Google Dataset Search: Visão geral e perspectivas para indexação e disponibilização de conjuntos de dados científicos abertos

Adilson Luiz Pinto
Pós-Doutorado pelo Institut de Recherche en Sciences de l´Information et de la Communication (IRSIC) - França. Doutor em Documentação pela Universidad Carlos III de Madrid (UC3M) - Espanha. Professor da Universidade Federal de Santa Catarina (UFSC) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/4767432940301118
E-mail: adilson.pinto@ufsc.br

Eduardo Diniz Amaral
Doutorando em Ciência da Informação pela Universidade Federal de Santa Catarina (UFSC) – SC - Brasil. Mestre em Biotecnologia pela Universidade Estadual de Montes Claros (Unimontes) - Montes Claros, MG - Brasil. Professor da Universidade Estadual de Montes Claros (Unimontes) - Montes Claros, MG - Brasil.
http://lattes.cnpq.br/1241483438206633
E-mail: eduardo.diniz@unimontes.br

Submetido em: 25/11/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Com o intuito de colaborar com a produção científica na área de ciência de dados, especificamente em ferramentas de armazenamento e recuperação de conjuntos de dados pela internet, este artigo tem como propósito obter uma visão geral do funcionamento, padrões e perspectivas sobre a ferramenta *Google Dataset Search* –lançada em 2018 com a proposta de identificar, indexar e disponibilizar pela internet *datasets* (conjuntos massivos de dados) - instrumentos salutares para a comunidade científica. A metodologia utilizada foi descritiva, de caráter exploratório e bibliográfica sobre o tema. Foi realizado levantamento bibliográfico sobre a plataforma, identificando funcionamento interno, padrões, diretrizes, formatos e instituições de padronização que norteiam a plataforma, além de estatísticas atuais de dados indexados. Em seguida, foram executados testes práticos de utilização, usabilidade e funcionamento da ferramenta, conforme documentação disponível. Os resultados obtidos mostraram uma plataforma promissora, com índice satisfatório de usabilidade, alinhada com padrões internacionais de interoperabilidade de dados e com volumes consideráveis de *datasets* já disponíveis, em sua grande maioria no idioma inglês. Observou-se ainda, após os testes, que já existem diversos repositórios brasileiros de dados indexados pelo *Google Dataset Search*. Entretanto, alguns deles, mesmo adotando iguais padrões de metadados desta ferramenta, ainda não estão disponíveis. A conclusão é que se trata de um sistema criado pela Google, com alta capacidade de rastreamento, identificação, indexação, interoperação e disponibilização de conjuntos de dados disponíveis na internet utilizando padrões internacionais e, por isso, apresenta expressivo potencial. Este trabalho contribui para a grande área que está inserido reduzindo a escassez de publicações científicas acerca de ferramentas de disponibilização de conjuntos de dados, especificamente sobre o funcionamento, protocolos, mecanismos e interface da ferramenta em questão.

Palavras-chave: Conjuntos de dados. Interoperabilidade. Acesso aberto. Padrões de metadados. *Google Dataset Search.*

Google Dataset Search: Overview and perspectives for indexing and availability of open scientific datasets

ABSTRACT

In order to collaborate with scientific production in the field of data science, specifically in tools for storage and retrieval of data sets over the internet, this article aims to obtain an overview of the functioning, standards and perspectives on the Google Dataset Search tool - launched in 2018 with the proposal of identifying, indexing and making available internet datasets (massive sets of data) - essential instruments for the scientific community. The methodology used was descriptive, exploratory and bibliographic. A bibliographic survey was carried out on the platform, identifying internal functioning, standards, guidelines, formats and standardization institutions that guide the platform, in addition to current statistics of indexed data. Then, practical tests of use, usability and operation of the tool were performed, according to available documentation. The results obtained showed a promising platform, with a satisfactory usability score, aligned with international data interoperability standards and with considerable volumes of datasets already available, mostly in the English language. It was also observed, after the tests, that there are already several brazilian data repositories indexed by Google Dataset Search. However, some of them, even adopting the same metadata standards as this tool, are not yet available. The conclusion is that it is a system created by Google, with a high capacity for tracking, identification, indexing, interoperation and making available data sets available on the internet using international standards and, therefore, has significant potential. This work contributes to the large area that is inserted, reducing the scarcity of scientific publications on tools for making data sets available, specifically on the functioning, protocols, mechanisms and interface of this current tool.

***Keywords**: Data sets. Interoperability. Open access. Metadata standards. Google Dataset Search.*

Google Dataset Search: descripción general y perspectivas para indexar y poner a disposición conjuntos de datos científicos abiertos

RESUMEN

Para colaborar con la producción científica en el campo de la ciencia de datos, específicamente en herramientas para el almacenamiento y recuperación de conjuntos de datos a través de Internet, este artículo tiene como objetivo obtener una descripción general del funcionamiento, los estándares y las perspectivas de la herramienta Google Dataset Search, lanzada en 2018 con la propuesta de identificar, indexar y poner a disposición conjuntos de datos de Internet (conjuntos masivos de datos), instrumentos saludables para la comunidad científica. La metodología utilizada fue descriptiva, exploratoria y bibliográfica sobre el tema. Se realizó un relevamiento bibliográfico, identificando funcionamiento interno, estándares, lineamientos, formatos e instituciones de estandarización que orientan la plataforma, además de estadísticas actuales de datos indexados. A continuación, se realizaron pruebas prácticas de uso, usabilidad y funcionamiento de la herramienta, según documentación disponible. Los resultados obtenidos mostraron una plataforma prometedora, con un índice de usabilidad satisfactorio, alineada con los estándares internacionales de interoperabilidad de datos y con volúmenes considerables de conjuntos de datos ya disponibles, en su mayoría en idioma inglés. También se observó, después de las pruebas, que ya existen varios repositorios de datos brasileños indexados por Google Dataset Search. Sin embargo, algunos de ellos, incluso adoptando los mismos estándares de metadatos que esta herramienta, aún no están disponibles. La conclusión es que se trata de un sistema creado por Google, con una alta capacidad de seguimiento, identificación, indexación, interoperación y puesta a disposición de conjuntos de datos en Internet utilizando estándares internacionales y, por tanto, tiene un potencial significativo. Este trabajo contribuye a la gran área que se inserta, reduciendo la escasez de publicaciones científicas sobre herramientas para la puesta a disposición de conjuntos de datos, específicamente sobre el funcionamiento, protocolos, mecanismos e interfaz de la herramienta en cuestión.

