Chapter 3
How Open Are Public Government Data? An Assessment of Seven Open Data Portals

Sanja Bogdanović-Dinić, Nataša Veljković, and Leonid Stoimenov

Abstract Growing open data initiatives are offering different solutions for opening governmental data to the public. Open data platform solutions provide simple tools for enriching governmental portals with a data dimension. The new data-oriented shape of government inevitably imposes the need for the evaluation of government efficiency in light of open data. Regardless of the numerous initiatives, there is still no globally accepted open government evaluation framework. The purpose of the research presented in this chapter is to present and apply a model for assessing data openness, which relies on eight open data principles established by the Open Government Working Group. The model represents a new approach to the evaluation of open data with real-world application capabilities and is fully described throughout the chapter. As a confirmation of this model's capabilities, we illustrate the results of its application on seven data portals along with analyses, comparisons, and conclusions regarding the results.

3.1 Introduction

The concept of open government has been covered extensively in the academic literature over the past several years (Bertot et al. 2010; Veljković et al. 2012; Di Maio 2010; Gustetic 2010). Open data, transparency, participation, and collaboration are enumerated as the main attributes behind the concept of open government. The initiatives for introducing open government bring revolutionary changes into the traditional e-Government model, forcing a transition from service-oriented to data-oriented government. However, open government does not neglect e-services,

S. Bogdanović-Dinić (✉) • N. Veljković • L. Stoimenov
Faculty of Electronic Engineering, University of Niš, A. Medvedeva 14, 18000 Niš, Serbia
e-mail: sanja.bogdanovic.dinic@elfak.ni.ac.rs; natasa.veljkovic@elfak.ni.ac.rs; leonid.stoimenov@elfak.ni.ac.rs

M.P. Rodríguez-Bolívar (ed.), *Measuring E-government Efficiency*, Public Administration and Information Technology 5, DOI 10.1007/978-1-4614-9982-4_3,

which are still an important part of a successful online government, but rather extends the traditional approach with open data. Governments around the globe have recognized the advantages of opening internal data and information flows to the public and started to embrace the initiative by introducing strategies for the successful implementation of openness.

The benefits of making data public and freely available are numerous and are reflected in many different areas of application. Charlotte Alldritt, a public policy and transparency specialist and the current advisor in the UK Deputy Prime Minister's Office, noted the significance of openly available governmental data in delivering *innovative products*, *services and networks created by the public*, *private and civil society sectors* (Alldritt 2012). The 2009 Digital Britain Report described data as *new currency for the digital world* and *the lifeblood of the knowledge economy*, referring to the great potential of open data applications in business and the economy (Carter 2009). Open data offer new job openings; enormous information growth is followed by equal market growth and thus an increased need for specialists. Environmental challenges can be managed easily with the help of data. Public environmental data could be essential in making predictions, tracking behaviors, and making inferences based on recognized event patterns. That approach is the foundation for producing new knowledge and vital conclusions. Publicly available data are crucial for scientific activity that relies heavily on global collaboration based on large data collections. As the 2010 report "riding the wave" emphasizes, open scientific data have enormous potential to change the nature of scientific processes (High level expert group on scientific data 2010).

Many initiatives around the world have focused on defining open data catalogues and open data portals (data.gov.uk, digitaliser.dk, data.gouv.fr, etc.) (European Commission 2011). Some of them are already providing significant results and creating a path for others to follow. A direct consequence of the increased number of data portals has been the generation of large data piles, but the question is whether all published data are open data. What make data open? What are the features that separate open data from online data? What are the rules that need to be followed to distinguish open data? These are only some of the questions that call for an evaluation framework that can determine the extent of openness of published data. There are already several evaluation proposals and initiatives for assessing open data, which focus on data usage and some openness features (Osimo 2008; Berners-Lee 2010; Lee and Kwak 2011). They certainly represent a noteworthy source of experience and ideas and a strong foundation for further research in this area. We will review these initiatives throughout the chapter and emphasize their advantages and disadvantages. We will also debate the reasons behind our choice to develop a new data openness (DO) evaluation model rather than adopt the existing ones.

The focus of this chapter is evaluating data openness in the context of open government. The second section will introduce the reader with definitions of open data, as observed by the open government, and will provide an overview of the existing initiatives regarding the evaluation of data openness. The third section acts as an introduction to the model for evaluating the openness feature against eight openness characteristics defined by the Open Government Working Group (OGWG).

The data openness (DO) model is part of a larger e-Government Openness Index framework (eGovOI) intended to evaluate the level of a government's openness in the context of open data, transparency, participation, and collaboration features (Veljković et al. 2011a). The framework will only briefly be mentioned to provide the reader with the context of the openness component. The application potentiality of the proposed model is demonstrated on selected open data portals, and the results are presented. The assessment is performed automatically via Web tool, which is an implementation of the proposed model, but the tool itself will not be thoroughly explained because this task exceeds the scope of the presented research.

3.2 Open Data: Definition and Evaluation

The definition of open government data has been the subject of many academic and public debates. A precise definition of open government data is needed because it will ensure interoperability between different piles of government data (Gottschalk 2009) and enable their evaluation. To understand the *why* and *how* of open government data, we first need to discuss *what* open data mean.

Open data are data that are available for anyone to use and reuse without any restrictions and at no cost. As stated by Costa et al. (2012), the underlying rationale of open data is that *promoting unconstrained access to raw information enables its reuse and knowledge creation*. The Open Knowledge Foundation (OKF), as the world's best known promoter of open knowledge, has issued an *open definition* (Open Knowledge 2013):

> Open data is data that can be freely used, reused and redistributed by anyone—subject only, at most, to the requirement to attribute and share alike.

The OKF defines open data in general, but can we apply the "open" part to government-held data and thus define *open government data*? The British Government has done so. As stated in the British Government's Open Data Whitepaper (HM Government 2012), open government data are public sector information that is available as open data, which further implies that open government data meet the following three conditions: (1) accessible via the Internet at no more than the cost of reproduction and without limitations based on user identity or intent, (2) published in a digital, machine-readable format, and (3) free of restriction on use or redistribution in its licensing conditions. By contrast, Tauberer (2012) considers OKF's open data definition too weak for defining open government data because it allows the government to require attribution for data reuse. He looks at open government data as raw material that can be transformed and shaped into something different and more powerful.

