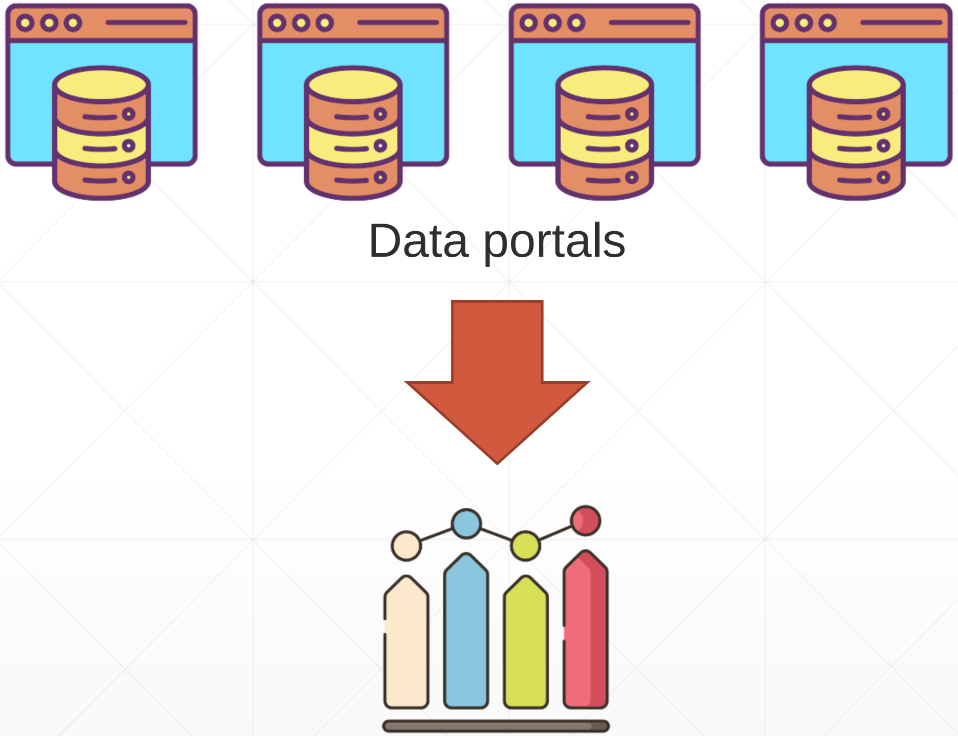
Laying the foundations for benchmarking open data automatically: a method for surveying data portals from the whole web

20th Annual International Conference on Digital Government Research (dg.o 2019)

Dubai, UAE, June 18-20, 2019
Presenter: Andreiwid Sheffer Correa – http://andreiwid.info

Research problems

- Growing number of data portals worldwide
- Benchmarking: demands for evidence
- Hardworking process to find data portals
- Manage fast changing context of open data
- Suggests a whitelist of healthy data portals

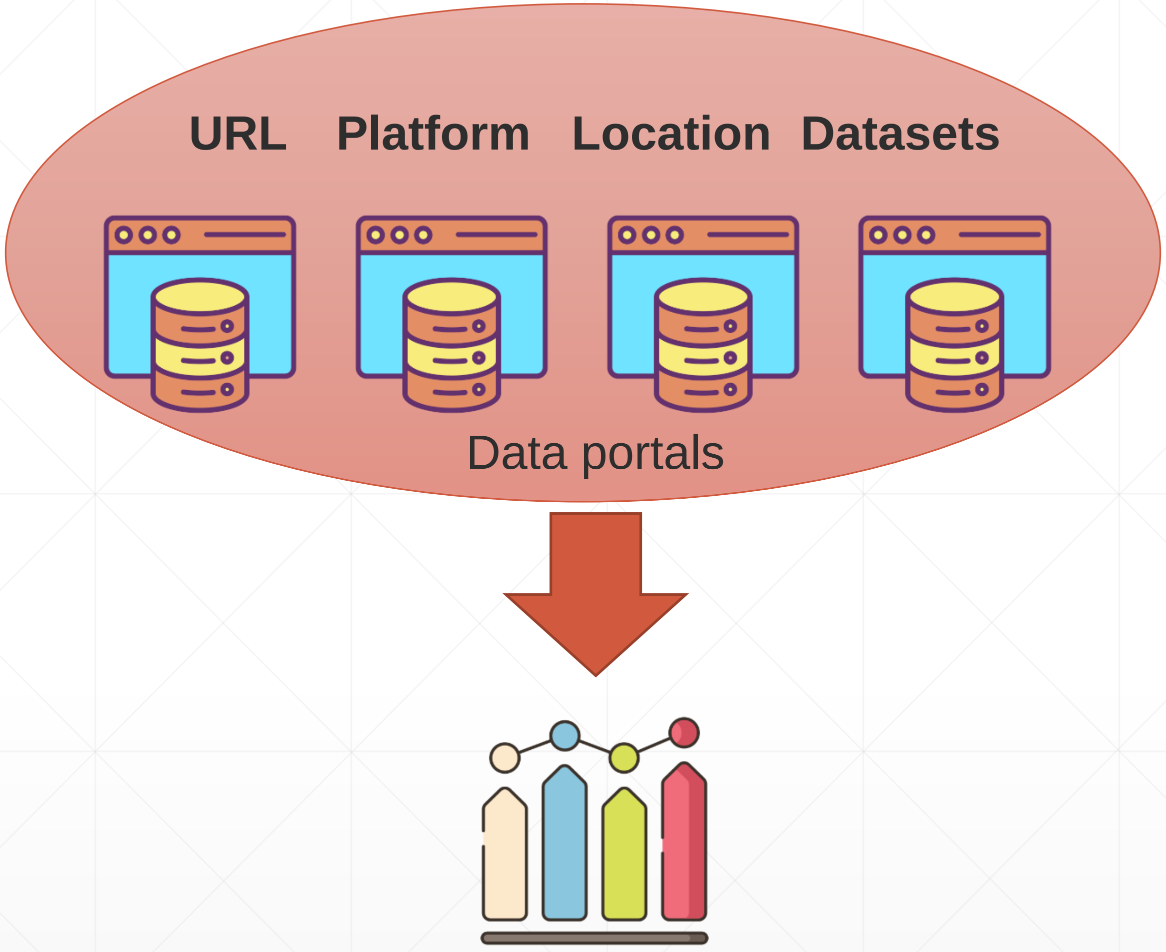


Data portals

Benchmark exercises

Research problems

- How about a global repository?
- We could know:
  - Data portal web address (published? working?)
  - Software platform in use (standardized?)
  - Geographic location (country?)
  - Dataset inventory (how many? are they updated? …)



