International Journal of Artificial Intelegence Research
Vol 6, No 2, December 2022
ISSN: 2579-7298

Open Health Data Development for Machine Learning-based Resource Sharing: Indonesia Case Study

Vitri Tundjungsari

Fakultas Ilmu Komputer, Universitas Esa Unggul, Jakarta, Indonesia
vitri.tundjungsari@esaunggul.ac.id
** corresponding author*

ARTICLE INFO

Article history:
Received 31 Sep 2022
Revised 06 Nov 2022
Accepted 13 Dec 2022

Keywords:
Data sharing,
Health data,
Open data,
Open health data,
Resource sharing

ABSTRACT

Currently, data has a role as a vital requirement in research, especially in the health domain. Unfortunately, Indonesia does not yet have many quality datasets that can be obtained easily. Even though the central and regional governments in the Republic of Indonesia have started implementing Open Data Initiatives, the fact is that the data obtained is still not good enough to be used as datasets. This is where open data is required to enable researchers to perform data analytics and visualization, especially research that uses techniques in machine learning. This research fills the gap to solve the above-mentioned problems by providing resource-sharing datasets. We have built an Open Data Health portal from various trusted data sources so that 200 datasets have been collected. The architecture used in our proposed Open Health Data is a single model. This model is suitable for relatively small datasets and is controlled by a single agency. Our test results show that the portal we build can function properly, but lacks data quality and usability. This requires further work, especially in terms of improving the quality and adequacy of the amount of data to be used as datasets, especially for the purposes of resource sharing and machine learning-based research.

I. Introduction

Open data provides data that can be freely accessed, reused, and distributed by anyone [1]. Research in health sectors requires large datasets to be accounted for properly. However, one of the obstacles in research in the health sector is the difficulty of obtaining related data for reasons of bureaucracy and privacy. As a result, the openness of the government's local health development is not optimal. The data has no level of detail, is not updated, and is not published in open format [2].

Kostkova mentions some benefits of Open Health Data for increasing health qualities for citizens and providing better health policies for a country [3]. The health sector needs data that is accurate and easy to obtain for research and educational purpose. We have investigated some benefits and challenges in developing Open Data related to the health domain, as follows:

a. Nayek in his paper investigates several Open Data initiatives in six countries, i.e., India, USA, UK, Kenya, Australia, and New Zealand [1]. The result of the research shows that many of the metadata are not sufficient for datasets. Other problems are data presented by various government data repositories without standards to achieve interoperability.
b. Culnane et al. in their papers improve the accuracy and confidence of patients' data in Australia by doing patient reidentification. Their research was performed on a 10% sample dataset and results in high confidence that patients in the sample have been properly identified [4].
c. Seastedt et al. mention in their paper that large health datasets through Open Data enable innovation by employing machine learning (ML) to increase understanding of patients' data and their diseases. However, data privacy and sharing should be regulated by determining acceptable risk thresholds while doing data sharing [5].
d. Heijlen and Crompvoets in their paper investigate and map the ecosystem of open health data. The open health data ecosystem contains stakeholders, interests, and information policies by performing data preparation activities. They use socio-technical environment-related data, such

as wealth data managed by governments. Finally, the open health data ecosystem is tested via a case study concerning the Data for Better Health initiative from the government of Belgium [6].

e. Piasecki and Cheah in their research find out in their research that data ownership, whether private or public, has no relevance to data accessibility and availability [7].
f. Hurbean et al. conduct a systematic literature review to investigate open data-based machine learning applications in six different areas of smart cities, i.e., smart governance, smart economy, smart mobility, smart environment, smart people, and smart living [8].
g. Chen et al. mention three categories of open data for smart cities and health-related for open data-based machine learning, i.e., sensor data, image and video data, and text data [9].

In Indonesia, there are several initiatives to develop Open Data by various institutions, public and private. Indonesia as one of the founding countries of the OGP (Open Government Partnership) has directed public policy focus on data disclosure. Still, in its implementation, Indonesia lacks the supply and utilization of open data. This can be seen from Indonesia's low position in several organizations working in the open data sector rankings. The Open Data Institute Open Data Barometer ranked Indonesia at 38 out of 100 [10]. In another ranking issued by Open Data Watch, Indonesia ranked 33 of 187 countries [11]. Indonesia's low ranking is caused by several things: the provision of data that is not yet comprehensive, the lack of data in a machine-readable format, data licenses, the readiness of the government, and the lack of social, economic, and political impact.