Considering government-held data from the aspect of open government, we can talk about data relevant to government transparency, innovation, participation, and collaboration. In this regard, there is often confusion between open government data and data transparency. Open government data are related to government transparency,

but the transparency of government data should not be considered as an openness feature. Data can be open but not transparent. The aim of transparency for government data is enabling access to government-held data in a uniform way, making sure that data are well known, comprehensible, easily accessible, and open to all (Jaeger and Bertot 2010). We strongly distinguish these two features of data, which is why we have defined distinct indicators for their evaluation in the eGovOI framework.

If we agree to apply the open definition for government-held data, calling them open government data, the next step of our evaluation of data openness requires determining mandatory open data characteristics. Open data evaluation approaches and the selected open data characteristics found in the literature are given in Table 3.1.

Some of the given approaches include accessing a set of chosen open data characteristics to determine aspects of data quality or transparency (Ren and Glissmann 2012; European Commission 2011), whereas others are more oriented towards the evaluation of specific open data aspects, mostly data availability (Osimo 2008; Berners-Lee 2010; Socrata 2011). For example, Ren and Glissmann (2012) propose a five-phase process for identifying information assets as open data: (1) define business goals and develop business architecture, (2) identify stakeholders and prioritize information needs, (3) identify potential information assets for open data, (4) assess the quality of information assets, and (5) select information assets for open data initiatives. By going through these phases, stakeholders should be one step closer to the identification of open data. In the fourth phase of this approach, the authors apply quality assessment on open data. They have developed a questionnaire for the evaluation of data quality based on the fulfillment of six open data features, as shown in Table 3.1. Lee and Kwak (2011) propose a framework for open government maturity assessment. Within the framework, they evaluate the transparency of open data through assuring data quality in terms of accuracy, consistency, and timeliness.

Tauberer (2012) notes *defining* open government data qualities, namely, being open (accessible) and large (analyzable), and *desired* ones, being open, accurate, and authentic. Defining qualities can be observed as the minimal set of features that open government data must satisfy, whereas the desired ones represent optional features that, if implemented, make data even more open. In accordance with such an approach, Tauberer creates 17 openness principles and classifies them into five distinguished categories: the basic principles, data format, universality of use, data publishing, and the openness process (2012). The basic principles acknowledge the availability, primary, timeliness, and accessibility features of published data. Data format is concerned with the need for providing data in a machine-readable format. Universality of use assembles requirements related to license-free, nonproprietary, and nondiscriminatory data usage. The data publishing category focuses on features such as data permanency, promoting analysis, safe file formats, and provenance and trust, whereas the openness process category gives general recommendations on how to decide what to open using public input, public review, interagency coordination, endorsements of technology, and prioritization as guidelines.

David Osimo (2008) proposes a five stage model for measuring the availability feature of open data. If no data are available, the availability is considered *stage 0*.

Table 3.1 Overview of the different approaches of open data evaluation

Open data evaluation	Which aspects of open data to measure?
Quality aspect of open data (Ren and Glissmann 2012)	• Accessibility and availability • Understandability • Completeness • Timeliness • Error-free • Security
Open government maturity model (Lee and Kwak 2011)	• Accuracy • Consistency • Timeliness
Open government data principles (Open Government Working Group 2007)	• Complete • Primary • Timely • Accessible • Machine processable • Nondiscriminatory • Nonproprietary • License-free
Open government data principles (Tauberer 2012)	• Free access to data • Primary • Timely • Accessible • Machine processable • Nondiscriminatory • Nonproprietary • License-free • Permanent • Promote analysis • Safe file formats • Provenance and trust • Public input • Public review • Interagency coordination • Prioritization
Four-stage model of open data availability (Osimo 2008)	• Availability
Five-star model of open data availability (Berners-Lee 2010)	• Availability
Open data impact (European Commission 2011)	• Number of open datasets available • Timeliness • Data format • Reuse conditions • Pricing • Institutional positioning of the portal governing body • Accessibility • Take-up by citizens • Take-up by app developers • Number of applications developed on open data
Open data benchmark (Socrata 2011)	• Accessibility • Availability

If data are obtainable, availability reaches *stage 1*. When data are available in a nonreusable and non-machine-readable format, the availability is a *stage 2*. If data are in reusable and machine-readable formats, the availability reaches *stage 3*. Finally, if the stage 3 conditions are fulfilled and data are visualizable, the availability is stage *4*. Although this model lacks in the assessment of data quality and does not consider linked data, it still represents a solid foundation for future initiatives and a strong starting point for the development of open data evaluation.

Sir Berners-Lee (2010) proposes a star rating system for assessing the extent of public data availability. This model focuses on linked open data and is intended for wide application. According to the rating system, the data receive one star if they are available on the Web with an open license. If data are published as machine-readable-structured data, they receive two stars. Three stars are appointed to data published in nonproprietary formats. If data comply with all of the above rules and additionally use Semantic Web standards to identify things, they receive four stars. If all of the above rules are met and links to other people's data exist to provide context, the data receive five stars. As Berners-Lee noted, to apply this model to government data, a new requirement should be added: published metadata about the datasets.

The first three levels of the five-star model match stages 1–3 from Osimo's model, whereas the latter two focus on the linked features of data. A higher value is given to data that can be easily reused and whose context is well described through linked information, thereby promoting the need for efforts towards data structuring and formatting rather than simply publishing PDF files. Both approaches, the five-star and four-stage model, focus on only one open data feature: data availability. Although it is one of the key features that defines open data, it is not the only one; therefore, neither of the mentioned evaluation models could be used alone to measure the level of openness of public data.

