Towards an automated method to assess data portals in the deep web

Andreiwid Sheffer Correa[a,b], Raul Mendes de Souza[a], Flavio Soares Correa da Silva[b]

[a] Federal Institute of Education, Science and Technology of Sao Paulo - IFSP, Rodovia D. Pedro I (SP-65), Km 143,6 Campinas, Sao Paulo (SP) CEP 13069-901, Brazil
[b] Institute of Mathematics and Statistics, University of Sao Paulo, Brazil

This is the author's version of Accepted Manuscript* of an article published by Government Information Quarterly – ISSN: 0740-624X (copyright Elsevier). Please refer to the formal publication at https://doi.org/10.1016/j.giq.2019.03.004.

Please cite this article as:
Correa, A. S., Souza, R. M. de, Silva, F. S. C. (2019). Towards an automated method to assess data portals in the deep web. Government Information Quarterly, 36(3), 412–426.
https://doi.org/10.1016/j.giq.2019.03.004

Abstract
The rising number of data portals has been increasing demand for new techniques to assess data openness in an automated manner. Some methods have emerged that presuppose well-organized data catalogs, the availability of API interfaces and natively exposed metadata. However, many data portals, particularly those of local governments, appear to be misimplemented and developed with the classic website model in mind, which provides access to data only through user interaction with web forms. Data in such portals resides in the hidden part of the web, as it is dynamically produced only in response to direct requests. This paper proposes an automated method for assessing government-related data in the deep web on the basis of compliance with open data principles and requirements. To validate our method, we apply it in an experiment using the government websites of the 27 Brazilian capitals. The method is fully carried out for 22 of the capitals' websites, resulting in the analysis of 5.6 million government web pages. The results indicate that the keyword search approach utilized in the method, along with the checking of web pages for multifield web forms, is effective for identifying deep web data sources, as 1.5% of web pages with potential government data that are analyzed are found to contain data stored in the deep web. This work contributes to the development of a novel method that allows for the continuous checking and identification of government data from surface web data portals. In addition, this method can be scaled and repeated to assure the widest possible content coverage.

Keywords: Deep Web; Data Portals; Assessment; Open Government Data; Benchmarking

1. Introduction

The rise in use of Information and Communication Technologies (ICTs) has revolutionized the way citizens interact with governments and take part in their decisions. In the last decade, a movement has emerged and gained increasing importance that calls for the unrestricted access and consumption of data through a specific infrastructure conceptually called "Open Government Data" or simply "open data."

Open data establishes a set of requirements for institutions to follow that are conveyed in the

form of guidelines and principles for the opening of public records, mainly accomplishing this task through the use of ICTs. In essence, to comply with open data principles is to meet certain conceptual and technical requirements allowing data to be freely used, reused and redistributed to anyone for any purpose (Open Knowledge Foundation, 2012; Tauberer, 2014).

Ensuring the availability of open data increases transparency, accountability and value creation by making government data available to everyone, including to machines through automated data processing. This makes it possible for members of the public – typically journalists, economists, political scientists and other experts with critical views – to become more involved in the operations of governments.

Though open data principles have been adopted around the world through policies and practices, often driven by legislation, the problem of developing automatic assessment methods for open data is an emerging area of research that practitioners in the field have been struggling to make progress in. At present, it is not enough that governments merely introduce their own open data initiatives: citizens have the right to confirm whether or not these initiatives are truly designed to enhance social, environmental and economic outcomes through measures for benchmarking open data (Caplan et al., 2014; Ulrich, Tom, & Jamie, 2015). The import of such measures can be inferred from Janssen et al. (2012) who, in investigating several of the myths regarding open data, stress the need for new types of governance mechanisms and policies that counteract the idea, for example, that "the publicizing of data will automatically yield benefits."

Such new types of governance mechanisms and policies seem to be justified given the ways in which the open data portals that governments rely on are often misimplemented, especially those spread across regional and local levels of government. These data portals are usually developed with the classic website model in mind where access to data is guided by human interaction among web forms.

As there are countless such data portals worldwide, demand for large-scale, high-frequency and low-cost automatic benchmarking assessment methods has become increasingly pronounced (Ulrich et al., 2015). Moreover, as there is no guarantee that government agencies are properly implementing their data infrastructure with sustainable open data software platforms, much of the data within these data portals is often found in areas of the Internet that are behind web forms and thus not registered with any search engine – in typical deep web contexts, in other words (Bergman, 2001). Developing automated benchmark exercises on such data portals thus becomes a complex computational problem that, for data openness to be assessed, must ultimately be addressed.

That research regarding automated assessment methods is lacking has been confirmed in a paper by Ulrich et al. (2015), in which the authors explore how feasible it would be to conduct automated assessments based on a generalized framework. The authors clearly recognize that not all their suggestions are feasible, given that the assessment methods that are most widely used today are entirely manual-based, requiring significant amounts of human interaction and reasoning.

In addition, tools and methods that have been developed that attempt an automated approach

have assumed the existence of standardized data catalogs in which metadata is the primary artifact on which the assessment process is conducted. Standardizing data catalogs in this way, however, would require agencies to have a highly informed understanding of open data and skilled IT staff capable of implementing specialized software platforms as a central component of their open data infrastructure.

This paper proposes an automated method for assessing the technical aspects of open data by evaluating the data's compliance with open data requirements derived from the well-established open data principles. The proposed method accounts for data in deep web contexts, and is carried out following three basic steps: 1. Access, 2. Classification, and 3. Decision-making. The paper proceeds to describe an experiment carried out on the data portals of all 27 Brazilian capitals, in which the first two steps of the proposed method are applied to yield consistent results.

The contributions made in this work are twofold. First, we fill a gap in the literature by assessing government-related data stored in the deep web, which has previously challenged practitioners seeking to perform benchmarking exercises. Furthermore, we contribute to the development of a novel approach that allows for the continuous checking and identification of such data in the deep web through a process that can be scaled and repeated to assure the widest possible content coverage. Gaining access to deep web content sources is proven to be feasible from the surface web in this method, which simply requires a single known URL to serve as a seed to carry out the overall process.

The remainder of this paper is organized as follows. Section 2 presents relevant background knowledge, including fundamental information regarding open data compliance and the newly-introduced deep web in which government-related data is stored, while Section 3 presents related research in the field. Section 4 details our proposal in the form of a method lifecycle, explaining how the steps of the process were implemented in experiment carried out for the project. Section 5 discusses the findings of this experiment, and Section 6 concludes the paper with reflections and closing remarks.

2. Background

In this section, we briefly explain the significance of compliance with open government data principles in order to clarify what the assessment method proposed by this work is designed to accomplish. We also outline the basic information needed to understand (open) data portals and their relation to the recently created deep web of government-related data.

2.1 Open Government Data compliance

The term "Open Government Data" became popular after a working group in 2007[1] defined a set of eight principles expounding a philosophy regarding the production and commission of data by

[1] https://opengovdata.org

public bodies that is based on the idea of openness, or the free availability of data for use, reuse or redistribution by anyone for any purpose (Open Knowledge Foundation, 2012; Ubaldi, 2013).

Since then, the concept of Open Government Data has evolved to include a total of 14 principles reflecting a more robust understanding of open data (Tauberer, 2014). In 2015, open data experts from governments, multilateral organizations, civil society and the private sector drafted the International Open Data Charter,[2] which includes a list of six core principles meant to define this global movement that aims to generate significant social and economic benefits through civic engagement.

Though they are far from exhaustive, the principles associated with Open Government Data and the International Open Data Charter have since served as guidelines for data publishing practices and established criteria for the evaluation of open data initiatives. On the basis of these principles, the initiatives of various countries, organizations and projects have been and continue to be assessed (most often manually) across several dimensions related to data content, data manipulation, participation and engagement capabilities (Sayogo, Pardo, & Cook, 2014). Table 1 presents several examples of regularly utilized open data assessment methods that are described in the literature.

Table 1

Prevalent examples of current open data assessment methods

Study's promoting institution	Method and timeframe	Unit of analysis	Coverage and frequency
Open Data Barometer (Brandusescu, Iglesias, & Robinson, 2016) www.opendatabarometer.org The World Wide Web Foundation	A peer-reviewed expert survey carried out between May and September 2016.	Datasets submitted by national governments.	The 2016 fourth edition covers 155 countries. Previous editions published in 2013, 2014 and 2015.
Global Open Data Index https://index.okfn.org Open Knowledge	Domain expert reviewers responsible for checking data across all locations (countries). Data refers to the period from October 2016 to March 2017.	Datasets submitted by national governments.	94 countries in the 2016/2017 edition. Ongoing project with previous releases in 2013, 2014 and 2015.
Survey on Open Government Data (OECD, 2017; "Open Government Data," 2017; Ubaldi, 2013) OECD	Survey completed by public sector officials from OECD countries and partners with analysis from the OECD Secretariat. Survey conducted in November and December 2016.	Responses from central/federal governments.	35 OECD countries and 3 partners (Colombia, Lithuania and Peru). Pilot index launched in 2015 as part of the OECD Government at a Glance.

[2] https://opendatacharter.net

Open Data Inventory (Open Data Watch, 2017a, 2017b) http://odin.opendatawatch.com Open Data Watch	Research carried out by trained researchers. Assessments were carried out between June and October 2017.	Websites maintained by national statistical offices (NSOs).	The last inventory includes NSOs in 180 countries. Previous release in 2016.
E-Government Survey (United Nations Publications, 2016) United Nations	Desk research with assessment by at least two research studies. Collection of data spanned from May 2015 through July 2015.	National web portals.	193 countries in the in eight editions of the survey since 2003. Questions about open data were introduced in the previous 2014 edition.

The assessment methods listed in Table 1 are all clearly designed to benchmark open data. As these examples indicate, organizations around the globe appear to be in agreement about the need to produce quantitative evidence that the promised benefits of open data are being delivered. It is important to note, however, that generally only certain aspects of open data can be easily assessed and represented in quantitative terms, including technical features such as the format, completeness, accessibility and machine-readability of the data in question. Assessing other aspects of open data such as the impact it might have, whether it is up to date and its comprehensiveness requires significant human reasoning, making the process complex and time-consuming.

2.2 Data portals, open data platforms and the deep web

Data portals are a key component of any data infrastructure. To understand the role that data portals play in data infrastructure and the significance of data portals in the present paper, we need first to differentiate "data portals" from "open data portals," bearing in mind that not all data portals publish open data.

