Useful Information

- Current active team:
  - Andreiwid, Kellyton, Nick, Raoni
- Initial dataset: https://github.com/Andreiwid/2020elifesprint/tree/master/datasets
- Miro: https://miro.com/app/board/o9J_kmIOog4=/
- Interactive demo https://invis.io/8KYMTCPADM9#/430479160_map-3
- Code https://github.com/Andreiwid/2020elifesprint
- Sprint report https://docs.google.com/presentation/d/1oCKvHZoSii9xTH5HZYcLs9R53WpWs_ZK3vowt5tndJQ/edit?usp=sharing

Crawling data portals for improving research communication

LED BY ANDREIWID CORREA AND KELLYTON BRITO
{andreiwid@gmail.com; kellyton@kellyton.com.br}

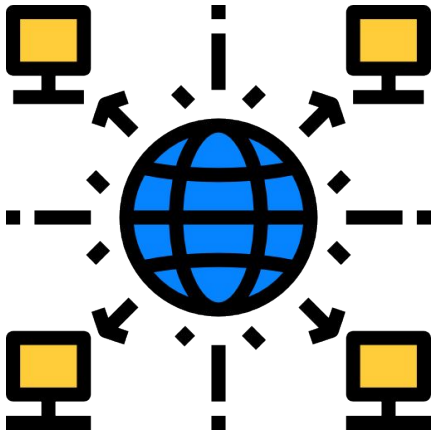
eLife Sprint Project
https://sprint.elifesciences.org/crawling-data-portals-for-improving-research-communication/

Problem

- Data portals appears/disappears everyday in the web
- It’s difficult for users to remain up to date about data portals availability
- There is no global repository of open data portals
- Benchmarking studies demand manual efforts

The solution

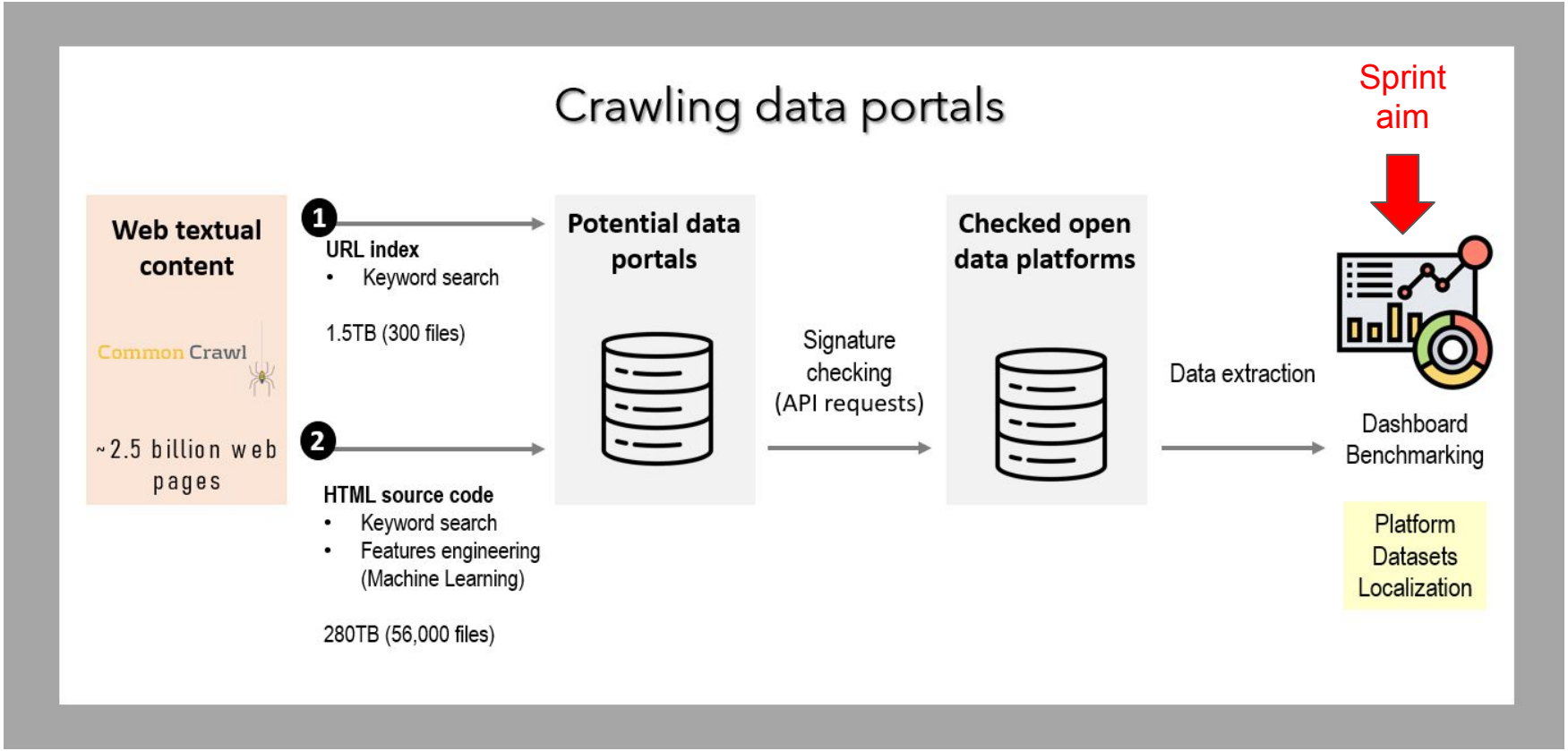
- A single, up-to-date, and reliable source (as a repository) of data portals available worldwide
- No need for manual intervention for data portals discoverability



