CENTRO DE COMPUTAÇÃO CIENTÍFICA E SOFTWARE LIVRE

Disponibilizando dados abertos educacionais

Marcos Didonet Del Fabro
marcos.ddf@inf.ufpr.br
www.inf.ufpr.br/didonet
www.c3sl.ufpr.br

Departamento de Informática
Universidade Federal do Paraná
Curitiba – PR

SABER 2019

Dados abertos

- Dados disponíveis para uso e publicação
  - Utilização livre, mas com direito autoral (quem produziu o dado original)
- Ciência: INPE, Projeto Genoma
  - Difusão da pesquisa
- Governo
  - Dados educacionais, econômicos, folha de pagamento, etc.
  - Possibilidade de monitoramento, auditoria, melhora, etc.
- Empresas privadas: balanços

- **PONTOS CHAVE:**
  - dados abertos de interesse público
  - respeito à privacidade

Princípios para dados abertos (https://opengovdata.org/)

- *Completos*
- *Primários*
- *Atuais*
- *Acessíveis*
- *Processáveis por máquina*
- *Acesso não discriminatório*
- *Formatos não proprietários*
- *Licenças livres*

Dados abertos no Brasil

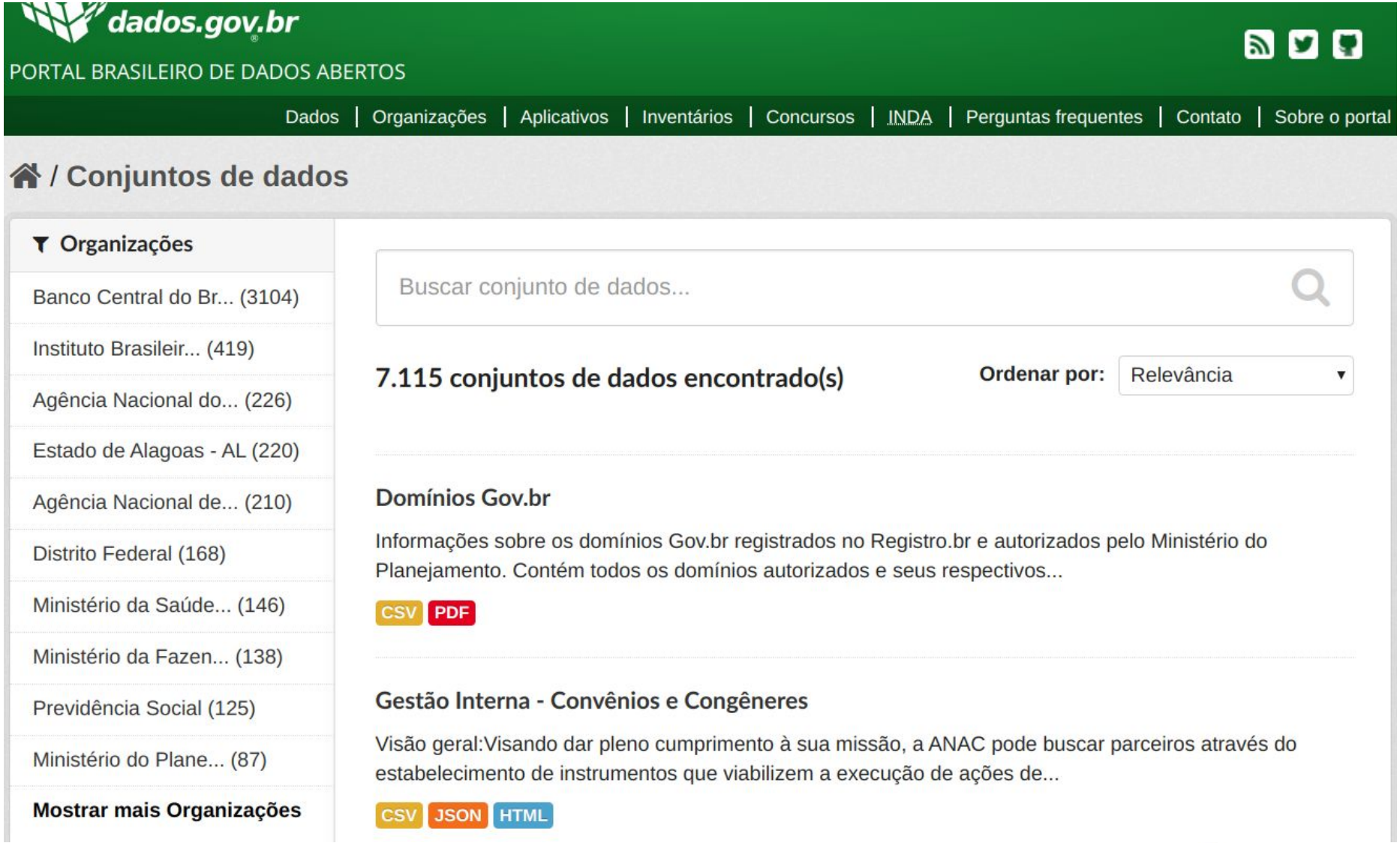
Legislação

por Cintia de Freitas Rodrigues Loureiro — publicado 17/10/2017 11h57, última modificação 17/10/2017 11h57