The European Commission (2011) has also showed an interest in assessing open data and performed a study on open data portals' impact through a Web survey of selected portals in Europe and elsewhere in the world and in-depth interviews with government representatives (European Commission 2011). During the analysis of the gathered results, they applied the Berners-Lee five-star model to measure the level of data availability and defined more detailed sub-indicators for clearly expressing each result. However, the study did not go any further than listing the obtained results. They did not define any calculation to classify analyzed portals on a scale of openness or the impact of open data. Therefore, this study is an excellent resource regarding benchmark methodology, but it lacks in processing methods, which are essential for assessing, categorizing, and comparing different open data initiatives.

Socrata Company (2011) took a different approach and performed a study on open government data through three independent surveys of government, citizens, and developers. The surveys were conducted in the form of questionnaires with the goal of broadly assessing open data not only from the perspective of the government but also from the perspectives of its data consumers and contributors. The results were organized into five categories: attitudes and motivation, current state of open data initiatives and programs, current state of data availability and accessibility, high value data, and engagement and participation. Although this extensive study is of

significant importance because it directly reflects opinions, attitudes, and motivations behind three major stakeholder groups, it is solely based on interviews and pure trust in respondents' answers and does not reflect the state of data analyzed from different points of view, using different techniques. Therefore, this study cannot be observed as sufficient, but it does impose some significant aspects that should not be neglected in the process of open data evaluation.

The Open Government Working Group (2007) has defined a set of eight principles of open government data. These are primary, complete, timely, accessible, machine processable, nonproprietary, nondiscriminatory, and license-free and are now globally accepted as guidance for opening governmental data.

Open Data Principles Adopted for DO Evaluation: Based on the analysis of open data requirements and evaluation initiatives, which is thoroughly presented above, and after carefully reviewing the cited sources, we embrace the OGWG's eight open government data principles as a foundation for our DO evaluation model. We found the OGWG's definition to most clearly reflect government requirements for open data. Other initiatives focus on open data in general (Ren and Glissmann 2012; Lee and Kwak 2011), not specifically government-held data. The recognized indicators in other analyzed benchmarks can be mapped onto these eight characteristics. For example, the *quality aspect of open data* addresses error-free and security features. The error-free feature can be observed as part of the *primary* data feature because original data are expected to be accurate. By contrast, security relates to accessing data. Because openness implies free access to anyone, security can be analyzed as part of data accessibility. Accuracy and data consistency, which are emphasized as indicators in the *open government maturity model*, can be observed as part of the primary feature for the same reasons as the previously explained error-free feature. The European *open data impact* framework provides many indicators that are similar to the OGWG's eight characteristics (timeliness, data format as machine processable, pricing and reuse conditions as license-free, accessibility), but they also go beyond the scope of the OGWG and define additional indicators, such as the number of datasets, institutional positioning, the number of applications, and take-up by citizens and developers. As will be seen later in the chapter, some of these indicators can be observed as possible extensions of our DO model. Tauberer gives a very detailed analysis of the principles of open government data, but he seems to overlap the principles of open data (the first eight principles) with the principles of open government (the last eight principles). If we exclude the principles of open government from Tauberer's proposal, the result would match the OGWG's definition.

3.3 Benchmark Model for Evaluating Data Openness

We are addressing data openness in terms of eight openness characteristics established by the Open Government Working Group (2007). We present an evaluation method that is based solely on information made available via governments' data portals. The method is implemented as a Web-based assessment tool.

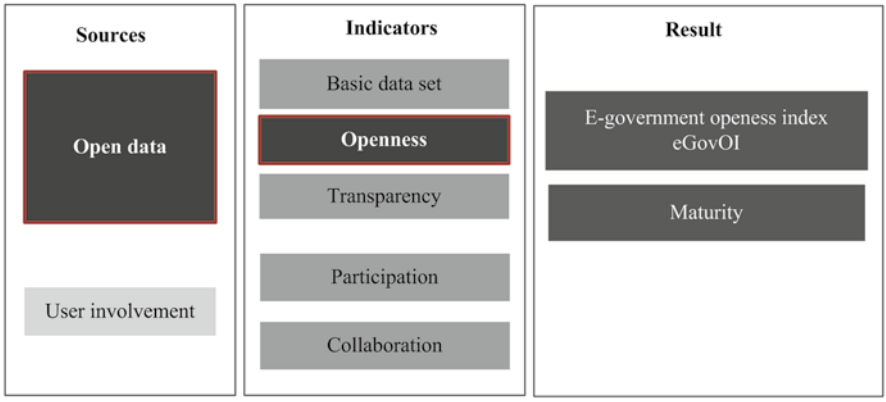

Fig. 3.1 Data openness evaluation: components and position in e-Government Openness Index framework

The general idea behind this approach is to enable openness assessment at any time and, more importantly, to automate the entire process by exploiting available open data portals' APIs. Keeping in mind the increased popularity of the Comprehensive Knowledge Archive Network (CKAN) open data platform among governments and, consequently, its increased utilization, we have developed a Web tool that utilizes CKAN API and enables data openness index calculation relaying solely on API calls and data published on the portal. We will further present our data openness model, starting with the more general scope of the e-Government Openness Index Benchmark and later discussing the aspects of assessment and explaining how these aspects have been evaluated based on open data meta-descriptions. At the end of this section, we will provide the results from the application of model to seven open data portals.

3.3.1 Open Government Benchmark: Data Openness Indicator

Our research regarding the evaluation of open government has resulted in a benchmark model for assessing the extent of governments' openness in accordance with well-defined and globally embraced openness principles. The benchmark is fully described in Veljković et al. (2011a) and is intended for the exploration of government openness boundaries and determination of the extent of fulfillment of open government's main goals.

Figure 3.1 depicts the main benchmark's building blocks as well as the relationships between them. There are five indicators that reflect the main open government concept features: a basic dataset, openness, transparency, participation, and collaboration. These measures are calculated based on available sources and used for computing the final benchmark results: the e-Government Openness Index and maturity.