***Palabras clave**: Conjuntos de datos. Interoperabilidad. Acceso abierto. Estándares de metadatos. Google Dataset Search.*

INTRODUÇÃO

De acordo com Gavron e Canto (2017), o acesso aberto à informação científica exerce importante influência no desenvolvimento da ciência, pois, por meio do acesso aberto é possível conhecer o que está sendo realizado pelos pesquisadores em todas as partes do globo. Quanto mais atualizado, mais relevante será para os pesquisadores, promovendo um melhor diálogo e intercâmbio informacional entre eles. Neste contexto, os *datasets* (ou conjunto de dados) representam alta relevância. Trata-se de coleções brutas de dados organizados sobre um tema ou contexto específico, geralmente dispostos em colunas (como atributos) e linhas como dados individuais, elementos ou unidades, nos mais diversos formatos (planilhas, arquivos texto, listas, tabelas etc).

No que tange o universo científico, os conjuntos de dados coletados, organizados e armazenados em experimentos, por exemplo, fazem parte do que chamamos de cauda longa de pesquisa, e são fundamentais para replicação, comprovação e novas análises destes. São informações valiosas que, se bem trabalhadas, podem gerar novos caminhos para pesquisas científicas.

Existem disponíveis na internet milhares de repositórios de dados, provendo acesso a milhões de *datasets* (NOY, 2020). Assim, dada a importância destes repositórios, os esforços de instituições nacionais e internacionais para indexar e organizar conjuntos de dados abertos ao público é cada vez maior.

No Brasil, por exemplo, temos, dentre estas iniciativas, o Portal Brasileiro de Dados Abertos, plataforma disponibilizada pelo governo brasileiro para que todos possam encontrar e utilizar dados e informações públicas (BRASIL, 2019). No cenário internacional a empresa *Google* – uma das maiores empresas da indústria informacional eletrônica contemporânea – lançou, em setembro de 2018, a ferramenta *Google Dataset Search* (referenciado pela empresa com a sigla GOODS), que se propõe a localizar e indexar *datasets* disponíveis na internet, promovendo a descoberta e catalogação destes repositórios (através de *harvesting*, inteligência artificial, big data e outras tecnologias de dados) bem como a disponibilização destes através de interfaces de consulta, desde que estejam de acordo com os padrões de dados e metadados estabelecidos internacionalmente.

Assim, dada a relevância das ferramentas disponibilizadas pela referida empresa, bem como a sua notável predominância no mercado informacional, o escopo e objeto de estudo deste trabalho orbita o buscador *Google Dataset Search*.

Configura-se como objetivo principal obter uma visão do GOODS, identificando aspectos funcionais e técnicos, bem como dos padrões de dados utilizados e por fim perspectivas acerca da ferramenta. Como objetivos específicos, lista-se:

a) Mapear estrutura de funcionamento e funções da ferramenta através da aplicação de testes;

b) Identificar procedimentos e padrões de interoperabilidade entre *datasets* e a plataforma;

c) Realizar buscas de testes em sites de domínios .br, conforme metodologias específicas e;

d) Conjecturar sobre as perspectivas da ferramenta e seus impactos informacionais para a comunidade científica.

METODOLOGIA

De acordo com Gerhardt e Silveira (2009), esta pesquisa caracteriza-se como natureza aplicada, descritiva e de caráter exploratório. É também bibliográfica, pois fez uso de artigos, manuais e outros documentos a respeito ferramenta, disponibilizados pela própria empresa, disponíveis no site oficial, https://developers.google.com, e também por outros autores especialistas nesta temática.

Inicialmente, foi realizado um levantamento bibliográfico sobre a ferramenta, com o intuito de descrever seu funcionamento, estrutura tecnológica e padrões de interoperabilidade. Tal estudo foi realizado em meados do mês de agosto de 2019 através do buscador google.com e scholar.google.com, aplicando o termo "*google dataset Search*", e atualizado em setembro de 2020. Após o levantamento bibliográfico iniciou-se o processo de testes do GOODS.

Como o sistema funciona em ambiente web, os testes de interface e usabilidade foram realizados aplicando-se as 10 heurísticas propostas por Jakob Nielsen (ROSA; VERAS, 2013), específicas para este fim. São elas: Visibilidade do status do sistema; Compatibilidade entre o sistema e o mundo real; Controle e liberdade para o usuário; Consistência e padronização; Prevenção de erros; Reconhecimento ao invés de memorização; Eficiência e flexibilidade de uso; Estética e Design minimalista; Ajude os usuários a reconhecerem, diagnosticarem e recuperarem-se de erros; e Ajuda e documentação. Quanto às funcionalidades, aplicou-se testes funcionais de unidade no método caixa-preta - técnica de teste software em que, de acordo com Myers, Sandler e Badgett (2012), o objeto a ser testado é abordado como se fosse uma caixa-preta, ou seja, dados de entrada são inseridos, o teste é executado e o resultado do processamento obtido é comparado ao resultado esperado previamente conhecido, considerando-se o desconhecimento do funcionamento interno do sistema. Haverá sucesso na aplicação do teste se o resultado obtido for de acordo com o resultado esperado.

Por fim, com base nos levantamentos e análises práticas, além de informações e opiniões sobre a ferramenta, recursos das últimas versões e notícias, foram realizadas conjecturas sobre as perspectivas do GOODS para a organização e disponibilização dos conjuntos de dados para a comunidade científica.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

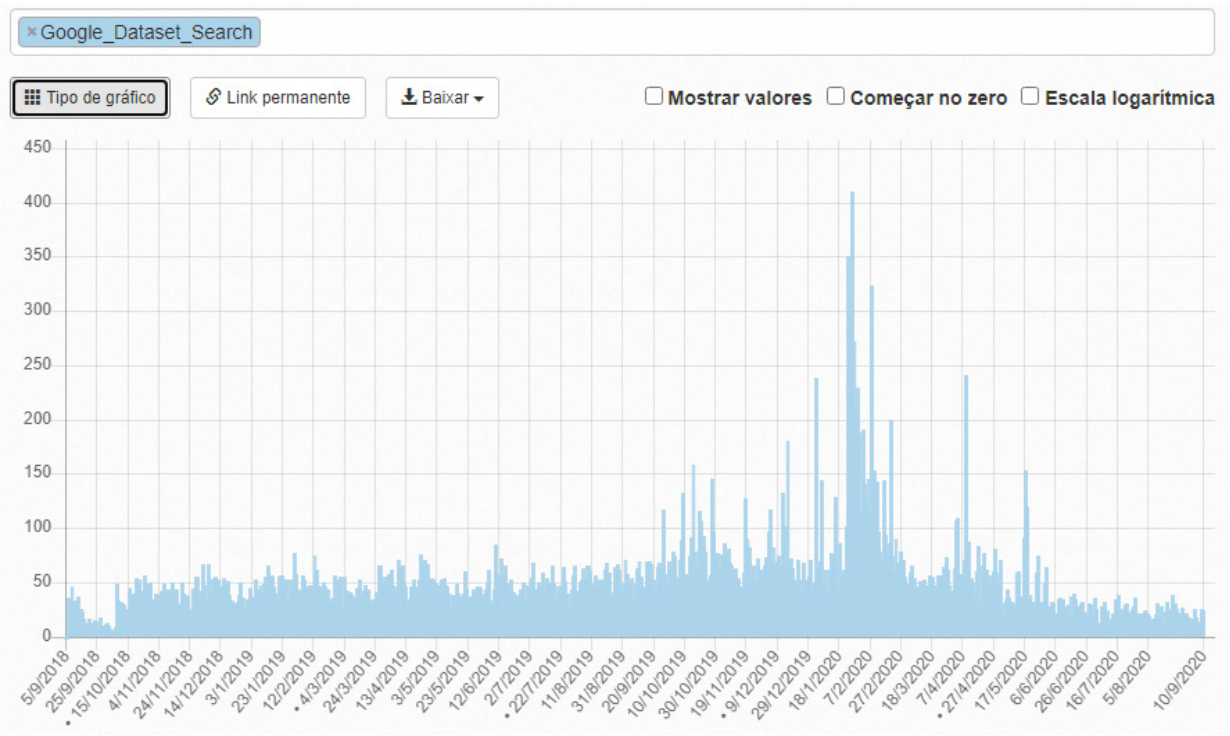
Serão apresentados a seguir os dados obtidos através dos levantamentos bibliográficos e, em seguida, os resultados dos testes realizados.

