SPRINT REPORT
Crawling data portals for improving research communication
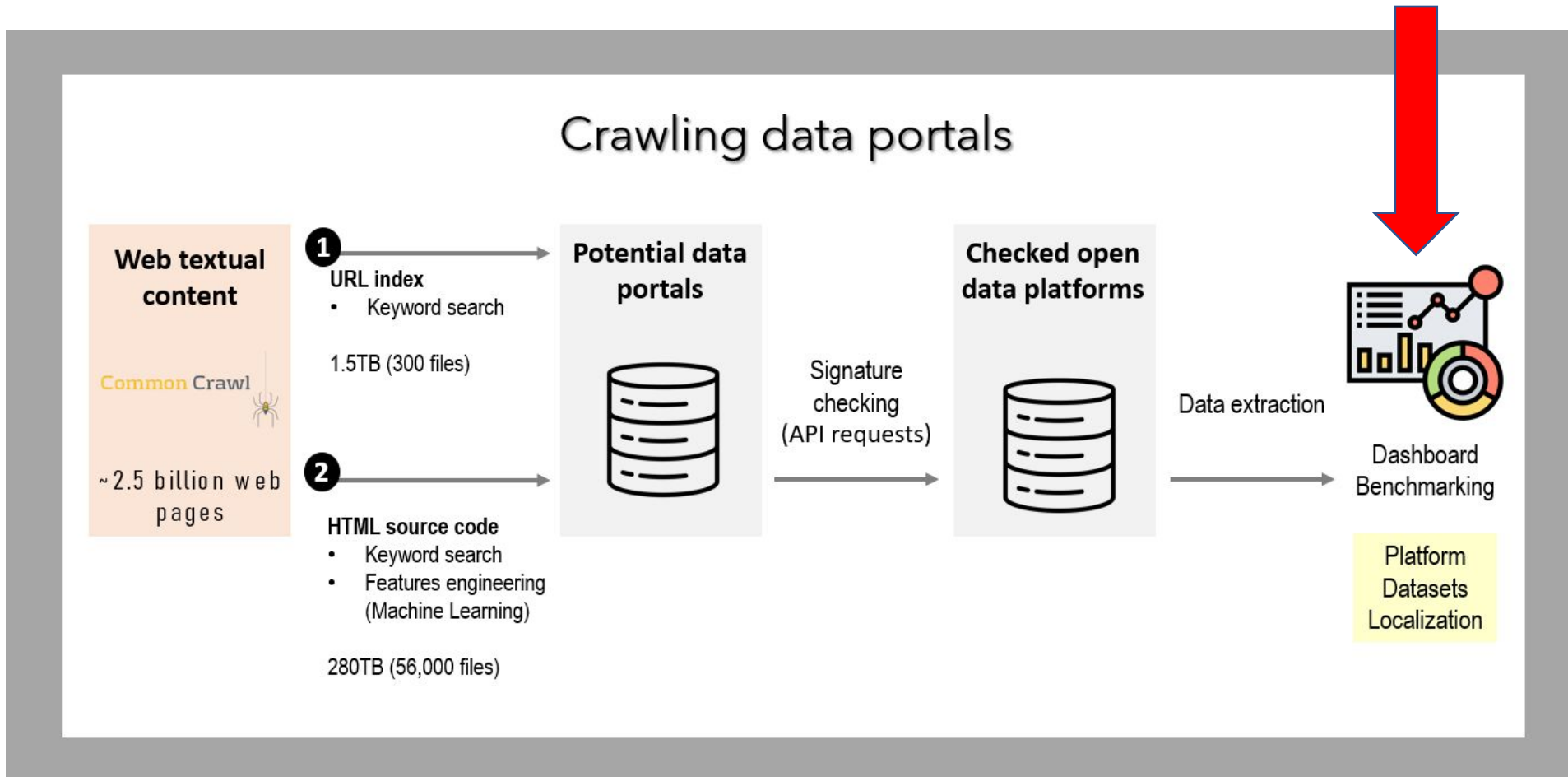
LED BY ANDREIWID CORREA AND KELLYTON BRITO