URL Platform Location Datasets

Data portals

Benchmark exercises

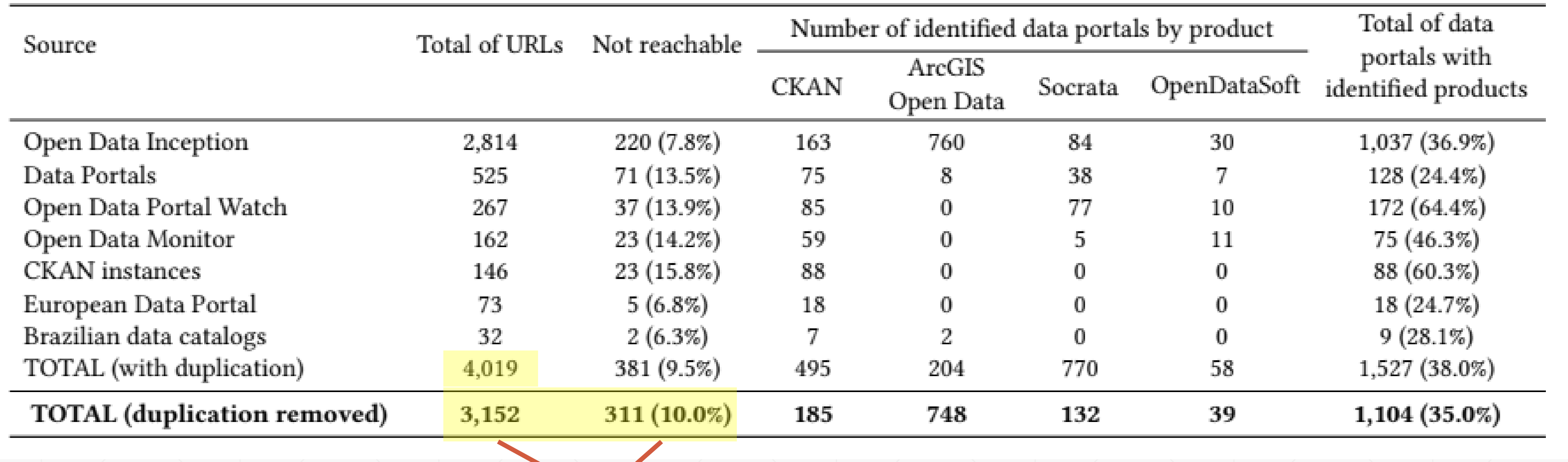
Related work

- Data Portals Repository (http://dataportals.org/)
    - Since 2011 – 587 data portals as of June 18, 2019
- Open Data Inception Project (https://data.opendatasoft.com/explore/dataset/open-data-sources@public/)
    - Since 2015 – 3,140 data portals as of June 18, 2019
- Current issues/challenges
    - Redundancy - duplicated entries
    - Discoverability - handle new entries
    - Updateability – constantly check if data portals are working
    - Traceability – keep track of e.g. software platform in use

Related work

- Need of an **independent**, **trustable** and **updatable** repository of data portals around the world

- Discussed by Correa, Zander & Silva (2018)

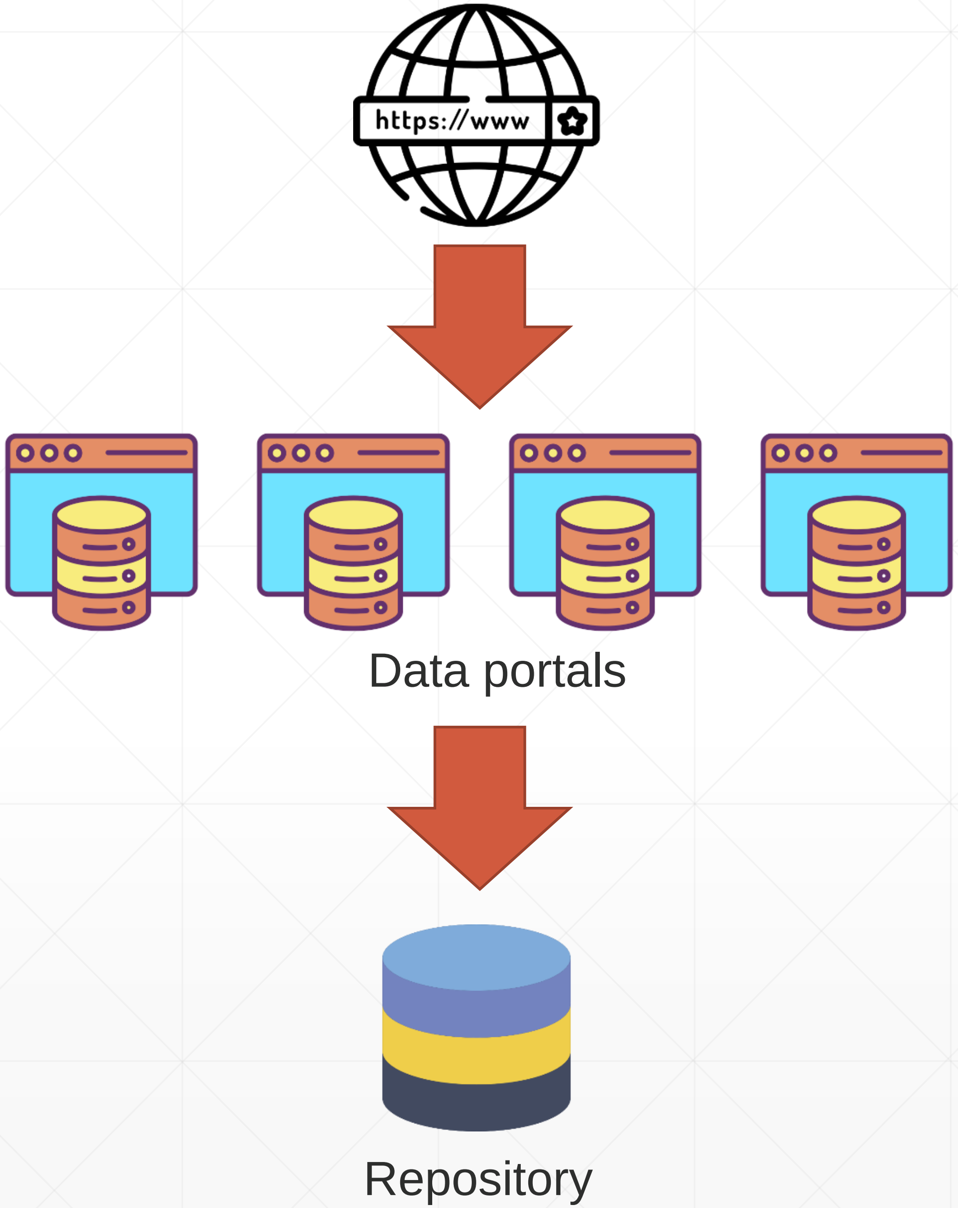


Source	Total of URLs	Not reachable	Number of identified data portals by product				Total of data portals with identified products
			CKAN	ArcGIS Open Data	Socrata	OpenDataSoft	
Open Data Inception	2,814	220 (7.8%)	163	760	84	30	1,037 (36.9%)
Data Portals	525	71 (13.5%)	75	8	38	7	128 (24.4%)
Open Data Portal Watch	267	37 (13.9%)	85	0	77	10	172 (64.4%)
Open Data Monitor	162	23 (14.2%)	59	0	5	11	75 (46.3%)
CKAN instances	146	23 (15.8%)	88	0	0	0	88 (60.3%)
European Data Portal	73	5 (6.8%)	18	0	0	0	18 (24.7%)
Brazilian data catalogs	32	2 (6.3%)	7	2	0	0	9 (28.1%)
TOTAL (with duplication)	4,019	381 (9.5%)	495	204	770	58	1,527 (38.0%)
TOTAL (duplication removed)	**3,152**	**311 (10.0%)**	**185**	**748**	**132**	**39**	**1,104 (35.0%)**

How can we solve this?