Table 3.2 Data openness levels

DO value (%)	DO level	Description
0–5	*0—cradle*	The government has only started to publish data on a data portal. The majority of mandatory data categories are still empty, and data are not entirely described
6–35	*1—basic openness*	Data are published under open licenses, and meta-descriptions are made available; however, not all required descriptions are present. The majority of data are published in DOC, XLS, or other non-processable and/or proprietary formats
36–75	*2—average openness*	Data are published in original form and are regularly updated. They are mostly published in TXT, PDF, CSV, and other processable and nonproprietary formats. However, there is no semantics attached to the data descriptions, and the majority of data are not linked to other data
76–90	*3—openness*	The majority of data are published in RDF, XML, and other semantic formats, available to anyone and linked to other data
>90	*4—high openness*	Data are complete and in accordance with all 8 data openness principles

The focus of this chapter is on open data assessment; therefore, we will further address only the parts of the benchmark that are concerned with this aspect, *open data* sources and *openness* indicators, which are indicated in orange in Fig. 3.1. We will reference this approach as the data openness (DO) model. The aim of the DO model is to evaluate the degree of openness of government data made publicly available on an open data portal. The model relies solely on data features' descriptions collected directly from the open data portal. We aim to develop a tool that can estimate governments' data openness online, without human intervention. To perform this deep openness estimation, the model would have to be expanded with other assessment tools, such as questionnaires, which would provide a wider range of information and address different stakeholders.

3.3.2 Data Openness Model

After examining the existing data evaluation models, we have developed a DO evaluation model that perceives the openness of government data through the following indicators: *complete*, *primary*, *timely*, *accessible*, *machine processable*, *nondiscriminatory*, *nonproprietary*, and *license-free*. These indicators match the OGWG's eight open data characteristics (Open Government Working Group 2007). The calculation of DO is performed by grading each indicator with a maximal score of 1 and finding the average value of all indicators. We named the final DO value the DO Index (DOI), which has a range of (0, 1).

Table 3.2 presents data openness levels based on the achieved overall score, expressed as percentages. We define five openness levels: *cradle*, *basic openness*, *average openness*, *openness*, and *high openness*. Cradle openness is intended to

simply acknowledge the existence of an openness initiative and recognize the efforts towards embracing openness principles. As the government progresses in the development of open data, it will advance through the defined levels. The category *high openness* is the most sophisticated level of open data, to which every government strives.

To establish a standard evaluation measurement model, it was necessary to establish some standard set of mandatory categories that each portal should implement, considering the fact that online data are organized into various data categories or tags. We have performed an analysis on the available open data portals around the world from the aspect of the supported data categories. Based on the gathered results, we have defined a basic dataset comprised of nine data categories: Finance and Economy, Environment, Health, Energy, Education, Transportation, Infrastructure, Employment and Population (Veljković et al. 2011b). Each category is comprised of datasets with sizes ranging from one to hundreds or even thousands of pieces of data, depending on the publishing sources. Therefore, sample sizes for datasets in the categories were needed, imposing as additional aspect of the DO model.

The aspects of the DO model can be divided into two categories: measurement indicators and the data categories' sample size. Measurement indicators directly reflect the level of satisfaction about openness principles for a data portal, whereas choosing a proper sample size is vital for obtaining the most accurate results. Considering their importance for the implementation of the DO model, these two categories of aspects are further analyzed separately.

3.3.2.1 DO Measurement Indicators

The indicators of the DO model are presented in Table 3.3 along with a brief insight into their structure and grading.

The CKAN platform enables entering and providing meta-descriptions of datasets in the form of structured documents, where each feature is described via a pair [tag: value]. *Tag* is a feature name (notes, relationships, url, etc.), and *value* is usually a textual value for a feature. Our model relies on these descriptions during the evaluation process. A detailed description of tags is not presented here because it exceeds the scope of this research. However, we will provide short explanations for each used tag to better explain the indicators' contexts (Fig. 3.2).

Open data are *complete* if the following conditions are satisfied: they are published with available meta-description, in a machine-readable format, linked to other data and directly downloadable. Assessment of these conditions is performed based on meta-tags describing each particular dataset: the description is available if a tag [notes] contains some text and if, for each available resource, tag [description] contains some text. However, we cannot evaluate whether the contained text makes sense or whether it is actually related to a particular dataset. Data can be downloaded if, for each available resource, there is a [url] tag containing a download link. A dataset is machine readable if its resources are published in formats that allow computer processing. A dataset is linked to other datasets if there are listed links in the [relationships] tag.

Table 3.3 DO model indicators

Indicator	What to measure?	How to measure?	Score
Complete	1. Description is available—0.25	1. [notes] + resource => [description]	(0, 1)
	2. Can be downloaded—0.25	2. Resource => [url]	
	3. Machine readable—0.25	3. 0.25*MachineProcessable	
	4. Linked—0.25	4. [relationships]	
Primary	Are data provided raw, in original form?	[format] ∈ {CSV, TXT, XML, RDF} => score 1	(0, 1)
		[format] ∈ {XLS, DOC} => score 0.5	
Timely	1. Time period—0.3	1. extras => [temporal_coverage_from] and [temporal_coverage_to]	(0, 1)
	2. Update frequency—0.4	2. extras => [update_frequency] or [frequency_of_update]	
	3. Last update—0.3	3. resource => [last_modified]	
Accessible	Are data accessible to anyone for any purposes?	License_free + Resource => [url]	(0, 1)
Machine processable	PDF/XLS—0.2	[format] ∈ {PDF, XLS} => score 0.2	(0, 1)
	CSV/HTML/TXT—0.5	[format] ∈ {CSV,HTML,TXT} => score 0.5	
	XML/RDF—1	[format] ∈ {XML,RDF} => score 1.0	
Nondiscriminatory	Are data available to anyone?	Accessible + MachineProcessable	(0, 1)
Nonproprietary	Are data available in nonproprietary formats (not DOC/XLS/CDR/PSD)?	[format] ∈ {XLS, DOC, CDR, PSD, NULL} => score 0	(0, 1)
License-free	Are data published under open license?	[is_open]	(0, 1)
DO			(0, 8/8)