{andreiwid@gmail.com; kellyton@kellyton.com.br}

The whole project's big picture

Sprint aim

Crawling data portals

Web textual content

Common Crawl

~2.5 billion web pages

1 URL index
- Keyword search

1.5TB (300 files)

2 HTML source code
- Keyword search
- Features engineering (Machine Learning)

280TB (56,000 files)

Potential data portals

Signature checking (API requests)

Checked open data platforms

Data extraction

Dashboard Benchmarking

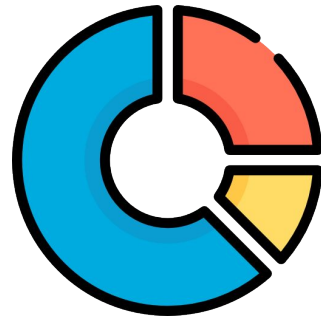Platform Datasets Localization

Project's aim

Produce open, accessible, and user-friendly data ready for reuse by the community
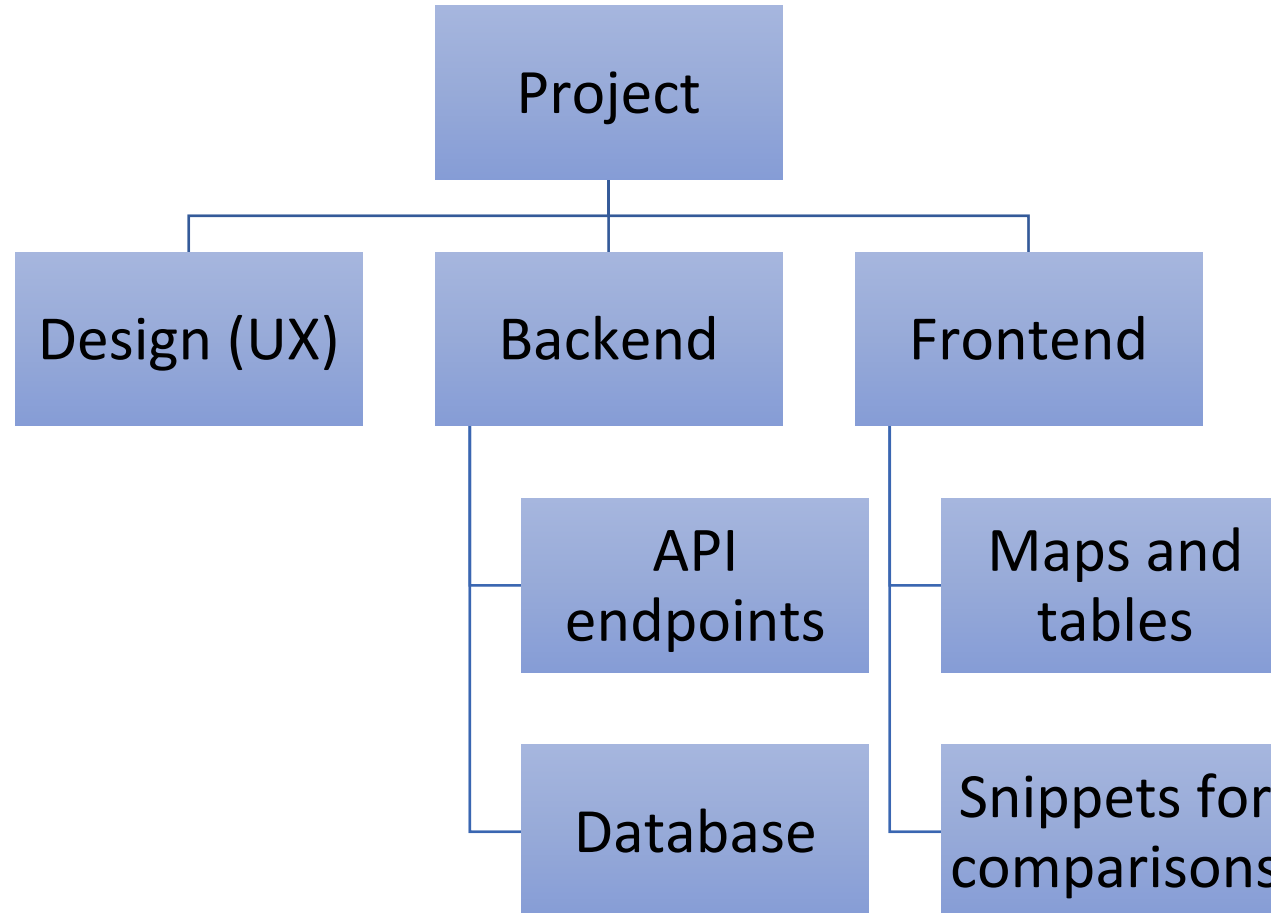
REST APIs Charts Comparisons Data reuse