SOBRE O *GOOGLE DATASET SEARCH*

Na data da submissão deste trabalho foram encontrados aproximadamente 37.300 resultados para "*google dataset search*" no buscador da *Google*, incluindo notícias e postagens em fóruns técnicos especialistas. Efetuando a mesma busca no Google Acadêmico (scholar.google.com) foram localizados aproximadamente 335 resultados.

Assim, em comparação com a busca convencional, poucos artigos e documentos científicos foram encontrados, o que aparenta ser justificado pelo relativo espaço de tempo entre a data de lançamento da ferramenta e a data atual. Grande parte destes foram escritos por desenvolvedores, funcionários e cientistas da própria *Google*. Lançado em 2018, em 23 de janeiro de 2020 o *Google Dataset Search* (GOODS) deixou de ser beta, passando a integrar o rol de produtos oficiais da empresa, oferecendo cerca de 25 milhões de *datasets* indexados, além de novas funcionalidades (NOY, 2020). A página no idioma inglês sobre esta ferramenta na enciclopédia *Wikipedia* aponta cerca de 38.200 visualizações, com auge na ocasião do evento de lançamento oficial, conforme apontado na figura 1.

Figura 1 – Visualização de acessos da página do Google Dataset Search na Wikipedia

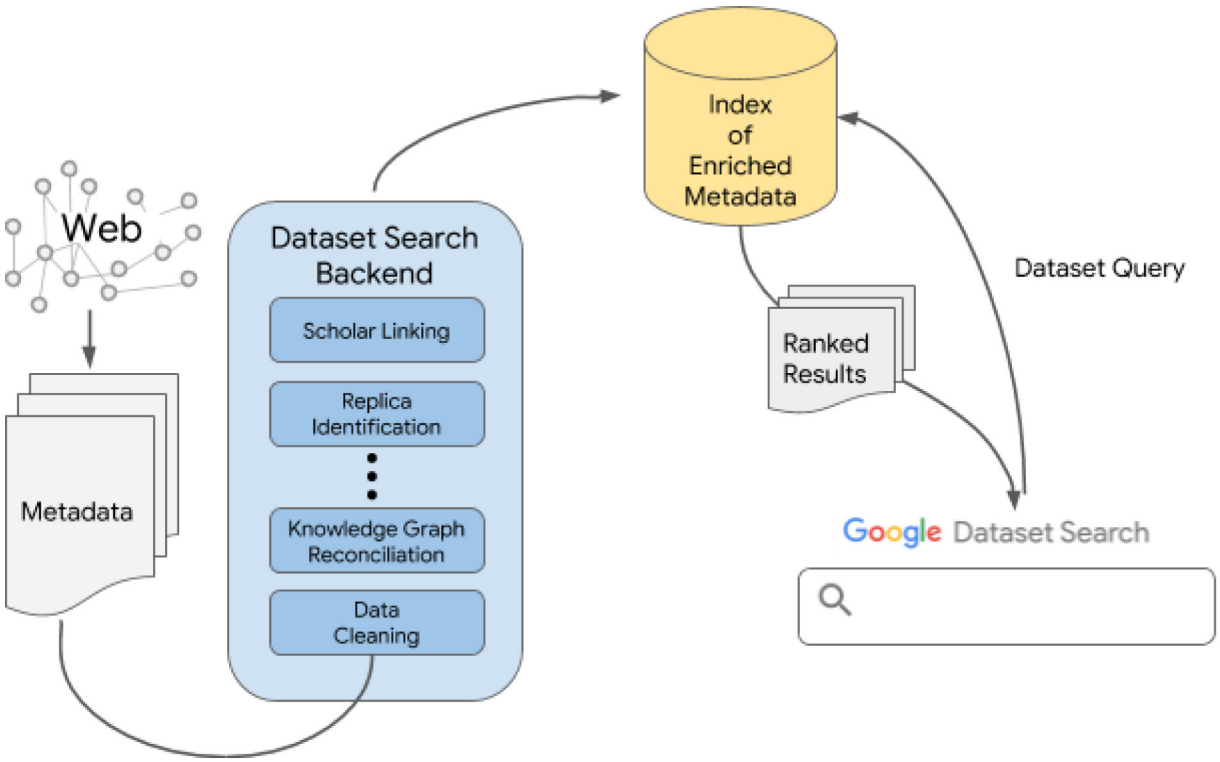


Fonte: (WIKIPEDIA, 2020).

Realizadas as buscas e levantamentos bibliográficos, prossegue-se com as definições e contextualizações. De acordo com Google (2019a), exemplos do que pode ser qualificado como um conjunto de dados representáveis por *metadados* são: uma tabela ou um arquivo CSV com alguns dados; um conjunto organizado de tabelas; um arquivo em formato proprietário que contenha dados; uma coleção de arquivos que unidos formam um conjunto de dados significativo; um objeto estruturado com dados em algum outro formato para processamento; imagens que capturam dados; arquivos relacionados ao aprendizado de máquina, como parâmetros treinados ou definições de estrutura de rede neural ou qualquer outro conjunto de dados representável e quantificável, em seus mais diversos formatos (HALEVY *et al.*, 2016): texto puro, planilhas, tabelas gigantes, sistemas de arquivos em nuvem, bases de dados relacionais etc. naturalmente ocasionando em uma ampla diversidade de metadados.

