Open science-based framework to reveal open data publishing: an experience from using Common Crawl

Andreiwid Correa and Israel Fernandes

Introduction

1 Scientific research has been part of human nature because it tries to answer all the questions we have about how the world works. With an underlying process called scientific method, researches observe their surround to test hypotheses by experimentation, so they can prove theories. Open Science emerged in this context as a movement to make scientific research more collaborative, encompassing e.g. open access and open data in its essence (Wikipedia contributors). The idea behind open science is to make scientific knowledge widely available without barriers that impede broad dissemination of scientific papers and data.

2 But assuring solely the broad dissemination (of data) seems not to be enough to promote the social benefits expected from open science and data. To support this affirmation, we recall the wide availability of data portals and all the concerns related to the discoverability and the validity of those sources to make them useful.

3 Discoverability is the quality of being easy to find a data portal, no matter where and how it is published. In this regard, a previous study (Correa et al.) estimated the existence of +3,000 data portals worldwide. These sources often publish data for public sector and government interests, at all levels from local to federal. In addition, it is possible to see the grow beyond its original application in government, as private entities are increasingly adopting open data (Harper and Hughes). A major example in this regard is Uber Movement (https://movement.uber.com) that provides anonymized data for non- commercial re-use from over billions of transportations made every day.

This grow in the number of data sources shows as a complexity for anyone who are interested in open data, because the first step always involves finding a data source to consume.

4 Validity is the quality of being able to check if a data portal is implemented or based on standardized open data software platforms. The implementation and use of such platforms are a step forward to materialize Open Data Principles (Tauberer), as platforms offer infrastructure, processes and tools to fully accomplish data publishing by providing e.g. dataset and metadata management, multiple formats and API support, all these essential to the sustainability of open data-related initiatives (European Union).

5 Civil society, policymakers and businesses often need evidence on how well organization adopt and use data. Such a demand requires from open data practitioners the proposition of methods that have in their essence a way to reveal data publishing without worrying about workloads increased by traditional desk research methods. Thus, searching for datasets is gaining importance worldwide and has been an area of attention by academic and industry players. An example of this is an initiative from Google that launched Dataset Search (https://datasetsearch.research.google.com) to help researches, scientists and data journalists find data available on known sources.

6 This paper reports the development of a framework to reveal open data publishing from an open science perspective, serving as an answer to the discoverability and the validity issues. The framework considers collecting data from the Common Crawl open science project to first identify data portals and then gather information about their availability, having in its essence an iterative and differential process. A model for the historical data repository and how it can be published are proposed in the way consuming of produced data are feasible by the interested community. This work involves both use and creation of open science and open data to branch new sort of research possibilities based on publishing of derived data.

The Common Crawl Project

7 The Common Crawl project (https://commoncrawl.org) can be similarly described as an open version of the Google search engine. Instead of providing an out-of-the-box mechanism like the giant of searches does, it makes available a database of raw data containing a copy of the textual web freely available for everyone use, mainly for research purposes.

8 The project is funded by a non-profit organization with the same name and aims at crawling the whole web content by following web pages links and repeating this process until reaching the maximum allowed depth level. Common Crawl makes a copy of the data contained in each HTML source code available on the web as it stood on a given point in time, including textual data and metadata but without the images, CSS stylesheets, JavaScript files and other non-HTML content (Leeward). The copy of the web is performed once a month and the generated data is estimated roughly 230TB/ month of uncompressed content.

9 What makes the project unique is the fact it handles the enormous size of the web and its continuous and fast changing characteristics, making Common Crawl an open science choice. The interested community has nothing to worry about technical