After the concept of open data initially entered onto the scene, promising benefits in a variety of areas, various governments around the world rushed to implement their own data infrastructures to permit the consumption of data by their citizens. Others have done the same, but in a more reactive way, driven usually by the enforcement of laws or simply the understanding they ought to have such a data infrastructure. The implementation of data infrastructures arising from both scenarios has raised concerns about data openness and the sustainability of portals.

An open data portal is usually built upon an open data platform, sometimes known also as an "open data catalog" or "open data repository." Both open data portals and open data platforms use software engines that permit integrated open data management and include features such as metadata support and management, basic visualizations, user management tools, data publishing, data storage capabilities and natively-exposed API support. The implementation of open source software solutions has been often recommended as a means to make portal architecture more sustainable

(European Union, 2017). In this way, one of the software platforms most frequently implemented by high-load open data portals like the European Data Portal[3] and the data portal[4] of the US government is the open source Comprehensive Knowledge Archive Network (CKAN). Other widely implemented proprietary solutions include Socrata, OpenDataSoft and ArcGIS Open Data (Correa, Zander, & da Silva, 2018).

Starting in 2007, CKAN has been maintained by Open Knowledge International, a worldwide non-profit network whose work focuses on openness and knowledge-sharing through the use of technology. CKAN has an active network of developers who work constantly to improve the platform so that it can serve as an affordable out-of-the-box solution for any type of organization. A study by Osagie et al. (2015) comparing CKAN to 11 other open data platforms currently available on the market concluded that the platform fulfilled 9 out of 12 criteria defined by the study to indicate the overall quality and reliability of the platform, and that its main strength consisted in the collaborative community of developers who support the product. Some weaknesses of CKAN include its data analysis and visualization tools, which are still developing in relation to those of its counterparts.

The two most important features of an open data platform are its metadata and API interfaces. Metadata is a structured description of content (or of the data itself) including basic information, for instance, about the authorship, category, provenance and licensing of the data in question, all of which is essential to describing the data in an accurate way that facilitates its discovery by consumers. Among a myriad of existing metadata standard proposals, Data Catalog Vocabulary (DCAT) (Maali & Erickson, 2014) in particular is recommended by the World Wide Web Consortium (W3C) for the web-based publishing of data and has been used as a model for the homogenization of varying metadata sources with heterogeneous schemas within existing open data platforms (Neumaier, Umbrich, & Polleres, 2016). Meanwhile, API interfaces extend metadata by allowing agents (most often programmatically) to retrieve data descriptions in a structured format that insures interoperability across different types of requests, no matter whether the request comes from a web browser or a programming language.

When a data portal does not implement an open data platform, its specifications are made precisely to meet the needs of the institution. In this case, there is no guarantee that the data portal will provide features to support data openness, as these data portals usually require one to fill particular parameters in web forms before being able to access or download data in formats such as HTML, PDF, CSV, or Excel Spreadsheets. Such practices defy the two principles of accessibility and machine-processability that are both essential requirements for the implementation of open data. Ensuring the availability of metadata and the use of API interfaces is therefore critical to permitting the discoverability and accessibility of data.

Such shortcomings, particularly those related to file formatting issues, can be found in data

[3] www.europeandataportal.eu
[4] www.data.gov

portals used at every level of government (Bunyakiati & Voravittayathorn, 2012; Davies & Frank, 2013; Helbig, Cresswell, Burke, & Luna-Reyes, 2012; Machado & Oliveira, 2011; Ribeiro, Matheus, & Vaz, 2011; Veljković, Bogdanović-Dinić, & Stoimenov, 2014). But as local governments tend to work in a less centralized manner, and their IT staff are usually free to plan, acquire and implement the data infrastructures they prefer, openness seems to be lacking most frequently in the data portals used by local governments. In particular, a specific type of data portal has come to be widely used in local governments that adopts a classic website approach: a non-open data portal, essentially. Multiple studies have investigated the relation between the use of this type of data portal and the lack of data openness in local governments. Surveys (Andreiwid Sh. Corrêa, Paula, Corrêa, & Silva, 2017; Andreiwid Sheffer Corrêa, Corrêa, & Silva, 2014) conducted in Brazilian municipalities have revealed that HTML is the most frequently used format for data publishing in data portals across the country. Likewise, Lourenço et al. (2013) have conducted an assessment of data portals for 94 municipalities in Portugal and Italy, finding that these municipalities generally did not disclose data properly, as their data lacked visibility and proper format and structure.

In this context, the present work specifically considers government-related data in the deep web. The term "deep web" became famous following the publication of a white paper by Bergman (2001) in which the author articulated the differences between the deep web and surface web, explaining that "deep web sources store data in searchable databases that only produce results dynamically in response to a direct request." This definition is adapted to the aims of the present work by treating query parameters as direct requests that are input into web forms prior to the production of the databases' dynamically generated data. Open data portals, in contrast, make data natively available and discoverable from the surface web, usually through the use of metadata and API interfaces. Figure 1 provides an illustration that compares open data and non-open data portals and indicates the relation of each to the deep web.

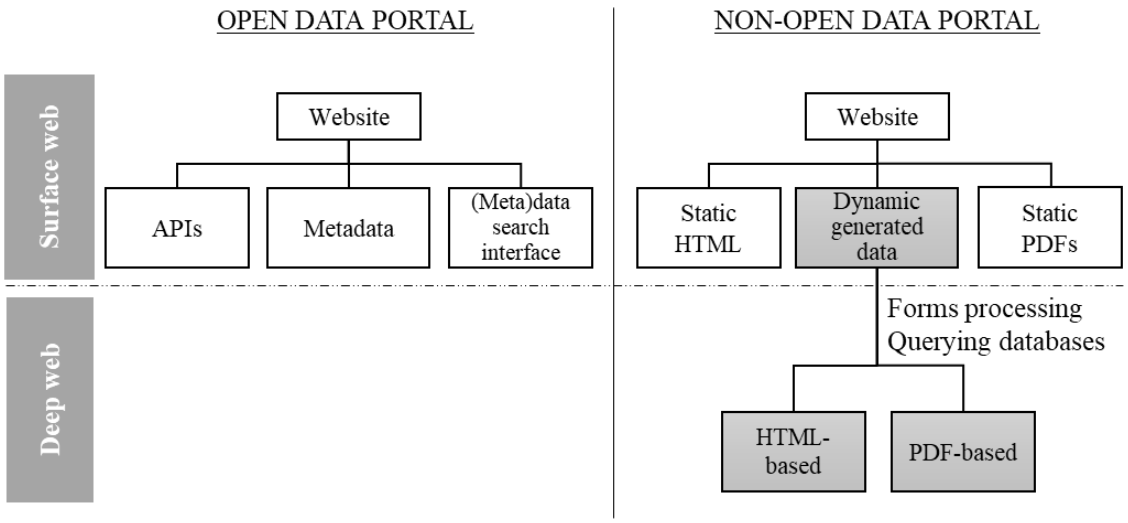

Figure 1. Comparison of open data and non-open data portals and their relation to deep web

Before going on to discuss the topic of government-related data in the deep web, it is important to distinguish the deep web from the dark web. The former consists of Internet web pages that are publicly accessible but not registered with any search engine; as web crawlers typically do not index such content, these web pages are only accessible through specific user query databases. On the other hand, the dark web is often publicly accessible on the Internet, but the communities associated with dark web pages use an extra layer of protection to preserve their anonymity and autonomy, the most famous example being Silk Road, an online black market used to trade illicit goods and services (Bradbury, 2014). As a result of this extra layer of protection, dark web pages can only be accessed by people with clear intention to illegality.

Government-related data in the deep web is publicly available on the Internet, but only accessible through means other than one would expect to access an open data portal. To illustrate the process by which data is obtained from the deep web, Figure 2 presents a typical example of a non-open data portal that requires form processing and web database querying to access the data it contains. Apart from being difficult to access and download, data in this type of data portal is only generated after a manual intervention (form processing) is carried out, thereby prevent the data from being discovered beforehand. Moreover, no exposition of metadata is provided in such portals, a resource essential for describing the data contained within.

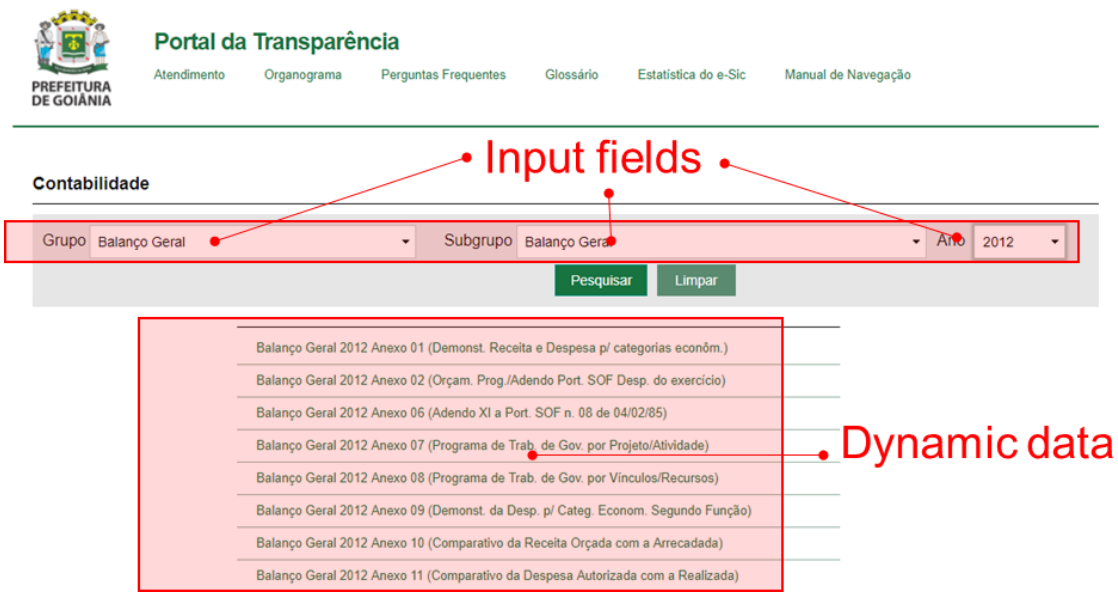

Figure 2. A non-open data portal that requires form processing and the querying of web databases to access data in the deep web. Extracted from: http://www10.goiania.go.gov.br/transweb/Contabilidade.aspx. Accessed: 03/12/2018 10:20am

In Figure 2, one can see that the web page provides three dropdown lists as input fields and a

submit button marked “Pesquisar” (meaning “Search”). Users need to select options from dropdown lists and click the submit button to access the underlying dynamically generated data. Otherwise, the dynamic data remains hidden in the deep web.