G+ Tweet

- DECRETO Nº 8.777, DE 11 DE MAIO DE 2016 – Institui a Política de Dados Abertos do Poder Executivo federal;
- DECRETO DE 15 DE SETEMBRO DE 2011 - Institui o Plano de Ação Nacional sobre Governo Aberto e dá outras providências;
- Instrução Normativa SLTI nº 4, de 12 de abril de 2012 – Institui a Infraestrutura Nacional de Dados Abertos – INDA;
- Lei nº 12.527, de 18 de novembro de 2011 – Lei de acesso à informação;
- Decreto nº 7.724, de 16 de maio de 2012 – Regulamenta a Lei nº 12.527/2011;
- Decreto nº 6.666, de 27 de novembro de 2008 – Institui a Infraestrutura Nacional de Dados Espaciais – INDE.

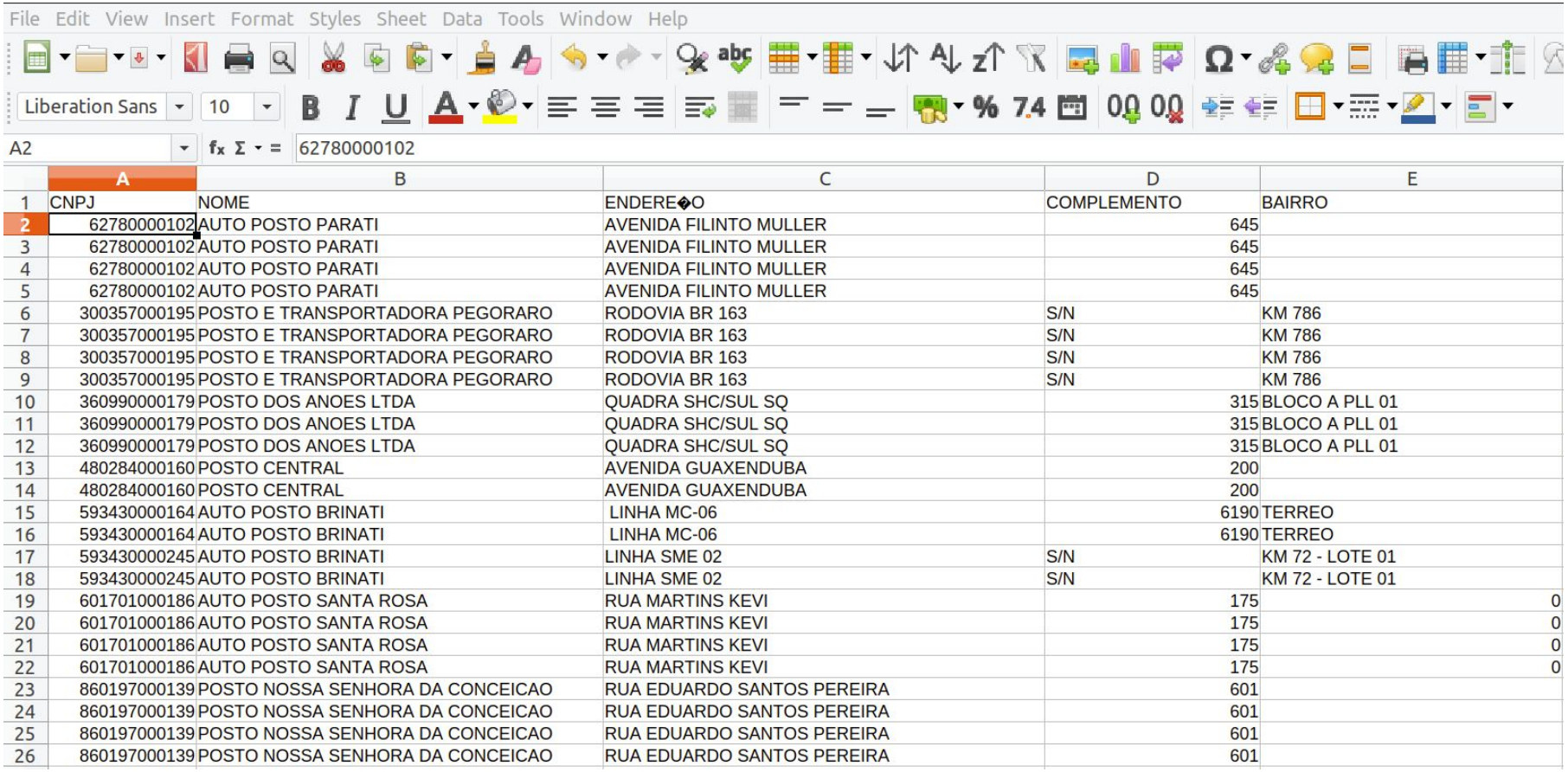
Dados abertos no Brasil (http://dados.gov.br)



Dados abertos no Brasil

- Muitos dados abertos disponíveis
  - Ordenar por **nome, relevância**???
- Muitas fontes
  - Ministérios, agências governamentais, universidades, órgãos públicos.
- Muitos formatos
  - CSV (Comma-Separated-Format), PDF, JSON, XLS, HTML, DOCX, etc.
- Muitos tamanhos
  - Alguns Kbs
    - Infopreço: 112Kb
  - Vários Gbs:
    - dados educacionais: arquivos com mais de 4Gb