Colocadas estas considerações, o GOODS, segundo Noy (2020), surgiu partindo da dificuldade em encontrar repositórios de dados na internet. A autora afirma que, nos últimos anos, houve um aumento significativo em quantidade, volume e tamanho, além da proliferação e expansão de dados desestruturados na web, que afetou também o mundo científico e as ferramentas de busca, que não conseguem localizar dados no espectro chamado "cauda longa da pesquisa" disponíveis na internet. A proposta, portanto, foi a criação de um buscador, segundo (HALEVY *et al.*, 2016), que fosse capaz de coletar, organizar e indexar de acordo com os padrões vigentes metadados de *datasets* acessíveis pela internet. Com base nestas premissas, o GOODS foi concebido em meados de 2015, lançado em 2018 em versão beta e em 2020 como versão oficial. A figura 2 mostra uma visão geral da ferramenta.

Figura 2 – Visão geral do Google Dataset Search (GOODS)



Fonte: (BURGESS; NOY, 2018, on-line).

O protocolo "robots.txt", ou Protocolo de Exclusão de Robôs, é um conjunto de comandos que são interpretados por robôs de busca e indicam quais os diretórios e páginas de seu site não devem ser acessados por eles. Geralmente é disponibilizado na raiz das páginas web e contém comandos específicos sobre as URLs e conteúdos específicos do domínio. Tradicionalmente, por meio deste protocolo, 'robôs' digitais da *Google* coletam metadados de páginas web. Estes metadados são organizados, normalizados, indexados e organizados por prioridades para serem localizados por usuários através da interface de consulta (BURGESS; NOY, 2018). Graças aos padrões de metadados, a plataforma consegue identificar os *datasets*, conectar com outras ferramentas (como o *Google Scholar* e o *Google Knowledge Graph*) e assim extrair de forma otimizada a informação desejada. A indexação dos metadados permitem ainda eliminar *datasets* duplicados ou disponibilizados em lugares diferentes.

Com a saída da plataforma da versão BETA, novas funcionalidades foram incorporadas, além da possibilidade de democratizar o acesso a *datasets* para qualquer tipo de usuário (como por exemplo buscar "esqui" e encontrar *datasets* que abrangem desde a velocidades dos esquiadores mais rápidos às receitas dos resorts de esqui). Dentre as novas funcionalidades, estão a filtragem dos resultados de busca de acordo com o tipo de *dataset* que se deseja (ex.: tabelas, imagens, arquivos texto, se são gratuitos ou pagos etc), a possibilidade de acesso através de dispositivos móveis e melhorias de interface (NOY, 2020).

Segundo Canino (2019), é importante ressaltar a diferença entre o GOODS e outras ferramentas da *Google*. O primeiro provê uma busca mais profunda e detalhada em fontes distintas de dados, comparado com outras ferramentas da mesma empresa, como por exemplo o "*Google Public Data Explorer*", que suporta apenas formatos XML e CSV, ou do "*Google Knowledge Graph*", específico para visualização e conexão entre metadados. Ainda assim, a empresa afirma promover a interação entre os produtos com o intuito de aprimorar e testar algoritmos de busca já utilizados nas plataformas (NOY, 2020).

PADRÕES, DIRETRIZES E FORMATOS ACEITOS PARA INTEROPERABILIDADE ENTRE *DATASETS*

As orientações gerais e formatos para desenvolvedores, segundo *Google* (2019), dizem que é possível processar dados estruturados em páginas da Web sobre conjuntos de dados das seguintes formas: ou usando a marcação de conjunto de dados do *schema.org* ou estruturas equivalentes representadas no formato de vocabulário do catálogo de dados (DCAT, na sigla em inglês) do W3C (páginas em inglês). Também estão sendo testadas suportes experimentais para dados estruturados com base no CSVW do W3C. Para auxiliar a compreensão desta terminologia, o quadro 1 traz definições sobre as siglas utilizadas neste trabalho e encontradas nas diretrizes de interoperabilidade do GOODS. No entanto, a abordagem do *Google* para a descoberta de conjuntos de dados recomenda fortemente o uso das diretrizes da *schema.org*, que podem ser adicionados a páginas que descrevem conjuntos de dados.

Quadro 1 – Informações básicas sobre as siglas de formatos de dados utilizadas neste trabalho

SIGLA	DESCRIÇÃO
RDF	Resource Description Framework: representa meta dados no formato de sentenças sobre propriedades e relacionamentos entre itens na web.
JSON	Formato de interoperabilidade de dados entre sistemas, independente da linguagem de programação.
JSON-LD	JSON Linked data: é a maneira na qual a internet usa para conectar dados relacionados.
DCAT	Data catalog vocabular. Esquema de dados para facilitar a interoperabilidade entre dados de catálogos publicados na web.
URI	Uniform resource identifier: permite obter um identificador único para qualquer recurso na internet através de uma URL inteligente
Microdados	Conjunto de etiquetas de organização de conteúdos que são legíveis por computadores e pessoas.
CSVW	CSV on the web working group: utiliza o padrão RDF para dados tabulares

Fonte: Dados da pesquisa, 2019.

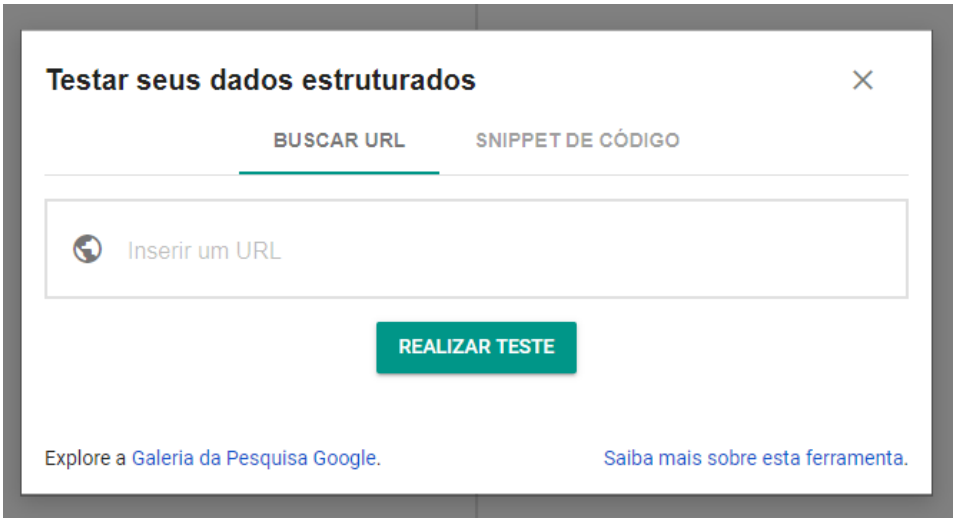
Criada pelo *Google*, Microsoft, Yahoo e Yandex, a *schema.org* é uma comunidade colaborativa com a missão de criar, manter e promover esquemas de dados estruturados na internet: em páginas web, mensagens de e-mail, conjuntos de dados e afins. O vocabulário *schema.org* pode ser utilizado em diversas codificações (como RDFa, Microdados e JSON-LD), e seu vocabulário compreende entidades, relacionamentos entre entidades e ações entre elas para estruturar e organizar dados. Segundo o site oficial, mais de 10 milhões de páginas utilizam *schema.org*, dentre elas aplicações da empresa *Google*. As orientações e vocabulários para estruturar conjuntos de dados do *schema.org* estão disponíveis em *https://schema.org/Dataset* e incluem variáveis como autor, data de publicação, palavras chave, link de acesso, codificação, licença, versão, linguagem etc.