3. Related research

Given the high demand for implementations of open data as a means to improve the accountability of and publicity surrounding government projects and initiatives, we identified several works adopting similar theoretical and practical approaches to our own that contribute to our understanding of the growing importance and urgency of developing methods to assess open data on a larger scale, at higher frequency and with lower costs.

To the best of our knowledge, however, no empirical research as of yet has been conducted on automated methods to assess technical aspects of Open Government Data that specifically focuses on data in the deep web. As we discuss below, the studies that have been conducted and the tools and methods that have been developed assume that the assessments will be carried out on open data platforms in which metadata is the primary artifact, such that the metadata can be freely used to conduct the automated assessment process. As we have illustrated in previous sections, this is not the case when dealing with non-open data portals with data in the deep web.

A paper concerning the automatic benchmarking of open data (Ulrich et al., 2015) has explored the potential feasibility of conducting automated assessments using a methodological framework called “Common Assessment Methods for Open Data” (CAF), the first version of which was developed by the World Wide Web Foundation in a workshop held in June 2014 (Caplan et al., 2014). The CAF framework, however, only provides a standardized conceptual overview of four high-level dimensions of open data that can vary widely in their potential for automation. Of these four dimensions, the data dimension is most relevant to our interests, as it concerns the technical openness, relevance and quality of the data. Ulrich et al. (2015) argue that the data dimension has the highest potential for automated assessment, despite providing only a few idealized metrics with which automated assessments could be carried out and emphasizing that automation requires high-quality metadata, which is normally accessed through specialized software such as open data catalogs. Ultimately, then, the capabilities of the CAF framework do not allow for the automated assessment of the data dimension outside of these ideal circumstances.

The Open Data Monitor[5] is an online tool that generates overviews of available open data resources, focusing on regional, national and pan-European open data repositories. It provides a system platform that is used to collect metadata from the well-known open data catalogs CKAN and Socrata. The project also harvests data from HTML pages with specific metadata from the W3C’s DCAT (Maali & Erickson, 2014). The metrics that the Open Data Monitor provides relate to the existence and availability of open licenses, the machine-readability of datasets and metadata

[5] http://opendatamonitor.eu

completeness. There is no reason to view Open Data Monitor as an automated assessment tool, however, with respect to discoverability because, as the methodology page of the project makes clear, agencies must apply for registration via email prior to gaining access to the tool dashboard of the project's platform. This platform also requires agencies to collect and organize their metadata from CKAN-based catalogs or Socrata platforms located in the surface web, for which agencies must have skilled IT staff capable of carrying out the task.

Open Data Certificate[6] is another online tool that formally recognizes sustainable publications of quality open data. The tool issues badges identifying levels of achievement of particular open data publication (thus the title "certificate"). Open Data Certificate seems to employ a broader concept of open data than is typical, as it is uncommon to see private companies whose data repositories have received such certificates. To be issued a certificate, institutions must fill out a form to request it; the Open Data Certificate system then checks whether the institution meets the requirements using established DCAT metadata or open data catalogs.

A series of studies (Neumaier et al., 2016; Umbrich, Neumaier, & Polleres, 2015) have been conducted focusing on automated quality assessment methods which primarily use metadata to monitor and assess the quality of open data portals. The authors of these studies first reported on the automated monitoring of 82 CKAN portals, which provided several interesting findings, such as the observation of metadata heterogeneity across portals, a growth in the overall number of datasets and a majority presence of open formats and open-license exposing datasets. Later they improved upon their work, using a generic model to map metadata from the three most widely used data catalogs (CKAN, Socrata, and OpenDataSoft). At the time of this writing, Neumaier et al. (2016) have made available an online tool called "Open Data Portal Watch"[7] that provides reports on the monitoring of 261 data portals. The tool features a dashboard user interface that presents gathered data for selected periods of time. A more recent tool that is derived from Open Data Portal Watch (Kubler, Robert, Neumaier, Umbrich, & Le Traon, 2018) compares 250 open data portals in 43 different countries, seemingly using the same framework as in Neumaier et al. (2016).

The last three initiatives mentioned above (Open Data Monitor, Open Data Certificate and Open Data Portal Watch) each utilize a dashboard that provides users an integrated perspective of quantitative evidence, with the ultimate aim of aiding public awareness about the development of open data. These initiatives only become effective when interaction with them is stimulated through communication, by allowing feedback from the public, for example (Matheus, Janssen, & Maheshwari, 2018); this is inconceivable, however, if the relevant data is hidden behind forms and rendered undiscoverable by its location in the deep web.

Of the tools and studies that are discussed above, none take into account the existence of data in the deep web where metadata is not available at all; such deep web resources are specifically designed not to allow agents to interact with them or their underlying data to be automatically

[6] https://certificates.theodi.org
[7] http://data.wu.ac.at/portalwatch/portals

described. As we have indicated previously, the data publishing practices of government agencies frequently involve the construction of silos of data relying on dynamically generated content. Following the principles of the open data movement, there is an urgent need for assessment methods that can be used to evaluate the technical aspects of data in such portals and the degree to which they are compliant with open data principles.

4. Proposal of method lifecycle

The considerable number of websites that currently serve as data portals at all levels of government would make it impossible to conduct any sort of manual assessment frequently on a large scale at relatively low cost. As a result, it is necessary that automation techniques be used to carry out such assessments in an efficient manner.

Efficiency, in this case, means being able to check a large number of governmental websites relatively frequently. The continuous checking of data availability across websites in search of data portals is the first step of the process, preceding the assessment itself. Once the websites are checked, potential data portals are identified to start the assessment process. This step should be repeated regularly given the possibility of changes and developments, and as the only limitations to carrying out this process are the computational resources at hand. To design this stage of the method, we have relied on techniques for extracting web content with efficiency that are found in the literature.

One source of inspiration was an extensive survey conducted by Ferrara et al. (2014) concerning techniques and applications for web content extraction. The authors highlighted the use of web wrappers, which are defined as procedures involving one or potentially many classes of algorithms designed to search for and find data in semi- or unstructured web sources. In a web wrapper process, algorithms containing regular expressions are generated to form a basis for the finding of specific HTML elements. The authors also describe hybrid approaches (Crescenzi, Mecca, & Merialdo, 2001; Della Penna, Magazzeni, & Orefice, 2010) that achieve higher levels of automation by using models for decision making, an approach that is highly suitable to our aims in the present work.

Figure 3 illustrates the overall method lifecycle we propose implementing for the continuous monitoring and classification of data portals and assessment of Open Government Data.

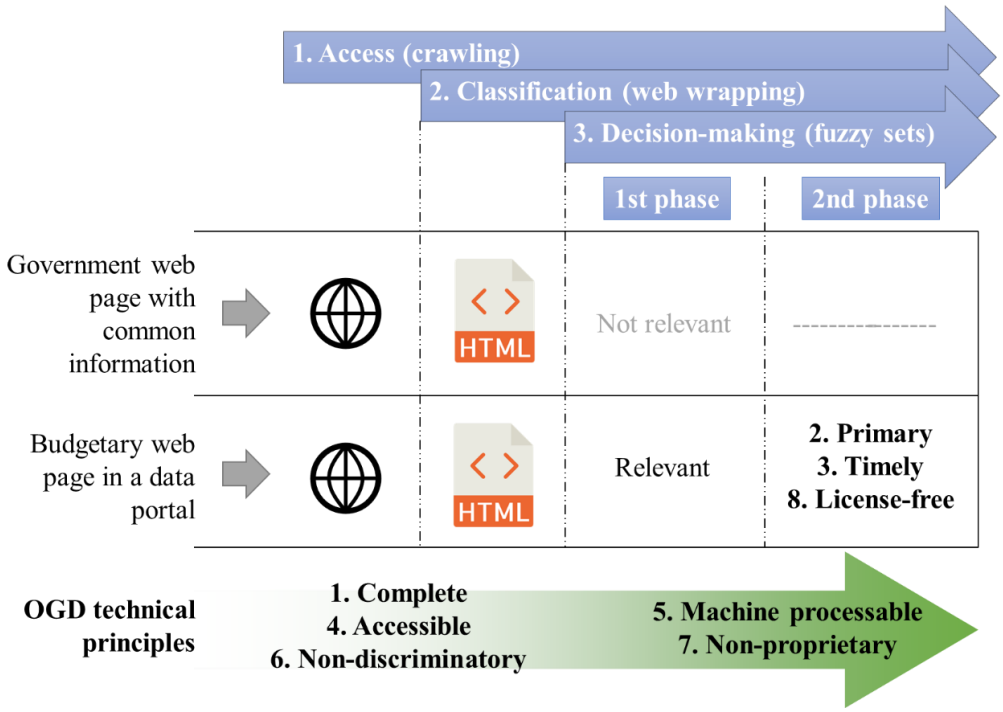

Figure 3. Proposed method lifecycle for continuous monitoring and classification of data portals and assessment of Open Government Data

As seen in this visualization, the method lifecycle is divided into three basic steps that can be summarized as follows:

1. **Access**: This step requires a known URL to be used as a seed for input into the web crawling algorithm. The URL is generally a root address representing the website of a government agency, such as www.london.gov.uk, which directs to the city of London's main website. The algorithm follows all the hyperlinks found in the initial page of the website to find its subpages. This is a recursive process that is repeated until the system crawls the entire website, storing the absolute URLs in a database for later wrapping.
2. **Classification**: Once a database with the absolute URLs of an agency website has been constructed, this step involves the employment of web wrapper techniques. First, HTML source code for each web page is downloaded whenever possible and analyzed to detect specific keywords that identify web pages with potential government data – that is, to detect typical data portals. Second, the algorithm checks the web pages for the presence of web forms, and their input field configurations are determined. The classification step aims to assign a specific weight to each web page with content in the deep web. There is no guarantee that the government data can be found, however, as the target content may require users to fill out web forms before returning dynamic data from web databases (a scenario typical of deep web content).
3. **Decision-making**: There are two phases in the decision-making step. In the first phase, the

algorithm is expected to automatically decide whether a web page's content is relevant to Open Government Data. In other words, after finding an HTML page with the expected dynamically generated data, the algorithm distinguishes whether this HTML is being used to display budgetary data, which is of interest to Open Government Data, or disclose other agency information that is not relevant. In the second phase, the algorithm is expected to assess the compliance of this data with open data principles after querying the underlying database through HTML forms. The automatic processing of web forms to reach dynamic data in the deep web and execution of both phases of the decision-making step involves various experimental techniques that, as indicated below, we are still in the process of developing (Khurana & Chandak, 2016).