CSV é um formato comum (*análise orientada a planilha...*)



	A	B	C	D	E
1	CNPJ	NOME	ENDERE◆O	COMPLEMENTO	BAIRRO
2	62780000102	AUTO POSTO PARATI	AVENIDA FILINTO MULLER	645	
3	62780000102	AUTO POSTO PARATI	AVENIDA FILINTO MULLER	645	
4	62780000102	AUTO POSTO PARATI	AVENIDA FILINTO MULLER	645	
5	62780000102	AUTO POSTO PARATI	AVENIDA FILINTO MULLER	645	
6	300357000195	POSTO E TRANSPORTADORA PEGORARO	RODOVIA BR 163	S/N	KM 786
7	300357000195	POSTO E TRANSPORTADORA PEGORARO	RODOVIA BR 163	S/N	KM 786
8	300357000195	POSTO E TRANSPORTADORA PEGORARO	RODOVIA BR 163	S/N	KM 786
9	300357000195	POSTO E TRANSPORTADORA PEGORARO	RODOVIA BR 163	S/N	KM 786
10	360990000179	POSTO DOS ANOES LTDA	QUADRA SHC/SUL SQ	315	BLOCO A PLL 01
11	360990000179	POSTO DOS ANOES LTDA	QUADRA SHC/SUL SQ	315	BLOCO A PLL 01
12	360990000179	POSTO DOS ANOES LTDA	QUADRA SHC/SUL SQ	315	BLOCO A PLL 01
13	480284000160	POSTO CENTRAL	AVENIDA GUAXENDUBA	200	
14	480284000160	POSTO CENTRAL	AVENIDA GUAXENDUBA	200	
15	593430000164	AUTO POSTO BRINATI	LINHA MC-06	6190	TERREO
16	593430000164	AUTO POSTO BRINATI	LINHA MC-06	6190	TERREO
17	593430000245	AUTO POSTO BRINATI	LINHA SME 02	S/N	KM 72 - LOTE 01
18	593430000245	AUTO POSTO BRINATI	LINHA SME 02	S/N	KM 72 - LOTE 01
19	601701000186	AUTO POSTO SANTA ROSA	RUA MARTINS KEVI	175	0
20	601701000186	AUTO POSTO SANTA ROSA	RUA MARTINS KEVI	175	0
21	601701000186	AUTO POSTO SANTA ROSA	RUA MARTINS KEVI	175	0
22	601701000186	AUTO POSTO SANTA ROSA	RUA MARTINS KEVI	175	0
23	860197000139	POSTO NOSSA SENHORA DA CONCEICAO	RUA EDUARDO SANTOS PEREIRA	601	
24	860197000139	POSTO NOSSA SENHORA DA CONCEICAO	RUA EDUARDO SANTOS PEREIRA	601	
25	860197000139	POSTO NOSSA SENHORA DA CONCEICAO	RUA EDUARDO SANTOS PEREIRA	601	
26	860197000139	POSTO NOSSA SENHORA DA CONCEICAO	RUA EDUARDO SANTOS PEREIRA	601	

Mas é necessário conhecimento para encontrar a informação relevante

Dados educacionais

- INEP é um grande produtor de dados: http://portal.inep.gov.br/microdados
  - Censo Escolar, Educação Superior, Enad, ENEM, etc.
- Arquivo de matrículas
  - ~4Gb
  - + 50 milhões de linhas, por ano
  - Tabelas com mais de 100 colunas