O DCAT (*data catalog vocabulary*) é um vocabulário RDF criado pela W3C para facilitar a interoperabilidade entre catálogos de dados publicados na internet. O W3C (*World Wide Web Consortium*) atualmente é uma organização composta por 450 membros, dentre eles órgãos governamentais, empresas, organizações independentes e comunidades científicas, com a finalidade de estabelecer padrões para criação e interpretação de conteúdo na internet (W3C, 2019).

Assim, para serem inseridos no GOODS, os sites precisam seguir as diretrizes de dados estruturados (estar em JSON – padrão recomendado -, DCAT ou microdados). Além dessas diretrizes, o site oficial indica as práticas recomendadas de *sitemap* (arquivos que auxiliam o *Google* a encontrar as URLs do site) e origem e procedência – sobretudo quando conjuntos de dados abertos são republicados, agregados e baseados em outros conjuntos de dados.

Com o objetivo de facilitar a adesão aos padrões internacionais, a *Google* disponibiliza, no endereço eletrônico *https://search.google.com/structured-data/testing-tool*, uma ferramenta para realização de testes de dados estruturados, conforme ilustra a figura 3. A *Google* recomenda ainda a utilização de procedimentos como a Ferramenta de inspeção de URL (disponível no próprio console de busca fornecido pela empresa), para testar como o *Google* vê a página, o que acelera ainda mais o processo de indexação e disponibilização dos metadados nos resultados de busca do GOODS (caso existam *datasets* eleitos e disponíveis) e em outros serviços da *Google*.

Figura 3 – Ferramenta de teste de dados estruturados disponibilizada pela Google



Fonte: (GOOGLE, 2019b, on-line).

RASTREAMENTO E INSERÇÃO DE *DATASETS* NO *GOOGLE DATASET SEARCH*

O rastreamento dos sites que contém *datasets* é feito da mesma forma que as outras ferramentas da *Google*. Através do protocolo "robots.txt", os robôs buscadores ao acessarem o endereço do domínio podem ser orientados quanto ao controle de acesso e indexação de arquivos de imagem, a conteúdos em arquivos em geral (entram aqui os *datasets*), a arquivos de programação da própria página, o acesso aos mapas do site (sitemaps), dentre outros. Uma boa configuração do "robots.txt" e programação da página que contém o dataset conforme as diretrizes da schema.org é essencial para a indexação e disponibilização dos conjuntos de dados que se deseja inserir na plataforma de busca (GOOGLE, 2020).

TESTES: VISÃO GERAL DA FERRAMENTA

No tocante a utilização prática da ferramenta, o acesso pode ser feito pelo site https://datasetsearch.research.google.com. Ao acessar a página, percebe-se o mesmo padrão visual do buscador da empresa. O usuário conta com quatro opções distintas: 1) um campo de busca para realizar sua pesquisa por *datasets;* 2) um link "saiba mais" sobre como incluir conjuntos de dados nos resultados de busca; 3) um botão "sobre" com informações gerais acerca do GOODS e 4) um botão de feedback, para que os usuários possam informar eventuais problemas técnicos ou dificuldades de acesso.

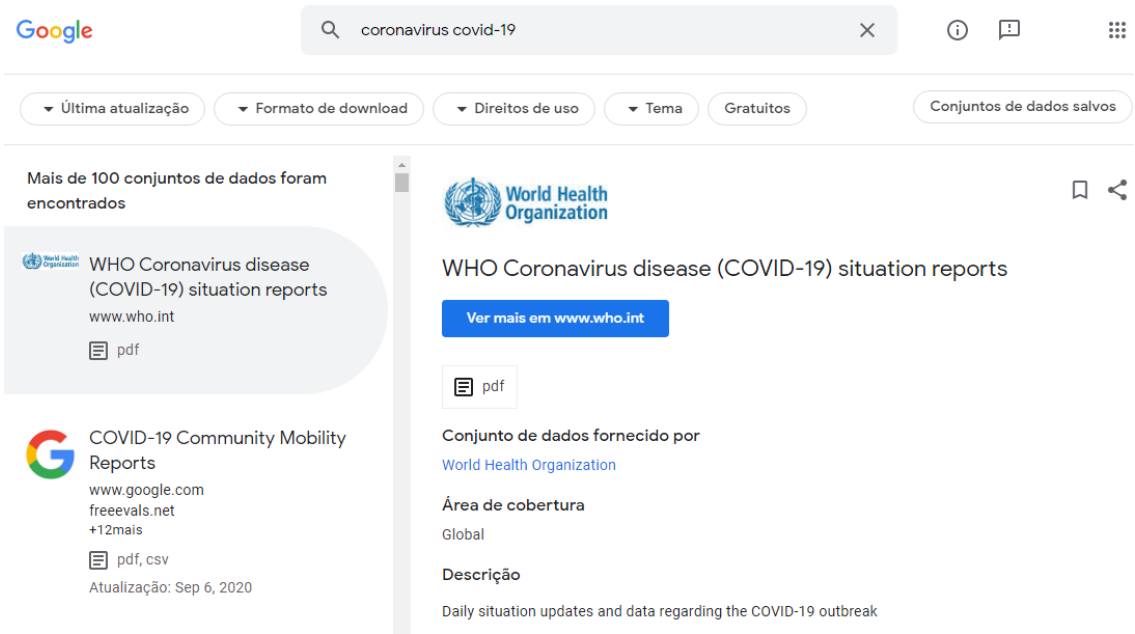
A página inicial permite ao usuário selecionar o idioma de utilização, tendo como padrão a linguagem definida pelo sistema operacional do usuário.

Os resultados de busca são dispostos da seguinte forma: à esquerda, com barra de rolagem, estão listados os *datasets* localizados conforme o critério de busca, sendo os mais relevantes primeiro e os menos relevantes em seguida. Ao clicar em um deles, são exibidos à direita da tela, nesta ordem (de cima para baixo): o título do *dataset*, o link para acesso aos dados, informações gerais como data de publicação e atualização, o fornecedor do conjunto de dados, a área de cobertura, licença, formato e descrição. A figura 4 ilustra a tela inicial do *Google Dataset Search* e a figura 5 mostra os resultados de busca. Ao visualizar os resultados de busca (figura 5), ficam visíveis uma das novidades da versão oficial (pós-beta), que são os filtros dos resultados de busca. São eles: por tempo de atualização/publicação, por formato de download, por direitos de uso, por tema e por tipo de acesso (gratuito ou não), além da possibilidade de salvar conjuntos de dados e vinculá-los à conta autenticada no momento para posterior análise.