Each step illustrated in Figure 3 corresponds to one or more checkpoints corresponding to the Open Data Principles. Technical principles are straightforward to check within the method lifecycle we have proposed because their assessment is highly suitable to automation. On the other hand, non-technical principles involve decision-making activities that are usually made by an expert, thus posing challenges to automation attempts. Table 2 details the Open Data Principles (Tauberer, 2014) and their corresponding requirements, explaining how these are checked within the proposed method lifecycle.

Table 2

Open Data Principles, their requirements and checking measures within the proposed method lifecycle

Principle	Requirements	Way to check within the proposed method
Technical principles		
1. Complete 4. Accessible	Complete and accessible data must be freely available on the Internet to the widest possible range of users for any purpose, including to software for the purpose of data collection and decoding.	Principles 1 and 4 can be checked in Step 1 through web crawling. Negative evaluations in these principles are produced when a web page cannot be reached or a server refuses the use of the crawlers.
6. Non-discriminatory	Data must be available to anyone with no requirement of registration; this includes avoiding discriminatory practices in API terms of services agreements.	Principle 6 can be checked in Step 2. A negative evaluation is produced when a web form requires pre-registration to use data or includes password-typed input fields.
5. Machine-processable	Data must be reasonably structured to allow automated processing guided by the choice of file format.	Principles 5 and 7 can be checked only at the end of Step 3, as it is necessary to first reach the data to check its file format and

7. Non-proprietary	Data must be available in a format over which no entity has exclusive control, mainly through the use of open formats, such as CSV for tabular data.	openness. A negative evaluation is produced when data is encoded through PDF, scanned images or non-open file formats or using free-form text.
Non-technical principles		
2. Primary	Collected data is just like that found at the source, with the highest possible level of granularity and not in aggregate or modified forms.	Principles 2, 3 and 8 are non-technical principles, and constitute the most challenging part of the process due to uncertainties they involve that would ordinarily require manual assessment by an expert. For this purpose, we propose using fuzzy sets with templates of data disclosures to assess data portals and deal with uncertainty. Negative evaluations for these principles are produced when a data portal does not comply with desired templates as defined by the fuzzy sets.
3. Timely	Data is made available as quickly as is necessary to preserve the data's value.	
8. License-free	Data is not subject to any copyright, patent, trademark or trade secret regulation.	

In the following subsections, we detail the experiment we carried out that involved the implementation of steps 1 and 2, i.e. Access and Classification. The procedure for step 3, Decision-making, is still being developed; in the subsection dedicated to this step, we thus present the theoretical approach we are taking to this step's development**.** In these subsections, we describe the various tools and procedures that were used in these steps to obtain the experiment's results. All collected data is openly provided in this paper's data source.

4.1 Access

This step essentially involved applying web crawling techniques to analyze the surface web, which is the portion of the web that can be discovered by following hyperlinks. The process involved in this step consists of using a particular seed URL to check the entire website of government agencies.

Here we relied on a tool called "GNU Wget" ("Wget," 2017). Wget was introduced in 1996 and is widely used among Unix users and in Linux distributions. This tool was originally designed for downloading web resources or mirroring entire websites to allow them to be accessed locally. With some configuration adjustments, Wget can be made to work in spider mode, which means that it will not locally store pages but instead simply check them and log essential information such as the date, the time of access and the absolute URL that was reached. This feature is combined with application of the recursive option to allow the tool to follow and check all the hyperlinks on a page, repeating the process on all underlying subpages in a recursive loop until the entire website has been crawled

or a maximum application of recursion has been reached.

For this step in our experiment, we ran Wget on a 64-bit Windows environment using version 1.19-1 of the tool. We configured Wget according to the parameters and values detailed in Table 3. Besides specifying the parameters used, this table provides a brief description of each particular configuration and explains the reason for applying it in this experiment. An exhaustive list of all possible parameters and how to use them can be found in the Wget official user manual.[8]

Table 3

Wget configuration parameters used in experiment

Parameter and value	Description
execute robots=off	Instructs Wget to ignore robots' exclusion list, which is a configuration that be set up on a website to prevent crawlers from accessing its web pages. As a data portal is supposed to be accessed through automatic means to assure machine-processable data consumption, there is no reason to exclude robots in government websites.
user-agent="Mozilla/5.0 (Windows NT 6.1; WOW64; rv:40.0) Gecko/20100101 Firefox/40.1"	Configures Wget to act on behalf of a Mozilla browser and behave like a user graphically browsing an agency website. Some agencies somehow avoid agents other than known web browsers. We understand this to be a mistake, given that open data portals must be available to the widest range of users with no restrictions.
spider	Instructs Wget to act as a crawler and not download or mirror website content. This configuration is useful when one only wishes to follow hyperlinks.
recursive	Turns on recursive retrieving. This means that Wget first accesses the seed web page, then the web pages linked from that web page, then the web pages linked from each of those, and so on, until it reaches the desired depth (depth=5 by default).
no-verbose	Configures Wget to record only essential information of the process, namely date and time of access and the accessed URL. As a government agency website can contain millions of URLs, activating this option may help reduce the amount of data retrieved, thus simplifying post-processing.
local-encoding=UTF-8	Instructs the encoding system for URLs to use the most dominant character encoding for the world wide web.
output-file=*results.txt*	Configures Wget to record output results in a text file called "results.txt," which can be renamed later on.
span-hosts	Instructs Wget to span across any of the hosts of an agency's website domain. A government agency website domain such as london.gov.uk, for example, may have countless host names listed on the left side of the domain that can be crawled in this way. In this experiment, we used this configuration in combination with

[8] https://www.gnu.org/software/wget/manual/wget.html

	the "domains" parameter (see below) to restrict the hosts crawled to those associated with the government agency, as this can help to avoid a scenario in which the tool crawls the entire web.
domains=*brasil.gov.br*	Configures Wget to crawl within an agency's domain boundaries and avoid crawling the entire web. In one case in our experiment, Wget was configured to crawl within the domain of "brasil.gov.br," the Brazilian federal government's domain website.
no-host-directories	Disables the creation of any directory structure locally in the user's operating system due to the spider mode.
no-directories	As above, disables the creation of any directory structure locally in the user's operating system due to the spider mode.
no-check-certificate	Ignores server certificate warnings against available certificate authorities. Several government agencies employ their own certificate authority hierarchy due to the costs involved in this kind of acquisition.
random-wait	Instructs Wget to wait a random amount of time between requests. The purpose of this is to prevent agency websites that perform log analyses to search for statistically significant similarities in the times between crawling requests to identify that a retrieval program such as Wget is being used.
reject=js,css,ico,txt,gif,jpg,jpeg,bmp,tif,png,avi,mpeg,xml,mp4	Instructs Wget not to record particular web resources other than HTML-like web pages. In our experiment, we observed that some additional resources were retrieved despite the configurations of these parameters and values.

In order to get Wget running, we provided seed URLs corresponding to the main Internet address of each of the government websites of the 27 Brazilian capitals. When Wget tried to access these websites at the time it was run, the list of reached hyperlinks regularly varied due to connection quality issues and the technical availability of each website. Many websites did not respond as expected, or raised timeouts that influenced the collected results. We also noticed that due to the way some websites were built, Wget could not properly access all the website's hyperlinks. In light of these initial results, we decided to run nine instances of Wget using different machines at different dates and times while keeping track of all accessed hyperlinks in each instance. Table 4 lists these instances, providing their ID information and the date range of each Wget run.

Table 4

Wget run instance details

Wget instance	File ID in the data source	Date range of run
Instance 1	WGET_v1.7z	From 9/30/2017 to 10/02/2017
Instance 2	WGET_v2.0.7z WGET_v2.1.7z WGET_v2.2.7z	From 10/02/2017 to 10/11/2017
Instance 3	WGET_v3.0.7z WGET_v3.1.7z	From 10/02/2017 to 10/07/2017
Instance 4	WGET_v4.7z	From 10/10/2017 to 10/10/2017

Instance 5	WGET_v5.7z	From 10/16/2017 to 10/19/2017
Instance 6	WGET_v6.7z	From 10/16/2017 to 10/19/2017
Instance 7	WGET_v7.7z	From 10/17/2017 to 10/19/2017
Instance 8	WGET_v8.7z	From 10/17/2017 to 10/25/2017
Instance 9	WGET_v9.0.7z WGET_v9.1.7z	From 10/25/2017 to 10/29/2017

Each instance is associated with at least one compressed zip file containing a sequence of Wget log records, which in turn consist of the records of a batch of tries to access the 27 government websites. A total of 243 distinct log files were produced. Once this step was complete, we merged all the log files corresponding to the websites of each capital into a single file to produce a list of unique hyperlinks indicating web resources that were utilized in the following step. The number of unique hyperlinks can be checked in Table 9 in the column "Number of hyperlinks crawled."

4.2 Classification

This step involved using web wrappers to access the native sources of file formats (e.g. HTML) and capture the information in a machine-readable structure. An HTML web page was treated as an XML schema, making it possible to parse into elements in order to find specific terms.

To determine whether a given web page might disclose government data, we adopted the method of searching for specific keywords and web forms, then pulling data out of HTML files using the Python library Beautiful Soup (Mitchell, 2015). In the first part of this step, an algorithm was used to search for specific keywords to obtain a list of candidate data portals; in the second part, the candidates were analyzed to find web forms used to build database queries for the filling of dynamic web pages.

For the first part of the step, keywords were identified on the basis of a list of words that are frequently associated with government data and found in typical data portals. It is worth noting that the best keywords for a given context depend on the government whose data is being assessed and the native language of the area. In this study, we made use of the 2011 Brazilian Access to Information Law number 12.527 to compile a list of words relevant to data practices at subnational levels of government. This law established a legal framework of guidelines for the opening of data across all levels of government in the country, including the governments of the states in which the 27 capitals are located. In addition, we manually analyzed a number of websites in order to understand the ways governments tended to design their data portals and identify the words they frequently employed.

