WIDaT 2019

GOOGLE DATASET SEARCH (BETA): Visão geral e perspectivas para indexação e disponibilização de conjuntos de dados científicos abertos

Adilson Luiz Pinto[1], Eduardo Diniz Amaral[2]

(1) Universidade Federal de Santa Catarina – UFSC

(2) Universidade Estadual de Montes Claros - UNIMONTES

UnB

PGCIn Programa de Pós-Graduação em Ciência da Informação

FCI FACULDADE DE CIÊNCIA DA INFORMAÇÃO

UNIVERSIDADE FEDERAL DE SANTA CATARINA

Unimontes Universidade Estadual de Montes Claros

O ACESSO ABERTO À INFORMAÇÃO CIENTÍFICA

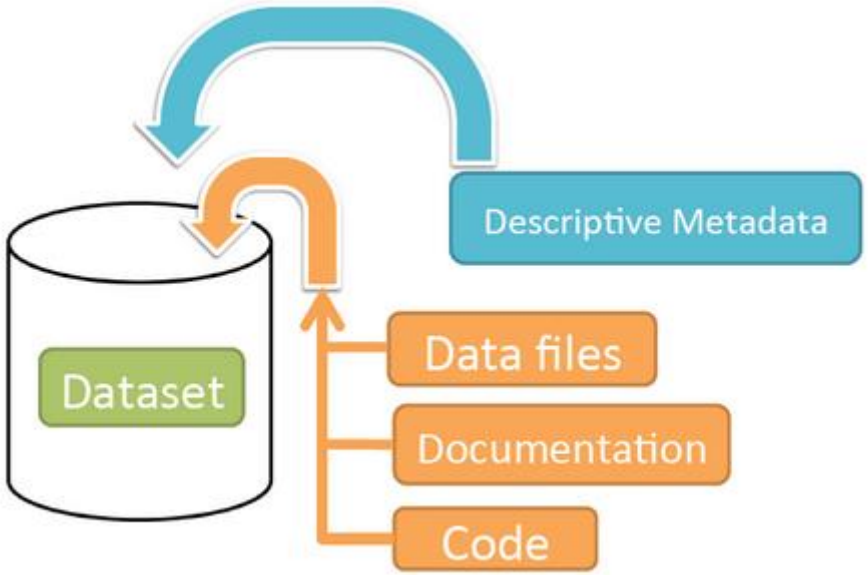
- O que é acesso aberto (open access)?

- O acesso aberto à informação científica exerce importante influência no desenvolvimento da ciência. (GAYRON, 2017)

- Diálogo e intercâmbio informacional entre pesquisadores e seus objetos de pesquisa.

Fonte: wikipedia.org

CONJUNTOS DE DADOS: DATASETS

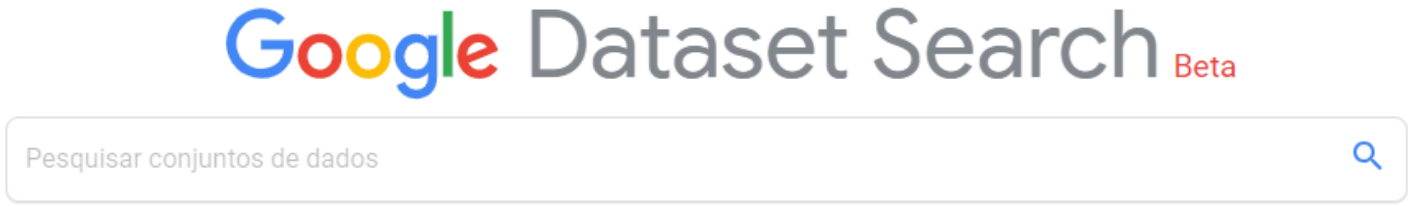


Fonte: dataverse.org

- DATASET: Definição geral
- Esforços para disponibilização são cada vez maiores
- Situação no Brasil e no mundo
  - Diversas iniciativas específicas
  - Portal Brasileiro de Dados Abertos, DATASUS, Portal da Capes, RFB etc etc

GOOGLE *DATASET SEARCH*

- Lançado pela Google oficialmente em setembro *de 2018.*
- Proposta: um buscador capaz de localizar e indexar conjuntos de *datasets* a partir de diversas fontes e instituições.
- Vínculos podem ser criados de maneira automática (robôs/IA) ou manual, através do cadastro de instituições e seus *datasets*.
- Padrões de dados específicos para interoperabilidade.