Figura 4 – Tela inicial do *Google Dataset Search*

Fonte: (GOOGLE, 2019b, on-line).

Figura 5 – Resultados de busca pelo termo "coronavirus covid-19", sugerido pela página inicial



Fonte: Captura de tela da busca dos termos coronavírus covid-19 (GOOGLE, 2019b).

TESTES DE FUNCIONALIDADE

Para o teste de funcionalidade do tipo caixa-preta, e acatando a sugestão da própria ferramenta em sua página inicial, foi feita uma busca com o termo "*coronavirus covid-19*". Foram obtidos mais de 100 resultados indicando *datasets* contendo estes termos. No entanto, ao realizar outras buscas, percebemos que a ferramenta limita a quantidade de resultados a "mais de 100" não mostrando de fato o número real de itens obtidos. Um aspecto importante a ser mencionado é a eficiência e rapidez de indexação e disponibilização de *datasets*: diante da sugestão da própria ferramenta (covid-19) foi possível perceber que conjuntos de dados publicados em questão de poucos dias já estavam indexados e disponíveis para consulta.

Observou-se que o buscador também permite alguns comandos de busca avançada do buscador tradicional da *Google*, como operadores "site:" para restringir a busca a um domínio específico, ou "inurl:" para identificar um termo dentro de uma URL específica, os caracteres asterisco (*) para substituir por qualquer conteúdo e aspas duplas para encontrar frase exata.

Os cliques nas funções de visualização do conjunto de dados, do fornecedor e da visualização externa (no site de origem) dos 20 primeiros resultados de busca funcionaram adequadamente.

Dessa forma, os resultados obtidos no teste caixa-preta demonstraram que a ferramenta, de acordo com suas funções, entrega o que propõe, sendo capaz de buscar, listar e viabilizar conjuntos de dados já indexados pela plataforma.

TESTES DE INTERFACE E USABILIDADE

Observou-se que o padrão de interface e interação com o usuário segue o padrão Material Design, linguagem de design e padrão visual desenvolvido em 2014 pela *Google*, utilizando layouts baseados em grids, animações e transições responsivas, preenchimentos, e efeitos de profundidade como luzes e sombras, nas cores predominantemente azul, vermelho e branco.

O quadro a seguir mostra o resultado da aplicação, a título de visão geral, das 10 heurísticas de Nielsen (2014) no *Google Dataset Search*, considerando a letra C para alto grau de violação (heurística comprometida), a letra B para violação parcial e A nenhuma violação (heurística preservada):

Quadro 2 − Aplicação das 10 heurísticas propostas por Nielsen aplicadas ao GOODS

HEURÍSTICA	GRAU	OBSERVAÇÃO
Visibilidade do status do sistema	B	Não mostra o total exato de datasets obtidos
Compatibilidade entre o sistema e o mundo real	A	Familiaridade com buscador Google
Controle e liberdade para o usuário	A	Permite abrir links em novas janelas, cancelar buscas e baixar diretamente os arquivos etc.
Consistência e padronização	A	Segue os mesmos padrões visuais de todos os produtos da empresa.
Prevenção de erros	A	Não foram encontrados erros durante os testes.
Reconhecimento ao invés de memorização	A	Interface simples e intuitiva.
Eficiência e flexibilidade de uso	B	Os testes demonstraram rapidez e agilidade para retornar os resultados. Porém, não há campos de busca avançada ou instruções explícitas para utilização de operadores de busca.
Estética e Design minimalista	A	Apresenta conteúdo relevante e funcional para o contexto.
Auxilia os usuários a reconhecerem, diagnosticarem e recuperarem-se de erros	A	Não foram encontrados erros, mas existe botão de feedback nas telas.
Ajuda e documentação	A	A ferramenta disponibiliza links de ajuda e instruções gerais de uso.

Fonte: Dados da pesquisa, 2019.

A aplicação das heurísticas obteve 80% de não violação, apontando preliminarmente que a interface e usabilidade do *Google Dataset Search* atende de maneira satisfatória aos seus usuários. O resultado não poderia ser diferente, uma vez que o modelo mental proporcionado pelo padrão *material design* já é bastante difundido e reconhecido por grande parte das pessoas que utilizam as ferramentas *Google*, incluindo o sistema operacional *Android*, presente em 85,4% dos smartphones em operação até a presente data (IDC, 2020).

TESTES DE BUSCA EM DOMÍNIOS E ÁREAS ESPECÍFICAS

De acordo com Noy (2020) as áreas que possuem mais volumes de *datasets* são: geociências, biologia e agricultura. A área governamental também merece destaque, visto que grande parte dos governos publicam dados de acordo com os padrões estabelecidos pela schema.org. Mais de 2 milhões de *datasets* governamentais norte-americanos estão disponíveis. A pesquisadora ainda ressalta que o tipo de *datasets* mais presentes na plataforma são tabelas. As figuras 6, 7 e 8 trazem mais informações estatísticas sobre o GOODS.

Uma vez mencionados dados governamentais e, para fins de testes exploratórios, foram realizadas, em meados de setembro de 2020, consultas de testes em sites e domínios governamentais do Brasil para se ter uma noção geral da quantidade de conjuntos de dados mapeados no espaço eletrônico brasileiro. Os termos utilizados, a quantidade de resultados encontrados e os itens de maior relevância (segundo ranking do GOODS) foram listados no quadro 3.

Figura 6 – Estatísticas do GOODS. Em a, domínios com maior número de datasets, responsáveis por 65% dos conjuntos de dados disponíveis. Em b, quantitativos de datasets por domínios e em c quantidade de datasets por idiomas

a)

Domain	Datasets
ceicdata.com	3.7M
data.gov	3.1M
hikersbay.com	2.3M
tradingeconomics.com	2.2M
knoema.com	1.7M
figshare.com	1.3M
stlouisfed.org	1.2M
datacite.org	1.1M
thermofisher.com	1.0M
statista.com	0.9M

b)

Top-level domain	Number of datasets
.com	14,956K
.org	4,696K
.gov	3.386K
.at	819K
.net	760K
.es	524K
.de	366K
.edu	293K
.fr	281K
.eu	263K

c)

Language	Number of datasets	% increase
English	18,650K	67%
Chinese	1,851K	82%
Spanish	1,485K	70%
German	743K	74%
French	492K	76%
Arabic	435K	75%
Japanese	404K	72%
Russian	354K	65%
Portuguese	304K	69%
Hindi	288K	70%

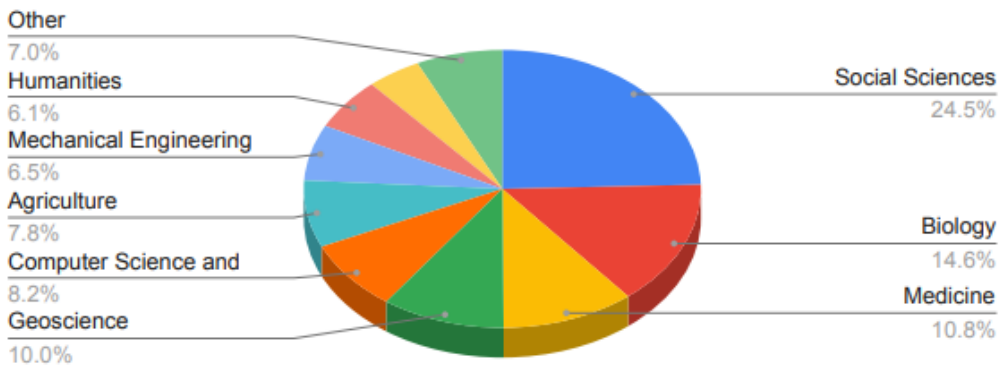
Fonte: (BENJELLOUN; CHEN; NOY, 2020).