The algorithm used to find keywords ignored letter cases (upper/lower) and the accents used in the Portuguese language; however, it did consider string variations due to the composition of phrases and the use of punctuation marks, acronyms and other variants such as the singular and plural forms of words. Table 5 shows a list of examples of the most common keywords and the variations that were considered in this step. The entire list of keywords used in our experiment can be found in this

paper's data source.

Table 5

List of native words related to Brazil's open government legal framework and data practices

Words	String variations	Description
"acesso à informação"	"acesso a informações"	These terms are associated with the Brazilian Access to Information Law and serve to identify data portals that disclose data according to law requirements.
"12.527"	"12.527/2011," "12527," "12527/2011"	
"informações ao cidadão"	"sic," "e-sic," "serviço de informações ao cidadão"	
"transparência"	"Portal de transparência," "portal de informações," "portal do cidadão"	
"dados abertos"	"portal de dados abertos," "catálogo de dados," "armazém de dados"	These terms are associated with ordinary denominations of data portals. The variations were identified by noting typical practices and usage in subnational data portals.
"prestação de contas"	"Prestando contas," "contas públicas," "orçamento," "finanças públicas," "execução orçamentária," "despesas"	These terms are associated with government accountability. These variations were also identified by noting typical practices and usage in subnational data portals.

We configured Beautiful Soup to find keywords within any HTML elements. The algorithm used in this phase was run between 11/09/2017 and 11/15/2017, and each hyperlink that was analyzed which potentially contained open government data was recorded in a text file. This process produced 24 different text files grouped by capital that can be found in this paper's data source. The number of unique hyperlinks with potential government-related data can be checked in Table 9 in the column "Number of candidate web pages with data."

For the second part of this step, the algorithm used to find and analyze web forms identified *form* tags that were present. If form tags were found, the algorithm checked for *input*, *textarea* and *select* HTML tags to establish the number of input fields in each form, and then determined whether each web form was a candidate for building queries against a database. The input fields within a web form were counted according to their input type. Table 6 lists the HTML input types considered in this experiment and indicates whether they were counted as input fields.

Table 6

List of input types counted as input fields within web forms

Counts as an input field	Does not count as an input field
text	hidden
search	submit
date	image

datetime-local	button
email	radio
month	checkbox
number	color
range	file
time	reset
url	
week	
textbox	
tel	

The algorithm used in this phase was run between 11/15/2017 and 11/30/2017. The hyperlinks that were analyzed were recorded in CSV files. Each hyperlink was written using two lines, in which the first line identifies the analyzed hyperlink itself, and the second line contains data for classifying web forms according to their input field configurations. Raw outputs from this step can be found in this paper's data source, which contains a total of 22 CSV files named according to the initials of the corresponding Brazilian capital. Figure 4 illustrates the pattern used to delimit CSV files including the analyzed hyperlinks and the classification of web forms according to their input fields.

```
hyperlink
timestamp,#forms,#=1-field,#>1-field,#pwd-field
```

Where,

`hyperlink` – Indicates the full URL

`timestamp` - Indicates when the hyperlink was analyzed

`#forms` - Indicates number of forms found within a web page by the hyperlink

`#=1-field` - Indicates number of forms with only one field

`#>1-field` - Indicates number of forms with two or more fields

`#pwd-field` - Indicates number of forms with password-typed fields

Example:

```
http://intranet.riobranco.ac.gov.br/sistemas/telefone/
2017-11-21 17:46:41.974233,1,1,0,0
```

Figure 4. Pattern used to delimit CSV files, with analyzed hyperlinks and classification of web forms according to their input fields.

Web forms were ultimately classified into three categories that are represented by the Table 9 columns "=1 field," ">1 field" and "Password-type." The column "=1 field" indicates the number of

forms found with only a single input field; we took these to be web forms inviting users to search the entirety of a given website. The column ">1 field" indicates forms with more than one input field; we assumed these web form to have a high potential for use supporting query building mechanisms to be sent to database systems. Finally, the column "Password-type" indicates a special type of web form that requires a masked input and which is normally used to hide underlying content with a password or passcode that must be validated before the user can gain access. It is important to observe that a single web page may contain multiple web forms, such that the total number of web forms falling into these three categories may be higher than the number of web pages containing HTML forms, the latter of which is specified in the column "Number of pages with web form" in Table 9.

4.3 Decision-making

As we briefly mentioned above, this decision-making step involves form processing and querying of web databases in order to access dynamic data in the deep web. To make this possible, our proposed method begins with carrying out the Access and Classification steps described above, which were developed by adapting the domain-specific approach proposed by Wang et al. (2008), and which involve crawling potential data portals according to specific keywords (step 1) and then selecting candidate web forms based on the configuration of their input fields (step 2). Step 3 of our method thus begins by creating a subset of pages and web forms with the aim of dramatically reducing the computational power required for their processing.

To begin this step, we have sought to adapt a method developed by Zheng, Wu, Cheng, Jiang and Liu (2013) that involves the algorithm's learning through reinforcement the keyword queries that yield results rather than repeatedly conducting full-text searches, as is typical. In this way, the algorithm is designed to distinguish rewarding keywords from non-rewarding ones through experience. If we consider a website like that illustrated in Figure 2, for instance, in which inputs are selected from a dropdown list, we can randomly select values from those available and analyze the dynamic data that is retrieved to decide whether the selection is rewarded (if it returns a non-empty dataset) or not (if it returns an empty dataset).

Once a non-empty dataset is retrieved, the first phase of the decision-making process can be carried out, which consists of checking whether the content is relevant or not. If it is relevant, we move on to the second phase of this step, which aims to evaluate whether the open data requirements yet to be evaluated are fulfilled, namely those associated with technical principles 5 (Machine-processable) and 7 (Non-proprietary) and non-technical principles 2 (Primary), 3 (Timely) and 8 (License-free).

As mentioned earlier, the decision-making process involves various uncertainties that would ordinarily need to be handled by an expert. However, as we aspire to an automated approach, we wish to establish alternative models and techniques that facilitate an automated decision-making process, mainly by taking into account artificial intelligence techniques.

We envisage implementing mechanisms inspired by fuzzy logic or fuzzy set theory (Zadeh, 1965) within the decision process to define and evaluate criteria for the compliance of datasets with open data principles and their relevant requirements. This approach draws upon several previously developed tools and methods that employ multicriteria analysis methods (Opricovic & Tzeng, 2003) and fuzzy set theory for the modeling of particular metrics or criteria for a number of applications (Mardani, Zavadskas, Govindan, Amat Senin, & Jusoh, 2016) including the evaluation of website content (Bilsel, Büyüközkan, & Ruan, 2006; Büyüközkan, Ruan, & Feyzioğlu, 2007; Chou & Cheng, 2012).

Making use of fuzzy set theory, Bilsel, Büyüközkan and Ruan (2006), for instance, have described a model for measuring the performance of websites of Turkish hospitals. Ruan and Feyzioğlu (2007), meanwhile, have proposed ways of measuring performance in distance education websites. Finally, Chou and Cheng (2012) have developed a hybrid approach rooted in fuzzy set theory for evaluating the quality of websites of accounting consulting companies, and their results indicate that it is possible to classify in this way the positive and negative features of each website, thereby motivating companies to take actions to improve their web pages.

The process of acquiring knowledge in order to model a paradigm and represent it in fuzzy sets can be carried out in several ways. The most common is the soliciting of expert opinion (Negnevitsky, 2005). In the present work, website models are treated as the paradigm and basis for evaluation. Figure 5 provides an illustration of how fuzzy sets are developed through the use of website models to evaluate compliance with open data principles.

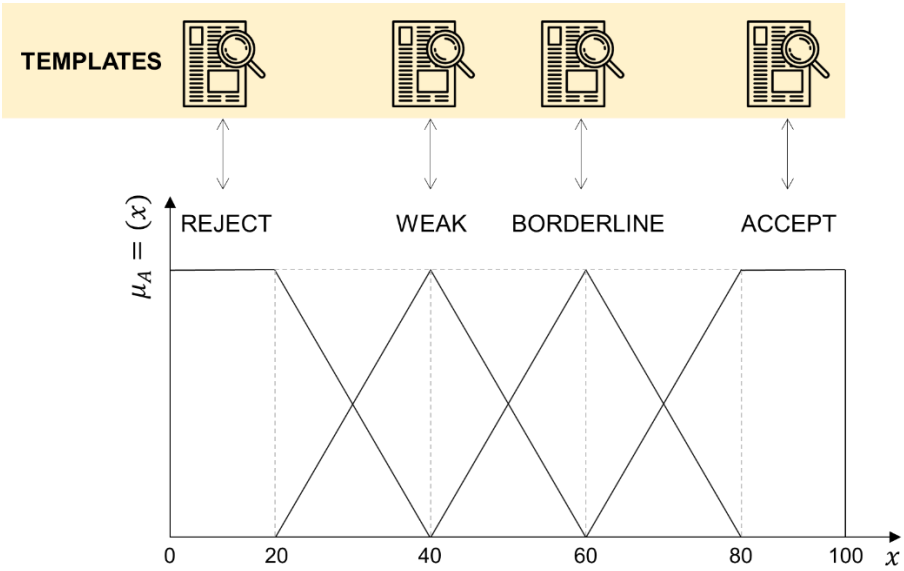

Figure 5. Representation of a fuzzy set according to website templates with open data principles requirements

The membership function $\boldsymbol{\mu_A(x)}$ that is illustrated in Figure 5 represents a fuzzy set defined by the linguistic variables "REJECT," "WEAK," "BORDERLINE" and "ACCEPT" in the universe of discourse represented by crisp values from 0 to 100. The composition of the set is determined by

website templates, which serve as a paradigmatic model for evaluation.

It should be noted that the templates to be developed represent both positive and negative models for evaluating compliance with open data principles and requirements. The precise features of the model may vary according to the preconceptions of each specialist or institution; the uncertainties arising from these variable conceptions, however, can be modeled using fuzzy scales that are represented in the sets through the use of linguistic variables.

Table 7

Gathered results grouped by Brazilian capitals.

