



VIII ISKO BRASIL 2025



Modelo de classificação automática de publicações do OpenAlex com base nas áreas do conhecimento do CNPq e da CAPES

Fabio Canto (IBICT)
Patricia Neubert (IBICT)
Washington Segundo (IBICT)
Thiago Dias (IBICT)



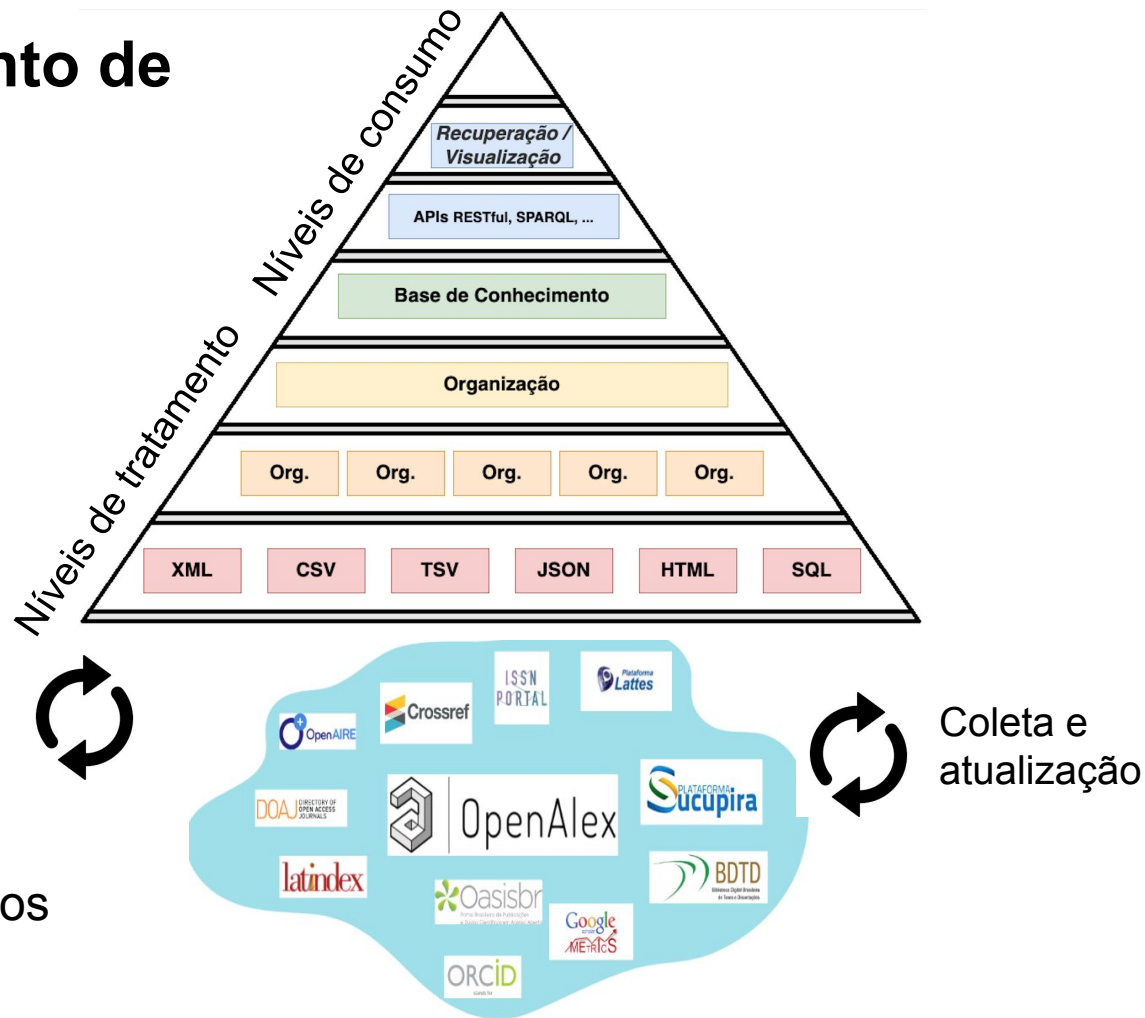


Pesquisa desenvolvida no projeto Laguna

- Coleta de dados de fontes abertas de CT&I
- Armazenamento em lago de dados
- Processamento em larga escala
- Prestação de serviços informacionais
- Criação de modelos de interoperabilidade
- Compartilhamento de dados enriquecidos