Figura 7 – Percentuais de conjuntos de datasets agrupados por formatos

Category	Number of datasets	% of total	Sample formats
Tables	7,822K	37%	CSV, XLS
Structured	6,312K	30%	JSON, XML, OWL, RDF
Documents	2,277K	11%	PDF, DOC, HTML
Images	1,027K	5%	JPEG, PNG, TIFF
Archives	659K	3%	ZIP, TAR, RAR
Text	623K	3%	TXT, ASCII
Geospatial	376K	2%	SHP, GEOJSON, KML
Computational biology	110K	<1%	SBML, BIOPAX2, SBGN
Audio	27K	<1%	WAV, MP3, OGG
Video	9K	<1%	AVI, MPG
Presentations	7K	<1%	PPTX
Medical imaging	4K	<1%	NII, DCM
Other categories	2,245K	11%	

Fonte: (BENJELLOUN; CHEN; NOY, 2020).

Figura 8 – Percentuais de datasets organizados por áreas de conhecimento



Fonte: (BENJELLOUN; CHEN; NOY, 2020).

Quadro 3 − Amostra de testes em sites de domínio.br realizados no GOODS

	TERMO DE BUSCA	RESULTADOS	ITENS MAIS RELEVANTES
A	site:*.gov.br	> 100	ana.gov.br e hub.arcgis.com (dados ambientais, geografia e agricultura)
B	inurl:*gov.br	> 100	data.wu.ac.at (indicadores de gestão e infraestrutura)
C	inurl:*edu.br	> 100	cloud.csiss.gmu.edu (informações ligadas a graduação e pós-graduação)
D	inurl:*com.br	> 100	dados comerciais
E	site:"*.org.br"	57	Dados de eventos e monitoramentos em geral
F	universidade	> 100	Dados biológicos - GBIF
G	"são paulo"	> 100	Diversos (dados governamentais e repositórios em universidades).

Fonte: Termos de pesquisados em 2020 no buscador Google Dataset Search (GOOGLE, 2019b).

Após análise, os resultados de busca em (A e B) apontam que grande parte dos dados indexados pelo GOODS são provenientes de portal de dados abertos[1] da Agência Nacional de Águas (ANA), que oferece ferramentas para utilização de dados públicos sobre recursos hídricos no Brasil. Dados geográficos, mapeamentos e afins também estão disponíveis em hub.arcgis.com. Outro fato interessante é que muitos conjuntos de dados já estão indexados e disponíveis na plataforma csiss.gmu.edu, que pertence à GEOSS *Information Exchange DataHub*. Tal ferramenta informa que existem, na data deste trabalho, 15.223 conjuntos de dados etiquetados com "BRASIL" e 7.852 com a etiqueta "IBGE", mostrando forte relevância para a indexação de conjuntos de dados no Brasil, especialmente em itens relacionados a gestão e infraestrutura. Em C nota-se que a ferramenta indexou majoritariamente dados de instituições de ensino superior ligados à graduação e pós-graduação. Observou-se que em D o GOODS indexou tabelas de preços, listas de clientes e até especificações de produtos disponíveis em e-commerce. Em E foram localizados dados ligados a organização de cursos e eventos, mapas e resultados de monitoramentos para fins específicos (geográficos e biológicos).

Em F, observou-se que existem diversas universidades brasileiras que disponibilizam seus dados pela plataforma internacional GBIF - Sistema Global de Informação sobre Biodiversidade. Em G tentou-se identificar *datasets* relacionados ao Estado mais populoso do Brasil. Foram encontrados repositórios governamentais ligados a ANA, Embrapa, em plataformas como a *Zenodo, Kaggle* e *ReSearchGate*, dados comerciais isolados e alguns repositórios de universidades públicas.

Merece destaque o fato de que os testes realizados confirmam as publicações oficiais da *Google*, que apontam a maioridade em volume de dados de cunho social, nas áreas de geociências e biologia, com destaque para dados governamentais.

Com relação ao formato dos *datasets* encontrados, os que mais se sobressaíram nos resultados de busca expostos pelos termos do quadro 2 foram *xls, csv e json*, além de API's próprias para consulta e exploração dos dados, confirmando as estatísticas descritas por Benjelloun, Chen e Noy (2020).

É interessante notar também que já se encontram indexados e disponíveis vários *datasets* provenientes do Portal Brasileiro de Dados Abertos (dados.gov.br), uma vez que seguem os mesmos padrões e formatos seguidos pelo GOODS.

[1] Disponível em: dadosabertos.ana.gov.br

Por fim, outro fator relevante é que, naturalmente, os itens catalogados em português ainda são poucos (BENJELLOUN; CHEN; NOY, 2020) se comparados ao idioma inglês, embora muitos conjuntos de dados brasileiros estejam de acordo com os padrões exigidos pelo GOODS para indexação e disponibilização.

CONSIDERAÇÕES FINAIS

Os levantamentos bibliográficos apontam que as ferramentas *Google* desempenham importante papel como agentes das transformações informacionais que a sociedade de hoje vivencia. Desde o tradicional buscador às ferramentas voltadas para a comunidade científica, a *Google* vem, entre prós e contras (CANINO, 2019), galgando espaços significativos como instrumento de acesso e difusão da informação.

A proposta da *Google* para indexação e organização de *datasets* por meio do GOODS apresenta-se como promissora, considerando o histórico de ferramentas lançadas pela referida empresa e sua participação no mercado de tecnologia, padrões internacionais de dados e a notável capacidade de seu aparato computacional e tecnologias, como infraestrutura em nuvem, mineração de dados e inteligência artificial.

Partindo desta premissa, acredita-se num não tardio aumento de integrações com bases nacionais de conjuntos de dados, universidades e indústrias, sobretudo as que não possuem ferramentas para armazenar e distribuir seus *datasets*. Esta será uma grande contribuição para expor parte da cauda longa de pesquisa, tão importante para a comunidade científica. Quanto aos desafios e oportunidades, Goben e Sandusky, (2020) apontam que ainda existem diversos entraves técnicos, financeiros einformacionais no provimento de acesso a dados abertos, sendo para isso fundamental a participação de profissionais como bibliotecários, cientistas da informação, analistas e políticos, bem como a disponibilização de recursos financeiros por parte de instituições credenciadas, a fim de fazer com que os *datasets* – importantes ferramentas para a comunidade científica - possam ser compartilhados de maneira apropriada e assim possam chegar aos respectivos interessados.

