OpenAlex como fonte de dados para sistemas nacionais de informação científica: a experiência do projeto Laguna

OpenAlex as a data source for national scientific information systems: the Laguna project experience

OpenAlex como fuente de datos para los sistemas nacionales de información científica: la experiencia do proyecto Laguna

Patricia da Silva Neubert
Doutora em Ciência da Informação, Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brasil.
Docente, Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brasil.
http://lattes.cnpq.br/8506732139258131
https://orcid.org/0000-0002-8909-1898

Fábio Lorensi do Canto
Doutor em Ciência da Informação, Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brasil.
Bibliotecário, Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brasil.
http://lattes.cnpq.br/5914776544385758
https://orcid.org/0000-0002-8338-1931

Adilson Luiz Pinto
UFSC, ORCID: 0000-0002-4142-2061,
Doutor em Documentação, Universidad Carlos III de Madrid, Getafe, Madrid, Espanha.
Docente, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, Brasil.
http://lattes.cnpq.br/4767432940301118
https://orcid.org/0000-0002-4142-2061

Daniel Sundfeld Lima
Doutor em Informática e Engenharia de Software, Universidade de Brasília (UnB), Brasília, Distrito Federal, Brasil.
Docente, Universidade de Brasília, Brasília, Distrito Federal, Brasil.
http://lattes.cnpq.br/2619423058109475
https://orcid.org/0000-0002-5147-3698

Flávio Roberto Cruz Silva
Graduado em Ciência da Computação, Universidade Católica de Pernambuco (UniCaP), Recife, Pernambuco, Brasil.
Bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil.
http://lattes.cnpq.br/4875792884731951

Anais do **VII Workshop de Informação, Dados e Tecnologia**, Porto Velho, 25-27 jun. 2024.
https://doi.org/10.22477/vii.widat.184

Resumo

Introdução: O OpenAlex é utilizado como uma das principais fontes de dados no Laguna de dados, projeto para a criação de uma Infraestrutura Informacional Aberta para processamento e integração de dados ao sistema de informação científica nacional. Este trabalho apresenta a experiência da equipe do Laguna na extração e no tratamento dos dados do OpenAlex. **Metodologia:** O processo de extração de dados é realizado mensalmente por meio de *download* dos *dumps* públicos disponibilizados pelo OpenAlex. É realizado a carga completa dos dados de cada entidade, que posteriormente são processados e, conforme o caso, cruzados com dados de outras fontes. Para os processamentos mais complexos, são realizados testes com amostras menores, a fim de estimar o tempo de processamento e nível de precisão. **Resultados:** A integração dos dados do OpenAlex deve observar as particularidades do sistema de informação científica brasileiros, incluindo a diferença de atributos, discrepância entre a cobertura da produção científica e a não equivalência de metadados entre os registros das mesmas entidades em diferentes fontes de dados. Como resultado, este processo requer o estabelecimento de procedimentos sistematizados, como meio de estabelecer uma metodologia para manutenção e atualização do lago de dados, além do desenvolvimento de soluções tecnológicas específicas que auxiliem na resolução das incompatibilidades encontradas. **Conclusão:** OpenAlex contribui para o mapeamento e relacionamento de entidade, atores e manifestações do sistema de informação científica na web, não substituindo outras ferramentas e/ou fontes sobre a atividades científicas, sendo necessário o cruzamento e compatibilização com dados de outras fontes.

Palavras-chave: fontes de informação científica; sistema de informação científica; dados abertos; ciência Aberta; OpenAlex.

Abstract

Introduction: *OpenAlex is used as one of the main data sources in Laguna de dados, a project to create an Open Information Infrastructure for processing and integrating data into the national scientific information system. This paper presents the Laguna team's experience in extracting and processing OpenAlex data.* ***Methodology:*** *The data extraction process is carried out monthly by downloading the public dumps made available by OpenAlex. The data for each entity is loaded in full and then processed and, as appropriate, cross-referenced with data from other sources. For more complex processing, tests are carried out with smaller samples in order to estimate the processing time and level of accuracy.* ***Results:*** *The integration of OpenAlex data into the IAA developed observes the particularities of the Brazilian scientific information system, including the difference in attributes, discrepancy between the coverage of scientific production and the non-equivalence of metadata between the records of the same entities in different data sources. As a result, this process requires the establishment of systematized procedures as a means of establishing a methodology for maintaining and updating the data lake, as well as the development of specific technological solutions to help resolve the incompatibilities found.* ***Conclusion:*** *OpenAlex contributes to mapping and relating entities, actors and manifestations of the scientific information system on the web, but it does not replace other tools and/or sources on scientific activities, and it is necessary to cross-reference and make compatible data from other sources.*