OBJETIVOS DESTE TRABALHO

a) Mapear estrutura de funcionamento e funções da ferramenta;

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

c) Realizar buscas de testes em sites de domínios **.br,** e;

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

PROCEDIMENTOS METODOLÓGICOS

- Levantamento bibliográfico e documental sobre a ferramenta;
- Observação e documentação dos aspectos funcionais através da realização de buscas específicas, de acordo com a documentação levantada;
- Testes com resultados de busca específicos e resultados emitidos pela ferramenta;
- Conjecturas, contribuições e perspectivas sobre o Google DATASET

RESULTADOS OBTIDOS

https://toolbox.google.com/datasetsearch

RESULTADOS OBTIDOS – VISÃO GERAL

- Levantamento preliminar: 40.200 resultados de busca encontrados
- Motivações:
  - Dificuldades em fornecer datasets confiáveis aos usuários pelos mecanismos de busca (GOOGLE, 2018).
  - Proliferação de dados desestruturados pela web e suas consequências
- Concepção da ferramenta: 2015. Lançamento: 2018.
- Considera como datasets desde conjuntos de dados em geral (CSV, planilhas, listas, bancos de imagens) até recursos avançados (registros de treinamento de IA, bases de dados de produção, sistemas em nuvem etc)

RESULTADOS OBTIDOS – VISÃO GERAL

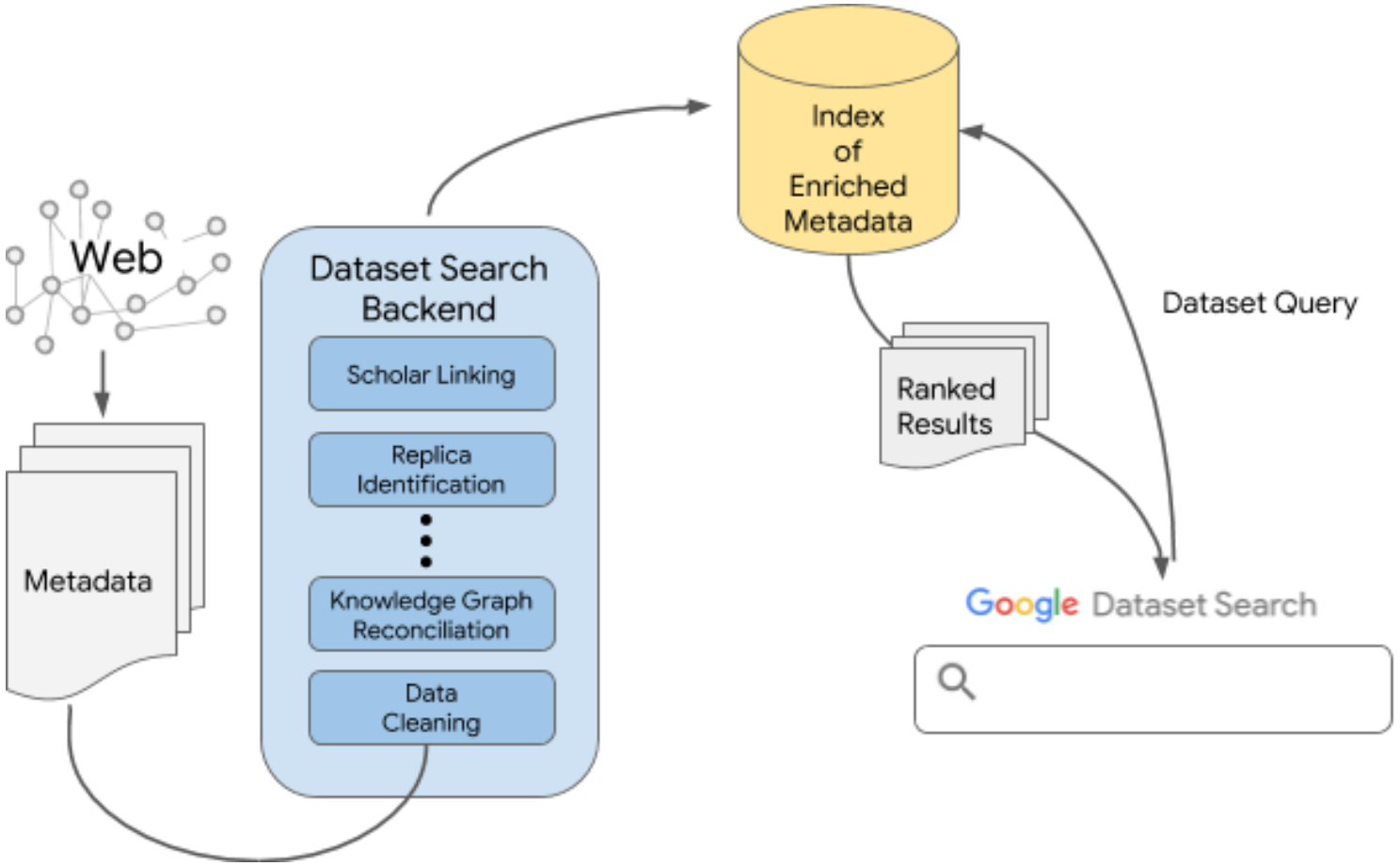


Figura 1: visão geral do Google Dataset Search (GOODS). Fonte: https://ai.googleblog.com/2018/09/building-google-dataset-search-and.html

RESULTADOS OBTIDOS – VISÃO GERAL

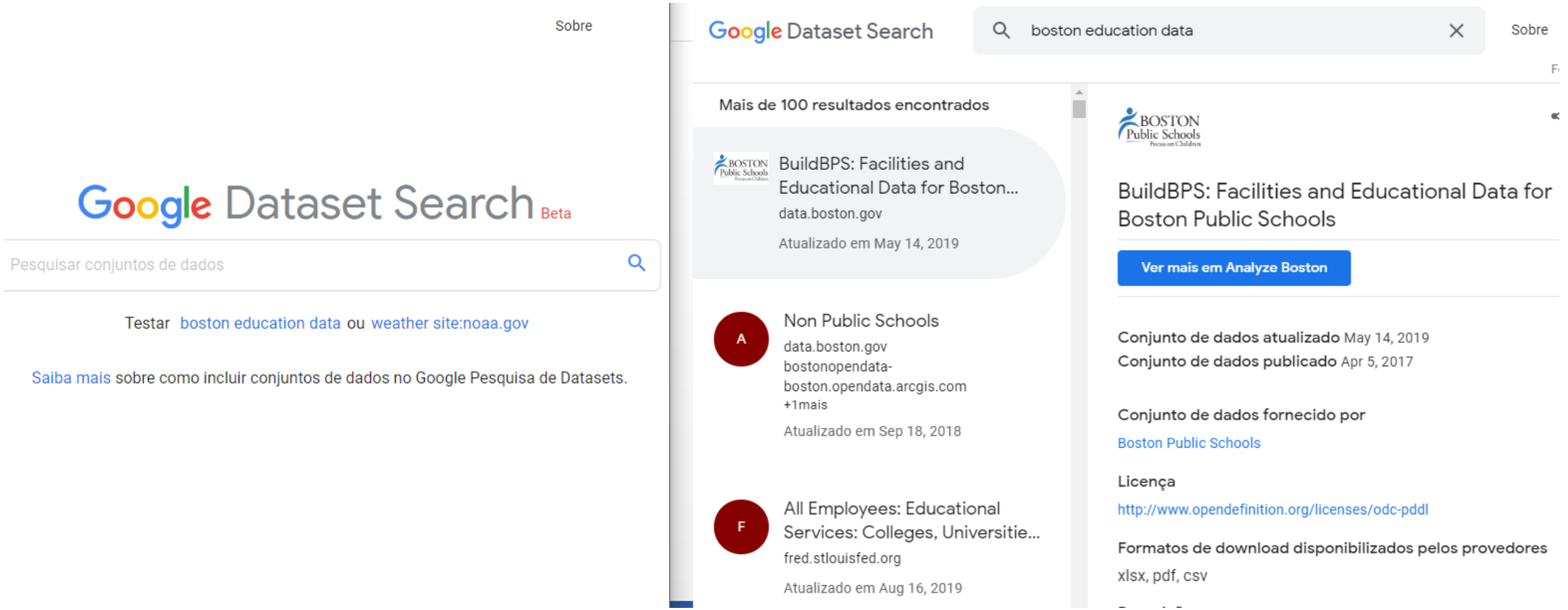


Figura 2: Visão geral da tela de busca do Google Dataset Search. Fonte: Própria, 2019

RESULTADOS OBTIDOS – PADRÕES

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
Micro dados	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