Especificamente quanto ao GOODS, Canino (2019) conjectura que o futuro da ferramenta ainda é impreciso, considerando ou uma ascensão meteórica, ou apenas mais um dos muitos produtos *Google* que não vingaram, não passando nem mesmo da fase BETA. No entanto, podemos perceber que o estágio atual do GOODS já não é mais beta, indicando pela quantidade ascendente de *datasets* catalogados que a ferramenta está em pleno crescimento. Canino (2019) ressalta que uma interface simples de busca, interoperável com os principais padrões de dados e suportado pela *Google* tem total potencial para tornar a descoberta de relevantes conjuntos de dados uma tarefa simples, rápida e eficiente, transformando assim de modo positivo o jeito de lidar com dados brutos em pesquisas científicas e tornando-se referência nesta tarefa.

Por fim, este breve estudo, ao oferecer uma visão geral e perspectivas para indexação e disponibilização de conjuntos de dados científicos abertos através da plataforma *Google Dataset Search*, cumpre os objetivos a que foi proposto.

REFERÊNCIAS

BENJELLOUN, O.; CHEN, S.; NOY, N. Google Dataset Search by the Numbers. *arXiv:2006.06894 [cs]*, 11 jun. 2020. Disponível em: http://arxiv.org/abs/2006.06894. Acesso em: 10 mar. 2021.

BRASIL. *Portal Brasileiro de Dados Abertos.* 2019. Disponível em: http://dados.gov.br. Acesso em: 13 set. 2019.

BURGESS, M.; NOY, N. *Building Google Dataset Search and Fostering an Open Data Ecosystem*. 26 set. 2018. *Google AI Blog: The latest news from Google AI.* Disponível em: https://ai.googleblog.com/2018/09/building-google-dataset-search-and.html. Acesso em: 10 set. 2019.

CANINO, A. Deconstructing Google Dataset Search. *Public Services Quarterly*, v. 15, n. 3, p. 248–255, 3 jul. 2019. DOI 10.1080/15228959.2019.1621793. Disponível em: https://www.tandfonline.com/doi/full/10.1080/15228959.2019.1621793. Acesso em: 10 mar. 2021.

GAVRON, E. M.; CANTO, F. L. do. Análise da utilização dos periódicos de acesso aberto de uma base de dados assinada pela biblioteca universitária da UFSC. *In*: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO (CBBD), 27., 2017. *Anais* [...]. Fortaleza: FEBAB, 2017. v. 27, p. 1–6. Disponível em: https://portal.febab.org.br/anais/article/view/1787. Acesso em: 13 set. 2019.

GERHARDT, T. E.; SILVEIRA, D. T. (Orgs.). *Métodos de pesquisa.* Porto Alegre: Editora da UFRGS, 2009. Disponível em: http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf. Acesso em: 12 set. 2019.

GOBEN, A.; SANDUSKY, R. J. Open data repositories: Current risks and opportunities. *College & Research Libraries News*, v. 81, n. 2, p. 62, 4 fev. 2020. DOI 10.5860/crln.81.1.62. Disponível em: https://crln.acrl.org/index.php/crlnews/article/view/24273. Acesso em: 10 mar. 2021.

GOOGLE. *Conjuntos de diretrizes e orientações sobre o Google Dataset Search.* 2019a. *Google Search Central*. Disponível em: https://developers.google.com/search/docs/data-types/dataset?hl=pt-br. Acesso em: 13 set. 2019.

GOOGLE. *Dataset Search.* 2019b. *Google*. Disponível em: https://datasetsearch.research.google.com/. Acesso em: 12 set. 2019.

GOOGLE. *Rastreamento e indexação: manual de orientações técnicas para criação de metadados para rastreio de páginas web*. 2020. *Google Search Central.* Disponível em: https://developers.Google.com/Search/reference/robots_meta_tag. Acesso em: 20 abr. 2020.

GOOGLE. *Testar seus dados estruturados.* 2019c. **G***oogle: Ferramenta de teste de dados estruturados.* Disponível em: https://search.google.com/structured-data/testing-tool. Acesso em: 12 set. 2019.

HALEVY, A.; KORN, F.; NOY, N. F.; OLSTON, C.; POLYZOTIS, N.; ROY, S.; WHANG, S. E. Goods: Organizing Google's Datasets. *In*: SIGMOD/PODS'16: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 14 jun. 2016. *Proceedings of the 2016 International Conference on Management of Data* [...]. San Francisco California USA: ACM, 14 jun. 2016. p. 795–806. DOI 10.1145/2882903.2903730. Disponível em: https://dl.acm.org/doi/10.1145/2882903.2903730. Acesso em: 10 mar. 2021.

INTERNATIONAL DATA CORPORATION (IDC). *Smartphone Market Share.* 15 dez. 2020. *IDC*. Disponível em: https://www.idc.com/promo/smartphone-market-share/os. Acesso em: 2 set. 2020.

MYERS, G. J.; SANDLER, C.; BADGETT, T. *The art of software testing*. 3rd ed. Hoboken, N.J: John Wiley & Sons, 2012.

NIELSEN, M. A. *Reinventing discovery: the new era of networked science.* [*S. l.*: *s. n.*], 2014. Disponível em: http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781400839452. Acesso em: 9 set. 2020.

NOY, N. *Discovering millions of datasets on the web.* 23 jan. 2020. *Google: The keyword.* Disponível em: https://blog.google/products/search/discovering-millions-datasets-web/. Acesso em: 20 abr. 2020.

ROSA, J. M.; VERAS, M. Avaliação heurística de usabilidade em jornais on-line: estudo de caso em dois sites. *Perspectivas em Ciência da Informação*, v. 18, n. 1, p. 138–157, mar. 2013. DOI 10.1590/S1413-99362013000100010. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362013000100010&lng=pt&tlng=pt. Acesso em: 10 mar. 2021.

WIKIPEDIA. *Visualizações da página: comparação das visualizações entre várias páginas.* 10 set. 2020. *Visualizações.* Disponível em: https://en.wikipedia.org/wiki/Google_Dataset_Search. Acesso em: 10 set. 2020.

WORLD WIDE WEB CONSORTIUM (W3C). *Current Members.* 2019. *World Wide Web Consortium.* Disponível em: https://www.w3.org/Consortium/Member/List. Acesso em: 9 set. 2019.

W3C - World Wide Web Consortium. *Data Catalog Vocabulary* (DCAT). 2014. Disponível em: https://www.w3.org/TR/vocab-dcat/. Acesso em: 09 set. 2019.