Toward a Reference Architecture for User-Oriented Open Government Data Portals

Ahmad Luthfi[1(✉)] and Marijn Janssen[2]

[1] Department of Informatics, Universitas Islam Indonesia, Jalan Kaliurang KM. 14,5 Sleman, Yogyakarta 55584, Indonesia
ahmad.luthfi@uii.ac.id
[2] Delft University of Technology, Delft, The Netherlands
m.f.w.h.a.janssen@tudelft.nl

Abstract. Governments have established Open Government Data Portals (OGDP) to open various types of datasets that can be used to increase transparency, accountability, and innovation. OGDP is becoming a strategic program for citizen engagement and empowering users. Nevertheless, many OGDP architectures focus merely on publishing data and do not support the actual data use. Therefore, this paper aims to develop a reference architecture (RA) that takes a broader set of requirements aimed at enabling the use of open data into account. The RA consists of recommended structures and integrations of the end-to-end user interactions and services. In this research, we use the DKAN open data management platform as the basis to design a full suite of cataloguing and visualising the end-to-end user interactions. Five layers are proposed providing functionalities for using data. Whereas most portals are focused on releasing data, our RA is focused on empowering users by providing functionalities for the use of data.

Keywords: Reference architecture · Open data · Portal · DKAN · End-user

1 Introduction

Governments provide and maintain vast amounts of datasets in the Open Government Data Portal (OGDP). The government has started implementing open data initiatives and also setting up open data portals to make these data available in the open by default and in reusable formats [1, 2]. The OGDP is an online management platform that assists end-to-end users in accessing the categorical datasets provided by the government organisation [3, 4]. These categorical datasets such as transportation, education, funding, geospatial, city taxes, energy, and COVID-19 outbreaks present the data openly and freely to the end-users. The OGDP encloses information of interest to the end-users, such as data scientists, data stewards, researchers and university students, business owners, non-profit organisations, and journalists. Theoretically, the simplest OGDP is a dataset catalogue with instructions for how end-users can access, search, download, and use the data in various file formats, such as CSV, PDF, XML, and JSON [5, 6].

B. Shishkov (Ed.): BMSD 2022, LNBIP 453, pp. 259–267, 2022.
https://doi.org/10.1007/978-3-031-11510-3_17

Yet, the usage of OGDP lags behind as the primary focus on opening data and the user view is given less attention [7]. In order to build successful open data portals, systematic evaluation is necessary to understand them better, assess the types of value they generate and identify what services need to be made to improve them. However, many non-trivial procedures must be considered in the course of execution of the (OGD) movement, including using the OGDP management platform and classifying the end-to-end user's interactions. Several characteristics of good OGDP refers to (1) the use of open standards and access to the dataset without human complicity [8, 9], (2) the straightforward and unsupervised portal to understand for the end-to-end users [10], and (3) data should be presented in clear structures and metadata [4, 10].

Ideally, OGDP should not be a practical or technical burden for end-to-end users. Therefore, the objective of this paper is to develop a RA for OGDP that aims to facilitate the use of OGD. The RA consists of five main layers which focus on different aspects of enabling the use of OGD. The five main layers of the RA developed in this paper include data collection, data lake, data management, data analysis and modelling, and data visualisation. These five major layers are supposed to support and relate to each other to orchestrate the use of OGDP. Furthermore, to ensure the complicity of end-to-end users in the implementation of the OGDP, we included the involvement of several open data stakeholders, such as data scientists, data engineers, researchers, data stewards, business owners, communities, data enthusiasts, and parents. These types of OGDP stakeholders represent the end-user-oriented in collecting, managing, analysing, and using open data.

In general terms, a RA in the field of software architecture refers to a list of functions and several indications and their interactions with each other [11]. Reference architectures for a domain capture the fundamental sub-systems shared by all systems within that domain as well as their interrelationships [12]. A RA is useful for both maintenance and design. It can improve understanding of a system, and it can be used to set up new systems and re-engineer existing systems [13]. In this study, an OGDP reference architecture can be defined as a fundamental architecture that captures relationships between the five main layers of developing OGDP, including the involvement of representative stakeholders.