Islami in her paper, mentions One Data Indonesia or Satu Data Indonesia (SDI) as an effort to provide credible, accountable, and reliable data that can be used as a reference in any policy-making and its implementation. The initiative is a presidential mandate stipulated in Presidential Regulation No. 39 of 2019. However, in practice, there are still many challenges in data planning, collection, checking, and dissemination. Challenges and problems in implementing SDI are identified to determine the Critical Success Factors (CSFs) for implementing SDI [12].

There are some drawbacks to implementing Open Data in Indonesia, such as differences in platforms and standards for sharing data between electronic systems, both within and between government agencies. It is an important issue in implementing an electronic-based government system [13]. Other problems are related to platform differences, sectoral egos, and data interoperability mechanisms. All those problems require not only policy to support intended activities but also relevant technology [14].

The issue of human resource capability [15], which is a challenge for adopting and implementing open government and open data, is more commonly found in local governments that have lower capabilities and resources in comparison to the Central government. Agencies within the local government environment must be able to share data and information, both technically and organizationally [16].

On the other hand, the Health Law of Republic Indonesia No. 36 of 2009 about the confidentiality of personal data and patient data transmission states that health data can be shared based on public interest or in response to an emergency [17].

This is where this research fills the gap in the demand for open health data resources. This paper aims to design and develop an Open Health Data-based machine learning that can serve the public for resource sharing and research purposes. Hence, it contributes to improving health quality by facilitating the public in providing, obtaining, and reusing health data.

II. Methods

We adopt concept of COMSODE methodologies [18] to develop Open Health Data. There are four main phases of the open data publication process proposed in the COMSODE methodology [18]:

1. (P01) Development of open data publication plan,
2. (P02) Preparation of publication,
3. (P03) Realization of publication,
4. (P04) Archiving.

Figure 1 shows how COMSODE methodologies is used to develop open data publication.

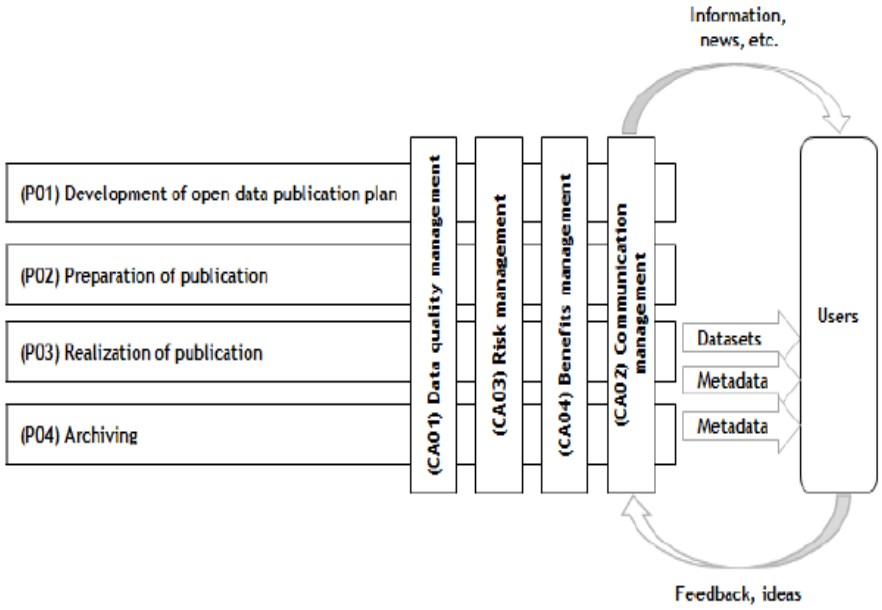

Fig. 1. COMSODE Methodology [18]

Based on the COMSODE methodology above, we define these four main stages to develop our proposed Open Health Data, i.e.:

a. Data Collection and Categorization
b. Data Preparation and Analysis
c. Design and Development of the Portal
d. Testing and Evaluation.

Figure 2 shows the stages used in our research.