Main purpose

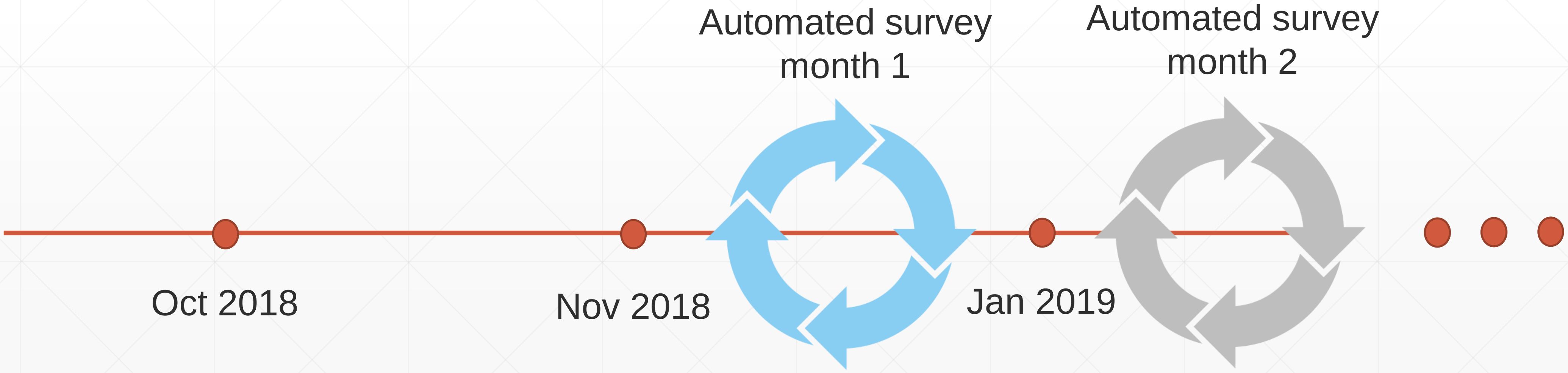
- Survey data portals automatically
- Whole web as the main source
- Method mainly considers:
  - Data portals availability
  - Software platform in use



Common Crawl

Background

- Common Crawl Project (http://commoncrawl.org/)
  - Makes a “copy” of the textual web every month (Nov 2018 – 220TB)
  - Freely available to everyone via Amazon Public Datasets
  - **We used URL Index** (3.3 billion entries / ~1.5TB)