Keywords: *scientific information sources; scientific information system; open data; Open Science; OpenAlex.*

Resumen

Introducción: *OpenAlex se utiliza como una de las principales fuentes de datos en Laguna de dados, un proyecto para crear una Infraestructura de Información Abierta para procesar e integrar*

datos en el sistema nacional de información científica. Este artículo presenta la experiencia del equipo de Laguna en la extracción y procesamiento de datos de OpenAlex. ***Metodología:*** *El proceso de extracción de datos se realiza mensualmente mediante la descarga de los dumps públicos puestos a disposición por OpenAlex. Los datos de cada entidad se cargan en su totalidad y, a continuación, se procesan y, en se fuera el caso, se cruzan con datos de otras fuentes. Para los tratamientos más complejos, se realizan pruebas con muestras más pequeñas con el fin de estimar el tiempo de tratamiento y el nivel de precisión.* ***Resultados:*** *La integración de los datos de OpenAlex debe observar las particularidades del sistema de información científica brasileño, incluyendo la diferencia de atributos, la discrepancia entre la cobertura de la producción científica y la no equivalencia de metadatos entre los registros de las mismas entidades en diferentes fuentes de datos. Como resultado, este proceso requiere el establecimiento de procedimientos sistematizados como medio para establecer una metodología de mantenimiento y actualización del lago de datos, así como el desarrollo de soluciones tecnológicas específicas para ayudar a resolver las incompatibilidades encontradas.* ***Conclusión:*** *OpenAlex contribuye a mapear y relacionar entidades, actores y manifestaciones del sistema de información científica en la web, pero no sustituye a otras herramientas y/o fuentes sobre actividades científicas, y es necesario cruzar y compatibilizar datos de otras fuentes.*

Palabras clave: *fuentes de información científica; sistema de información científica; datos abiertos; Ciencia Abierta; OpenAlex.*

1 INTRODUÇÃO

O OpenAlex[1] é uma fonte de dados sobre a produção científica mundial, que atende aos princípios FAIR (*findability, accessibility, interoperability and reusability*) - segundo os quais os dados devem ser localizáveis, acessíveis, interoperáveis e reutilizáveis - sendo 100% aberta (open data, open API, open-source code), incluindo em seu catálogo metadados sobre autores, instituições, publicações e conceitos[2] (Priem; Piwowar; Orr, 2022).

A partir dos dados sobre a atividade científica brasileira, disponibilizado em variadas fontes de informação, o Laguna de dados, projeto desenvolvido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) em parceria com instituições de ensino superior, objetiva a criação de uma Infraestrutura Informacional Aberta (IAA) para alimentação e integração de dados sobre a