State	Capital	Seed URL	(1) Number of hyperlinks crawled	(2) Number of candidate web pages with data	Web form and input field analysis			
					(3) Number of pages with web form	Number of pages with input fields		
						(4) =1 field	(5) >1 field	(6) Password-typed
Acre (AC)	Rio Branco	www.riobranco.ac.gov.br	182,623	27,568 (15.00%)	15,478 (56.00%)	15,468 (99.90%)	**115 (0.70%)**	5 (0.03%)
Alagoas (AL)	*Maceió*	*www.maceio.al.gov.br*	*518,899*	*N/A[2]*	-	-	-	-
Amazonas (AM)	Manaus	www.manaus.am.gov.br	217,373	131,824 (61.00%)	47,541 (36.00%)	47,461 (99.80%)	13,244 (27.90%)	56 (0.12%)
Amapá (AP)	Macapá	www.macapa.ap.gov.br	9,363	3,143 (34.00%)	2,314 (74.00%)	2,253 (97.40%)	**22 (1.00%)**	39 (1.69%)
Bahia (BA)	Salvador	www.salvador.ba.gov.br	231,158	62,771 (27.00%)	60,316 (96.00%)	3,216 (5.30%)	56,475 (93.60%)	8,784 (14.56%)
Ceará (CE)	Fortaleza	www.fortaleza.ce.gov.br	9,624	2,371 (25.00%)	2,343 (99.00%)	2,329 (99.40%)	**67 (2.90%)**	25 (1.07%)
Distrito Federal (DF)	Brasília	www.brasilia.df.gov.br	326,467	269,266 (82.00%)	79,222 (29.00%)	79,220 (100%)	**3,857 (4.90%)**	8 (0.01%)
Espírito Santo (ES)	*Vitória*	*www.vitoria.es.gov.br*	*235,538*	*N/A[2]*	-	-	-	-
Goiás (GO)	Goiânia	www4.goiania.go.gov.br	14,110	3,473 (25.00%)	2,136 (62.00%)	2,031 (95.10%)	**165 (7.70%)**	7 (0.33%)
Maranhão (MA)	São Luís	www.saoluis.ma.gov.br	747,910	733,520 (98.00%)	733,224 (100%)	732,857 (99.90%)	141,166 (19,30%)	360 (0.05%)
Mato Grosso (MT)	Cuiabá	www.cuiaba.mt.gov.br	583,040	261,404 (45.00%)	252,720 (97.00%)	252,620 (100%)	163,254 (64,60%)	24 (0.01%)
Mato Grosso do Sul (MS)	Campo Grande	www.campogrande.ms.gov.br	53,347	21,734 (41.00%)	21,718 (100%)	21,710 (100%)	**86 (0.40%)**	11 (0.05%)
Minas Gerais (MG)	*Belo Horizonte*	*www.prefeitura.pbh.gov.br*	*156,718*	*72,092 (46.00%)*	*N/A[3]*	-	-	-
Pará (PA)	Belém	www.belem.pa.gov.br	14,859	3,024 (20.00%)	2,017 (67.00%)	1,902 (94.30%)	**142 (7.00%)**	1 (0.05%)
Paraíba (PB)	João Pessoa	www.joaopessoa.pb.gov.br	139,278	34,124 (25.00%)	33,798 (99.00%)	33,281 (98.50%)	**495 (1.50%)**	26 (0.08%)
Paraná (PR)	*Curitiba*	*www.curitiba.pr.gov.br*	*582,418*	*379,717 (65.00%)*	*N/A[3]*	-	-	-
Pernambuco (PE)	Recife	www2.recife.pe.gov.br	422,264	389,798 (92.00%)	388,045 (100%)	387,052 (99.70%)	381,218 (98.20%)	363 (0.09%)
Piauí (PI)	Teresina	www.teresina.pi.gov.br	72,050	52,983 (74.00%)	52,823 (100%)	52,806 (100%)	**822 (1.60%)**	10 (0.02%)
Rio de Janeiro (RJ)	Rio de Janeiro	www.rio.rj.gov.br	29,178	7,838 (27.00%)	526 (7.00%)	296 (56.30%)	**225 (42.80%)**	25 (4.75%)
Rio Grande do Norte (RN)	*Natal*	*www.natal.rn.gov.br*	*N/A[1]*	-	-	-	-	-
Rio Grande do Sul (RS)	Porto Alegre	www2.portoalegre.rs.gov.br	160,740	79,971 (50.00%)	30,490 (38.00%)	30,179 (99.00%)	**339 (1.10%)**	1 (0.00%)
Rondônia (RO)	Porto Velho	www.portovelho.ro.gov.br	4,917	2,988 (61.00%)	601 (20.00%)	587 (97.70%)	**24 (4.00%)**	9 (1.50%)
Roraima (RR)	Boa Vista	www.boavista.rr.gov.br	19,818	17,112 (86.00%)	17,044 (100%)	240 (1.40%)	16,833 (98.80%)	2 (0.01%)
Santa Catarina (SC)	Florianópolis	www.pmf.sc.gov.br	153,270	101,494 (66.00%)	83,592 (82.00%)	54,812 (65.60%)	31,964 (38.20%)	38 (0.05%)
São Paulo (SP)	São Paulo	www.capital.sp.gov.br	180,738	166,374 (92.00%)	166,329 (100%)	166,329 (100%)	166,329 (100%)	8 (0.00%)
Sergipe (SE)	Aracaju	www.aracaju.se.gov.br	4,443	1,497 (34.00%)	1,188 (79.00%)	153 (12.90%)	**269 (22.60%)**	1,060 (89.23%)
Tocantins (TO)	Palmas	www.palmas.to.gov.br	592,604	14,682 (2.00%)	1,338 (9.00%)	35 (2.60%)	**1,272 (95.10%)**	41 (3.06%)
			5,662,747	2,840,768 (50.00%)	1,994,803 (70.00%)	1,886,837 (95.00%)	978,383 (49.00%)	10,903 (0.55%)

N/A[1] = Data not available data due to refused connection *N/A[2] and N/A[3] = Data not available data due to timeout and connection issues*

(1) Indicates the number of hyperlinks crawled based on seed URLs in step 1 (Access) of the proposed method.
(2) Indicates the number of candidate web pages with government-related data identified in step 2 (Classification) based on specific keywords. Percentages expressed are relative to values in column (1).
(3) Indicates the number of web pages containing HTML forms, regardless of how many forms a web page contains. Percentages expressed are relative to values in column (2).
(4) Indicates the number of web pages containing a form or forms with a single input field, likely indicating a search mechanism. Percentages expressed are relative to values in column (3).
(5) Indicates the number of web pages containing a form or forms with more than one input field, likely indicating a form to support query building. Percentages expressed are relative to values in column (3).
(6) Indicates the number of web pages containing a special type of field that requires a masked input, normally used to hide underlying content. Percentages expressed are relative to values in column (3).

5. Results

The entire data collection process, spanning the Access and Classification steps, was conducted between 9/29/2017 and 11/30/2017, resulting in a total of 5,662,747 hyperlinks crawled. Table 7 presents the results grouped by Brazilian capitals. Among the 27 capitals, the process was successfully completed and the results reported in 22 of them; in the case of the remaining five capitals, some access issues prevented the collection of complete data.

In particular, data corresponding to the column "Number of hyperlinks crawled" (denoted "N/A[1]") could not be reported in one capital (Natal), as the government website's system seemed to immediately refuse connection from spider-like algorithms. Data in the column "Number of candidate data portals found" (denoted "N/A[2]") could not be reported in two additional capitals (Maceio and Vitória). Although the first step applied to the websites of these two capitals' governments was successful, we experienced recurring timeouts and connection refusal in the second step, where it was necessary for the algorithm to download content prior to analysis. Likewise, data in the column "Number of pages with web form" (denoted "N/A[3]") could not be reported in two more capitals (Belo Horizonte and Curitiba) due to timeout and connection issues after only some dozens of hyperlinks were retrieved. The entire process and corresponding analyses were completed successfully for the remaining 22 capitals, however.

We thus begin this presentation of the results by analyzing some obvious outliers in order to refine the results. As signaled above, column (2) of Table 7 indicates the number of web pages that were probable candidates for containing government-related data, both as an absolute number and as a percentage of the total hyperlinks crawled as represented in column (1). Table 8 narrows these results to present the data of capitals for which more than 50% of crawled hyperlinks contained potential data.

Table 8

Capitals for which more than 50% of hyperlinks crawled were candidate web pages with data

Capital	Seed URL	(1) Number of hyperlinks crawled	(2) Number of candidate web pages with data
Manaus	www.manaus.am.gov.br	217,373	131,824 (61%)
Brasília	www.brasilia.df.gov.br	326,467	269,266 (82%)
São Luís	www.saoluis.ma.gov.br	747,910	733,520 (98%)
Recife	www2.recife.pe.gov.br	422,264	389,798 (92%)
Teresina	www.teresina.pi.gov.br	72,050	52,983 (74%)
Porto Velho	www.portovelho.ro.gov.br	4,917	2,988 (61%)
Boa Vista	www.boavista.rr.gov.br	19,818	17,112 (86%)
Florianópolis	www.pmf.sc.gov.br	153,270	101,494 (66%)
São Paulo	www.capital.sp.gov.br	180,738	166,374 (92%)
		2,144,807	1,865,359

The results collected in Table 8 are explained by the recurrence of specific keywords that indicate potential government-related data in the header and footer sections of nearly every page of some governments' websites. An example of a typical header found in these web pages is provided in Figure 6; in the website of this capital, a link to the main open data portal can be seen in nearly every web page's header section.

Figure 6. Specific keywords found regularly in web pages' header sections. Extracted from: http://www2.recife.pe.gov.br/noticias/22/12/2014/cultura-popular-anima-polos-do-recife. Accessed: 11/26/2017, 8:59pm

In Table 7, one can also find a high incidence of pages containing web forms, as shown in column (3), with 70% of the web pages with potential government-related data containing web forms. In turn, almost all of the pages with web forms (95%) contained one or more web forms with a single input field, as expressed in column (4). These results indicate the presence of search mechanisms, usually located in the header section of each web page.

Continuing on the topic of web forms, in 7 of 22 capitals' websites, a high percentage (more than 40%) of web pages contained forms with more than one input field. This fact is explained by the way some governments design their websites, putting multifield web forms in every header and footer section of each web page. An example of this website design is provided in Figure 7.