Automated survey month 1

Automated survey month 2

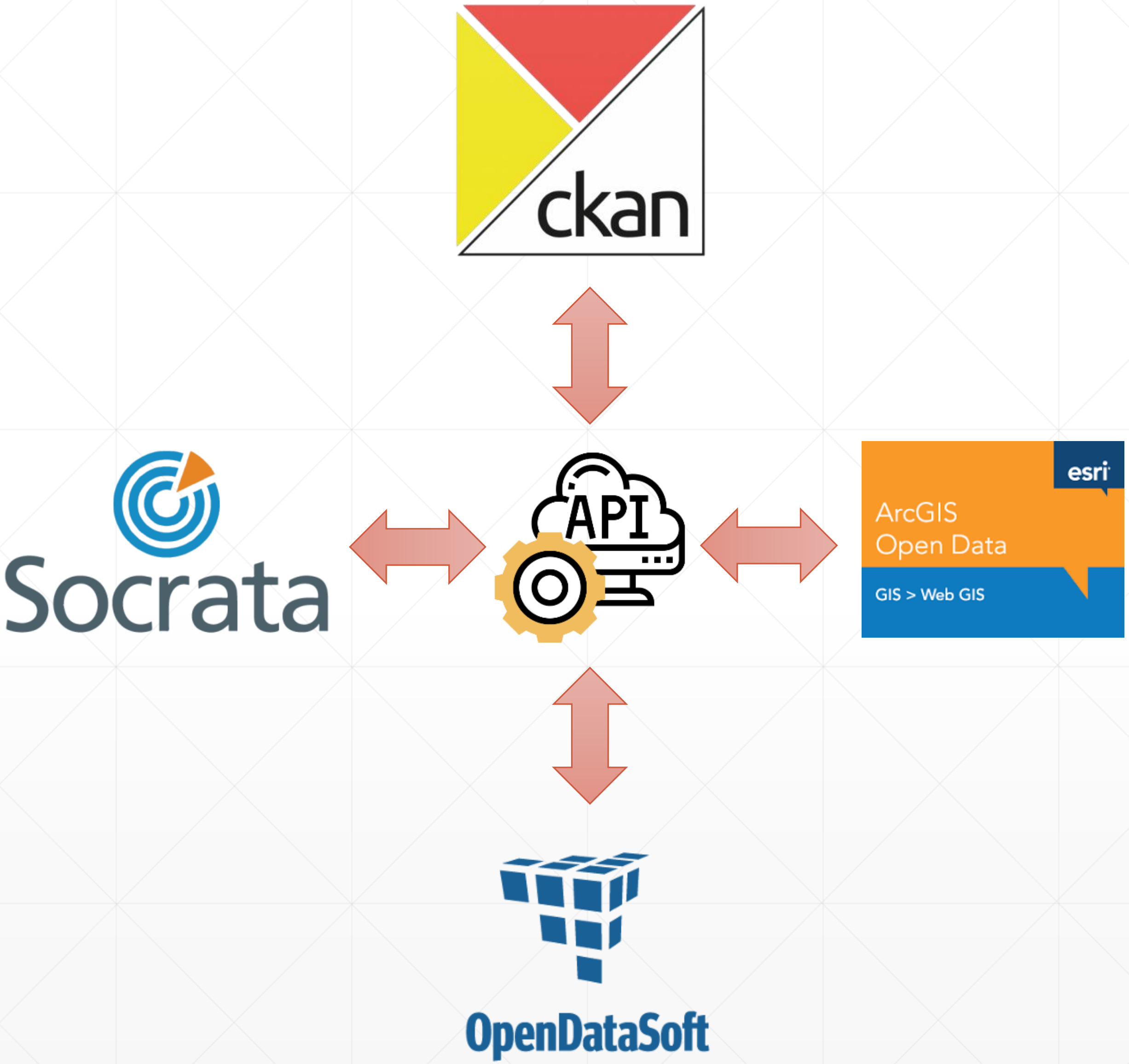
Oct 2018

Nov 2018

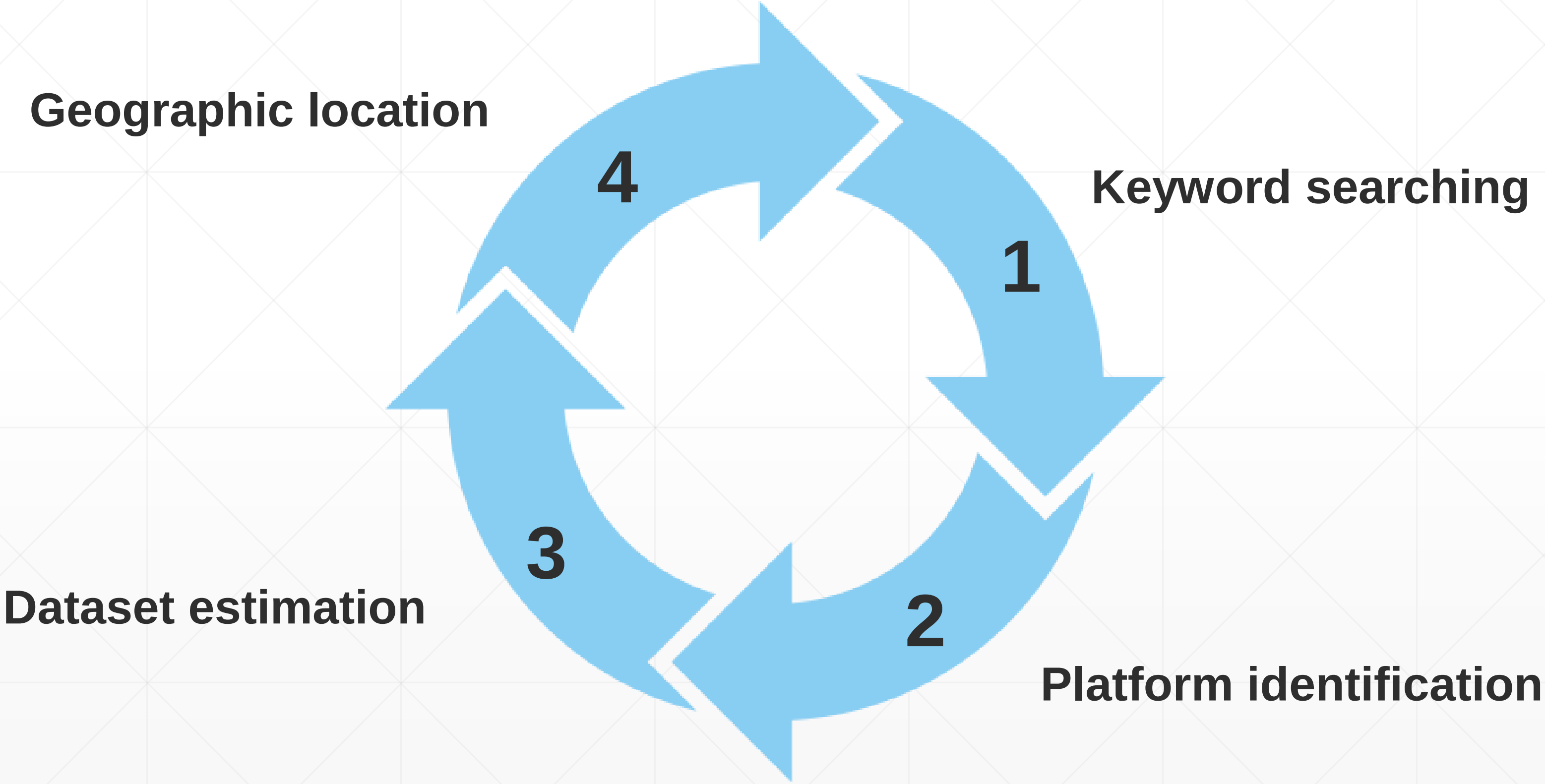
Jan 2019

Background

- Open data software platforms
  - Engine behind data portals
  - Store, publish and make data available
  - Main platforms:
    - CKAN – free and open source!
    - Socrata
    - OpenDataSoft
    - ArcGIS Open Data
  - **We developed a method to automatically identify them by API requests**



Method



1 Keyword searching

2 Platform identification

3 Dataset estimation

4 Geographic location

(1) Keyword searching

- Designed from previous findings with 1,104 identified data portals
- Term “data” was present in 90.6% of URLs
- The term used in 23 languages, according to translation services
- There are also variants in Spanish, Italian, Portuguese and German
- “data”, “datos”, “dati”, “dados”, and “daten” supposed to search in the URL

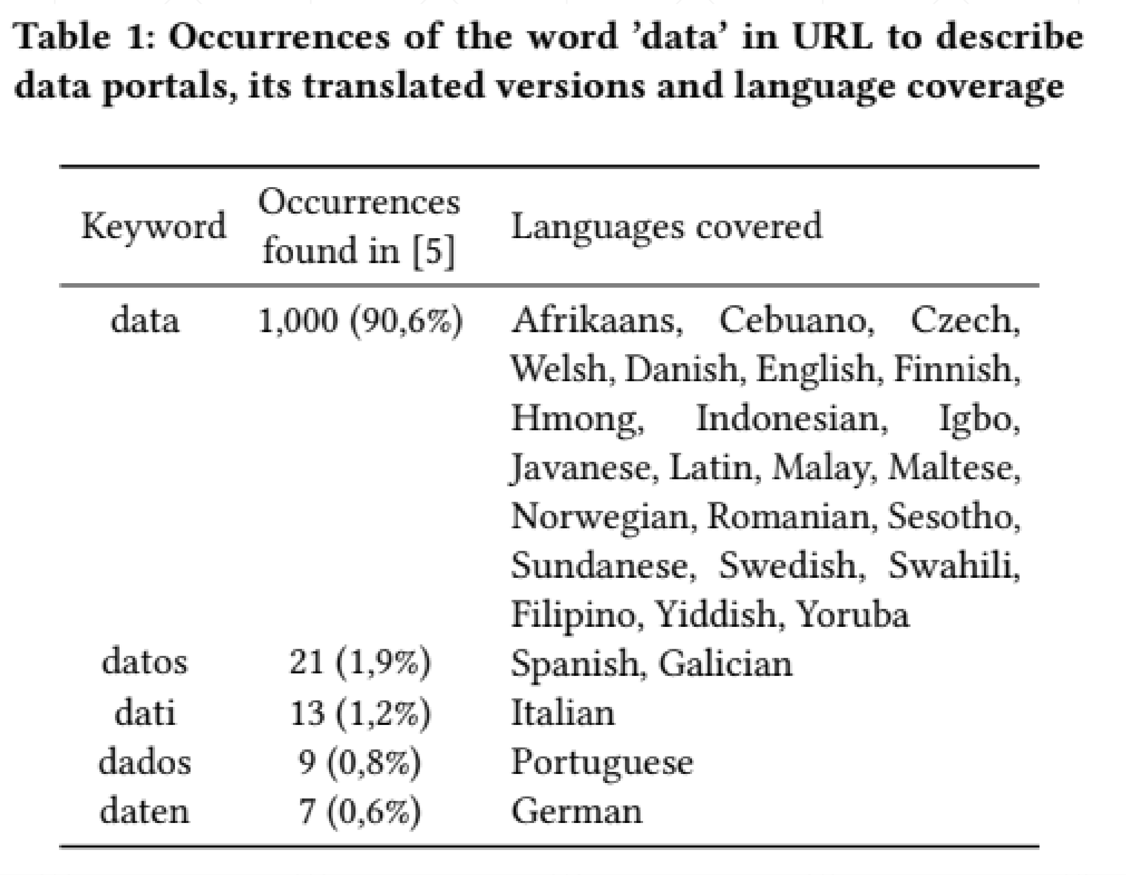


Table 1: Occurrences of the word ’data’ in URL to describe data portals, its translated versions and language coverage