TAG name	Content Description	Usage example
[notes]	Textual description	`notes: "The database contains satellite proprietary format on various types of ..."`
[description]	Textual description	`description: "2008 Report",`
[url]	Download link	`url: "http://www.ic.nhs.uk/statistics-and-data-collections",`
[relationships]	Links to other datasets	`relationships: [],`
[format]	Data format	`format: "CSV", format: "XLS",`
[temporal_coverage_from]	Temporal reference	`temporal_coverage-from:"["1972"]",`
[temporal_coverage_to]	Temporal reference	`temporal_coverage_to:"2008",`
[update_frequency] [frequency_of_update]	Textual description	`update_frequency:"Weekly"`
[last_modified]	Data and time	`last_modified:"2013-04-13T03:09:22.098385",`
[is_open]	Boolean true/false	`isopen: true,`

Fig. 3.2 Description of datasets' tags used in the evaluation process

Data are *primary* if they are published raw and in their original format directly from firsthand experience. If they are published in any pre-analyzed format, they are not considered primary. From the aspect of evaluation, the primary indicator is assessed based on the [format] tag for each available resource in dataset. If a resource is published in CSV, TXT, XML, or RDF format, it is most likely original because these formats allow the representation of structured data (the results of data collection processes, sensor readings, etc.). If a resource is published in XLS or DOC format, there is a significant possibility that the data were already processed and published in the form of a chart or graph. However, there is also the possibility that it is in its original format, which is why we have chosen to grade it with 0.5. In any other case, we consider data not primary and grade them as 0.

Data are *timely* if they contain information describing their timeliness (i.e., what period is covered by the data held in a dataset, how often the data are updated, and when the last update was). We give the highest sub-value to the update frequency feature because we consider it the most important in terms of keeping data as accurate as possible. The update frequency receives the highest score if the period covered by the dataset contains the present date and if the time interval that has passed since last update is smaller than the indicated update frequency value. We evaluate this feature by checking whether there is available information contained in the [update_frequency] or [frequency_of_update] tags for each resource. Time period is defined with the [temporal_coverage_from] and [temporal_coverage_to] tags, whereas the last update can be read from the [last_modified] tag.

Data *accessibility* imposes the rule that data should be accessible to everyone equally, regardless of the purpose. The data accessibility indicator has a maximum score if there is no policy regarding data usage. We evaluate this indicator through the *license-free* indicator and the downloadable feature of the *complete* indicator. If data are published under an open license, then they are accessible to everyone equally. If data are downloadable without additional conditions, they are also equally accessible to anyone.

The *nondiscriminatory* indicator reflects freely available data. We acknowledge that a dataset is nondiscriminatory based on its accessibility and machine processability, which means that a dataset is license-free, downloadable, available in machine-processable formats, and, consequently, ready for free usage among users. The nondiscriminatory indicator receives a maximum value of 1 if data are provided under the same conditions to each user. If, for example, user registration is required to download data, the indicated is scored as 0 and considered discriminatory.

Machine processable means that data are provided in a structured format that can be processed by a computer. The calculation recognizes three evaluation levels, which are actually adapted from the *5-star open-linked data model*: level 1, formats that are not machine processable (e.g., PDF, XLS): level 2, structured formats that can be automatically processed but do not contain any semantics (e.g., CSV, TXT, HTML); and level 3, structured formats that include meta-descriptions and semantics (e.g., XML, RDF). Level 1 receives the lowest score, 0.2, which simply gives credit for publishing data, even though they cannot be utilized for any type of processing. Level 2 receives a score of 0.5, considering that data in CSV format are of

a predictable structure and can very easily be further processed. Level 3 receives the highest score because in addition to properly structured data, additional information is provided that could enable highly sophisticated data processing. The calculation is performed by examining the [format] tag for each available resource within a dataset and finding the average score.

The *nonproprietary* feature relates to the previous one by considering data formats from the aspect of the supported processing programs; in that manner, for datasets available in a format that requires commercial Microsoft Excel or Microsoft Word programs for access, such as XLS or DOC, this feature is given a value of 0. For formats that do not require any specific, commercial program, such as CSV, XML, and RDF, this feature is given a value of 1. The estimation is, as for the machine-processable indicator is concerned, performed by examining the [format] tag for each available resource and calculating the average score.

Finally, the *license-free* feature relates to free access to data. It is scored 1 if data are published under an open license, which is found by examining whether the [is_open] tag for a dataset is set on true or false. If it is true, then the dataset is published under an open license. If it is false, then the dataset is not open.

3.3.2.2 Choosing a Relevant Data Subset

Keeping in mind that this issue is a statistical challenge, we have chosen a statistical approach to obtain a reliable method for determining the sample size with given restrictions such as the confidence level and the margin of error. Equations 3.1 and 3.2 represent the chosen formulas (NIST/SEMATECH 2012):

$$ss = \frac{Z^2 * p * (1-p)}{c^2} \tag{3.1}$$

$$ss = \frac{ss}{1 + \frac{ss-1}{pop}} \tag{3.2}$$

Equation 3.1 explains the process of calculating the sample size (ss) based on the confidence level (Z), margin of error (c), and expected accuracy (p). The margin of error indicates the precision of the chosen sample and the allowed deviation of the expected results. In our calculations, we used a 10 % value for the margin of error, which means that if 45 % of datasets in a chosen sample have demonstrated a specific feature, we can be "sure" that that feature has been demonstrated by the entire relevant datasets between $(45-10)$ and $(45+10)$% of the sample size. The confidence level tells how "sure" we can be (i.e., how often the true percentage of the sampled data satisfying the required condition lie within the confidence interval). Usually, Z is chosen to be 90 or 95 %. We have chosen a 95 % confidence, for which Z takes a value of 1.65 in the calculation according to the table of standard normal curve area values. This means that we can be 95 % "sure" that datasets from a chosen sample that satisfy the chosen condition are in the defined confidence interval,

which is between (45 – 10) and (45 + 10)%. Accuracy denotes the percentage of the sampled data that truly satisfy the required features. Because there is no trustworthy way to reliably predict such a percentage, we have used a value of 50 %.