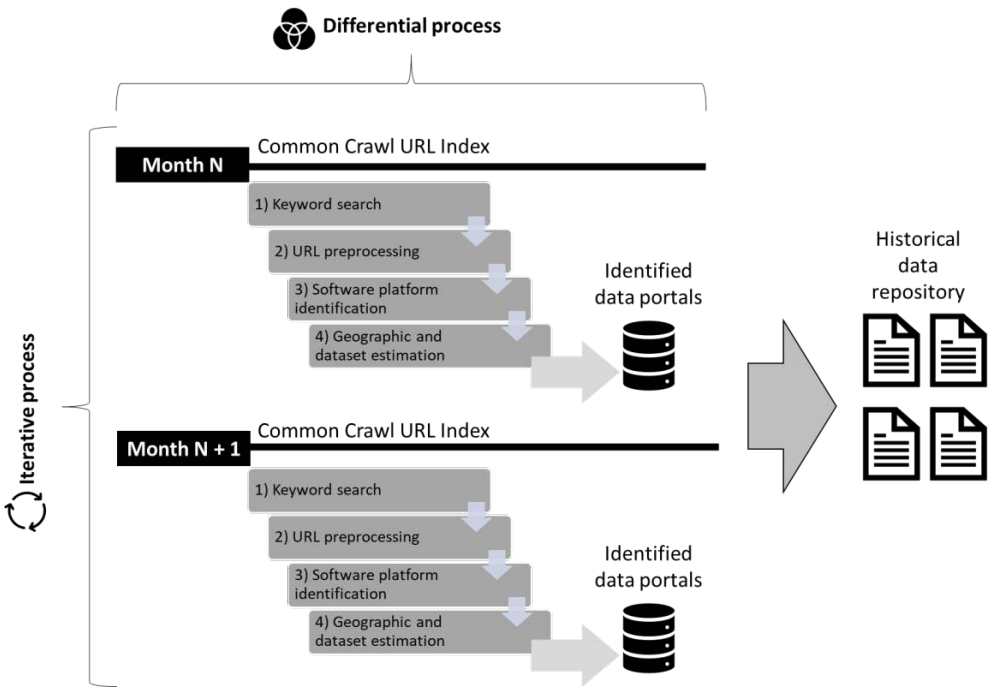
infrastructure necessary to explore textual content of web pages, as the project provides data ready-to-use even for large-scale data science exercises.

Framework

10 In previous work (Correa and da Silva) we developed a method for surveying data portals introducing the use of the Common Craw database. That work aimed to identify sources on the web responsible for disclosing open data through standardized open data software platforms, however, did not consider changes over time which is under consideration in this paper. Due to the rapid context of change, where open data sources come and go on the web, there is the need to keep track of past and existing data portals to store historical data with respective implemented software platform, number of datasets and geographic location. This information is intended to remain available to scholars and open data practitioners to conduct e.g. benchmark exercises which is an essential tool to evaluate and rank initiatives based on how well they use open data in different ways (Ulrich et al.).

11 The proposed framework is described in this section as an answer to the above-mentioned issues regarding discoverability and validity of data portals. This proposal differs from previous study (Correa and da Silva) in the way it features in its essence both iterative and differential processes to make the historical data repository feasible. The iterative process estimates recurring data collection events in a monthly basis, which is the actual frequency defined by the Common Crawl project. The differential process contemplates differences among data collection events, contributing to a baseline for making comparisons. A high-level representation of the framework is shown in the following figure.

Figure 1. Framework to reveal open data publishing with an emphasis on its iterative and differential processes



(The image copyright belongs to the authors)

12 As shown in the figure, within the framework data collection is made on Common Crawl URL Index database, which is the smallest database available as it contains only metadata for crawled web pages. Includes full URL (web address), MIME type (most common is 'text/html'), and a link to the full content (copy of HTML) stored in another database. Even though it is the smallest dataset available by the project, size sums 1.46TB of uncompressed content with roughly 3.3 billion of URL records, all split into 300 files for downloading purposes. The experience of handling such a content was reported in previous work (Correa and da Silva) and detail about it is out of the scope of this paper.

13 The iterative characteristics of the framework guarantees it occurs once a new URL index is made available by the Common Crawl. Process would start in a given month (marked as 'Month N') and subsequent data collection events are marked 'Month N + 1', 'Month N + 2', and so forth. For each data collection, there is a process with four steps that were developed in previous work and are essential for the understanding of this framework description. Then, a brief explanation is given as follows:

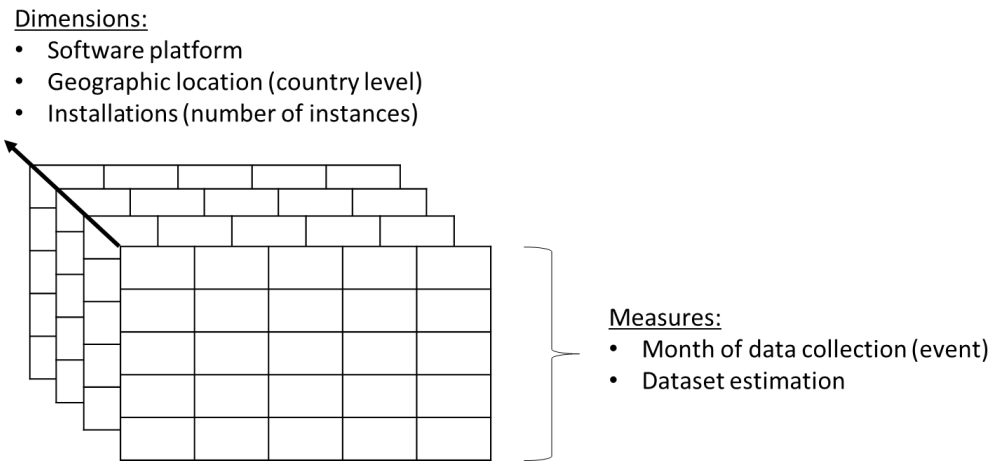
- Step '1) Keyword search'. Each URL contained in the Common Crawl URL Index is investigated for specific keywords that would identify potential data portals on the web. A data portal usually has in its URL e.g. the term 'data', so in this step the main idea is to search for such a keyword. At the current state of the research that inaugurated this work, language variations are under consideration to catch as many potential data portals as possible. According to our experiments, search for variations of the term in other languages such as Spanish (datos), Italian (dati), Portuguese (dados), and German (daten) increased the number of matched URLs with potential data portal, but also the number of false positives (looked like data portals but they were not). At the end of this step, a list of matched URLs with potential data portals is produced to be input into the next step.
- Step '2) URL preprocessing'. All matched URLs with potential data portals may contain redundant entries as many web addresses could point to the same data portal. Examples are URLs with prefixes 'http://' and 'https://' in the string, presence (or not) of 'www', and content in the path fragment of URL (text after first single slash '/'). These variations do not interfere in the way data portals work so redundant entries are removed by shortening web addresses until reaching the domain level (e.g. europeandataportal.eu).
- Step '3) Software platform identification'. Resulting URLs from last step are queried by sending API requests to them in order to identify the corresponding open data software platform. Once destination URLs reply with an expected JSON (JavaScript Object Notation), the platform is identified. Currently, it is possible to identify the following platforms: CKAN, DKAN, Junar, OpenDataSoft, Pmydata, Socrata, Udata, and ArcGIS Open Data.
- Step '4) Geographic and dataset estimation'. Once software platform is known, specific APIs are requested to return the quantity of datasets. Every considered platform has a basic structure responsible for storing data (called 'dataset') and underlying APIs to manipulate it. In addition, data portals are checked for their geographic location at the country level. This information is supposed to be obtained in two ways: 1) ccTLD method that is the identification of country through URL by the last characters e.g. '.fr' for France, '.jp' for Japan; 2) IP location method which tries to obtain the country by the place where a data portal is hosted.

14 Once an iteration of the four-step process is done, a list of identified data portals is ready to be stored in the historical data repository. This list solves the discoverability and validity issues as data portals in the list had their software platforms properly

identified. Records in the list reflect a given point in time, so additional iterations over time are expected to accomplish data collection events.

15 Each month of data collection represents an event to be consolidated in the historical data repository, which also builds the foundation of the differential process. The idea is to allow comparisons between events through a defined data model by providing a structure optimized to read, summarize and analyze numeric values obtained from identified data portals and underlying data. The figure in the following illustrates an intended multidimensional model for the historical data repository emphasizing its dimensions and measures.