Keyword	Occurrences found in [5]	Languages covered
data	1,000 (90,6%)	Afrikaans, Cebuano, Czech, Welsh, Danish, English, Finnish, Hmong, Indonesian, Igbo, Javanese, Latin, Malay, Maltese, Norwegian, Romanian, Sesotho, Sundanese, Swedish, Swahili, Filipino, Yiddish, Yoruba
datos	21 (1,9%)	Spanish, Galician
dati	13 (1,2%)	Italian
dados	9 (0,8%)	Portuguese
daten	7 (0,6%)	German

(2) Platform identification

- We designed API requests to uniquely identify software platforms
- Identification depends on the answer each one makes
- Every URL is checked 4 times

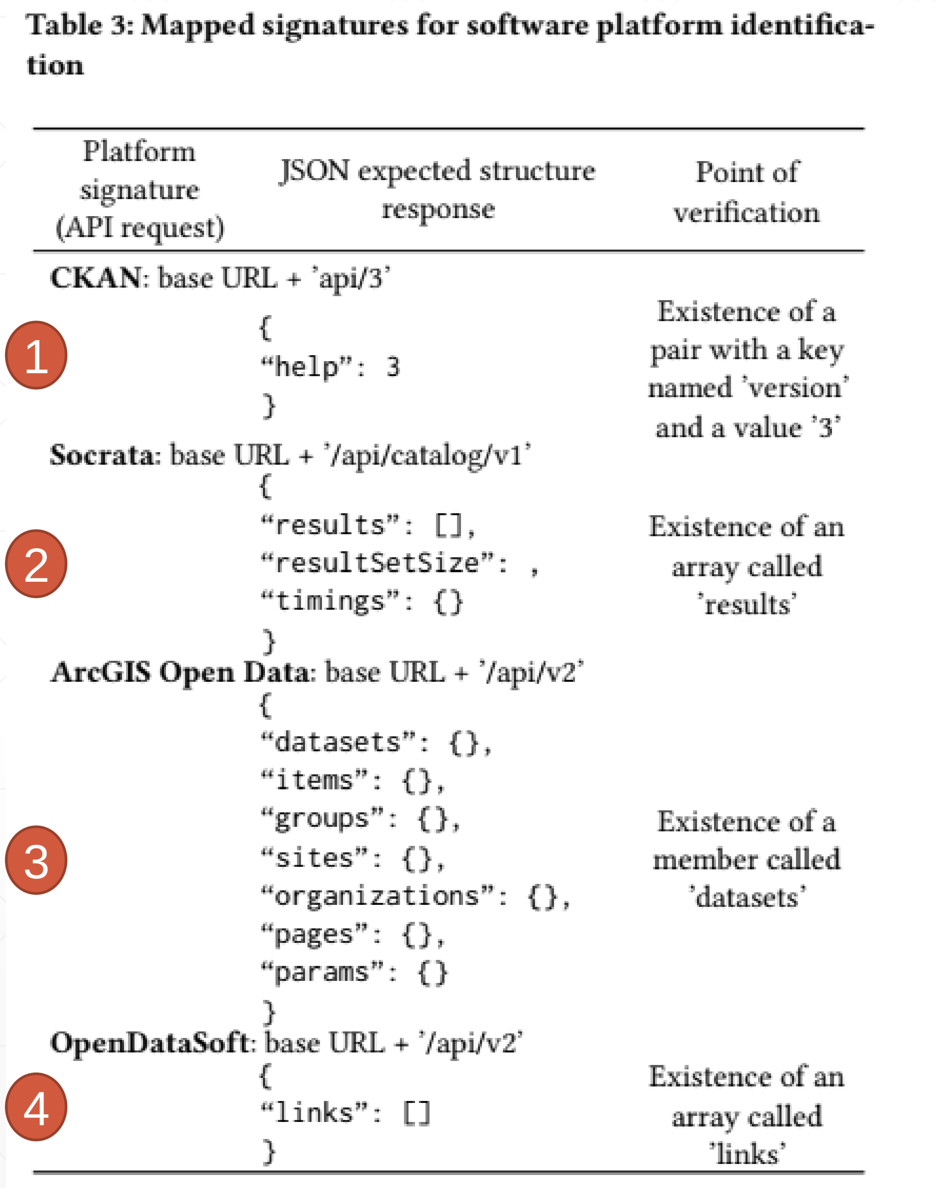


Table 3: Mapped signatures for software platform identification

Platform signature (API request)	JSON expected structure response	Point of verification

CKAN: base URL + 'api/3'	{
	"help": 3
	}
		Existence of a pair with a key named 'version' and a value '3'

Socrata: base URL + '/api/catalog/v1'	{
	"results": [],
	"resultSetSize": ,
	"timings": {}
	}
		Existence of an array called 'results'

ArcGIS Open Data: base URL + '/api/v2'	{
	"datasets": {},
	"items": {},
	"groups": {},
	"sites": {},
	"organizations": {},
	"pages": {},
	"params": {}
	}
		Existence of a member called 'datasets'

OpenDataSoft: base URL + '/api/v2'	{
	"links": []
	}
		Existence of an array called 'links'

(3) Dataset estimation

- Once identified, request APIs to get information about datasets
- Each one works differently
- Socrata and ArcGIS have “software as a service” model:
  - Share infrastructure among customers
  - Narrowing needed
- ArcGIS needs 3 requests!