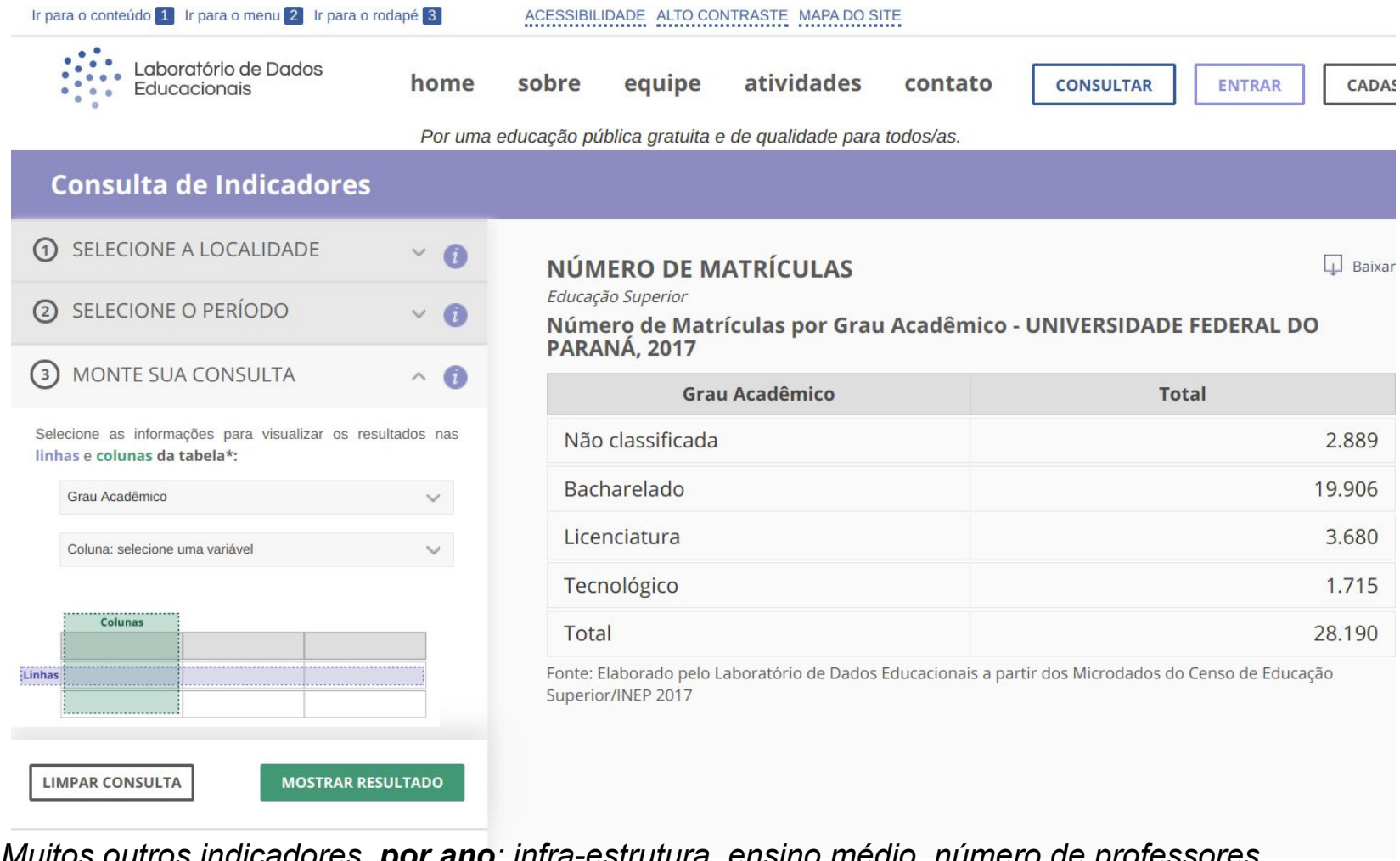
- Abrir em uma planilha não parece uma boa solução !!!

Projeto Laboratório de Dados Educacionais (LDE)

- Plataforma usando **dados educacionais**
  - Todos os estágios (creche até ensino superior)
  - Múltiplos segmentos interessados
- Indicadores educacionais para
  - Planejamento
  - Pesquisa
  - Criação de políticas públicas
  - Publicização da informação

LDE : número de matrículas no ensino superior

http://dadoseducacionais.c3sl.ufpr.br

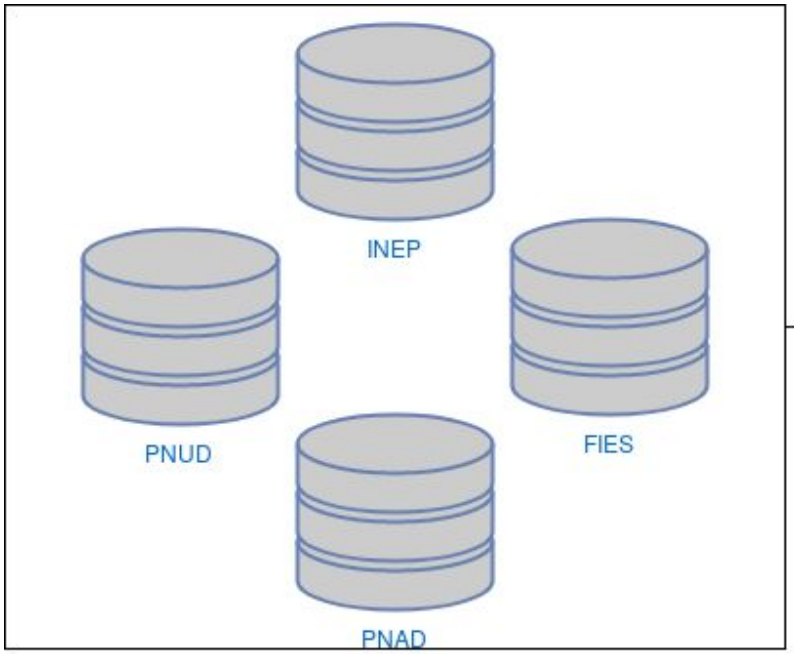


*Muitos outros indicadores, **por ano**: infra-estrutura, ensino médio, número de professores, escolas, etc.*

Projetos de dados abertos

- Esforços conjuntos
  - Especialistas do domínio
    - Laboratório de Dados Educacionais
  - Especialistas em computação
    - C3SL/UFPR
- Equipe
  - ~15 integrantes, entre alunos e professores
- Não é UM cientista de dados, mas vários atores para fazer ciência sobre os dados

Fluxo do laboratório



Extract the data

Transform/Evolve the data

Load

Model and develop metrics

Publish new metrics

Requisitos dos dados

Dados abertos:

- Várias fontes
- Dado atualizado anualmente
  - Algumas fontes são mensais
- Nomes das colunas e seus valores variam
  - Dado deve ser mantido consistente ao longo do tempo
- Tabelas com mais de 150 colunas e milhões de registros

- Como modelar os dados?
- Como evoluir os dados?
  - Compromisso entre normalização e evolução/manutenção

Requisitos da API

API Web

- Filtros e combinação de dimensões
- Geração de consultas
- Reutilização por diferentes clientes web
- Novos indicadores lançados periodicamente

Client web

- Usuários com objetivos diferentes
- Interface com usuário amigável
  - Diferentes resultados (somas, médias, porcentagens, etc.)