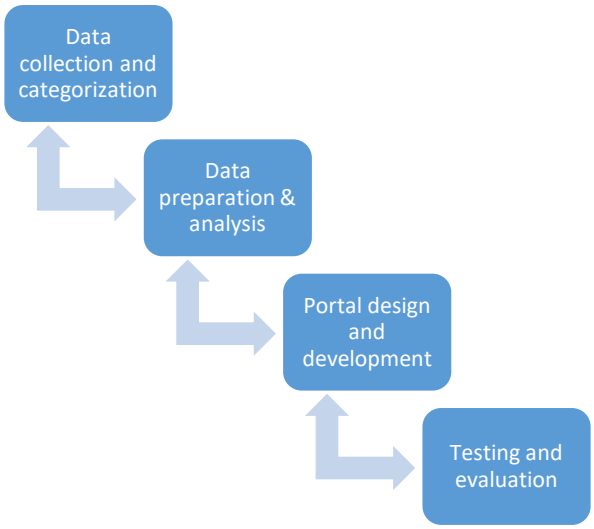

Fig. 2. Stages of proposed Open health data development

A. Data Collection and Categorization

This research was conducted using a mixed-method approach involving an online survey of health workers, interviews, and observations. Participants in this research are the stakeholders from the government, healthcare facilities, and the community involved in developing and using open data applications.

The study used a modified survey based on WHO's interim 2020 report. The survey is written in Indonesian, it included questions on using applications in data management and the impacts of open data on health services. It was conducted online between 10 May and 10 June 2021.

We conducted interviews with 7 participants to explore challenges from stakeholders' perspectives. Participants were chosen from three groups of relevant stakeholders, i.e., data management policymakers; workers either in the Health Office or Communication and Informatics Office in Indonesia; and application developers. In some cases, we interviewed more than one representative from the same institution to understand different perspectives. Observations were recorded via meetings with healthcare workers. The participants involved in this survey are very helpful in finding the source of the datasets. Several datasets from various government web and research organizations are used to develop this Open Data.

B. Data Analysis

This stage aims to identify functional and non-functional requirements which are the main components of this portal. The collection of health data is also taken from the data and information center of the Ministry of Health. This stage aims to change the data format previously only human-readable to become machine-readable. It is intended that the data used as sample data on the portal will be open data where one of the requirements is machine-readable. The data collected will be categorized according to its types.

Figure 3 shows the dimensions of Open Data from various aspects, i.e., strategic, economic, legal, conceptual, and technical [19]. In the strategic aspect, the data should be available; in the economic aspect, the data should be affordable; in the legal aspect, the data should be sharable. The conceptual aspect reflects that the data should be interoperable and primary, while the technical part provides that the data should have high quality, usable, and accessible. In this research, we focus on delivering datasets that qualify from a technical aspect (high quality, usable, and accessible).

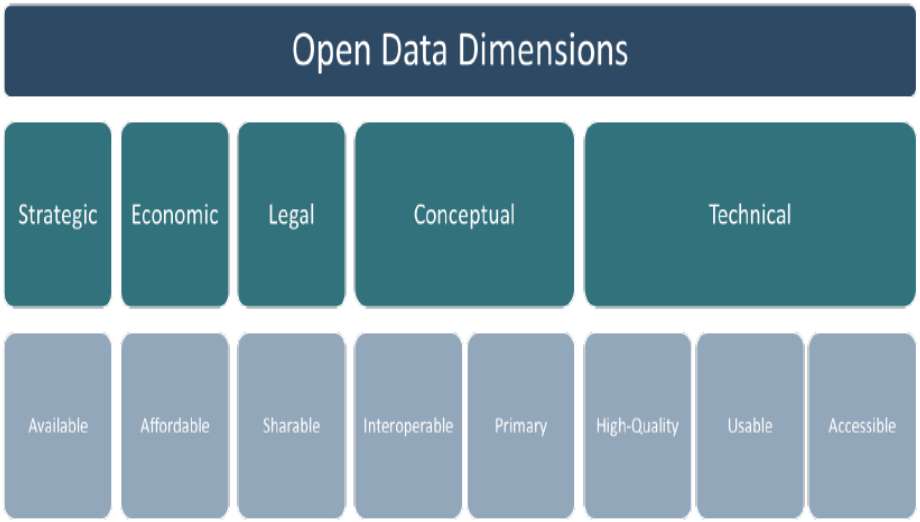

Fig. 3. Important factors determined in Open Data Dimensions [19]

C. Design and development of an Open Health Data Portal

We adopt an Open-data value chain, where data can be freely accessed, obtained, used, provided, modified, and shared by any (verified) organization and individual [20]. Therefore, the Open Health Data should be available under an open license and easily accessible in a machine-readable format. CKAN is used as an open-source data management system to develop the open health data portal.