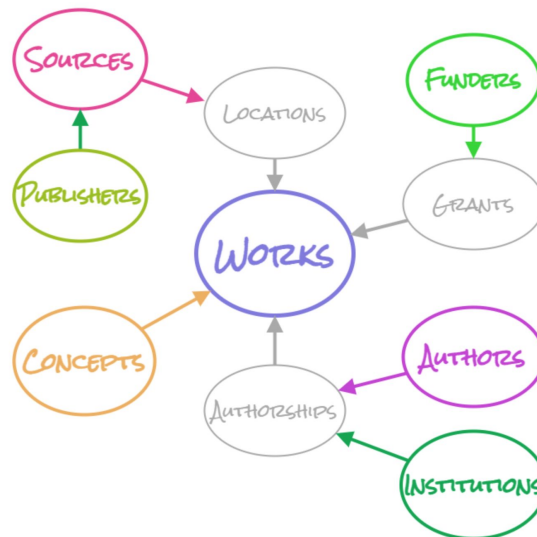


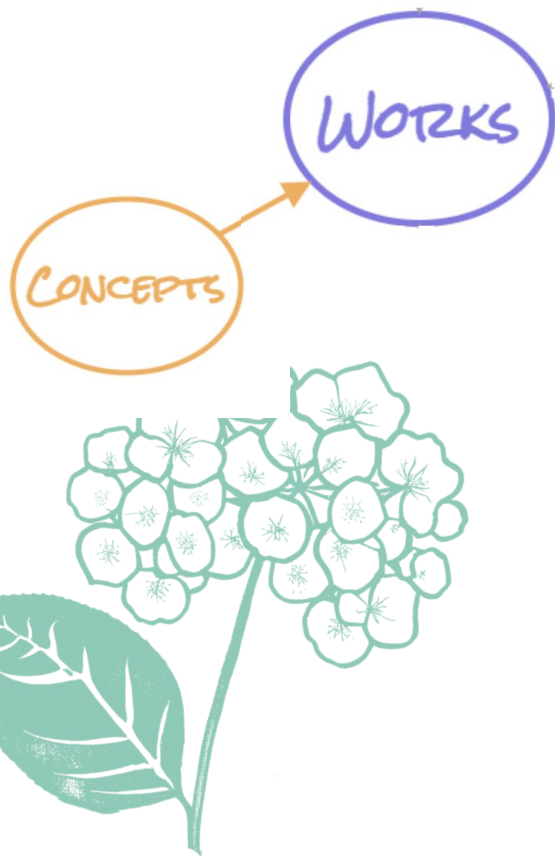
Modelo de tratamento de dados





É um amplo repositório de dados (260M works)
Criado a partir do último *dump* do MAG (2022)
Estrutura da forma de grafo





```
concepts: [  
  {  
    id: "https://openalex.org/C71924100",  
    wikidata: "https://www.wikidata.org/wiki/Q11190",  
    display_name: "Medicine",  
    level: 0,  
    score: 0.9187037  
  },  
  {  
    id: "https://openalex.org/C3007834351",  
    wikidata: "https://www.wikidata.org/wiki/Q82069695",  
    display_name: "Severe acute respiratory syndrome coronavirus",  
    level: 5,  
    score: 0.8070164  
  },  
]
```

Compatibilizar os sistemas de classificação (Wikidata, OpenAlex, CNPq e CAPES) ****Modelo de equivalência de dados****