Escolha do banco de dados e do modelo de dados

- Dado é extraído e incluído em um SGBD
- Todos os anos novos CSV são disponibilizados
- Dado não é normalizado
- Não há controle no nome dos campos e compatibilidade dos dados

Escolhendo o sistema de banco de dados

Benchmark TPC-H

Consulta	Tempo em ms		Desempenho do MonetDB comparado com PostgreSQL
	PostgreSQL	MonetDB	
1	14583	741	1968.02%
3	2056	2067	99.47%
4	942	234	402.56%
5	1406	1278	110.02%
6	1665	179	930,17%
10	3446	908	379.52%
12	2567	330	777.88%
14	1677	86	1950.00%
16	2047	204	1003.43%
19	2143	136	1575.74%

Evolução de dados e esquema

- Infraestrutura: a informação evolui

	2009	2010	2011	2012	2013
Internet Banda larga	???	65%	68%	73%	73%

	2019	2020	2021	2022	2023
Piscina	???	30%	27%	28%	28%

	2019	2020	2021	2022	2023
Internet discada	15%	8%	???	???	???

Evolução de dados e esquema

- Dados de-normalizados
- Mapeamentos diretos

```
NACIONALIDADE <- [2013-2017] NACIONALIDADE
```

- Mapeamentos com traduções

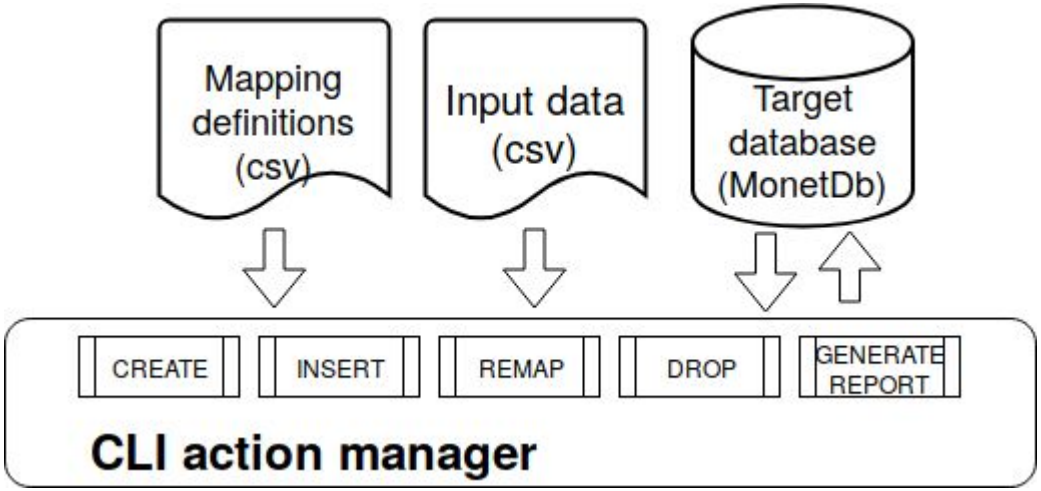
```
NECESSIDADES_ESPECIAIS <- [2013-2014] TEM_NECESSIDADE
NECESSIDADES_ESPECIAIS <- [2015-2017] NECISSIDADE_ESP
REGIAO <- [2013-2015] não-disponível
REGIAO <- [2016-2017] REGIAO
```