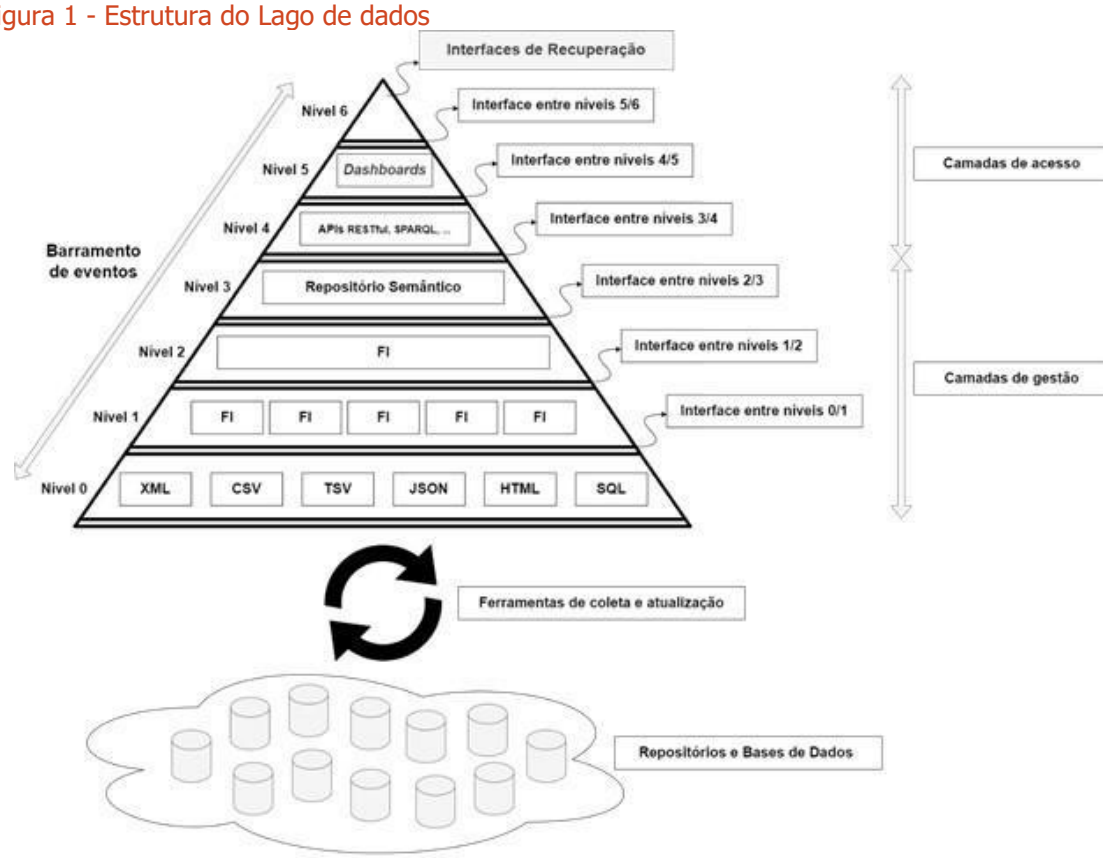
[1] Disponível em: https://openalex.org/.

[2] A partir dos Wikidata concepts, iniciativa da Wikimedia Foundation (WMF), associados a interoperabilidade semântica (Wikidata, 2024).

atividade científica ao ecossistema de informação científica brasileiro, a plataforma BRCRIS, também desenvolvido pelo IBICT (Carvalho Segundo, 2022, 2023).

O projeto é baseado na estrutura de um lago de dados, que inclui dados de diversas fontes e estrutura variada que são submetidos a diferentes métodos de tratamento (Carvalho Segundo, 2023).

Figura 1 - Estrutura do Lago de dados



Fonte: Carvalho Segundo (2023).

No Laguna, o OpenAlex é utilizado como uma das principais fontes de dados, que, no Laguna, serão processados e integrados aos dados do sistema de informação científica nacional, na plataforma BRCRIS[3], de modo a fornecer indicadores integrados sobre a atividade científica nacional e, a partir deles, a criação de serviços especializados (Carvalho Segundo, 2022, 2023). O Laguna visa, portanto, estudar e operacionalizar este processo.

[3] Disponível em: https://brcris.ibict.br/.

As particularidades em relação ao contexto brasileiro e a estrutura de seu sistema de informação científica, que inclui variados instrumentos e fontes de dados, especialmente em relação a compatibilidade e integração entre estas fontes de informação, devem ser considerados na exportação e uso dos registros do OpenAlex. Neste contexto, este trabalho tem por finalidade apresentar a experiência da equipe do Laguna de dados na extração e no tratamento dos dados do OpenAlex.

2 OPENALEX COMO FONTE DE DADOS PARA O SISTEMA DE INFORMAÇÃO CIENTÍFICA

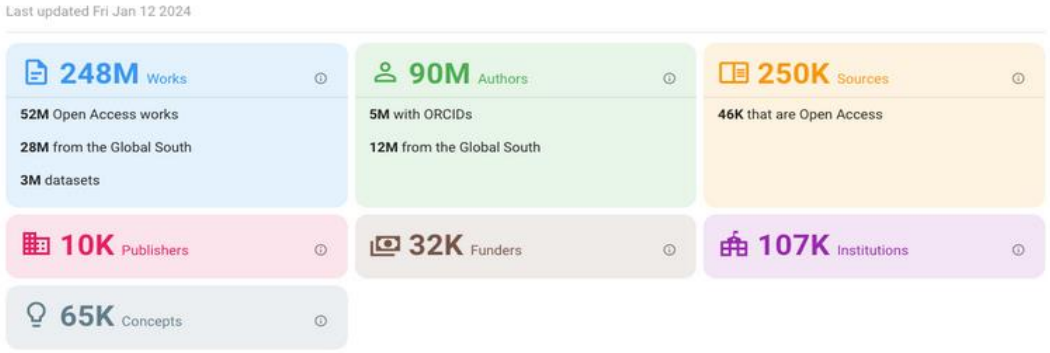
O aumento exponencial da atividade científica, e, no cenário digital, de sua complexidade, do estabelecimento de redes e da variedade de produtos da atividade científica tornam os estudos de produtividade cada vez mais necessários, e também cada vez mais complexos e dependentes de fontes estruturados, contemplando um grande conjunto de dados (Lin; Yan; Liu; Wang, 2023).