Equation 3.1 calculates the sample size for a very large population because the population size usually has no influence in statistics-related issues. However, when a population is finite, small, or relevant for the problem, it is important to obtain a size for the sample that is sufficient for the analysis. In the case of open data evaluation, we found that the size of data categories (which represent our population) is important for the final results; therefore, we have introduced Eq. 3.2, which performs corrections of the calculated sample size according to the true size of the data category, denoted as pop.

3.3.3 A Use Case Study: DO Tool in Action

For the purposes of testing the DO model's capabilities, we have performed an analysis of data openness for the following open data portals: the USA, the UK, European Union, Germany, Ottawa Canada, Austria, and Queensland data portals. We have chosen these portals because they all run on the CKAN open data platform, and our tool currently supports only CKAN[1]. The USA and the UK represent the oldest portals, launched in May and September 2009, respectively, and are the first initiators of the "open data portal" idea. By contrast, the European Union, Germany, and Ottawa Canada portals are the youngest, officially published in February (EU and Germany) and June (Ottawa) 2013. Austria and Queensland are in the middle, having been published in April and December 2012, respectively. It was interested to see how these portals compared to each other and whether their "age" and attained maturity had any influence on the final score. Table 3.4 gives an overview of the assessed data portals, along with information about the sizes of the nine mandatory categories for each portal. This information is important for analyzing the DO results from the aspect of the number of published datasets.

The process of calculating the DOI was performed automatically via our Web tool. For each portal, the tool first finds all available tags per category (subcategories) based on the provided keywords that describe those categories. For example, for Finance and Economy, we have provided two keywords: finance and economy in English and German. The second step is calculating a sample size for each tag based on the obtained information on the number of datasets per tag and randomly chosen datasets to form a sample. The third step is calculating the eight indicators for each dataset from a sample according to the rules explained in Table 3.3. The final DOI is calculated as the average of the DO indices for each data category. The data category obtains its DOI as the average of all its tags' DO indices.

[1] In addition to CKAN, other open data platforms are used around world governments, including Socrata (Kenya, State of Washington, City of Chicago, etc.), Junar (City of San Jose, City of Las Vegas, Government of Costa Rica, etc.), and the Open Government Platform (Ghana, Rwanda, India, etc.).

Table 3.4 Overview of analyzed data portals

			Number of tags (subcategories) per category								
Data portal	URL	Year launched	Finance and Economy	Environment	Health	Energy	Education	Transportation	Infrastructure	Employment	Population
USA	data.gov	2009	5	2	2	1	1	1	0	3	1
UK	data.gov.uk	2009	105	99	237	33	117	87	11	86	72
European Union	open-data.europa.eu	2013	5	4	17	15	13	0	0	14	10
Germany	govdata.de	2013	142	22	54	46	0	21	3	0	7
Ottawa, Canada	data.ottawa.ca	2013	1	1	2	0	0	1	0	0	0
Austria	data.gv.at	2012	11	4	1	6	0	2	2	1	5
Queensland	data.qld.gov.au	2012	2	2	2	1	4	2	1	2	1

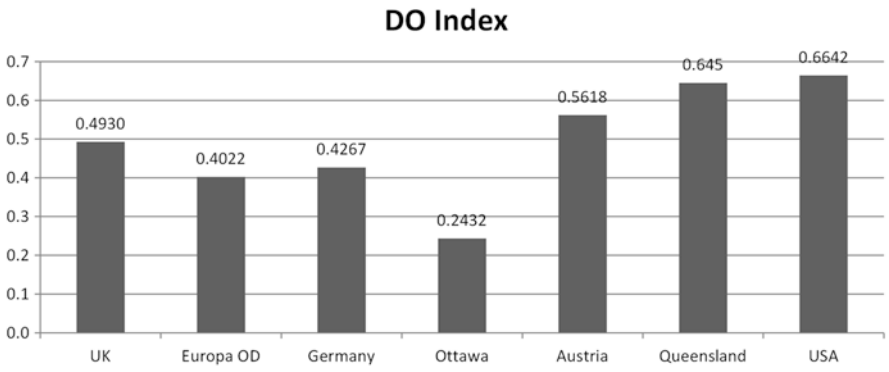

Fig. 3.3 Data openness index for analyzed data portals

Figure 3.3 illustrates the results of the DOI calculation. As can be observed, the highest score was achieved by the US data.gov portal, 0.6642, which indicates 66.42 % openness, belonging to the *average openness* category according to Table 3.2. The lowest score was achieved by Ottawa, Canada, 0.2432, which indicates 24.32 % openness and places Ottawa in the *basic openness* category. All other data portals place in the *average openness* category and achieved similar results. On average, the DOI was approximately 49.08 %. The UK, Austria, Queensland, and the USA scored higher than the calculated average; thus, we can consider them *high-average open*. Europe and Germany were below average, approximately 40 %; therefore, they can be considered *low-average open*. Ottawa is the only portal in the lower category, which points to the necessity for further openness improvements.

A closer look at the results provides information regarding the successful, less successful, and challenging aspects of each analyzed portal. For example, if we look