To ensure data quality from the public sector, we validate the data after it is collected. The data is then aggregated to be analyzed as the next step. Data service and production can be launched after the data is analyzed. These steps are required to ensure that the data available are sufficient as datasets. On the other hand, data from the private sector should be analyzed, aggregated, and validated to be readily available as datasets. Figure 4 shows the Open-data value chain [20].

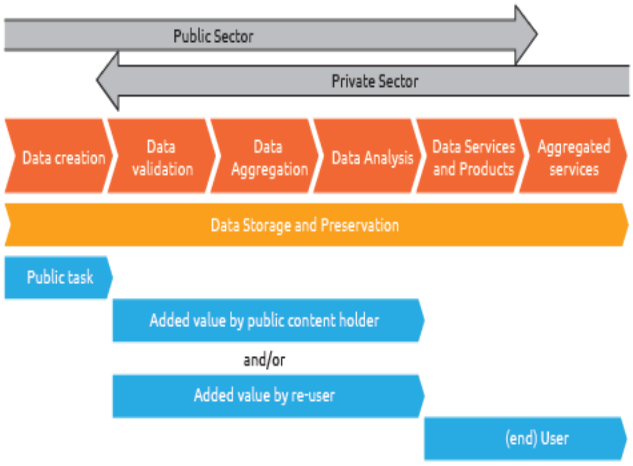

Fig. 4.Data value chain in Open Data [20]

D. Testing and Evaluation

This stage aims to test the portal that has been developed. We conduct requirement-based testing and usability testing using a Likert scale (1 to 5) on 7 participants, as mentioned in sub-chapter 2.1. There are three main components we test related to development goals, readability, and efficiency.

Usability testing result is very important to evaluate the proposed Open Health Data. Some challenges and barriers also can be identified from the evaluation stage.

III. Result and Discussion

A. Data Collection and Categorization

In developing this portal, the authors collect valid health data and used it as sample data. This data was obtained directly and indirectly from various credible and trusted sources, such as the Data and Information Center of the Ministry of Health of the Republic of Indonesia (pusdatin.kemkes.go.id) and the Indonesian Data Portal (data.go.id). The data we choose in TXT, XLS, and CSV format only.

B. Data Analysis

As a result of data collection, we categorize data into:

a. Data sources from the web of the Indonesian government institution, such as https://layanan-pusdatin.kemkes.go.id/, https://satudata.go.id/home
b. Data sources from the hospital, such as https://bdc-imeri-idealab.ui.ac.id/dataset/
c. Data sources from the web of provincial and local government, such as https://opendata.jabarprov.go.id/id, https://data.jakarta.go.id/
d. Data sources from other private organizations, such as https://dataportal.asia/dataset?groups=hlth&vocab_economy_names=Indonesia

As an experiment, this portal consists of two hundred datasets of TXT, CSV, and XLSX formats. The data set has been categorized into health facilities, widespread diseases, nutrition, mental health, chronic disease, and general health. In the data set, there is a menu related to matters where the user is given link to the relevant data set, such as articles, journals, and others.

C. Design and development of an Open Health Data System

In this stage, we design and develop a portal. The system specification is defined as follows:

a. An open health data portal was created to make it easier for the public, researchers, doctors, and health organizations to obtain valid data sets in the health sector.
b. The stakeholders involved when two or more organizations decide to share data. The data user is the person who requires the data for analysis. The data owner is the organization that owns the data. Both the data user and data owner should register with the data provider.
c. There is also a data center to provide data to more than one data user. The data center is also responsible for data access for a group of data users.

d. Valid data sets will be taken from various institutions (public and private), and health organizations, as mentioned in subchapter 3.2.
e. The portal will have data categories based on disease, health facilities, nutrition, etc. In addition, data can be visualized in graphical form to make it easy for users to understand.
f. Multi-layered architecture is used to model the architecture of the system. The User Interface on this portal will be user-friendly and informative by considering UI and UX design. We adopt the architecture of Open health data proposed by The World Bank, where the model is called a single platform [21]. Figure 5 provides the architecture of the Open Health Data used in our research. We choose this model because it has a simple infrastructure. The data catalog are hosted within a single server environment so that the server is easily managed by cloud hosting. The World Bank Open Data mentions that this single model is fit for Open Data with a data catalog consisting of a small number of datasets (less than 200), datasets are small in size (less than 100 Mb), and a single agency is a dominant role for data coordination and infrastructure management.