Neste contexto, ferramentas e recursos são desenvolvidos e implementados, como forma de estabelecer vinculação entre as redes, de estabelecer padrões que facilitem esse processo, contribuindo com a organização e recuperação da informação e facilitando o processo de obtenção de indicadores sobre o desempenho da Ciência.

O OpenAlex pode ser definido como um repositório de dados abertos relacionados à produção científica. Reúne documentos científicos eletrônicos, os works, que incluem teses e dissertações, conjuntos de dados, livros e capítulos, trabalhos em eventos e artigos em periódicos (Priem; Piwowar; Orr, 2022; OpenAlex, 2023). Está estruturado de maneira não só a listar este works, mas também realizando o processo de desambiguação destes registros ao conectar os works as outras entidades, instituições, fontes e autores, por exemplo (OpenAlex, 2023).

Atualmente, reúne registros sobre 248 milhões de documentos produzidos por 90 milhões de autores, vinculados a 107 mil instituições, 32 mil financiadores, publicados em 250 mil fontes e 10 mil editores (Figura 2)

Figura 2 - Dados da cobertura de OpenAlex



Fonte: Captura de tela (OpenAlex, 2023).

Quantitativamente possui cobertura superior a outras fontes de dados sobre a atividade científica, como Dimensions, Scopus e Web of Science, em número de documentos só é inferior ao Google Scholar (Figura 3). Embora seja considerado por alguns pesquisadores como uma alternativa aos recursos e indicadores produzidos pelas bases de dados (Culbert, 2024; Priem; Piwowar; Orr, 2022), não os substitui ou as suas funções no sistema de comunicação científica, especialmente aquelas associadas à avaliação e certificação do conhecimento científico (Codina, 2024). embora seja útil a produção de indicadores fora destes contextos.

Figura 3 - Comparação da cobertura de OpenAlex com outras fontes de informação científica

Text	Number of works	Open Access works	Citations
OpenAlex	248M	52M	1.9B
Scopus	90M	20.5M (ref)	1.8B
Web of Science (core)	89M (ref)	24M (ref)	1.8B
Dimensions	140M+	29M (ref)	1.7B
Google Scholar	389M (estimated)	?	?
Crossref	145M	20M	1.45B

Fonte: Captura de tela (OpenAlex, 2023).

Como fonte de dados sobre a atividade científica, sendo um recursos que atende aos princípios FAIR, alinhado às práticas de Ciência Aberta, é utilizado como fonte de dados tanto para o desenvolvimento de pesquisas quanto a criação de ferramentas específicas, como por exemplo SemOpenAlex e Equitable Science (Equitable Science, 2024; Färber et al., 2023; SemOpenAlex, 2024).