Some inspiring projects

- Community-driven projects
  - http://dataportals.org (manual-based)
  - https://opendatainception.io (manual-based)
  - https://datashades.info (only CKAN instances)
- Academic project
  - https://data.wu.ac.at/portalwatch (focused data portal harvesting)
- Funded project
  - https://www.opendatamonitor.eu (EU only, focused data portal harvesting)
- Repositories on Github
  - https://github.com/dadosgovbr/catalogos-dados-brasil
  - https://github.com/ckan/ckan-instances
  - https://github.com/sunlightpolicy/opendata

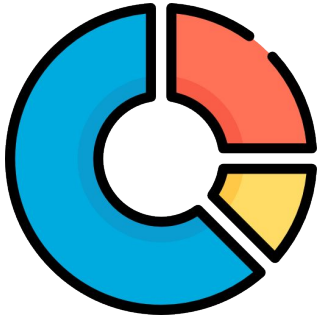
The whole project's big picture



Work at the Sprint

Produce open, accessible, and user-friendly data ready for reuse by the community

REST APIs



Charts

Comparisons

Data reuse

What data we have to work on the Sprint?

- Crawled data portals from 2019-8 to 2020-2
- Each CSV file represents a month (data is collected in a monthly basis)
- Data can show the evolution of data portals over the period considering:
  - The number of added and removed data portals
  - The increase in the number of datasets
  - The market share of each of the 8 considered open data software platforms
  - The country where data portals are used (when available)
- Raw data and metadata can be downloaded at Github
  - https://github.com/Andreiwid/2020elifesprint/tree/master/datasets
  - See README.md for info about each collected data (metadata)

What features can we develop?

- Client-side visualizations to answer e.g. the following questions:
  - What are *my* country's data portals?
  - How *my* country is evolving in terms of data portals?
  - How the world is evolving? Can we show a map where we can find more data portals?
  - What are the main used open data software platforms?
  - How my country compares to others?
- API endpoint to consume, reuse and republish:
  - Raw data to ensure interoperability
  - Data from visualization layers developed above
- (DREAM) Further data acquisition from cataloged data portals for:
  - Improving *my* research
  - Getting whatever data we want

Interested? Next steps

- Join our channel on Slack #crawling-dataportals
- Take a look at datasets
  - https://github.com/Andreiwid/2020elifesprint/tree/master/datasets
- See an inspiring similar project of what we can start from
- Starting talking to us!