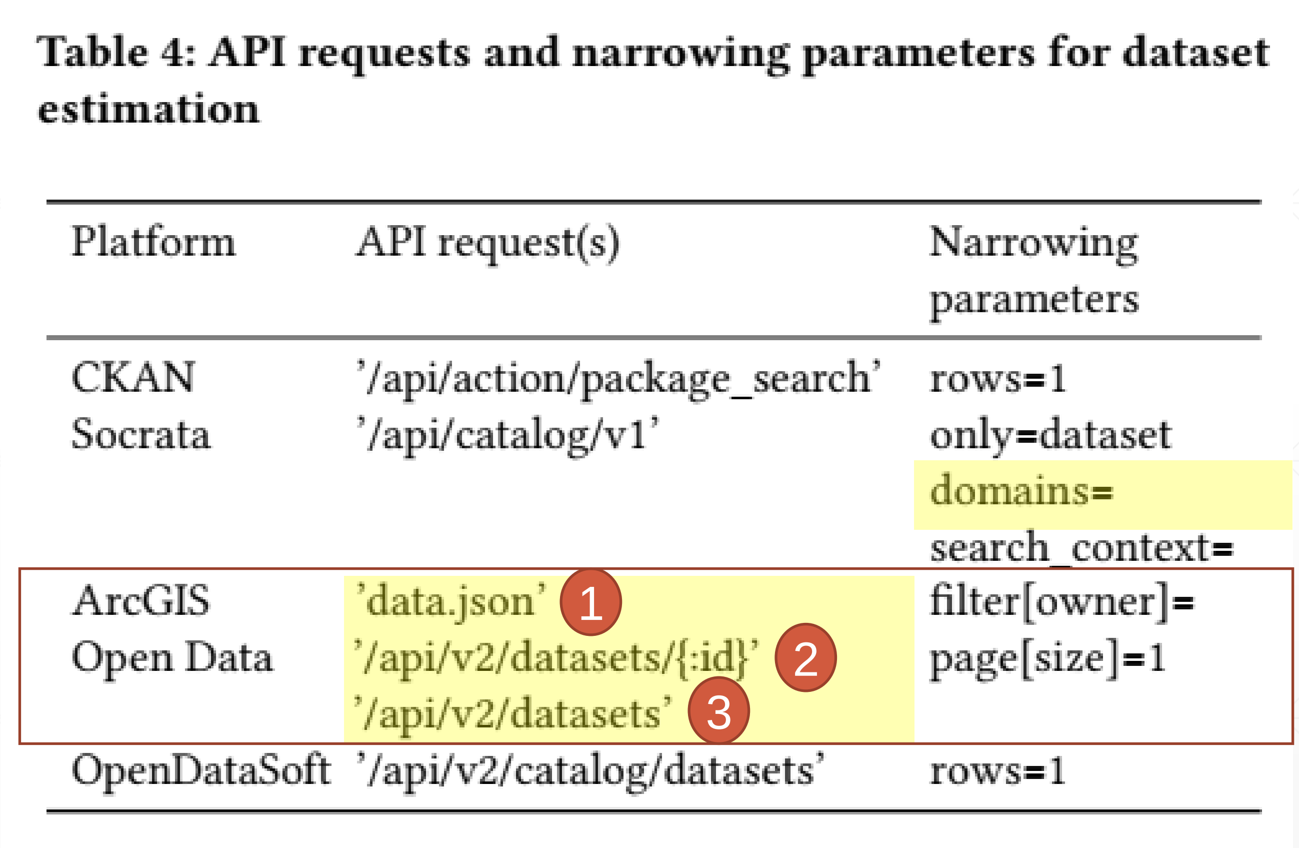


Table 4: API requests and narrowing parameters for dataset estimation

Platform	API request(s)	Narrowing parameters
CKAN	’/api/action/package_search’	rows=1
Socrata	’/api/catalog/v1’	only=dataset domains= search_context=
ArcGIS Open Data	’data.json’ 1 ’/api/v2/datasets/{:id}’ 2 ’/api/v2/datasets’ 3	filter[owner]= page[size]=1
OpenDataSoft	’/api/v2/catalog/datasets’	rows=1

(4) Geographic location

- Essential information to support benchmarking (e.g.: country)
- Designed 2 ways to get:
  - Country-Code Top-Level Domain (ccTLD) – more precise
  - IP country – where data portal is hosted (less precise)

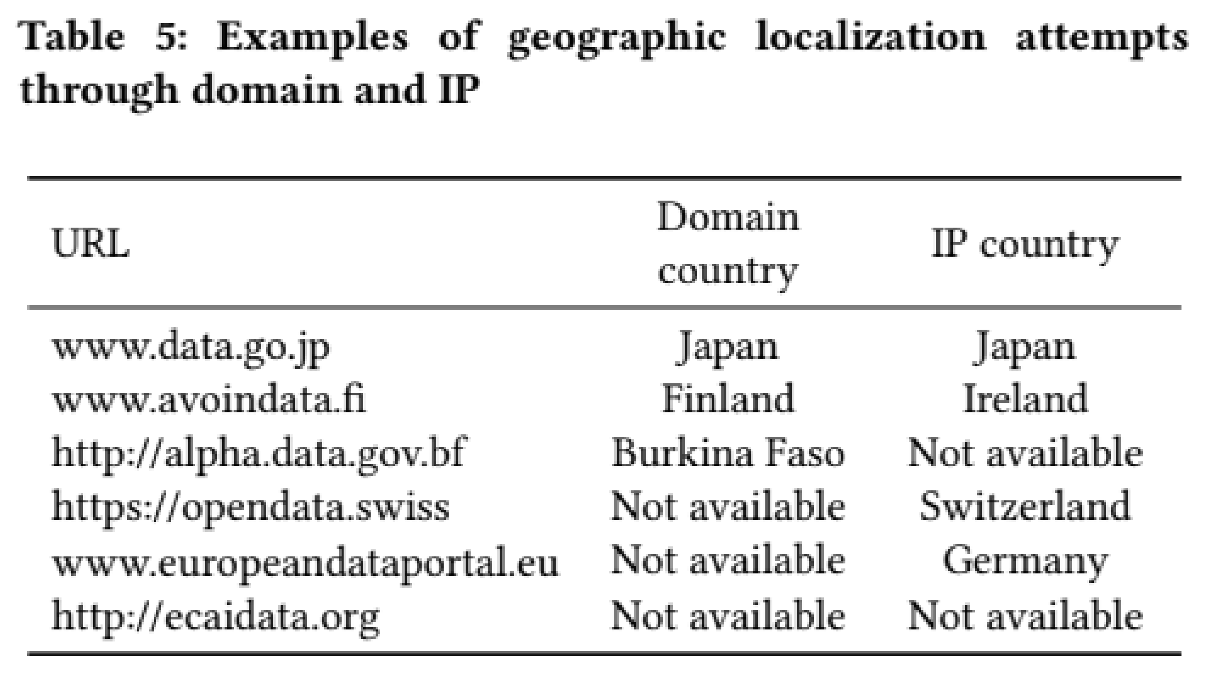


Table 5: Examples of geographic localization attempts through domain and IP

URL	Domain country	IP country
www.data.go.jp	Japan	Japan
www.avoindata.fi	Finland	Ireland
http://alpha.data.gov.bf	Burkina Faso	Not available
https://opendata.swiss	Not available	Switzerland
www.europeandataportal.eu	Not available	Germany
http://ecaidata.org	Not available	Not available

Results and discussion

- Algorithm execution times
  - Keyword searching **~16 hours**
  - Platform identification **~18 hours**
  - Dataset estimation and
  - Geographic location **~2 hours**

  Entire process takes ~36 hours to complete
- **We can repeat process in a time basis (e.g.: every month)**
- **Answer RQ1 about efficiency of keyword approach**

Results and discussion

- In terms of data portals found:
  - 2018: 1,104 data portals
  - 2019: 1,339 data portals (↑ ~21%)
  - Only 272 exist in both works!
- Scenario of change: increases manual efforts to handle changes
- Our method tries to solve this
- **Answering RQ2 about reducing efforts to support benchmarking**