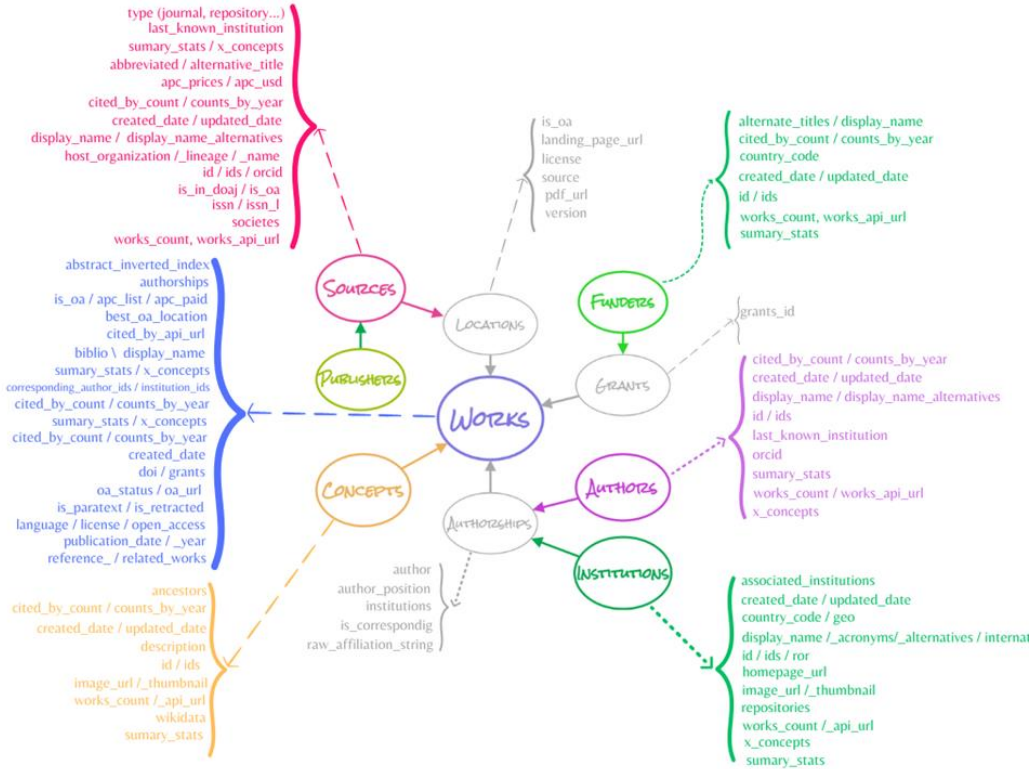
Portanto, no contexto de um sistema de informação científica nacional pode auxiliar a identificação da representação da produção científica, muitas vezes publicada e/ou depositada em fontes de limitada circulação, no contexto global; enriquecer os dados das entidades incluídas neste e/ou de seus atributos, tornando o sistema mais completo e robusto ao complementar os registros. Deste modo, que em conjunto com outras fontes sobre atividade científica, e do estabelecimento de associações e relações com outros autores e entidades do sistema possibilita que se obtenham mapas mais amplos e minuciosos sobre a atividade científica.

3 METODOLOGIA ADOTADA NO USO DE OPENALEX COMO FONTE PARA O LAGUNA DE DADOS

O processo de extração de dados é realizado mensalmente por meio de *download* dos *dumps* públicos disponibilizados pelo OpenAlex. É realizado a carga completa dos dados de cada entidade, que posteriormente são processados e, conforme o caso, cruzados com dados de outras fontes. Para os processamentos mais complexos de dados, são realizados testes com amostras menores, a fim de estimar o tempo de processamento e nível de precisão.

A partir do teste, são avaliadas se os atributos existentes nas entidades do conjunto de dados a serem exportados possuem equivalentes em outras fontes de dados, se existem identificadores persistentes que possam ser utilizados, se há necessidade de desambiguação, sobreposição e/ou equivalência entre estas fontes.

Figura 4 - Entidades, atributos e relacionamento dos works em OpenAlex



Fonte: Adaptado de OpenAlex (2023).

4 RESULTADOS PARCIAIS DO USO DE OPENALEX COMO FONTE DE DADOS PARA O SISTEMA DE INFORMAÇÃO CIENTÍFICA BRASILEIRO

A integração dos dados do OpenAlex a IAA desenvolvida no projeto Laguna de Dados observa as particularidades do sistema de informação científica brasileiros. Neste processo se observam a necessidade de atentar a:

a) diferença de atributos entre os registros das mesmas entidades em diferentes fontes de dados, como por exemplo sobre autores/pesquisadores que em OpenAlex inclui dados de identificação, afiliação e citações, enquanto a Plataforma de currículo lattes inclui ainda registros associados a formação e a genealogia do pesquisador (como área de formação, local de formação, orientações) e a vinculação ao uso do identificador persistente, o ORCId; este aspecto explicita a necessidade de integração entre os registros, com a sobreposição dos atributos oriundos de distintas fontes de dados sobre as entidades do sistema;

b) discrepância entre a cobertura da produção científica - os works de OpenAlex integram a produção indexadas em bases de dados variadas, de documentos que passaram pelo processo de avaliação pelos pares ou não, dentre as quais se incluem CrossRef e repositórios de pré-print (OpenAlex, 2023), enquanto a Plataforma Lattes inclui produções autodeclaradas pelo autor, publicadas e/ou aceitas para publicação (podendo incluir produções não indexadas em bases de dados e outros produtos da atividade científica, como produções técnicas); estas diferenças evidenciam a necessidade de cruzamento entre registros de produção para um panorama completo e preciso da produção científica brasileira;