This research may contribute to government organisations and respected researchers' understanding of the RA for the OGDP development and provide recommendations related to its RA development. The paper is structured as follows: Sect. 1 presents the current issues and problems, Sect. 2 reviews related literature background, Sect. 3 proposes the research approach, Sect. 4 summarises the RA. Finally, Sect. 5 concludes the paper.

2 Theoretical Background

Open data portals are a web-based system that collects data from a variety of sources in various forms and publish data to be used by users interacting with a user interface dashboard [14]. Open data portals must be considered rigorous architectures and infrastructures as interactions between governments and external users. They should be able to allow or disable operations and establish a range of possible data uses, and unlocking the promise of open data requires the capacity to locate relevant data [14, 15]. Therefore,

one endeavour to make these datasets more accessible and easier to reuse is to provide an open data gateway of available datasets [4, 9].

Despite the excitement generated by the availability of an ever-increasing amount of freely available data in the open data portal, several critical problems, such as unstructured metadata and data sources, have emerged to address the emerging issue of low quality in the available data portal, which is a severe disadvantage that might derail the open data portal development [2, 3, 16]. Similarly, there is a wide range of content, functionality, and technical standards among the various data management systems used by open data portals [14, 17]. For that reason, a re-designing and benchmarking framework or available data architecture are required to understand quality issues in open data portals better and investigate the impact of improvement approaches [9, 18].

A reference architecture frequently consists of a list of system functions and some indication of their end-user interfaces and their interactions with each other [13]. Reference architectures explain how to use specific patterns and methods to tackle specific types of challenges [19]. As a result, it can be used to reference the individual designs that businesses or organisations will use to solve their deficiencies [11, 13].

In the domain of OGDP, a RA is an abstract design that provides a frame of open data reference, common vocabulary, reusable structures, and end-user best practices compiled in rigorous layers for designing specific instances. Besides, multiple stakeholders of the OGDP, from data engineers to data analysts, should be able to access and comprehend the defined RA layers. Therefore, a RA should be explicit and technically tangible [13]. At the same time, the task at hand is to design a RA that is both generic and actual while still including specific information [11].

3 Research Approach

This research aims to develop a RA for OGDP. First and foremost, the challenge faced in this study was derived from the previous study literature. Therefore, we use a Systematic Literature Review (SLR), which follows three primary sequential steps: data collection and identification, screening for eligibility, and deductive and inductive coding [20]. The objective of this SLR process is to define some terminologies using three main Boolean keywords operator: “open government data portal” AND “reference architecture”. In this step, the inclusion was limited to English-language scientific journals using Scopus based gateway. Based on the data collection step using the Boolean operator, we acquired (n = 137 articles). Furthermore, we filtered our prior 137 articles by screening abstracts without explicit reference to “end-to-end user” (n = 88 articles). In the final step, we filtered the full-text version by excluding articles without comparing their research objective and empirical research. We finally found (n = 24) eligible papers by performing this PRISMA protocol.

The findings of this literature study were used to develop an OGDP RA. Thereafter, we also conducted content analysis for some references related to Open Data Portal management and platforms, such as Comprehensive Knowledge Archive Network (CKAN), Git Data Publisher, Socrata, OpenDataSoft, and Drupal-based Knowledge Archive Network (DKAN) Platform. This study aims to design interactions and relationships between open data portal communication and end-users. Therefore, we selected and followed the

DKAN Platform because: (1) it is an open data platform that enables data providers like governments to easily share data with the general public with a comprehensive set of cataloging, publishing, and visualization tools [21], and it is suitable to create an independent interconnection between the proposed RA and the roles of the OGDP users [21]. Although DKAN is criticised and has shortcomings, we opted for using DKAN as a starting point as many portals make use of this platform [22]. The RA can then be used to extend the DKAN platform. The DKAN platform that is used as the basis consist of three main elements [21, 23]: (1) DKAN Dataset Content Types, which contains the actual dataset and resource types and fields to develop layer 1 (data collection) and layer 2 (data lake), (2) The DKAN Dataset Representational State Transfer (REST) API and management, which defines the REST interface to integrate and protect data in the ODGP (layer 3), (3) The DKAN Dataset Groups, which combine both dataset analysis and site users including data visualisation (layer 4 and 5).