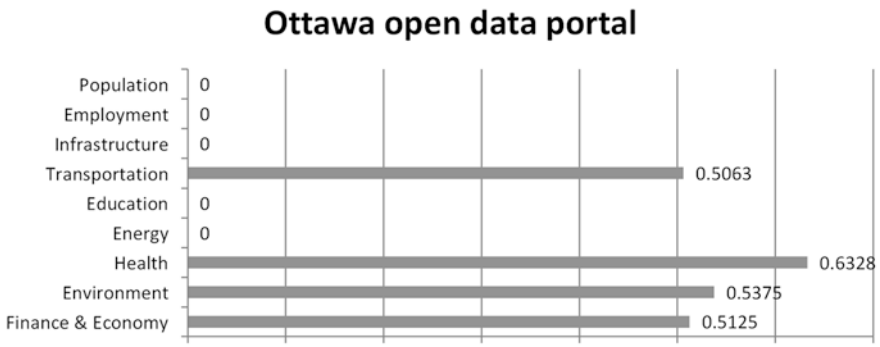

Fig. 3.4 Ottawa's data portal assessment results

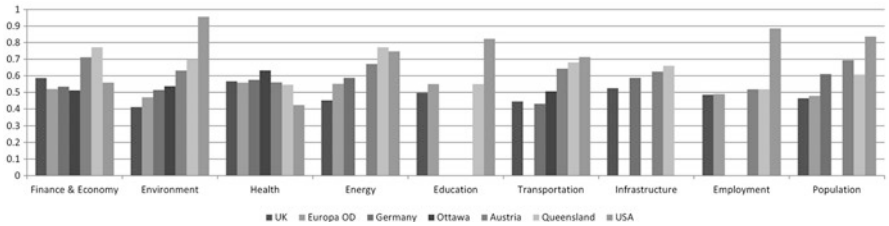

Fig. 3.5 Data.gov detailed categories' data openness assessment

at Ottawa's detailed information per category (Fig. 3.4), we can see that five out of nine mandatory categories are empty. This is certainly a significant cause for the low DOI. All other categories are above 50 %. Therefore, Ottawa should focus on providing data to the missing categories to improve its score. To improve the DO for each category, the detailed results per category should be analyzed from the aspect of the eight openness indicators.

Figure 3.5 provides a comparison of the DO indices among the analyzed data portals, achieved per category. We can see that US data.gov, as the overall highest scoring portal, achieved the best results in five out of nine categories, whereas Ottawa, the overall lowest scored, received the best score for the Health category. The worst graded category, if we exclude those with a score of 0, was the Transportation category on the German data portal, with 43.13 % DOI. The highest score, 95.59 % DOI, was attained by the US data portal for the Environment category. Education and Infrastructure are the least implemented categories, in only four out of the seven data portals, whereas Finance and Economy, Environment, and Health are present in all the analyzed portals. By analyzing in detail the results for the Transportation category of the German data portal, it was concluded that the critical indicators are timely, machine processable, and nonproprietary. The timely indicator achieved a score of 0, indicating a complete lack of timeliness for information for datasets (update frequency, publishing date, temporal coverage, last modified date). Machine processable and nonproprietary achieved scores of 0.24 and

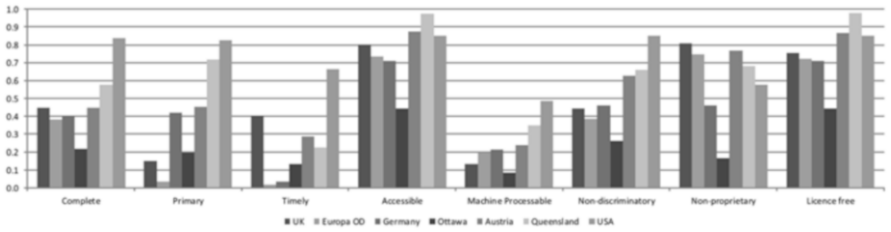

Fig. 3.6 Average scores of indicators

0.39, respectively, due to data being published mostly in PDF, DOC, XLS, and other non-processable and/or proprietary formats. Improvements in the machine-processable indicator, by making data available in processable and semantically enriched formats (XML, RDF, CSV, etc.), would also be reflected in the complete, accessible, and nondiscriminatory indicators because they are directly related to the machine-processable feature in the calculation process. Consequently, such modifications would inevitably lead to an improved overall DO score for the Transportation category of the German data portal.

Figure 3.6 gives an overview of the average scores of indicators per data portal. The best evaluated indicator is *license-free* on the Queensland's data portal, with 97.67 % DO. Close behind is the *accessible* indicator on the same portal, with 97.56 % DO. The US data.gov portal received the best scores for the accessible, nondiscriminatory, and license-free indicators, all approximately 85 % DO, and achieved the lowest score for the machine-processable indicator, with 48.82 % DO. Ottawa's highest scores were for the accessible and license-free indicators, both approximately 44 %, whereas timely and nonproprietary received the worst grades, approximately 13 % and 16 %, respectively.

The overall lowest score was achieved by the European open data portal for the timely (only 1.82 %) and primary (3.64 %) indicators. The detailed score of the European portal shows that both of these features are supported by only three out of the nine data categories and achieved very low scores in those three categories. The primary feature is related to data formats in the calculation process. The results show that these are available in XLS, DOC, or similar, and it cannot be ensured with 100 % confidence that the data are in their original form. A low score for the timely indicator is the direct consequence of a complete lack of timeliness-related information in the dataset's meta-descriptions. To improve low scores, it is necessary to provide detailed meta-descriptions and ensure that the data are published in appropriate formats. These findings draw attention to the necessity for improvements in these indicators for the EU data portal.

Younger data portals generally lack data, resulting in empty data categories and 0 scores for some indicators. This finding is expected but not sustainable. These portals should constantly work to improve their content regarding both the size of the data category and the dataset's meta-descriptions. However, the final results pinpoint a problem common for all portals, regardless of their experience: a low score for the

machine-processable indicator. The highest graded for this indicator is the US data.gov, with a score lower than 0.5, which means that every analyzed portal publishes data mostly in PDF, DOC, XLS, and other non-processable formats. Data publishers should be introduced to the benefits of semantically enriched data and encouraged to provide data in various formats, including XML, RDF, CSV, and similar formats.

3.4 Discussion and Future Work

The concept of open government has spread rapidly across the world's governments. Open data are a leading concept of Open Government. Publicly available governmental data mean more transparency, efficiency, and legitimacy, in addition to helping citizens build long and steady trust in their government. Many open data definitions have been created to establish a set of principles behind the development of open government. Although these definitions seem different, they are actually quite similar and point to some common defining features: completeness, timeliness, accessibility, machine readable, nonproprietary, nondiscriminatory, primary, and license-free. As a result, initiatives have been created measuring the extent of data openness. Although there have been several attempts at building openness assessment models, a standard and globally accepted evaluation approach that would enable estimating and comparing the openness advancements of the world's governments still does not exist.