RESULTADOS OBTIDOS – TESTES EXPLORATÓRIOS

	TERMO DE BUSCA	RESULTADOS	ITENS MAIS RELEVANTES
A	site:”*.gov.br”	91	ana.gov.br e cloud. csiss.gmu.edu
B	inurl:"*gov.br"	> 100	ana.gov.br e cloud. csiss.gmu.edu
C	inurl:"*edu.br"	0	-
D	inurl:"*com.br"	> 100	dados comerciais
E	site:"*.org.br"	3	-
F	universidade	> 100	GBIF
G	“são paulo"	> 100	Diversos

RESULTADOS OBTIDOS – TESTES EXPLORATÓRIOS

- ANA - Agência Nacional de Águas
- csiss.gmu.edu: GEOSS *Information Exchange DataHub*
- GBIF - Sistema Global de Informação sobre Biodiversidade
- Repositórios isolados ligados a ANA, Embrapa, e em plataformas como a *Zenodo, Kaggle* e *ResearchGate*, além de dados comerciais isolados.
- Encontramos *json, csv, pdf* e *xls*, além de API's próprias para consulta e exploração dos dados encontrados nas buscas
- Itens em português são poucos em comparação aos termos em inglês.
- Não foram localizadas ocorrências do portal dados.gov.br

CONSIDERAÇÕES FINAIS

- Ferramenta simples e intuitiva
- Similar ao Google Scholar (2004), a ferramenta ainda é tímida no que diz respeito a quantidade de *datasets* catalogados.
- Robôs de busca e indexação + padrões + poder financeiro e tecnológico = ferramenta promissora
- Acredita-se em uma crescente adesão nacional e internacional à ferramenta nos próximos anos

REFERÊNCIAS BIBLIOGRÁFICAS

- BRASIL. Portal Brasileiro de Dados Abertos. 2019. Disponível em: <http://dados.gov.br>. Acesso em: 13/09/2019.
- CANINO, Adrienne. **Deconstructing Google Dataset Search**. Public Services Quarterly, 15:3, 248-255, DOI: 10.1080 / 15228959.2019.1621793. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/15228959.2019.1621793>. Acesso em: 13/09/2019,
- FEBAB. 2017. Disponível em: <https://portal.febab.org.br/anais/article/view/1787>. Acesso em: 13/09/2019.
- GAVRON, E. M.; CANTO, F. L. **Análise da utilização dos periódicos de acesso aberto de uma base de dados assinada pela Biblioteca Universitária da UFSC**. In: Anais do Congresso Brasileiro de Biblioteconomia, Documentação e Ciência da Informação.
- GERHARDT E SILVEIRA (org.) **Métodos de pesquisa** / [organizado por] Tatiana Engel Gerhardt e Denise Tolfo Silveira. Porto Alegre: Editora da UFRGS, 2009. Disponível em: < http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf>. Acesso em: 12/09/2019.
- GOOGLE. **Conjuntos de diretrizes e orientações sobre o Google Dataset Search**. 2019. Disponível em: <https://developers.google.com/search/docs/data-types/dataset>. Acesso em: 13/09/2019.

REFERÊNCIAS BIBLIOGRÁFICAS

- HALEVY, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., and Whang, S. E. **Goods: Organizing Google's datasets.** Google, 2016. Disponível em: < https://static.googleusercontent.com/media/research.google.com/pt-BR//pubs/archive/45390.pdf>. Acesso em: 11/09/2019.
- NOY, Natasha. BURGESS, Matthew. BRICKLEY, Dan. **Google Dataset Search: Building a search engine for datasets in an open Web ecosystem.** WebConf'2019, May 2019, San Francisco, CA USA. Disponível em: <https://doi.org/10.1145/3308558.3313685>. Acesso em: 14/09/2019.
- NOY, Natasha. Burgess, Matthew. **Building Google Dataset Search and Fostering an Open Data Ecosystem.** Google AI Blog. 2018. Disponível em: <https://ai.googleblog.com/2018/09/building-google-dataset-search-and.html>. Acesso em: 10/09/2019.
- W3C - World Wide Web Consortium. **Data Catalog Vocabulary (DCAT).** 2014. Disponível em: < https://www.w3.org/TR/vocab-dcat/>. Acesso em: 09/09/2019.
- W3C - World Wide Web Consortium. **Current Members**. Disponível em: <https://www.w3.org/Consortium/Member/List>. Acesso em: 09/09/2019.