4 OGDP Reference Architecture

Implementing a RA within an organisation can boost productivity by using a proven solution and providing governance to ensure the consistency and applicability of the technology. Theoretically, a RA aids as an elementary guide to developing Information Technology (IT) systems and infrastructures. Therefore, in this paper, the RA proposed the highest level of abstraction and architectural guidance in developing OGDP rather than describing a system in detail or providing a detailed diagram of the interconnection among the layers and actors involved.

This study proposed the OGDP reference architecture consisting of five main layers that represent the data science life cycle: data collection, data lake, data management, data analysis and model, and data visualisation. In Fig. 1, we show the RA we derived. It consists of five major sub-systems plus their relationships focused on enabling the use of open data. These five layers were represented the DKAN open data management platform. Basically, there will be two primary roles of the representative stakeholders in this RA. First, the internal open data stakeholders such as data engineer, data steward, and data analyst are able to access, modify, and configure the selected datasets. In this part, the internal stakeholders can collect, store, manage, analyse, and visualise the dataset based on the specific role. For example, in the layer 1 (data collection), data engineer can extract, transform, and load the dataset to generate several specific file formats such as PDF, CSV, XLS, XML, and JSON. At the same time, data engineer can also manage the dataset to create a metadata, DKAN architecture, and ensure the security of the data. Second, the external open data stakeholders such as researchers, business owners, journalists, communities, and other potential public actors' domain, they all can access, download, and reuse the provided datasets in the OGDP. For example, in the layer 4 (data visualisation), journalists can find and view the selected categorical dataset related to the weather forecast to report the updated news. In this layer, OGDP provides a quantitative-based visualisation through the charts, graphs, and maps. At the same time, the OGDP also serves the dataset into a qualitative-based visualisation by figuring it via infographics, word clouds, and sentiment mapping.

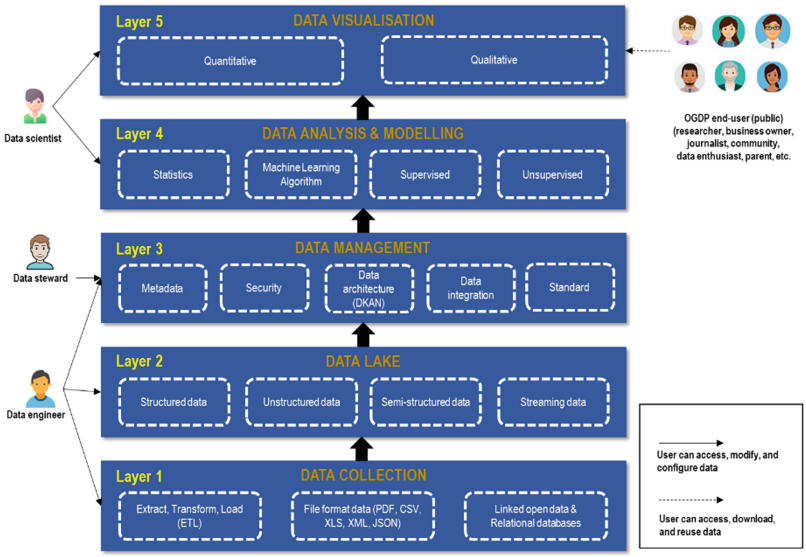

Fig. 1. Reference architecture of the OGDP

- **Layer 1. Data Collection**
  The primary layer of the OGDP reference architecture is data collection. In this layer, data providers collect, monitor, and analyse accurate data from various sources. As part of data collection, data providers have to specify what types of data will be collected, where they will obtain data, and how they will collect it. We defined three sub-systems in this first level: (1) extract, transform, load (ETL). The ETL process is an integration process that compiles data from multiple sources into a single, consistent data store that can then be uploaded to a data warehouse. Thus, in this layer, the objective of the ETL process is to enhance data quality by performing data cleansing before loading the data into the data lake; (2) file format data that refers to a standard of file extension to arrange logically within a file. File formats also describe the structure and type of data that can be stored in a file. There may be a header, metadata, and saved content in a typical file structure. In the case of OGDP, there have been several file formats structured by the data providers, such as Comma-separated Values (CSV), eXtensible Markup Language (XML), Portable Document Format (PDF), Microsoft Excel Format (XLS), Keyhole Markup Language (KML), JavaScript Object Notation (JSON), GeoJSON; (3) linked open data and relational database. In this sub-system, the OGDP should be able to integrate the data with other sources (other OGDPs) to provide specific context. The benefit derives from the open linked data that end-users like researchers and journalists can discover more expected and related data while accessing the OGDP.
- **Layer 2. Data Lake**
  In the second layer, data lake is used to collect and integrate data from the first layer. Essentially, a data lake is a central repository containing many raw data in its original format. In general, data lakes differ from traditional data warehouses in that