Figure 2. A multidimensional model for the historical data repository

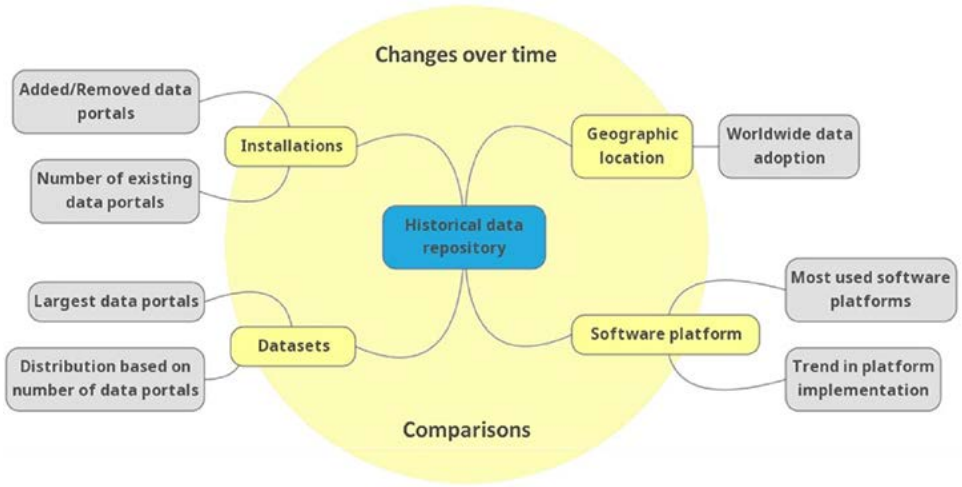


(The image copyright belongs to the authors)

16 Dimensions illustrated in the above model hold the context surrounding data portals. They consist of data regarding what open data software platform is implemented, the country where it is related to, and details about the instance run by the respective data portal. Measures hold facts about the data collection events and comprise the month when data is obtained and position about datasets within data portals on that point in time.

17 Data publishing through data portals changes over time. Thus, the idea of the historical data repository is basically to support comparisons between data collection events, in addition to keep track of the evolution of data openness worldwide. We envision a road map for many layers of data visualization once the repository is consolidated. The figure in the following explores these layers unfolded in a way that several visualizations are expected. Table in sequence describes these visualizations and emphasizes their contributions within the context of the proposed framework.

Figure 3. Data visualization layers from the proposed framework



(The image copyright belongs to the authors)

Table 1. Description of expected visualization

Layer	Visualization	Description
Geographic location	Worldwide data adoption	This visualization can show quantitative information on number of data portals found and datasets within them. Contributes to the understanding of how countries are managing their open data infrastructures based on the availability of data portals.
Software platform	Most used software platforms	This visualization can show which of the 8 open data software platforms considered in the study is implemented in each data portal found. Contributes to the understanding of how institutions choose their technology solution, especially if they opt for off-the-shelf software instead of making their owns tools.
	Trend in platform implementation	This visualization can show tendencies for open data software platform choosing, including preferences for open source products or 'software as a service' models. Information on this is based on the use of open data software platforms considered in this study, in addition to the evolution of number of instances and number of datasets. Contributes to the understanding of trends in platform use and implementation which help organizations to plan their data strategies besides necessary technology infrastructure.
Datasets	Largest data portal	This visualization can show the largest data portals in the world regarding the number of datasets within them. Contributes to the encouragement of organizations all around the world to do in the same way by empathizing the best practices in open data publishing.

	Distribution based on number of data portals	This visualization can show the concentration of data portals in ranges of available dataset numbers. This information is useful to figure out to what extent data portals are mature enough by analyzing the number of datasets in them in a clustered view.
Installations	Added/Removed data portals	This visualization is part of the core of the differential process. It makes evident which data portals are added or removed considering previous data collection events. Contributes to the understanding of how availability of data portals changes over time.
	Number of existing data portals	This visualization can show the last position of identified data portals. Contributes to a single, up to date and reliable source where community can find information about current data portals worldwide.

Validation perspectives

18 The framework developed herein encompasses processes prioritizing automation in a way manual intervention is drastically reduced. One issue that remains in mind is how to validate results, especially how to assure an identified data portal is intended to publish open data.

19 Within the framework, the checking for the existence of open data software platform is expected to occur automatically. In ‘Step 3) Software platform identification’ the availability of specific APIs with valid responses are used to make sure one of known software platforms is working behind data portals. However, there is no way to claim identified data portals publish open data just because the content (datasets) is not fully analyzed against open data principles. Also, open data software platforms are out-of-the- box tools intend for publishing open data, but adopters can use in the way they think right, including for publishing non-open data.

20 Validation perspectives include deep analysis of datasets assuming they fully comply with open data principles. Some principles are easier to check with automated approach, such as whether data are machine processable, available on-line, published with non- proprietary formats with open standards, available to anyone (non-discriminatory), and free of licenses. But other principles such as whether data is published with the highest possible level of granularity and if data is timely are much harder to accomplish, requiring a high level of human intervention and improved analysis skills.