Project description

Achievements

PROJECT PLANNING

Backend | 6

API Endpoint 3
Get all data portals only from latest crawling

Crawling: Get more data from portals

API Endpoint 4:
Get data regarding portals from one specific country

Andreiwid

Setup Flask & SQLite

Andreiwid

API Endpoint 1:
Get all data portals from the datasets
http://54.73.15.66/dataportals/

Andreiwid

API Endpoit 2
Get data portals of a specific year and month. 2019 means <year> and 8 means <month>
http://54.73.15.66/dataportals/2019/8

Andreiwid

Backlog | 5

Frontend | 9

Main page: number of portals by platform

Main page: Average number of datasets in country details

Main page: Max and Min number of datasets in country portals (country details)

In progress | 4

Main page: Refinements on contry details

Kellyton Brito

Integrating static pages with flask

Raoni Lourenço

Done | 9

Main Page: Clickable map with country details (number of portals and datasets)

Kellyton Brito

Main Page:
Bubble Map with Total of portals

Kellyton Brito

Setup Highcharts & Graphs (test)

Kellyton Brito
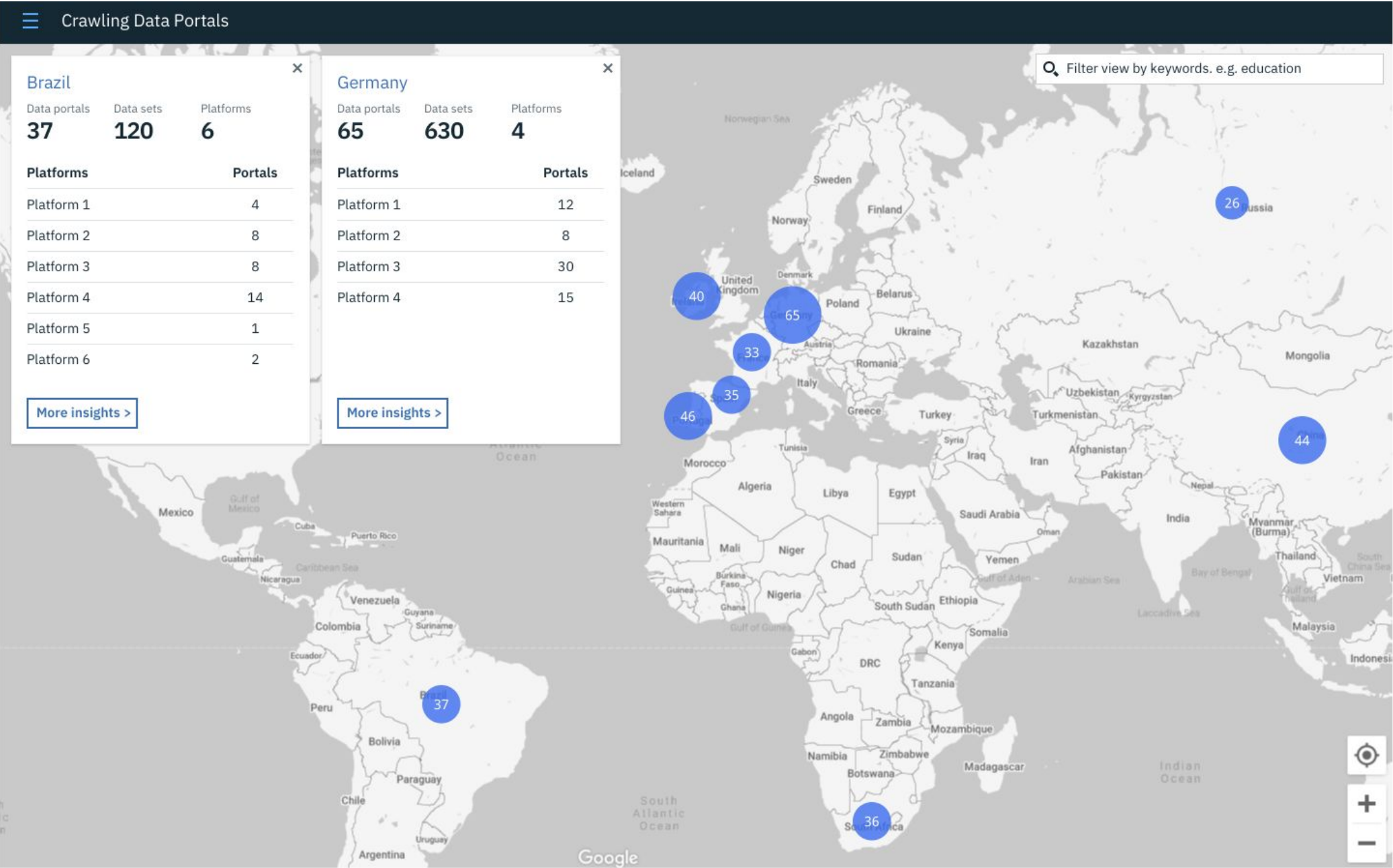
Setup Carbon charts & Graphs