they use flat architectures and object storage to store data [24]. The most important aspect of data lakes is open format, avoiding the lock-in that comes with proprietary systems like data warehouses. We identified four sub-systems in this data lake layer: structured, unstructured, semi-structured, and streaming data. Structured data refers to data that can be analysed effectively because its elements can be addressed easily. For instance, a structured table with rows and columns can store all the information that can be found in database SQL. Therefore, fields within the table can be mapped easily and have relational keys. Furthermore, unlike structured data, unstructured data lacks a predefined structure. Consequently, it's unsuitable for mainstream relational databases. However, since many data providers have unstructured data formats, such as PDF, DOCX, and Text, in this layer, we also considered keeping these types of data in the data lake. As such, unstructured data can be stored and managed on alternative platforms. It is more prevalent in IT systems, and businesses use it in various business intelligence and analytics application platforms. At this level, we also studied the semi-structured data. Semi-structured data has some organisation properties that make it easier to analyse, but it does not reside in a relational database. Some processes can be stored in relational databases (in XML format, for example), even if semi-structured data is challenging to store. Besides the three important sub-systems (structured, unstructured, and semi-structured data), we also provide the diverse data sources and format known as streaming data. Nowadays, the Internet of Things (IoT) connects devices in various situations, including cars, factories, homes, retail establishments, and wearables. These all things collected by sensors in IoT produce many types of streaming data, such as videos, images, geospatial, and e-commerce data. Therefore, we provide the streaming data sub-system to represent the need to store the stream processing the real-time data.

- **Layer 3. Data Management**
  The data management layer represents the techniques for organising, structuring, securing, integrating, and keeping an organisation's data. As government organisations generate and consume data at unprecedented rates, data management solutions are becoming increasingly important for making sense of massive amounts of data. After discussing the prior two layers (data collection and data lake), we now provided five main sub-systems in the data management layer: metadata, security, data architecture, data integration, and standard. In implementing the OGDP, technically, we entail establishing policies and procedures to ensure that information can be effectively integrated, accessed, shared, linked, analysed, and maintained throughout the government institutions. To do so, metadata management can help data engineers to provide the basic knowledge in identifying and classifying the collected data. At the same time, the security aspect is also an important issue in the OGDP. The OGDP is an open-access platform that anyone can access freely without restriction. Opening more data to the public domain can reap many advantages. However, protecting the data in the OGDP is not trivial. Therefore, we ensure that the security elements should be included in this data management layer. We need to configure and protect several high-level services for external access over the Internet at the hardware level, such as security firewalls, API services, and data connectors. In the meantime, we are also required to protect the data by performing pseudonymous algorithms on user access levels and proposing a regular assessment of the OGDP system. Furthermore, the data

management layer also deals with the data architecture for developing and organising the OGDP. In this sub-system, we used DKAN Open Data Platform as an open-source management platform with a comprehensive set of cataloguing, publishing, and visualisation tools, and therefore, government organisations can simply share data with the general public. DKAN is a Drupal-based open data portal built on Comprehensive Knowledge Archive Network (CKAN), the first extensively utilised open data portal with a mature and rigorous upgrade system. Therefore, DKAN can support the open data providers in developing an OGDP to organise and customise the collected data.