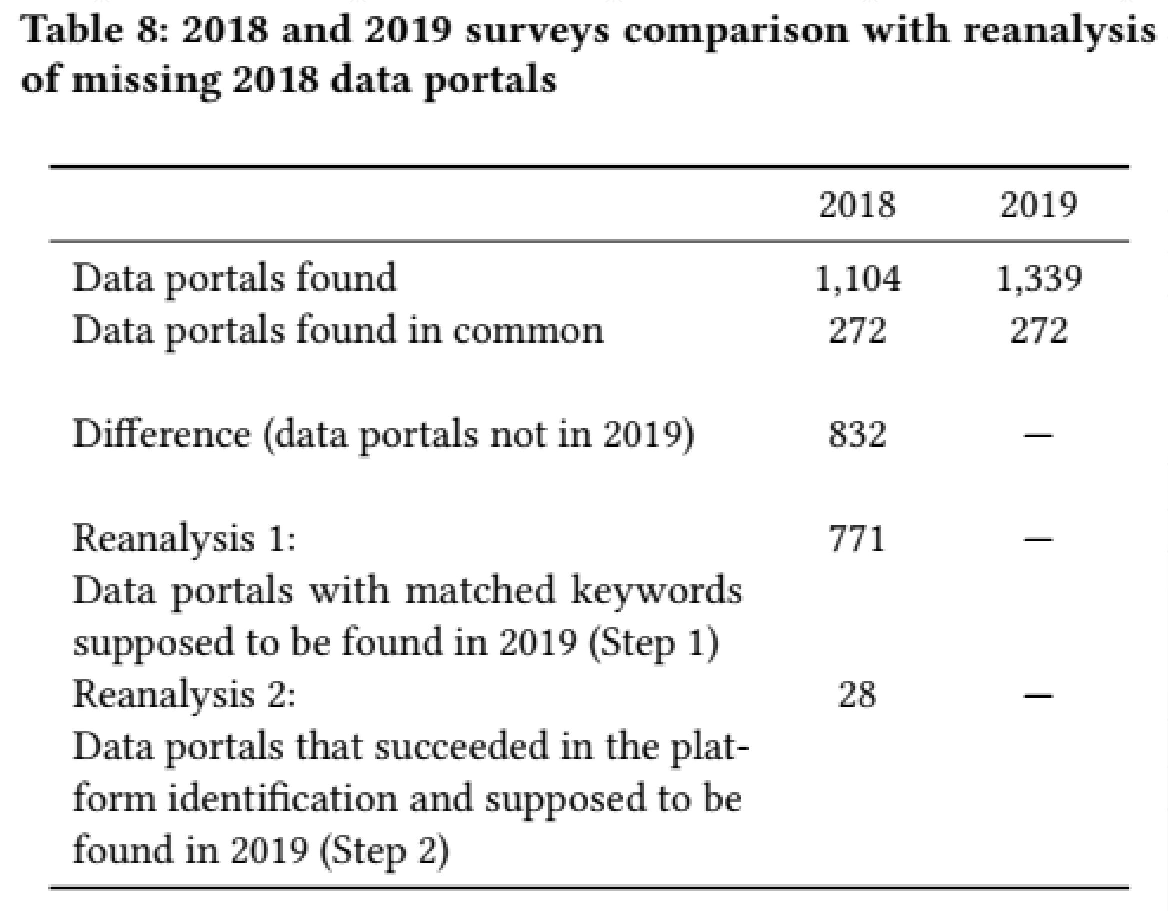


Table 8: 2018 and 2019 surveys comparison with reanalysis of missing 2018 data portals

	2018	2019
Data portals found	1,104	1,339
Data portals found in common	272	272
Difference (data portals not in 2019)	832	–
Reanalysis 1: Data portals with matched keywords supposed to be found in 2019 (Step 1)	771	–
Reanalysis 2: Data portals that succeeded in the platform identification and supposed to be found in 2019 (Step 2)	28	–

Results and discussion

Ranking 10 largest data portals

- 4 new large data portals found in 2019 (✳)
- On the other hand, 3 old data portals were not found by 2019 method
  - 2 without keyword “data”
  - 1 did not answer by the time of request (temporally)

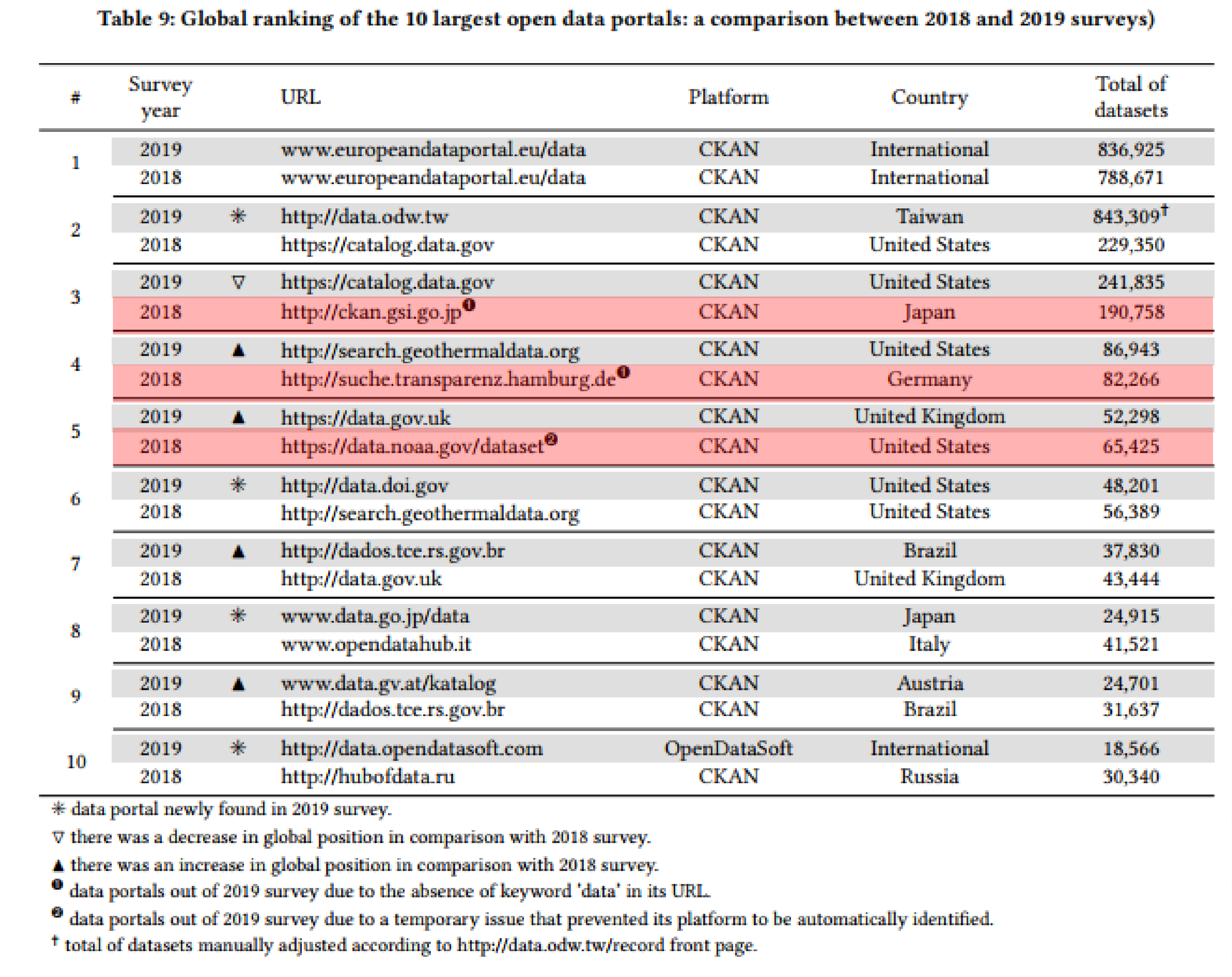


Table 9: Global ranking of the 10 largest open data portals: a comparison between 2018 and 2019 surveys)