- Evolução dos dados

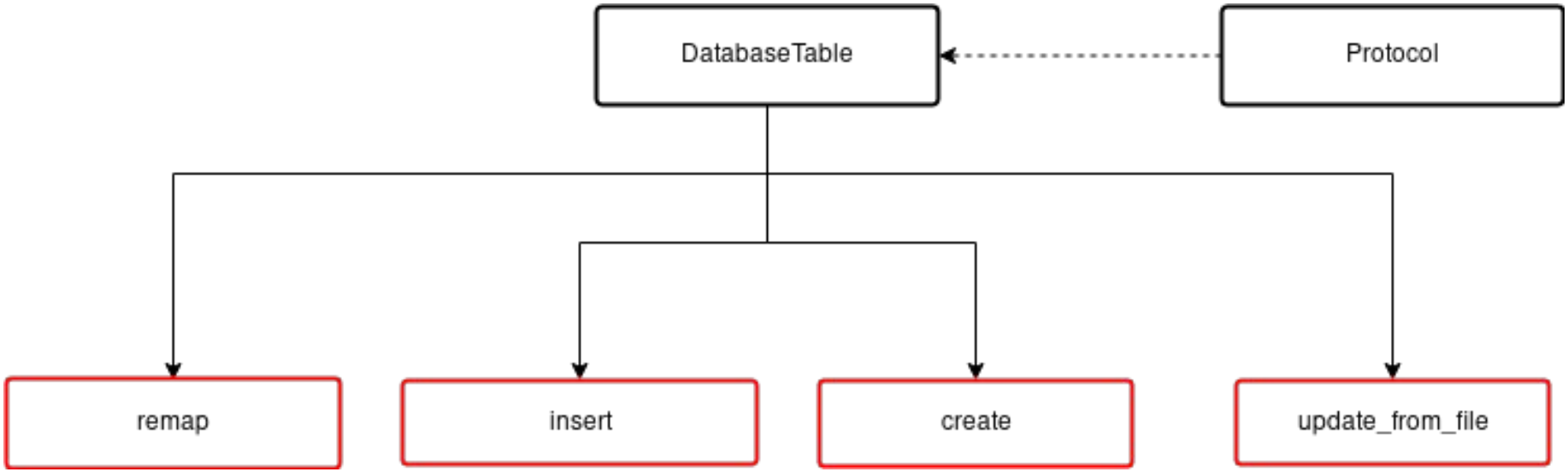
```
PROFISSIONALIZANTE <-[2013-2014]
    WHEN TIPO_ESTUDO between 30 and 40 THEN 1
    WHEN TIPO_ESTUDO between 41 and 50 THEN
```

Mapeamento através de scripts

- Mapeamentos em CSV
  - Subconjunto de SQL - expressões CASE
- CLI (Command Line Interface) fácil de usar
- Diferentes bancos podem ser plugados



Ações do CLI de mapeamento



Mapeamento: o ano é uma coluna

Nome	Nome padrão	Tipo de dado	2014	2015
ANO	NU_ANO_CENSO	INT	NU_ANO_CENSO	NU_ANO_CENSO
CEBES002N0	CO_ENTIDADE	INT	PK_CO_ENTIDADE	CO_ENTIDADE
CEBES003N0	NO_ENTIDADE	VARCHAR(256)	NO_ENTIDADE	NO_ENTIDADE

```
~CASE WHEN (cod_escolaridade = 1 OR cod_escolaridade = 2) THEN 1

WHEN (cod_escolaridade = 3) THEN 2 WHEN (cod_escolaridade = 4 OR
cod_escolaridade = 5) THEN 3

WHEN (cod_escolaridade = 6 AND "ID_DOUTORADO" = 1) THEN 8

WHEN (cod_escolaridade = 6 AND "ID_MESTRADO" = 1) THEN 7

WHEN (cod_escolaridade = 6 AND "ID_ESPECIALIZACAO" = 1) THEN 6

WHEN (cod_escolaridade = 6 AND ("ID_LICENCIATURA_1" = 1 OR
"ID_LICENCIATURA_2" = 1 OR "ID_LICENCIATURA_3" = 1)) THEN 5

WHEN (cod_escolaridade = 6) THEN 4 END
```

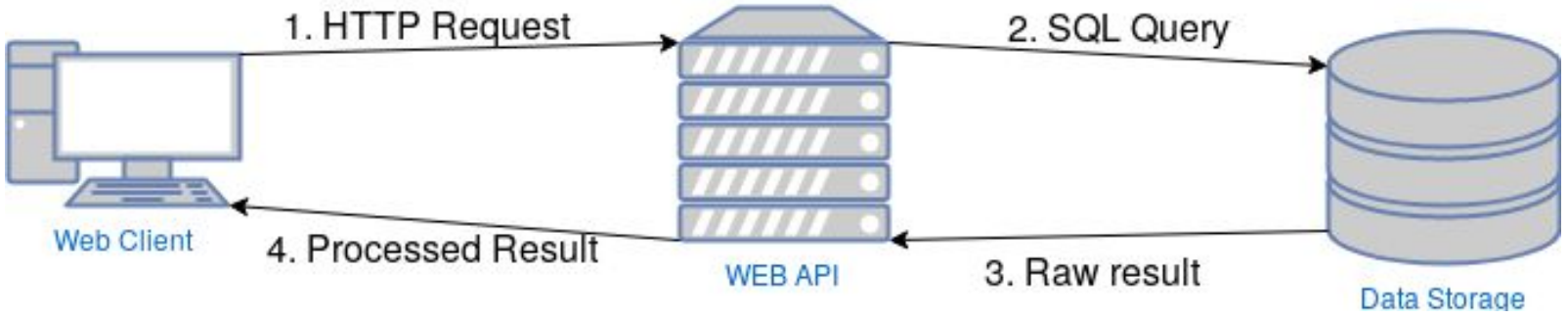
Descrição das métricas

- Comunicação com os especialistas do domínio
- Possibilidade de validação, auditoria e rastreamento
- A fonte sempre deve estar claramente descrita