- **Layer 4. Data Analysis and Modelling**
  In the fourth layer, we designed an essential step in developing the OGDP, namely data analysis and modelling. Data analysis is analysing, cleansing, manipulating and modelling data to identify usable information, inform conclusions, and assist decision-making [14]. This layer proposed four main sub-systems: statistics, Machine Learning (ML) algorithms, and supervised and unsupervised approaches. These data analysis methods aim to interpret and understand the results of such techniques and methods for data collection to make data analysis straightforward, more precise, or accurate, and all the equipment and outcomes. In the case of OGDP, statistical analysis can be used to generate predictive analysis using the collected data by performing text analytics, for instance. In the supervised machine learning, several algorithms can be used such as linear regression, classification, Naïve Bayesian Model, Random Forest Model, and Neural Networks. While in the unsupervised machine learning, the algorithms include exclusive, overlapping, hierarchical, and probabilistic clustering. These are all methods that can support the data analysis to become more accurate at predicting outcomes of the collected data. At the same time, this proposed RA can support semi-supervised learning and reinforcement learning. Therefore, this layer can promote a small amount of labelled dataset, and perform actions from trial-and-error of data analysis.
- **Layer 5. Data Visualisation**
  Data visualisation is the last layer of this study's proposed ODGPS's reference architecture. Data visualisation is the presentation layer for the external OGDP end-user, such as researchers, business owners, journalists, communities, and data enthusiasts. In this layer, the end-user can freely access, download, and reuse the provided data in the OGDP platform. Therefore, this layer is crucial and should be defined and designed comprehensively. The way OGDP developers approach data visualisation can greatly impact how thought-provoking and far-reaching the end-user's conclusions are. Data visualisation done correctly allows others to understand insights faster, simpler, and deeper, resulting in increased knowledge retention and a higher possibility of action being performed as a result. In this layer, we divided the sub-systems of the data visualisation into two main parts, namely quantitative and qualitative visualisation. Quantitative refers to numerical-based visualisation displayed through graphs, charts, tables, and maps. Meanwhile, qualitative data visualisation indicates the textual in building connections and landing context. Some example methods that can be selected to generate qualitative data visualisation are world clouds, sentiment mapping, infographics, and timelines graphics. Therefore, by using quantitative and qualitative data visualisation, the OGDP end-users are able to examine the data to obtain additional insights regarding the information or messages within OGDP.

Moreover, regarding the end-to-end user design that we approached in the prior introduction part of this paper, the involvement of the internal and external users can be seen in Fig. 1. Several studies about developing a RA in information systems and software domains did not explicitly include the end-to-end user as the primary component. The end-to-end user in the OGDP reference architecture presents the roles and responsibilities of each user, including their privilege in using the OGDP. This paper identified several actors like data engineers, data stewards, and data scientists (Fig. 1). For example, data engineers can access, modify, and configure the data in layer 1 (data collection), layer 2 (data lake), and layer 3 (data management). In contrast, data scientists are responsible for managing layer 4 (data analysis and modelling) and layer 5 (data visualisation). In addition, external end-users, such as researchers, business owners, journalists, community, data enthusiasts, and parents, can all access, download, and reuse the categorical dataset in the OGDP supported by the data visualisation layer.

5 Conclusion

We started our study from the issues of the quality and non-trivial procedures of the open data portals in general and their existing system architecture to maintain categorical datasets. OGDP should not be a practical or technical burden for end-to-end users. Therefore, the objective of this paper is to develop a RA for OGDP. The RA was based on DKAN as most of the existing portals are DKAN-based. This facilitates the extensions of existing OGDP to make them more user-oriented. Five main layers of the RA were formed in this paper, including data collection, data lake, data management, data analysis and modelling, and data visualisation. These major layers are expected to support and relate to each other to enable and orchestrate the use of open data. In addition, to make certain the involvement of end-to-end users in the implementation of the OGDP, we also involved a number of open data stakeholders, such as data scientists, data engineers, researchers, data stewards, business owners, communities, data enthusiasts, and parents.

Our next step will be to put the RA into practice. The proposed reference architecture of the OGDP in this study should be generalised with care, not only using a single DKAN open data platform. DKAN has shortcomings, and its metadata model is relatively simple. For further research, we recommend using and combining other open data platforms, such as Socrata, Git Data Publisher, and OpenDataSoft, to comprehensively capture the data management layers and their interaction among the stakeholders. Furthermore, we recommend developing more comprehensive meta-data models to advance the easy discovery and use of the OGDP.