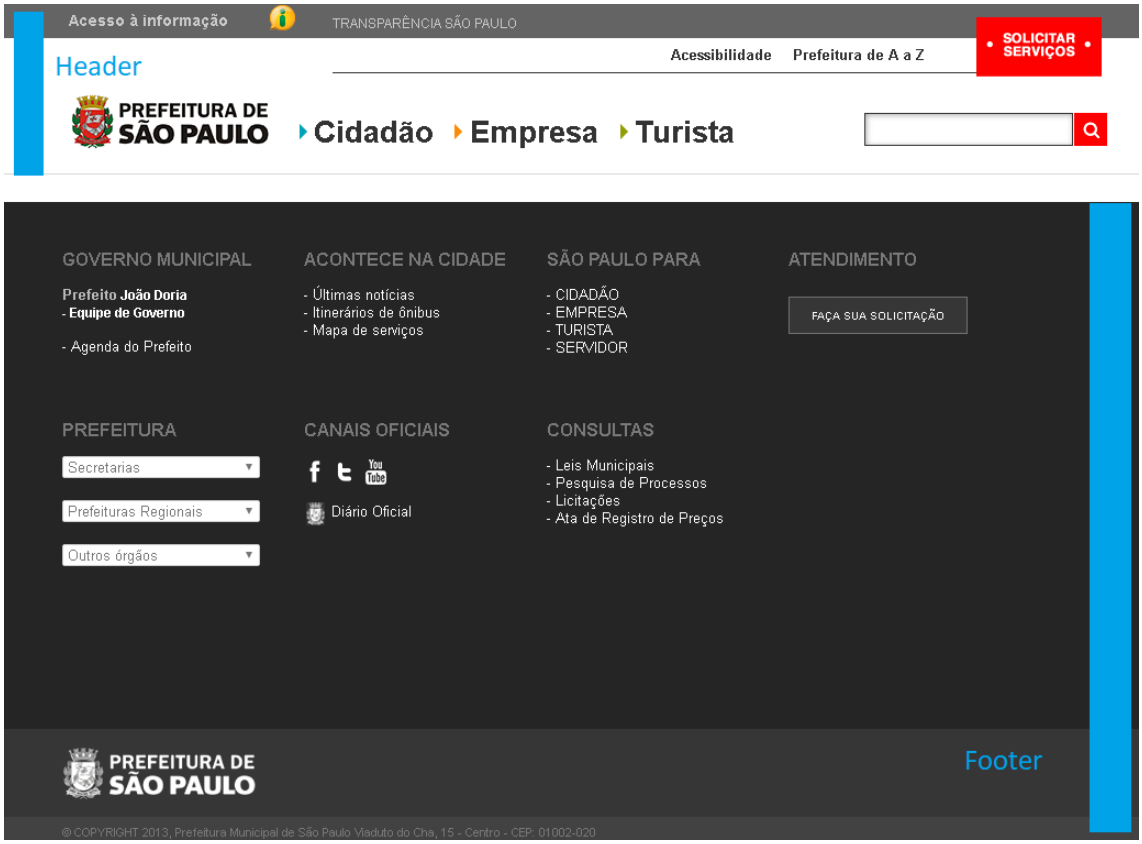

Figure 7. Header and footer from São Paulo website (adapted). Extracted from: http://www.capital.sp.gov.br/?p=1678. Accessed: 11/26/2017, 9:07pm

As for web forms with password-typed input fields, the website of Aracaju was a significant outlier, as 89.23% of its web pages included at least one such input field. This is due to the fact that the website of this capital has chosen to place login credential fields to their webmail system regularly in web pages' footer sections, as illustrated in Figure 8.

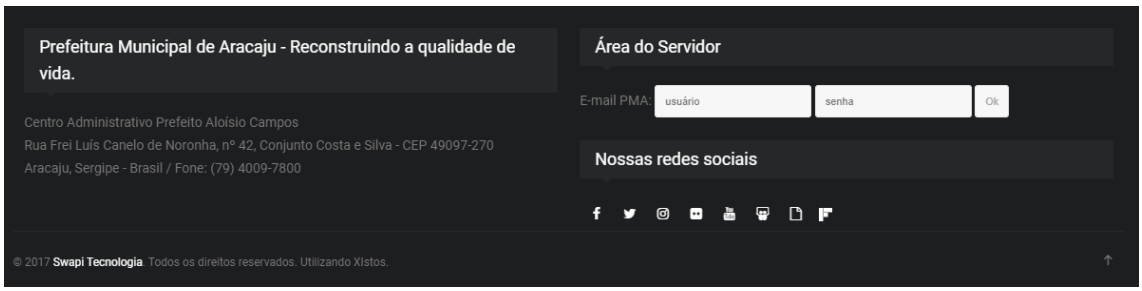

Figure 8. Aracaju website footer with a password-typed input field. Extracted from: http://www.aracaju.se.gov.br/transporte_e_transito/reclamacoes_e_solicitacoes. Accessed: 5/15/2018 10:46am

Websites such as in Figure 7 and 8 are examples that should be discarded due to the impossibility to conduct any web form and input field analysis. Therefore, concentrating on capitals with both a low incidence of web pages containing web forms with more than one input field (>1

field) and a substantial number of web pages with potential government-related data, we selected 14 capitals whose results suggested that the capitals may have sources of dynamic government data located in the deep web; these capitals are presented in Table 9.

Table 9

Capitals with high likelihood of having dynamic data located in the deep web

Capital	Seed URL	(2) Number of candidate web pages with data	(3) Number of pages with web form	(5) >1 field	(5) as % of (2)	(5) as % of (3)
Rio Branco	www.riobranco.ac.gov.br	27,568 (15%)	15,478 (56%)	115	0.4%	0.7%
Macapá	www.macapa.ap.gov.br	3,143 (34%)	2,314 (74%)	22	0.7%	1.0%
Fortaleza	www.fortaleza.ce.gov.br	2,371 (25%)	2,343 (99%)	67	2.8%	2.9%
Brasília	**www.brasilia.df.gov.br**	**269,266 (82%)**	**79,222 (29%)**	**3,857**	**1.4%**	**4.9%**
Goiânia	www4.goiania.go.gov.br	3,473 (25%)	2,136 (62%)	165	4.8%	7.7%
Campo Grande	www.campogrande.ms.gov.br	21,734 (41%)	21,718 (100%)	86	0.4%	0.4%
Belém	www.belem.pa.gov.br	3,024 (20%)	2,017 (67%)	142	4.7%	7.0%
João Pessoa	www.joaopessoa.pb.gov.br	34,124 (25%)	33,798 (99%)	495	1.5%	1.5%
Teresina	**www.teresina.pi.gov.br**	**52,983 (74%)**	**52,823 (100%)**	**822**	**1.6%**	**1.6%**
Rio de Janeiro	www.rio.rj.gov.br	7,838 (27%)	526 (7%)	225	2.9%	42.8%
Porto Alegre	www2.portoalegre.rs.gov.br	79,971 (50%)	30,490 (38%)	339	0.4%	1.1%
Porto Velho	**www.portovelho.ro.gov.br**	**2,988 (61%)**	**601 (20%)**	**24**	**0.8%**	**4.0%**
Aracaju	www.aracaju.se.gov.br	1,497 (34%)	1,188 (79%)	269	18.0%	22.6%
Palmas	www.palmas.to.gov.br	14,682 (2%)	1,338 (9%)	1,272	8.7%	95.1%
		524,662	218,551	7,900	1.5%	3.6%

As is seen in Table 9, only three capitals (Brasilia, Teresina and Porto Velho) were considered in which the number of candidate web pages with data, represented by column (2), was higher than 50%, the rest of those (seen in Table 8), being discarded. In these cases, the number of web forms with more than one input field (>1 field), particularly as a proportion of the number of web pages with potential government-related data, stood out. Palmas, meanwhile, is notable in this list for the proportion (95.1%) of its pages with web forms that include forms with more than one input field. However, only 2% of the web pages crawled from the Palmas website are candidates for containing government-related data, which motivated us to consider this capital as likely having dynamic data located in the deep web.

6. Conclusion

In this paper, we proposed an automated method to assess dynamically generated data located in the deep web, which is the hidden part of the web in which government-related data is often stored. The assessment process is designed mainly to evaluate compliance with technical requirements associated with the well-established open data principles. We applied this method in an experiment involving the government websites of the 27 Brazilian capitals; the process was successfully carried out for the websites of 22 of these capitals, with the result that 5,662,747 hyperlinks were crawled

and analyzed.

The proposed method implements an assessment process composed of three basics steps that are summarized as follow. In step 1 (Access), an algorithm crawls the entirety of governments' websites and stores all underlying web page hyperlinks to be used as input in the following steps, in the process allowing for the assessment of the open data principles 1 (completion) and 4 (accessibility). In step 2, (Classification) every web page is wrapped in search of keywords that indicate a potential web page with government data, and the presence of all web forms and their input field configurations are recorded; data principle 6 (non-discrimination) is thereby assessed. In step 3 (Decision-making), a proposal is made as to whether web page content is relevant to Open Government Data, and the content is finally assessed in terms of its compliance with the remaining open data principles.

In our experiment, we observed that connection quality and the technical availability of websites are critical to the effectiveness of this method. As step 1 requires the crawling of every web page associated with a government website, the process was naturally time-consuming. In addition, we noticed some websites delayed their fulfilling of our requests, particularly during working hours, and various problems occurred related to availability. This forced us to restart the process many times, which led to slightly different results being recorded according to the time of access. We thus recognize that this step of the process could be improved by having a greater quantity of dedicated computational resources made available for the task, or by taking advantage of existing projects such as Common Crawl[9] – an open and free repository of whole web-crawled data.

The results also revealed design issues in many web pages that made it difficult to automatically identify government-related data. One example is that of government websites which over-publicize their data-related practices and projects, advertising them in every web page's header or footer. As a result of such web page designs, the algorithm frequently found false positives that increased the recorded number of candidate web pages with government-related data. The subsequent analysis of web forms was compromised as a result.

In the analyses of web forms and their input fields, the high incidence of web pages containing forms with only one input field indicated that, while this type of form may be an essential tool for searching government websites, its presence is not helpful for identifying sources of dynamically generated data in the deep web. On the other hand, multifield web forms were found in a considerable number of web pages that are candidates for containing government-related data, and it is highly possible that computational efforts could be applied to these forms to reveal underlying dynamically generated data. In sum, after analyzing candidate web pages and identifying those with multifield web forms, our method was proven effective, as 1.5% (or 7,900) of web pages that were candidates to contain government data were found to be high-potential sources of dynamically generated data.

[9] http://commoncrawl.org

The number of such web pages and their characteristics provide evidence of the government-related data found in the deep web, which serves to confirm our stated hypothesis regarding the frequent misimplementation of open data portals that are developed with a classic website model in mind. These "open" data portals provide access to data exclusively through human interaction while forgoing API interfaces and natively-exposed metadata.