Classificar automaticamente as publicações do OpenAlex com base nas tabelas do CNPq e da CAPES



1ª Etapa - Modelo de compatibilização

- a) Correspondência automática entre os conceitos e as áreas CNPq e CAPES
- b) Classificação intelectual das áreas sem correspondência automática
- c) Validação final por dois bibliotecários
- d) Modelo de dados (JSON)

2ª Etapa - Sistema de classificação automática

- a) Classificador (requests, pandas, re, collections.Counter...)
- b) Interface web (Streamlit)
- c) Grafos (networkx, matplotlib, community (Louvain))



- 1 - Modelo de compatibilização (equivalência) de dados
- 2 - Classificador automático de publicações
- 3 - Interface web
- 4 - Análise de redes



```
{
  "category": "broad area",
  "cnpq_label": "Engenharias",
  "cnpq_label_eng": "Engineering",
  "cnpq_id": "3.00.00.00-9",
  "capes_label": "ENGENHARIAS",
  "capes_id": 30000009,
  "concepts": {
    "engineering": {
      "wikidata": [
        "https://www.wikidata.org/wiki/Q11023"
      ],
      "openalex": [
        {
          "id": "https://openalex.org/c127413603",
          "level": 0
        }
      ]
    }
  },
  "research_areas": [
    {
      "category": "research area",
      "cnpq_label": "Engenharia Civil",
      "cnpq_label_eng": "Civil Engineering",
      "cnpq_id": "3.01.00.00-3",
      "capes_label": "ENGENHARIA CIVIL",
      "capes_id": 30100003,
      "concepts": {
        "civil engineering": {
          "wikidata": [
            "https://www.wikidata.org/wiki/Q77590"
          ],
          "openalex": [
            {
              "id": "https://openalex.org/c147176958",
              "level": 1
            }
          ]
        }
      ]
    }
  ]
}
```

```
{
  "category": "sub area",
  "cnpq_label": "Anatomia",
  "cnpq_label_eng": "Anatomy",
  "cnpq_id": "2.06.04.00-9",
  "capes_label": "ANATOMIA",
  "capes_id": 20604009,
  "concepts": {
    "anatomy": {
      "wikidata": [
        "https://www.wikidata.org/wiki/Q514"
      ],
      "openalex": [
        {
          "id": "https://openalex.org/c105702510",
          "level": 1
        }
      ]
    }
  },
  "subjects": [
    {
      "category": "subject",
      "cnpq_label": "Anatomia Humana",
      "cnpq_label_eng": "Human Anatomy",
      "cnpq_id": "2.06.04.01-7",
      "capes_label": "ANATOMIA HUMANA",
      "capes_id": 20604017,
      "concepts": {
        "human anatomy": {
          "wikidata": [
            "https://www.wikidata.org/wiki/Q23852"
          ],
          "openalex": [
            {
              "id": "https://openalex.org/c2985650651",
              "level": 2
            }
          ]
        }
      ]
    }
  ]
}
```

CNPq/CAPES Classification System of OpenAlex Works

Search the OpenAlex works to be classified

Use the web search interface: <https://openalex.org/>

Or create an API request: <https://docs.openalex.org/how-to-use-the-api/api-overview>

Insert the OpenAlex API URL:

```
https://api.openalex.org/works?page=1&filter=title_and_abstract.search:isko&sort=relevance_score:de
```

Select the Brazilian classification system:

☒ CNPq

☐ CAPES

Compatibility data model loaded automatically.

Start Classification

Classification completed!

Unsaved search ▾



Show works where:

1 🔍 Title & Abstract includes ""Knowledge organization"" ✎ ✕

2 📅 and Year is 2024 ✎ ✕



API

https://api.openalex.org/works?page=1&filter=title_and_abstract.search:%22Knowledge+organization%22,publication_year:2024&sort=relevance_score:desc&per_page=10



Works



Stats



Knowledge organization

2024 · Rick Szostak · *Edward Elgar Publishing eBooks*

Cited by 6

Students' Satisfaction and Learning: Assessment of Teaching-Learning Process in Knowledge Organization

200 results



Open Access ⋮ ✕



Year



50%

100



2024

Classification completed!