Raoni Lourenço

Design | 3

Refinements on design

Sketch main page with a map

Sketch country details page

DESIGN

DESIGN

Brazil, filtered by education

View month

📅 September 2010

Data portals	Data sets	Platforms
37	**120**	**6**

Platforms	Portals
Platform 1	4
Platform 2	8
Platform 3	8
Platform 4	14
Platform 5	1
Platform 6	2

Six Month Data Portal Growth

Key

■ Portal growth

Data portals

🔍 Search data table

Portal url ▲	Portal description	Platform	Data sets	Status
https://data.nhm.ac.uk/	Data Portal. Explore and download the Museum's research and collections data. An open source...	Name	8	Active
https://data.nhm.ac.uk/	Data Portal. Explore and download the Museum's research and collections data. An open source...	Name	4	Active
https://data.nhm.ac.uk/	Data Portal. Explore and download the Museum's research and collections data. An open source...	Name	2	Active
https://data.nhm.ac.uk/	Data Portal. Explore and download the Museum's research and collections data. An open source...	Name	5	Active
https://data.nhm.ac.uk/	Data Portal. Explore and download the Museum's research and collections data. An open source...	Name	8	Active

APIs

http://54.73.15.66/dataportals/

http://54.73.15.66/dataportals/2019/8/

```
▼ [
    ▼ [
          "1337",
          "8",
          "2019",
          "2019-10-29 03:51:32.207",
          "0",
          "URL460364",
          "https://2020census-bucksgis.opendata.arcgis.com",
          "https://2020census-bucksgis.opendata.arcgis.com",
          "ARCGIS",
          "-1",
          "(not available)",
          "United States",
          "United States of America",
          "US",
          "2020census-bucksgis.opendata.arcgis.com",
          "1"
    ],
    ▼ [
          "1336",
          "8",
          "2019",
          "2019-10-29 03:51:29.767",
          "0",
          "URL460363",
          "http://2020census-bucksgis.opendata.arcgis.com",
          "https://2020census-bucksgis.opendata.arcgis.com/",
          "ARCGIS",
          "-1",
          "(not available)",
          "United States",
          "United States of America",
          "US",
          "2020census-bucksgis.opendata.arcgis.com",
          "2"
    ],
    ▼ [
          "1338",
          "8",
          "2019",
          "2019-10-29 03:51:34.730",
          "0",
          "URL460365",
          "http://aboutus-bucksgis.opendata.arcgis.com",
          "http://aboutus-bucksgis.opendata.arcgis.com",
          "ARCGIS",
          "-1",
          "(not available)",
          "United States",
```

Next steps

- Finish frontend development

- Implement missing APIs

- Go live with the project!

Contributors

- Andreiwid Correa
- Kellyton Brito
- Raoni Lourenço
- Nick Duffield