#	Survey year		URL	Platform	Country	Total of datasets
1	2019		www.europeandataportal.eu/data	CKAN	International	836,925
	2018		www.europeandataportal.eu/data	CKAN	International	788,671
2	2019	✳	http://data.odw.tw	CKAN	Taiwan	843,309[†]
	2018		https://catalog.data.gov	CKAN	United States	229,350
3	2019	▽	https://catalog.data.gov	CKAN	United States	241,835
	2018		http://ckan.gsi.go.jp[❶]	CKAN	Japan	190,758
4	2019	▲	http://search.geothermaldata.org	CKAN	United States	86,943
	2018		http://suche.transparenz.hamburg.de[❶]	CKAN	Germany	82,266
5	2019	▲	https://data.gov.uk	CKAN	United Kingdom	52,298
	2018		https://data.noaa.gov/dataset[❷]	CKAN	United States	65,425
6	2019	✳	http://data.doi.gov	CKAN	United States	48,201
	2018		http://search.geothermaldata.org	CKAN	United States	56,389
7	2019	▲	http://dados.tce.rs.gov.br	CKAN	Brazil	37,830
	2018		http://data.gov.uk	CKAN	United Kingdom	43,444
8	2019	✳	www.data.go.jp/data	CKAN	Japan	24,915
	2018		www.opendatahub.it	CKAN	Italy	41,521
9	2019	▲	www.data.gv.at/katalog	CKAN	Austria	24,701
	2018		http://dados.tce.rs.gov.br	CKAN	Brazil	31,637
10	2019	✳	http://data.opendatasoft.com	OpenDataSoft	International	18,566
	2018		http://hubofdata.ru	CKAN	Russia	30,340

✳ data portal newly found in 2019 survey.

▽ there was a decrease in global position in comparison with 2018 survey.

▲ there was an increase in global position in comparison with 2018 survey.

[❶] data portals out of 2019 survey due to the absence of keyword 'data' in its URL.

[❷] data portals out of 2019 survey due to a temporary issue that prevented its platform to be automatically identified.

[†] total of datasets manually adjusted according to http://data.odw.tw/record front page.

Results and discussion

- Increase in the number of CKAN installations:
  - 2018: 185 installations
  - 2019: 351 installations
- More utilization of ArcGIS Open Data
  - 2018: most installations with up to 10 datasets each
  - 2019: most installations with 11-100 datasets each

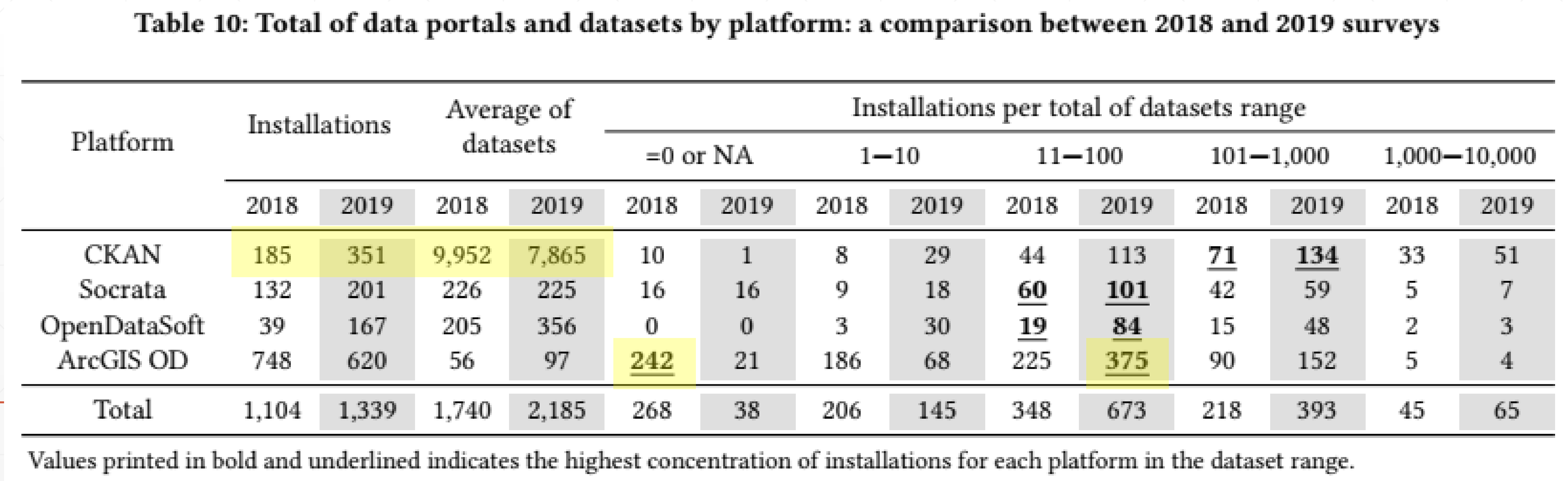


Table 10: Total of data portals and datasets by platform: a comparison between 2018 and 2019 surveys

Platform	Installations		Average of datasets		Installations per total of datasets range									
					=0 or NA		1—10		11—100		101—1,000		1,000—10,000	
	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019
CKAN	185	351	9,952	7,865	10	1	8	29	44	113	**71**	**134**	33	51
Socrata	132	201	226	225	16	16	9	18	**60**	**101**	42	59	5	7
OpenDataSoft	39	167	205	356	0	0	3	30	**19**	**84**	15	48	2	3
ArcGIS OD	748	620	56	97	**242**	21	186	68	225	**375**	90	152	5	4
Total	1,104	1,339	1,740	2,185	268	38	206	145	348	673	218	393	45	65

Values printed in bold and underlined indicates the highest concentration of installations for each platform in the dataset range.

Conclusion

- Merits:
  - Method is reproducible – high potential for automation
  - Based on a extensible list of keywords
  - Identify main open data software platforms (CKAN, Socrata, OpenDataSoft, ArcGIS OD)
- Limitations:
  - Data portals without the keyword “data” in the web address
  - Identification of software platforms other than main ones
  - Web pages overlooked by Common Crawl bots

Conclusion

- Findings include a fresh list of 1,339 healthy data portals (available on Github)
  - https://github.com/Andreiwid/wholewebdataportalsurvey
- Contribute to a **independent**, **trustable** and **updatable** repository of data portals
- Can reduce efforts to conduct benchmark exercises

Future work

- Look into the most detailed file available on Common Crawl
  - Do not limit by only URL index
  - Increase chance to find more data portals
  - Find keywords in the body of HTML, such as:
    - “Education”
    - “Open data”
    - “Access to information”
    - “Transparency”
    - “Accountability”, etc.

Thank you!

andreiwid@ifsp.edu.br

http://andreiwid.info

https://github.com/Andreiwid/wholewebdataportalsurvey