempre no idioma original dos respectivos registros recuperados. Adicionalmente, é possível selecionar um filtro para especificar um conjunto de dados (por exemplo, selecionar um repositório ou um periódico cientifico para realizar a busca no título).

O ***scan*COVID-19** guarda o histórico da coleta e, portanto, pode se comportar como um *hub* de dados. Assim, mesmo que as fontes primárias descartem os dados, altere seu endereço ou suas configurações, o sistema guardará os registros de metadados, e o dado da coleta.

Considerações finais

O ***scan*COVID-19** reúne um conjunto de métodos computacionais orientado à estratégia de monitoramento da produção de conhecimento no tema. Como uma plataforma que dispõe de um serviço escalável, híbrido e dinâmico, é possível ser utilizado para monitorar outros temas/temáticas, de quaisquer repositórios fontes, desde que seja configurável, para isso seguindo a mesma modelagem do banco de dados,

ou seja, o mesmo modelo e tipos de dados presentes no dicionário do banco de dados na documentação do código. Por trabalhar com métodos mistos, pode ser agregado outras fontes de dados (novos periódicos, *preprints* ou ensaios clínicos) ou configurada para capturar metadados ainda não coletados. Os dados de pesquisa, e as palavras-chaves para busca podem ser configurados por código, ou por interface gráfica. Todos os padrões de busca são métodos de agregação, do booleano "ou", e também pode ser configurado de forma flexível para trabalhar em alternativa ou adição a demais metodologias. Permite várias agregações (p.e., por países, por temas/termos, por recortes temporais) a partir dos metadados reunidos no banco de dados.

Além disso, seu banco de dados, apesar de ser do tipo estruturado, permite mineração de dados não estruturados e a configuração de sub *datasets* para posterior busca por métodos de download de arquivos *.csv* ou.*tsv* em seu modelo de guarda de dado, etapa a ser implementada em fase posterior do projeto.