		Broad Area (CNPq)	Area (CNPq)	Subarea (CNPq)
97	ologia;	Ciências Biológicas (0.50); Ciências Humanas (0.50)	Zoologia (0.33); Ciência da Computação (0.17)	Taxonomia dos Grupos Recentes (0.33); Zoologia (0.33)
181	litica	Ciências Humanas (0.40); Ciências Sociais Aplicadas (0.60)	Sociologia (0.10); Ciência Política (0.50)	Outras Sociologias Específicas (0.10); Sociologia (0.10)
175	litica	Ciências Humanas (0.45); Ciências Sociais Aplicadas (0.55)	Sociologia (0.09); Ciência Política (0.41)	Outras Sociologias Específicas (0.09); Sociologia (0.09)
159	isten	Ciências Humanas (0.83); Ciências Sociais Aplicadas (0.17)	Filosofia (0.33); Sociologia (0.50)	Outras Sociologias Específicas (0.33); Sociologia (0.33)
138	isten	Ciências Humanas (0.42); Ciências Exatas (0.58)	Matemática (0.25); Sociologia (0.25)	Outras Sociologias Específicas (0.25); Sociologia (0.25)
86	isten	Ciências Humanas (0.50); Ciências Exatas (0.50)	Filosofia (0.33); Sociologia (0.17)	Outras Sociologias Específicas (0.33); Sociologia (0.17)
23	isten	Ciências Humanas (0.50); Linguística, Letras e Artes (0.50)	Filosofia (0.17); Sociologia (0.33)	Outras Sociologias Específicas (0.17); Sociologia (0.33)
146	ginec	Ciências Humanas (0.50); Engenharias (0.50)	Sociologia (0.25); Educação (0.25)	Outras Sociologias Específicas (0.25); Sociologia (0.25)
107	; Poli	Ciências Sociais Aplicadas (0.33); Linguística, Letras e Artes (0.67)	Economia (0.17); Ciência Política (0.50)	Economia dos Recursos Humanos (0.17); Economia (0.17)
114	on); C	Engenharias (0.45); Ciências Sociais Aplicadas (0.55)	Engenharia Aeroespacial (0.10); Engenharia de Materiais (0.35)	Propulsão Aeroespacial (0.09); Propulsão Aeroespacial (0.09)