References

1. Ubaldi, B.: Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. OECD Working Papers on Public Governance, vol. 22, p. 60 (2013)
2. Lourenço, R.P.: Open Government Portals Assessment: A Transparency for Accountability Perspective. In: EGOV 2013, Koblenz, Germany (2013)
3. Ivanov, M., Varga, M., Bach, M.P.: Government open data portal: how government strategies should be more open. In: 7th International Conference: An Enterprise Odyssey: Leadership, Innovation and Development for Responsible Economy, Zadar, Croatia (2014)

4. Lněnička, M., Máchová, R.: Open (big) data and the importance of data catalogs and portals for the public sector. In: The 3rd International Global Virtual Conference (GV-CONF 2015) (2015)
5. Luthfi, A., Janssen, M.: A conceptual model of decision-making support for opening data. In: 7th International Conference, E-Democracy 2017, pp. 95–105. Springer CCIS 792, Athens, Greece (2017)
6. Luthfi, A., et al.: Bayesian-belief networks for supporting decision-making of the opening data by the customs. In: EGOV-CeDEM-ePart 2020. Linköping University, Sweden (2020)
7. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Inf. Syst. Manag. **29**(4), 258–268 (2012)
8. Lourenço, R.P.: An analysis of open government portals: a perspective of transparency for accountability. Gov. Inf. Q. **32**(3), 323–332 (2015)
9. Máchová, R., Lněnička, M.: Evaluating the quality of open data portals on the national level. J. Theor. Appl. Elect. Comm. Res. **12**(1), 21–44 (2016)
10. Alhawawsha, M., Panchenko, T.: Open Data Platform Architecture and Its Advantages for an Open E-Government, in Advances in Computer Science for Engineering and Education III (2021)
11. Cloutier, R., et al.: The concept of reference architectures. Syst. Eng. **13**(1), 14–27 (2010)
12. Grosskurth, A., Godfreyz, M.W.: A reference architecture for web browsers. J. Softw. Maint. Evol. Res. Pract. **1**(1), 1–7 (2006)
13. Martínez-Fernández, S., et al.: Aggregating empirical evidence about the benefitsand drawbacks of software reference architectures. In: ACM IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (2015)
14. Lněnička, M.: An in-depth analysis of open data portals as an emerging public e-service. Int. J. Soc. Behav. Educ. Econ. Bus. Ind. Eng. **9**(2), 589–599 (2015)
15. Kostovski, M., Jovanovik, M., Trajanovic, D.: Open data portal based on semantic web technologies. In: 7th South East European Doctoral Student Conference (2012)
16. Kučera, J., Chlapek, D., Nečaský, M.: Open government data catalogs: current approaches and quality perspective. In: International Conference on Electronic Government and the Information Systems Perspective, Prague, Czech Republic (2013)
17. Juana-Espinosa, S.: Open government data portals in the European Union: a dataset from 2015 to 2017. Data Brief **29** (2020)
18. Umbrich, J., Neumaier, S., Polleres, A.: Quality assessment and evolution of open data portals. In: 3rd International Conference on Future Internet of Things and Cloud (2015)
19. Wahyudi, A., Matheus, R., Janssen, M.: Benefits and challenges of a reference architecture for processing statistical data. In: 16th Conference on e-Business, e-Services and e-Society (I3E). Springer, Delhi, India (2017)
20. Biesbroek, R., et al.: Data, concepts and methods for large-n comparative climate change adaptation policy research: a systematic literature review. WIREs Clim. Change **9**, 1–15 (2018)
21. Aky¨urek, H., et al.: Maturity and usability of open data in North Rhine-Westphalia. In: International Conference on Digital Government Researc, Delft, the Netherlands (2018)
22. Zuiderwijk, A., Jeffery, K., Janssen, M.: The potential of metadata for linked open data and its value for users and publishers. eJ. eDemocracy Open Gov. (JeDEM) **4**(2), 222–244 (2012)
23. Seto, T., Sekimoto, Y.: The construction of open data portal using DKAN for integrate to multiple Japanese local government open data. In: Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings (2016)
24. Giebler, C., et al.: Leveraging the Data Lake: current state and challenges. In: International Conference on Big Data Analytics and Knowledge Discovery, Stuttgart, Germany (2019)