Although the conclusions of this work are supported by a considerable data sample (including roughly 5.6 million government web pages), the underlying method is subject to improvements based on our observations in the experiment. In particular, the crawling process of step 1 could be skipped and web-crawled data imported from existing projects that employ greater computational resources than are available to most academic researchers. In addition, the search for keywords indicating data portals that is part of step 2 could be facilitated by using a list of common roots and internationally-used words that can be applied to local contexts, as local governments are the major targets of the assessment process.

Acknowledgments

The authors would like to acknowledge Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Grant Project 402214/2017-0), Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – IFSP and University of Sao Paulo – USP for all the support provided.

References

Bergman, M. K. (2001). White Paper: The Deep Web: Surfacing Hidden Value. *The Journal of Electronic Publishing*, *7*(1). https://doi.org/10.3998/3336451.0007.104

Bilsel, R. U., Büyüközkan, G., & Ruan, D. (2006). A fuzzy preference-ranking model for a quality evaluation of hospital web sites. *International Journal of Intelligent Systems*, *21*(11), 1181–1197. https://doi.org/10.1002/int.20177

Bradbury, D. (2014). Unveiling the dark web. *Network Security*, *2014*(4), 14–17. https://doi.org/10.1016/S1353-4858(14)70042-X

Brandusescu, A., Iglesias, C., & Robinson, K. (2016). *Open Data Barometer. Global Report. Fourth Edition*. The World Wide Web Foundation. Retrieved from http://opendatabarometer.org/doc/4thEdition/ODB-4thEdition-GlobalReport.pdf

Bunyakiati, P., & Voravittayathorn, P. (2012). Dissemination formats and major statistic data sets of the AEC countries: A survey. In *Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in* (pp. 313–316).

Büyüközkan, G., Ruan, D., & Feyzioğlu, O. (2007). Evaluating e-learning web site quality in a fuzzy environment. *International Journal of Intelligent Systems*, *22*(5), 567–586. https://doi.org/10.1002/int.20214

Caplan, R., Davies, T., Wadud, A., Verhulst, S., Alonso, J., & Farhan, H. (2014). *Towards common methods for assessing open data: workshop report & draft framework*. New York, USA: The World Wide Web Foundation. Retrieved from http://opendataresearch.org/sites/default/files/posts/Common%20Assessment%20Workshop%20Report.pdf

Chou, W.-C., & Cheng, Y.-P. (2012). A hybrid fuzzy MCDM approach for evaluating website quality of professional accounting firms. *Expert Systems with Applications*, *39*(3), 2783–2793. https://doi.org/10.1016/j.eswa.2011.08.138

Correa, A. S., Zander, P.-O., & da Silva, F. S. C. (2018). Investigating Open Data Portals Automatically: A Methodology and Some Illustrations. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age* (pp. 82:1–82:10). New York, NY, USA: ACM. https://doi.org/10.1145/3209281.3209292

Corrêa, Andreiwid Sh., Paula, E. C. de, Corrêa, P. L. P., & Silva, F. S. C. da. (2017). Transparency and open government data: a wide national assessment of data openness in Brazilian local governments. *Transforming Government: People, Process and Policy*, *11*(1). https://doi.org/10.1108/TG-12-2015-0052

Corrêa, Andreiwid Sheffer, Corrêa, P. L. P., & Silva, F. S. C. da. (2014). Transparency Portals Versus Open Government Data: An Assessment of Openness in Brazilian Municipalities. In *Proceedings of the 15th Annual International Conference on Digital Government Research* (pp. 178–185). New York, NY, USA: ACM. https://doi.org/10.1145/2612733.2612760

Crescenzi, V., Mecca, G., & Merialdo, P. (2001). Roadrunner: Towards automatic data extraction from large web sites. In *VLDB* (Vol. 1, pp. 109–118).

Davies, T., & Frank, M. (2013). "There's No Such Thing As Raw Data": Exploring the Socio-technical Life of a Government Dataset. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 75–78). New York, NY, USA: ACM. https://doi.org/10.1145/2464464.2464472

Della Penna, G., Magazzeni, D., & Orefice, S. (2010). Visual extraction of information from web pages. *Journal of Visual Languages & Computing*, *21*(1), 23–32. https://doi.org/10.1016/j.jvlc.2009.06.001

European Union. (2017). *Recomendations for Open Data Portals: from setup to sustainability* (p. 76). Retrieved from https://www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, *70*, 301–323. https://doi.org/10.1016/j.knosys.2014.07.007

Helbig, N., Cresswell, A. M., Burke, G. B., & Luna-Reyes, L. (2012). The dynamics of opening government data. *Center for Technology in Government.[Online]. Available: Http://Www. Ctg. Albany. Edu/Publications/Reports/Opendata*.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268. https://doi.org/10.1080/10580530.2012.716740

Khurana, K., & Chandak, M. B. (2016). Survey of Techniques for Deep Web Source Selection and Surfacing the Hidden Web Content. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *7*(5). https://doi.org/10.14569/IJACSA.2016.070555

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly*, *35*(1), 13–29. https://doi.org/10.1016/j.giq.2017.11.003

Lourenço, R. P., Sá, P. M. e, Jorge, S., & Pattaro, A. F. (2013). Online Transparency for Accountability: One Assessing Model and two Applications. *Electronic Journal of E-Government (EJEG)*, *11*(2), pp279-291.

Maali, F., & Erickson, J. (2014). Data Catalog Vocabulary (DCAT). Retrieved October 15, 2017, from https://www.w3.org/TR/vocab-dcat/

Machado, A. L., & Oliveira, J. M. P. de. (2011). DIGO: An Open Data Architecture for e-Government. In *2011 IEEE 15th International Enterprise Distributed Object Computing Conference Workshops* (pp. 448–456). https://doi.org/10.1109/EDOCW.2011.34

Mardani, A., Zavadskas, E. K., Govindan, K., Amat Senin, A., & Jusoh, A. (2016). VIKOR Technique: A Systematic Review of the State of the Art Literature on Methodologies and Applications. *Sustainability*, *8*(1), 37. https://doi.org/10.3390/su8010037

Matheus, R., Janssen, M., & Maheshwari, D. (2018). Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly*. https://doi.org/10.1016/j.giq.2018.01.006

Mitchell, R. (2015). *Web Scraping with Python. Collecting Data from the Modern Web*. O'Reilly. Retrieved from http://shop.oreilly.com/product/0636920034391.do

Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems*. Harlow, England; New York: Addison-Wesley.

Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated Quality Assessment of Metadata Across Open Data Portals. *J. Data and Information Quality*, *8*(1), 2:1–2:29. https://doi.org/10.1145/2964909

OECD. (2017). *Government at a Glance 2017*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/gov_glance-2017-en

Open Data Watch. (2017a). *The Open Data Inventory 2017 Annual Report: A Progress Report on Open Data*. Open Data Watch. Retrieved from http://odin.opendatawatch.com/Downloads/otherFiles/ODIN-2017-Annual-Report.pdf

Open Data Watch. (2017b). *The Open Data Inventory 2017 Methodology Report*. Open Data Watch. Retrieved from http://odin.opendatawatch.com/Downloads/otherFiles/ODIN-2017-Methodology.pdf

Open Government Data. (2017). Retrieved October 13, 2017, from http://www.oecd.org/gov/digital-government/open-government-data.htm

Open Knowledge Foundation. (2012, November 14). Open Data Handbook Documentation. Retrieved from http://opendatahandbook.org/

Opricovic, S., & Tzeng, G.-H. (2003). Defuzzification within a multicriteria decision model. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *11*(05), 635–652. https://doi.org/10.1142/S0218488503002387

Osagie, E., Mohammad, W., Stasiewicz, A., Hassan, I. A., Porwol, L., & Ojo, A. (2015). *State-of-the-art Report and Evaluation of Existing Open Data Platforms* (No. 645860 H2020-INSO-2014). Retrieved from http://routetopa.eu/deliverable-2-1-state-of-the-art-report-and-evaluation-of-existing-open-data-platforms-now-available/

Ribeiro, M. M., Matheus, R., & Vaz, J. C. (2011). New perspectives for electronic government: the adoption of open government data in brazil. In *Anais do 8º CONTECSI*. Brasil. Retrieved from http://www.contecsi.fea.usp.br/envio/index.php/contecsi/8contecsi/paper/view/2875/1636

Sayogo, D. S., Pardo, T. A., & Cook, M. (2014). A Framework for Benchmarking Open Government Data Efforts. In *2014 47th Hawaii International Conference on System Sciences* (pp. 1896–1905). https://doi.org/10.1109/HICSS.2014.240

Tauberer, J. (2014). Open Government Data: The Book - Second Edition. Retrieved from https://opengovdata.io/

Ubaldi, B. (2013). *Open Government Data: towards empirical analysis of open Government Data Initiatives* (OECD Working Papers on Public Governance). Paris: Organisation for Economic Co-operation and Development. Retrieved from http://dx.doi.org/10.1787/5k46bj4f03s7-en

Ulrich, A., Tom, H., & Jamie, F. (2015). *Benchmarking open data automatically* (No. ADI-TR-2015-000). Open Data Institute. Retrieved from https://theodi.org/guides/benchmarking-data-automatically

Umbrich, J., Neumaier, S., & Polleres, A. (2015). Quality Assessment and Evolution of Open Data Portals. In *2015 3rd International Conference on Future Internet of Things and Cloud* (pp. 404–411). https://doi.org/10.1109/FiCloud.2015.82

United Nations Publications. (2016). *United Nations E-Government Survey 2016: E-Government in Support of Sustainable Development*. New York: United Nations. Retrieved from http://workspace.unpan.org/sites/Internet/Documents/UNPAN97453.pdf

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly*, *31*(2), 278–290. https://doi.org/10.1016/j.giq.2013.10.011

Wang, Y., Zuo, W., Peng, T., & He, F. (2008). Domain-Specific Deep Web Sources Discovery. In *2008 Fourth International Conference on Natural Computation* (Vol. 5, pp. 202–206). https://doi.org/10.1109/ICNC.2008.350

Wget. (2017). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wget&oldid=799950956

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. http://dx.doi.org/10.1016/S0019-9958(65)90241-X

Zheng, Q., Wu, Z., Cheng, X., Jiang, L., & Liu, J. (2013). Learning to crawl deep web. *Information Systems*, *38*(6), 801–819. https://doi.org/10.1016/j.is.2013.02.001