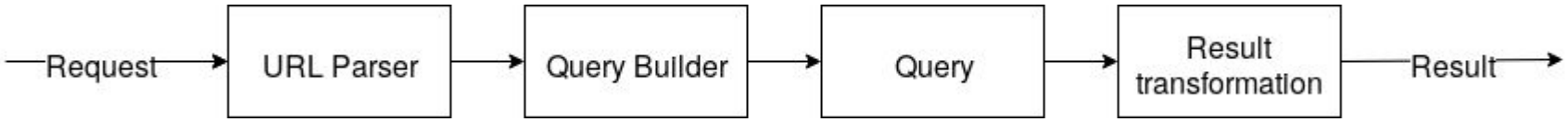
Nome	Dimensões & filtros	Periodicidade	Cálculo	Outros	Source
Número de matrículas	• Cidade, estado, região • Turma • Ano	Anualmente	Total de matrículas	Número de estudantes matriculados	Censo Superior/INEP (2010-2018)

WEB API

- NodeJs + ExpressRequisição para uma métrica
  - `HTTP GET site.com/api/enrollment?filter=min_year:2016`
- Tradução para consulta SQL -> tradução direta
- Requisição é enviada
- Retorno em JSON ou CSV



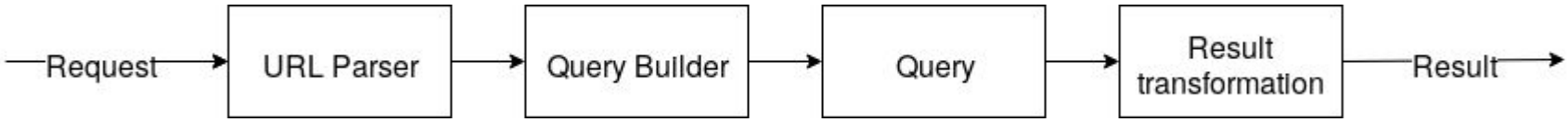
API flow



1. **Requisição**
   a. https://simcaq.c3sl.ufpr.br/api/v1/enrollment?dims=education_level,state&filter=min_year:2018,max_year:2018

2. **Parser:** identifica filtros e dimensões
   a. Dimensões: education_level and state
   b. Filtros: min_year (2018) and max_year (2018)

Definição dos elementos da consulta

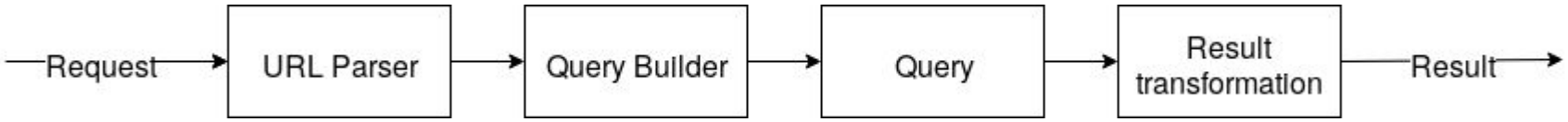


Definição dos elementos válidos: filtros e dimensões

```
query.addField({
    name: 'dims',
    field: true
}).addValue({
    name: 'education_level',
    table: 'matricula',
    tableField: 'etapa_ensino_id',
    resultField:
'education_level_id',
    where: {
        relation: '=',
        type: 'integer',
        field: 'etapa_ensino_id'
    }
)
```

```
{
    name: 'min_year',
    table: 'matricula',
    tableField: 'ano_censo',
    resultField: 'year',
    where: {
        relation: '>=',
        type: 'integer',
        field: 'ano_censo'
    }
}
```

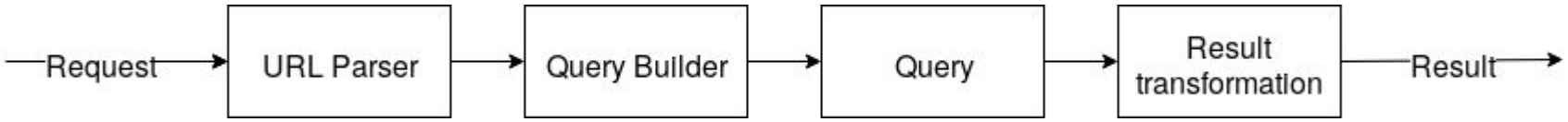
Construção da consulta



Construção da consulta em SQL

```
SELECT
    matricula.etapa_ensino_id AS "etapa_ensino_id",
    estado.nome AS "state_name",
    COUNT(*) AS "total",
    'Brasil' AS "name",
    matricula.ano_censo AS "year"
FROM matricula INNER JOIN
    estado ON (matricula.estado_id=estado.id)
WHERE (matricula.ano_censo >= 2018 )
    AND (matricula.ano_censo <= 2018 )
    AND (matricula.etapa_ensino_id=1)
GROUP BY matricula.etapa_ensino_id, estado.nome,
    matricula.ano_censo
ORDER BY matricula.etapa_ensino_id ASC, estado.nome ASC,
    matricula.ano_censo ASC;
```

WEB API response



Resultado em JSON, XML ou CSV

```
{
    “result”: [
        {
            “education_level_id”:1,
            “state_name”: “Acre”,
            “total”: 11749,
            “name”: “Brasil”,
            “year”: 2018,
            “school_year_name”: “Educação Infantil - Creche”
        },
        ...
    ]
}
```