c) não equivalência entre metadados das entidades, como por exemplo o uso do Wikidata concepts em OpenAlex para identificação das áreas de conhecimento, requerendo a compatibilização com as classificações utilizadas no Brasil, que também são variadas entre si, as áreas da Coordenação de Avaliação de Pessoal de Ensino Superior (CAPES), associadas aos cursos de pós-graduação e avaliação

da produção científica, ou seja, pela institucionalização do processo de formação dos pesquisadores no país, que utiliza uma classificação diversa da do Conselho Nacional de Pesquisa (CNPq), órgão federal responsável pelo financiamento da atividade de pesquisa no país. No cenário atual ambas as classificações, CAPES e CNPq, são necessárias aos diferentes atores científicos brasileiros, sendo necessária a compatibilização entre os registros e a manutenção das diferentes associações.

Como resultado, este processo de exportação e uso dos dados do catálogo do OpenAlex, requer que se estabeleçam procedimentos sistematizados, como meio de estabelecer uma metodologia para manutenção e atualização do lago de dados, além do desenvolvimento de soluções tecnológicas específicas, como apis e scripts que auxiliem na resolução das incompatibilidades encontradas. Os estudos e propostas de aplicações, assim como os testes realizados, fazem parte de todo este processo.

5 Considerações Finais

OpenAlex contribui para o mapeamento e relacionamento de entidade, atores e manifestações do sistema de informação científica na web, alinhado ao movimento de Ciência Aberta, contribuindo com o ideal de controle bibliográfico estabelecido na Biblioteca de Alexandria. Apesar disso, não substitui outras ferramentas e/ou fontes de dados sobre a atividades científicas, sendo necessário o cruzamento e compatibilização com dados de outras fontes - objeto da IAA desenvolvida no projeto Laguna de dados.

Assim, embora se utilize uma fonte de dados abertos para a obtenção de dados sobre a atividade científica de um país, a particularidade do sistema de informação científica nacional deve ser observada, requerendo o desenvolvimento de estudos e soluções específicas para o contexto nas quais se inserem. Neste cenário, existe a necessidade de pesquisa e desenvolvimento de técnicas e

tecnologias específicas para resolução das incompatibilidades encontradas nos diferentes ecossistemas de pesquisa.

REFERÊNCIAS

Carvalho Segundo, Washington Luís R. de. **Construindo uma Infraestrutura Aberta de Dados de Pesquisa no Brasil**. 10 ago. 2022. Disponível em: https://www.arca.fiocruz.br/handle/icict/54762

Carvalho Segundo, Washington Luís R. de. **BRCRIS**: ecossistema de informação da pesquisa científica brasileira. 2023. Disponível em: https://confap.org.br/news/wp-content/uploads/2023/12/Washington-Segundo-IBICT-F%C3%B3rum-CONFAP-2023.pdf

Codina, Lluís. **OpenAlex**: ¿una alternativa a Scopus y Web of Science? 2024. Disponível em: https://www.lluiscodina.com/openalex-scopus/

Culbert, Jack H. et al. Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus. **ArXiv**, 2024. Disponível em: https://arxiv.org/pdf/2401.16359.pdf

Equitable Science. **About Us.** 2024. Disponível em: https://equitablescience.com/

Färber, Michael; et al. SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples. IN: The Semantic Web – ISWC 2023. DOI: https://link.springer.com/chapter/10.1007/978-3-031-47243-5_6

Lin, Zihang; Yan, Yian; Liu, Lu; Wang, Dashun. SciSciNet: a large-scale open data lake for the science of science research. **Scientific Data**, 2023. Disponível em: https://doi.org/10.1038/s41597-023-02198-9

OpenAlex. How it works. 2023. Disponível em: https://help.openalex.org/how-it-works

Priem, J., Piwowar, H., & Orr, R. *OpenAlex:* A fully-open index of scholarly works, authors, venues, institutions, and concepts. **ArXiv**, 2022. Disponível em: https://arxiv.org/abs/2205.01833.

SemOpenAlex. **About**. 2024. Disponível em: https://semopenalex.org/resource/semopenalex:UniversalSearch

Vrandečić, D. and Krötzsch, M. Wikidata: a free collaborative knowledgebase. **Communications of the ACM**, v. 57, n. 10, p. 78-85, 2014. DOI: 10.1145/2629489.

Wikidata. **Wikidata Introduction.** 30 jan. 2024. Disponível em: https://www.wikidata.org/wiki/Wikidata:Main_Page