21 Apart from the limitation of this work, we believe full automation directly contributes to the purpose of this which is to reveal open data publishing by identifying data sources. This approach solely brings benefits to the interested community as it avoids manual tasks traditionally performed for this purpose.

Conclusion

22 This work proposed a framework to reveal data publishing from an open science perspective. Through both iterative and differential processes, data is collected on a monthly basis from the Common Crawl Project to identify data portals and to gather information about their availability. Within the framework, it is defined a model to accomplish the historical data repository that is the main element to allow comparisons considering metrics and dimensions of collected data.

23 This work contributes to a repeatable method of identification of data portals without or with a reduced human intervention, mainly providing a single, up to date and reliable source where information is shared. Leading open data benchmarking studies depend on such information and their approach are traditionally based on an extensive desk research process. Thus, community can find information not only about existing data portals worldwide, but also find past records of those that are not available anymore.

24 This proposal shows a different approach from dedicated data searches such as Google Dataset Search. Instead of considering the dataset within data portals as a single unit for providing data, this proposal gathers information about the data portals themselves. Type of software platform, geographic location and details of the installation of data portals are examples of information under analysis in the context of proposed framework. All gathered data are made available allowing several visualization layers to carry out comparisons.

25 We believe initiatives such as this one is an encouragement to branch new sort of researches based on publishing of open data and open science projects. Shows as a good candidate to promote social benefits expected from both broad availability and dissemination of data.

BIBLIOGRAPHY

Atz, Ulrich, et al. *Benchmarking Open Data Automatically. ADI-TR-2015-000*, Open Data Institute, 2015, theodi.org/article/benchmarking-open-data-automatically.

Correa, Andreiwid Sheffer, et al. "Investigating Open Data Portals Automatically: A Methodology and Some Illustrations." *Proceedings of the 19th Annual International Conference on Digital Government Research Governance in the Data Age - Dgo '18*, ACM Press, 2018, pp. 82:1–82:10. doi: 10.1145/3209281.3209292.

Sheffer Correa, Andreiwid, and Flavio Soares Correa Da Silva. "Laying the Foundations for Benchmarking Open Data Automatically: A Method for Surveying Data Portals from the Whole Web." *Proceedings of the 20th Annual International Conference on Digital Government Research: Governance in the Age of Artificial Intelligence*, ACM Press, 2019, pp. 287–96. doi: 10.1145/3325112.3325257.

European Union. *Recommendations for Open Data Portals: From Setup to Sustainability*. 2017, pp. 76, www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf.

Harper, Jim, & Adam Hughes. *The State of the Union of Open Data. Data Foundation*, 14 Jan. 2019, www.datafoundation.org/cover-page-ed-3-sotu.

Leetaru, Kalev. "Common Crawl And Unlocking Web Archives For Research." *Forbes*, 28 Sept. 2017, www.forbes.com/sites/kalevleetaru/2017/09/28/common-crawl-and-unlocking-web-archives-for-research/.

Tauberer, Joshua. *Open Government Data: The Book – Second Edition*. 2014, opengovdata.io/.

Wikipedia contributors. "Open Science." *Wikipedia*, 6 Jan. 2020, en.wikipedia.org/w/index.php?title=Open_science&oldid=934409107.

ABSTRACTS

The publishing of open data is considered a key element for civic participation paving the way to the 'public value', a term which underpins the social contribution. A result of that can be seen through the popularity of data portals published all around the world by governments, public and private organizations. However, the diffusion of data portals raises concerns about discoverability and validity of these data sources, especially to what extent they contribute to open data and open science. The purpose of this work is to develop a framework to reveal open data publishing with the use of a freely available open science project called Common Crawl. The idea is to identify open data-related initiatives and to gather information about their availability, having in the framework's essence an iterative and differential process. The main outcome is shown through a proposed model for the historical data repository which involves both use and creation of open science to branch new sort of research possibilities based on publishing of derived data.

INDEX

Keywords: open data, open science, common crawl, data portals

AUTHORS

ANDREIWID CORREA

Federal Institute of Sao Paulo,
andreiwid@ifsp.edu.br
(corresponding author)

ISRAEL FERNANDES

Federal Institute of Sao Paulo