Interface WEB

- React e Redux

Cliente apenas necessita montar a requisição para a API

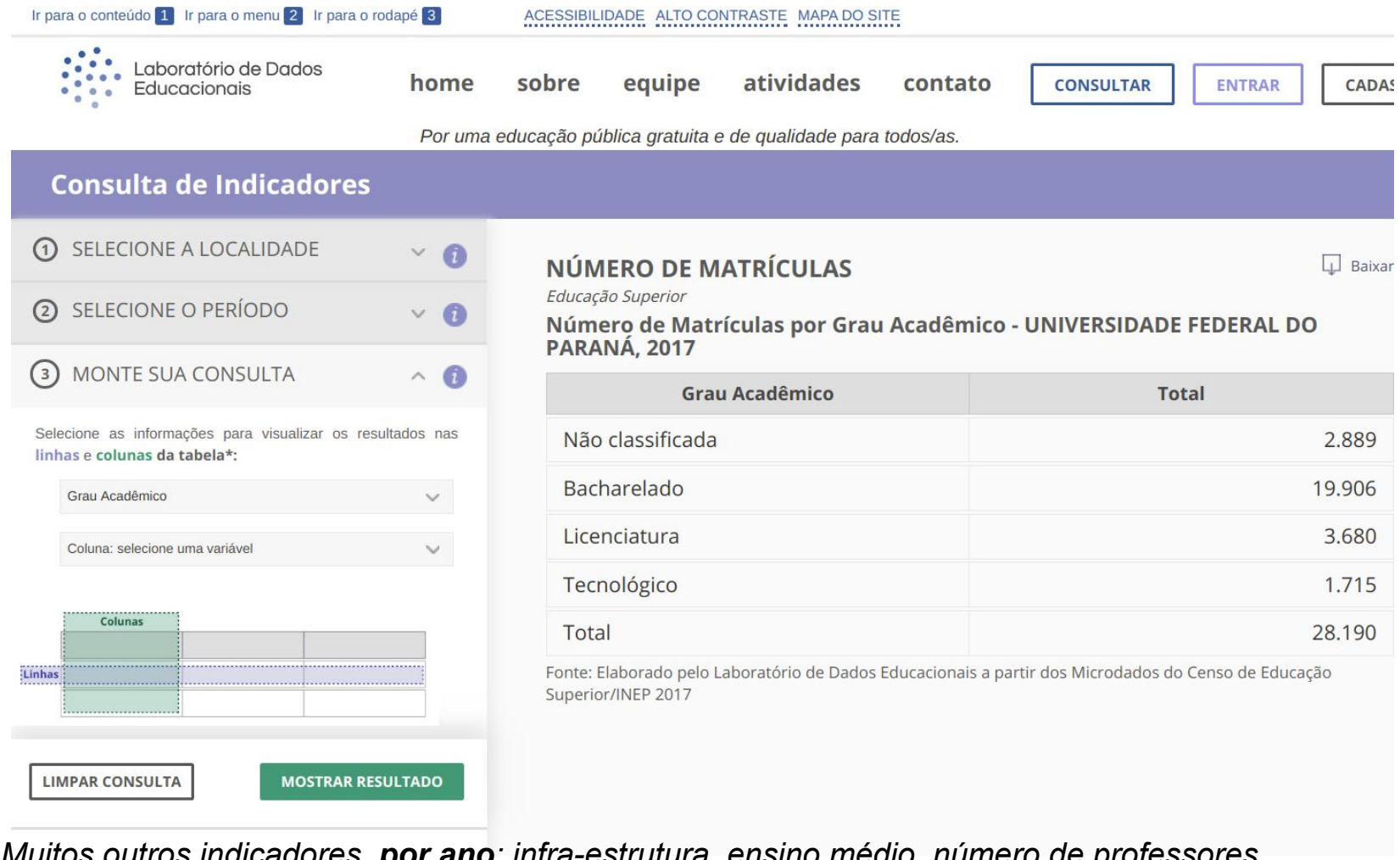
```
https://SITE/api/METRIC?dims=FIRST,SECOND,THIRD,N
    &filters=FILTER1:VALUE1, FILTER2:VALUE2,
FILTERN:VALUEN
    &format=JSON (or CSV or XML)
```

- Receber o resultado e montar a tela final

https://simcaq.c3sl.ufpr.br/api/v1/enrollment?dims=education_level,state&filter=min_year:2018,max_year:2018

LDE : número de matrículas no ensino superior

https://simcaq.c3sl.ufpr.br/api/v1//university_enrollment?dims=academic_level&filter=min_year:%222017%22,max_year:%222017%22



*Muitos outros indicadores, **por ano**: infra-estrutura, ensino médio, número de professores, escolas, etc.*

Conclusões

- Disponibilização de indicadores criados a partir de dados abertos é uma tarefa difícil
  - Muitas fontes de informação e formatos
  - Comunicação entre especialistas é **ESSENCIAL**
  - Não há garantia de consistência ao longo dos anos
  - Modelagem e evolução dos dados deve ser simples, para permitir atualização futura
    - Problema do “atualizado hoje”
- Uma única métrica pode ser simples, várias não
  - Indicadores novos, novos dados
- API REST separada do front-end
- Abordagem viável a atualmente funcional
  - http://dadoseducacionais.c3sl.ufpr.br
    - https://dadoseducacionais.c3sl.ufpr.br/#/indicadores

www.c3sl.ufpr.br