	work_id	concepts	↓ Broad Area (CNPq)	Area (CNPq)	Subarea (CNPq)	Subject (CNPq)
19	https://openalex.org/W439	Engineering; Civil engin	Engenharias (0.50); Ciências Agrárias (0.17);	Engenharia Agrícola (0.17); Ciência da	Engenharia de Água e Solo (0.17); Sar	Ecologia Aplicada à Engenhari
114	https://openalex.org/W440	Rocket (weapon); Com	Engenharias (0.45); Ciências Sociais Aplicad	Engenharia Aeroespacial (0.18); Econo	Propulsão Aeroespacial (0.09); Pesqui	Máquinas de Fluxo (0.09)
195	https://openalex.org/W440	Human capital; Econom	Ciências Sociais Aplicadas (1.00)	Economia (0.75); Administração (0.25)	Economia dos Recursos Humanos (0.1	Capital Humano (0.25); Cresci
29	https://openalex.org/W439	Marketing; Computer sc	Ciências Sociais Aplicadas (0.50); Ciências E	Administração (0.50); Ciência da Comp	Administração de Empresas (0.50)	Mercadologia (0.50)
71	https://openalex.org/W439	Computer science; Task	Ciências Sociais Aplicadas (0.50); Ciências E	Ciência da Computação (0.25); Engen	Engenharia do Produto (0.25); Teoria	Gerência do Projeto e do Prod
107	https://openalex.org/W440	Social capital; Political	Ciências Sociais Aplicadas (0.33); Lingüístic	Economia (0.17); Ciência Política (0.17	Economia dos Recursos Humanos (0.1	Capital Humano (0.17)
159	https://openalex.org/W440	Sociology; Epistemolog	Ciências Humanas (0.83); Ciências Sociais A	Filosofia (0.33); Sociologia (0.17); Ciên	Outras Sociologias Específicas (0.17);	Regionalização (0.17)
129	https://openalex.org/W440	China; Geography; Soci	Ciências Humanas (0.75); Ciências Sociais A	Demografia (0.25); Geografia (0.25); Sc	Política Pública e População (0.25); O	Política Populacional (0.25)
157	https://openalex.org/W440	Interpersonal communi	Ciências Humanas (0.75); Ciências Exatas e	Psicologia (0.75); Ciência da Computa	Psicologia Social (0.50)	Processos Grupais e de Comur
154	https://openalex.org/W440	Educational technology	Ciências Humanas (0.67); Ciências Exatas e	Educação (0.33); Ciência da Computaç	Ensino-Aprendizagem (0.33)	Tecnologia Educacional (0.33)
139	https://openalex.org/W440	Computer science; Soci	Ciências Humanas (0.67); Ciências Exatas e	Sociologia (0.22); Ciência da Computa	Outras Sociologias Específicas (0.11);	Regionalização (0.11)

Resultados



VIII ISKOBRA 2025



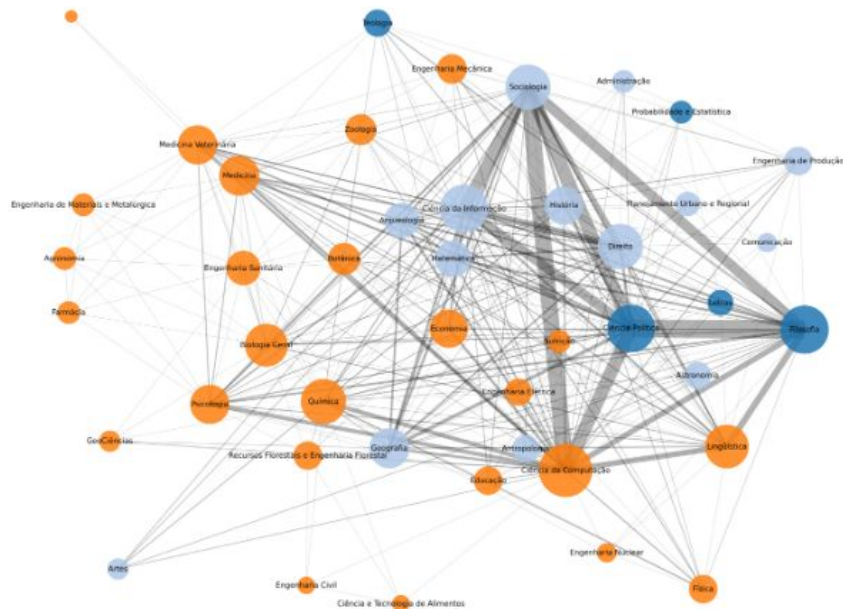
Select the level for the network:

Area

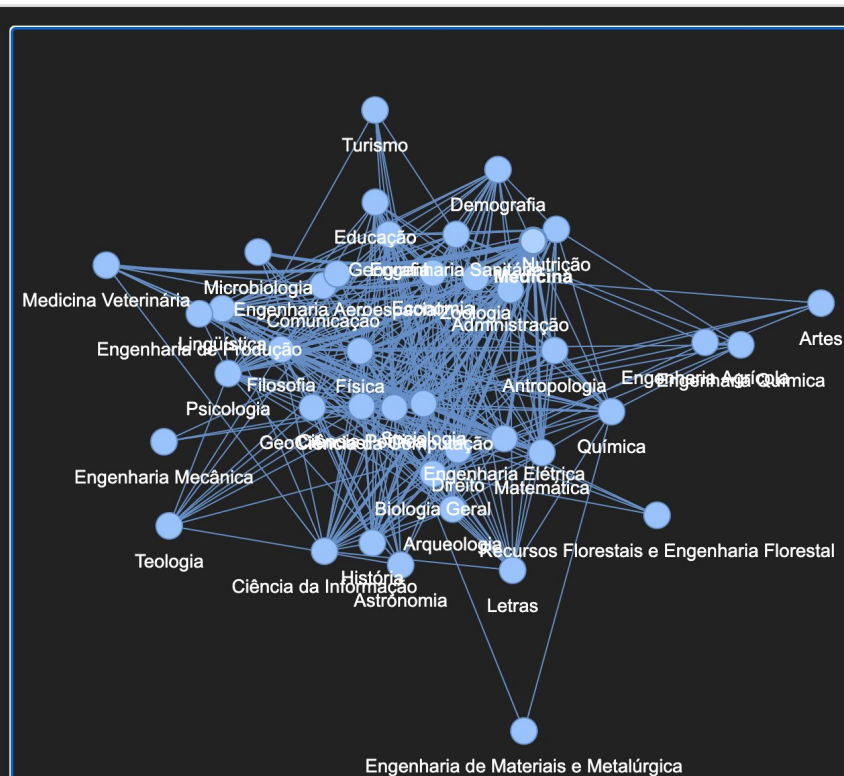


Create Classification Network

Co-occurrence Graph (Area) - System CNPq



Visualize Interactive Network (Zoom/Pan Enabled)



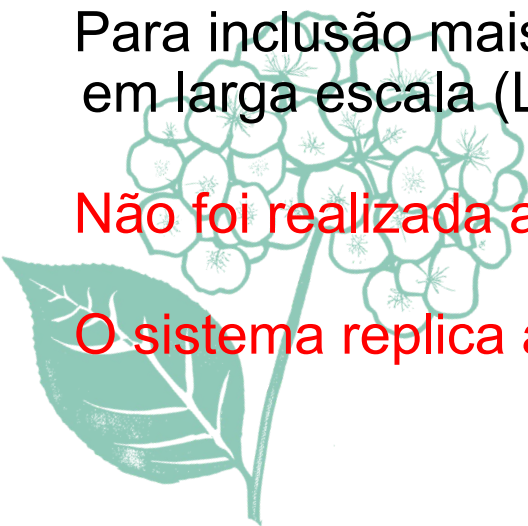
1ª fase limitada à construção do modelo de equivalência e do sistema de classificação

Foi garantida a equivalência de todas as áreas nos quatro níveis com pelos menos um conceito

Para inclusão mais ampla de conceitos é necessário um modelo mais em larga escala (LLM, ML...)

Não foi realizada avaliação (precisão, revocação...)

O sistema replica as limitações do modelo do OpenAlex!



Concepts foram substituídos por Topics no OpenAlex Modelo mais moderno e preciso (GPT + clusterização)

Leiden Madtrics

HOME CONTRIBUTORS ABOUT Q



Nees Jan van Eck
Senior Researcher and
Head of Data Science



Ludo Waltman
Professor of Quantitative
Science Studies



An open approach for classifying research publications

topics: [

{

```

    id: "https://openalex.org/T12028",
    display_name: "Knowledge Management and Sharing",
    score: 0.077,
    subfield: {
      id: "https://openalex.org/subfields/3315",
      display_name: "Communication"
    },
    field: {
      id: "https://openalex.org/fields/33",
      display_name: "Social Sciences"
    },
    domain: {
      id: "https://openalex.org/domains/2",
      display_name: "Social Sciences"
    }
  }

```

Fundep - pelo financiamento do projeto Laguna

CNPq - créditos AWS (Laguna Cloud)





Obrigado!
fabiocanto@ibict.br