Joining the openness pioneers, we have developed a model for evaluating the level of openness based on the information on open data available from open data portals. The model was implemented as a Web tool for the automated evaluation of openness and offers assistance in the process of building openness principles. As described throughout the chapter, it relies on eight open data principles and provides information on the level of openness of governments' data. The model was applied to seven selected data portals for to demonstrate its capabilities and possible results. Throughout the estimation process, we verified different types of analyses that could be performed on the resulting data and which presented the different aspects of the generated values.

Defining a new model and developing a tool for the automated assessment of data openness offer a significant advantage for the assessment process itself; now, the process can be performed at any time, without any type of human intervention, quickly and uniformly by following the predefined rules. This advancement is of great significance for governments, which can continually track their portal's performance regarding data openness, and for other stakeholders and policy makers, who can easily obtain information on what needs further improvements and what has achieved notable success. As the main strength of our DO model, we emphasize chosen indicators. By defining rules for their assessment, we have enabled a standardized application of the developed model on different governments' portals with the possibility of their comparison from various data aspects as well as comparative analyses of their current openness.

However, there are still some issues that need to be addressed in the future, which we will shortly summarize in the following paragraphs.

One important issue that has arisen concerns the DO model's scope: *Can data openness be measured only based on the chosen eight openness principles, or should data transparency be considered as well*? We see data transparency as an important aspect for analyzing open data and define it through data authenticity, understandability, and reusability. However, data can be open but not transparent. Although there are similarities between these two data aspects, we see them as separate data features and choose not to mix them. In our model for the evaluation of open government (Veljković et al. 2011a), we define openness and transparency indicators separately, with each dealing with different aspects of open data. The Transparency indicator considers data transparency and addresses related issues.

Throughout the development and application processes, we encountered some challenges that have imposed doubts related to some of the model's core features. The first and most obvious question asked was as follows: *Should the size of data categories have any impact on final DOI*? Indeed, if, for example, we take a look at the detailed results for the UK and Ottawa data portals, we can see that the UK scored 48.18 % DOI for the Health category with 237 tags, whereas Ottawa has scored 48.75 % DOI for the same category, with only two tags. The model places Ottawa higher than the UK because it now neglects the categories' sizes. Related to this issue, another question is logically raised: *Should we consider the portal's experience when calculating DOI*? In this way, we can acknowledge years of efforts and prevent the situation of newly built data portals with only a few tags and datasets receiving better scores than the more experienced portals. The UK's open data portal has more experience than Ottawa's because it has existed longer. If we calculate the UK's experience, would we obtain different DOIs, and would that new DOI better reflect the real state of openness of the data portal? We believe that experience and data categories' sizes should be involved in the calculation process as new indicators or simply as factors that would enhance/amplify portal's DOI. This area is one of our future model improvements, the research for which is already underway.

References

Alldritt C (2012) Open data case studies, data.gov.uk blog post. http://data.gov.uk/blog/open-data-case-studies

Berners-Lee T (2010) Linked data. http://www.w3.org/DesignIssues/LinkedData.html

Bertot JC, Shuler JA, Jaeger P (eds) (2010) Special issue: open/transparent government. Govern Inform Q 27(4):311–446

Carter L (2009) Digital Britain: final report, vol 7650. The Stationery Office, Norwich

Costa S, Beck A, Bevan AH, Ogden J (2012) Defining and advocating open data in archaeology. In: Proceedings of the 40th annual conference of computer applications and quantitative methods in archaeology (CAA), Southampton

Di Maio A (2010) How do open government and government 2.0 relate to each other?, Gartner Blog. http://blogs.gartner.com/andrea_dimaio/2010/09/03/how-do-open-government-and-government-2-0-relate-to-each-other

European Commission (2011) Pricing of public sector information study—open data portals: final report
Gottschalk P (2009) Maturity levels for interoperability in digital government. Govern Inform Q 26:75–81
Gustetic J (2010) E-gov versus open gov: the evolution of e-democracy. http://www.phaseonecg.com/docs/egov-opengov-whitepaper.pdf
High level expert group on scientific data (2010) Riding the wave—how Europe can gain from the rising tide of scientific data. European Union
HM Government (2012) Open data white paper unleashing the potential. Cabinet Office, London
Jaeger PT, Bertot JC (2010) Transparency and technological change: ensuring equal and sustained public access to government information. Govern Inform Q 27(4):371–376
Lee G, Kwak YH (2012) Open government implementation model: a stage model for achieving increased public engagement. In: Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times. ACM, pp 254–261
NIST/SEMATECH (2012) e-Handbook of statistical methods. http://www.itl.nist.gov/div898/handbook
Open Government Working Group (2007) 8 principles of open government data. http://www.opengovdata.org/home/8principles
Open Knowledge Foundation (2013). http://opendefinition.org/#sthash.48WfhjSB.dpuf
Osimo D (2008) Benchmarking e-government in the Web 2.0 era: what to measure and how. Eur J ePract 3(4):33–43, ISSN: 1988-625X
Ren GJ, Glissmann S (2012) Identifying information assets for open data: the role of business architecture and information quality. In: Commerce and enterprise computing (CEC), 2012 IEEE 14th international conference on. IEEE, pp 94–100
Socrata (2011) 2010 open government data benchmark study. http://www.socrata.com/benchmark-study
Tauberer J (2012) Open government data. http://opengovdata.io
Veljković N, Bogdanović-Dinić S, Stoimenov L (2011a). eGovernment openness index. In: Proceedings of ECEG 2011. Ljubljana, Slovenia, 16–17 June 2011, pp 571–577
Veljković N, Bogdanović-Dinić S, Stoimenov L (2011b) Municipal open data catalogues. In: Proceedings of CEDEM 2011, Krems au Donau, Austria, 25–26 May 2011, pp 195–207
Veljković N, Bogdanović-Dinić S, Stoimenov L (2012) Web 2.0 as a technological driver of democratic, transparent, and participatory government. In: Reddick CG, Aikins SK (eds) Web 2.0 technologies and democratic governance. Springer, New York, pp 137–151