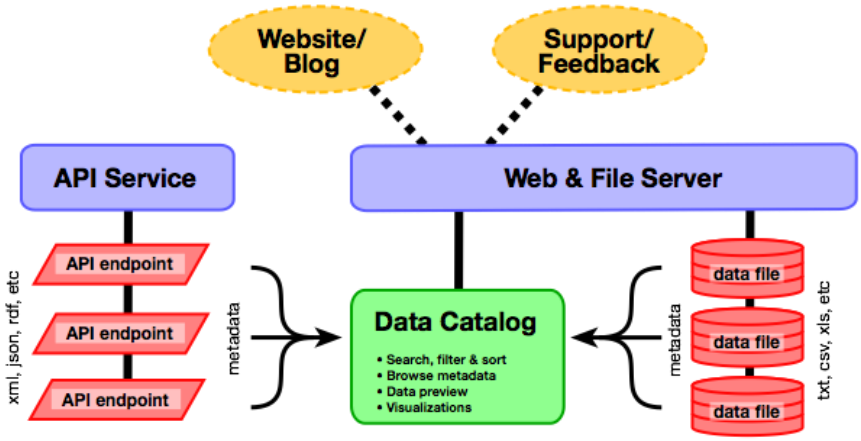

Fig. 5.The architecture of Open Health Data [21]

Our proposed Open Health Data has several main features, i.e.

a. Data Publication. This feature publishes data by portal users. This feature will consider data license, data level, data format, and classification.
b. Data Search. This feature is used to search for the desired data according to the keywords provided by the user.
c. Getting Data. Portal users use this feature to get the desired data.
d. Data Sharing. This feature allows users to share data on the portal with other organizations.
e. Authorization. This feature provides authorization to members in organizational features, i.e., Admin and User (data owner and data user).
    1) Admin: login, add, modify, delete datasets, and manage members in the organization.
    2) User as data owner: register, login, add datasets, download, upload, delete datasets.
    3) User as data user: register, login, download, view datasets.
f. Data privacy and sensitivity. This feature manages the data set, which can be viewed publicly or privately by the user.
g. Data Visualization. This feature is used so that users can view information clearly and efficiently with tables and graphs.

Figure 6 shows the design Interface of the Open Health Data System on the frontpage.

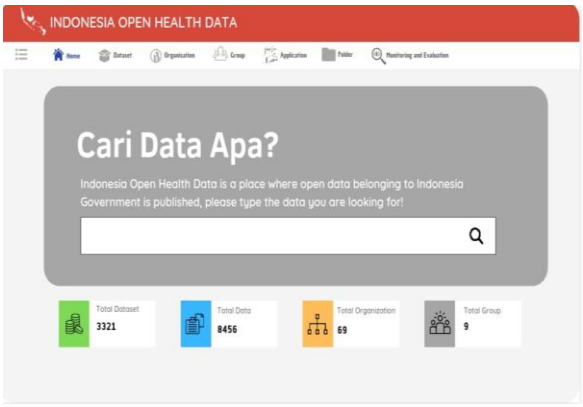

Fig. 6. The interface of Open Health Data

D. Testing and evaluation

For testing, we conduct several requirements-based testing, as follows:

a. The Open Health Data portal should be able to provide the result of all datasets.
b. The Open Health Data portal should be able to visualize the result in tabular and graphical form.

Figure 7 shows the result of all datasets that we have deployed for early testing. Figure 8 shows data visualization in tabular form, while figure 9 shows the data visualization in graphical form.

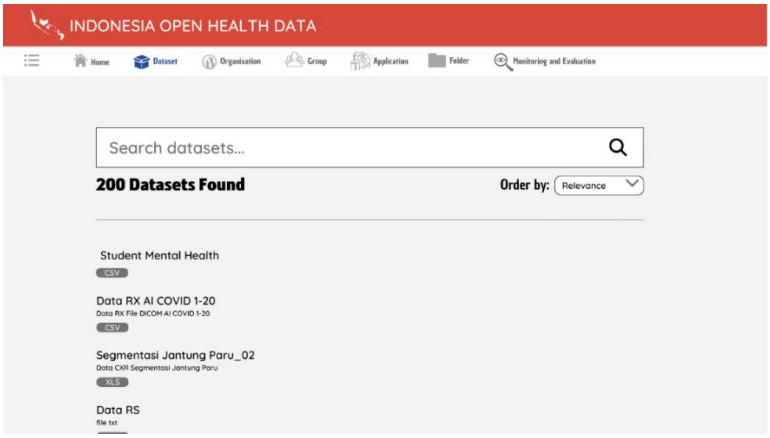

Fig. 7. Searching dataset of Open Health Data

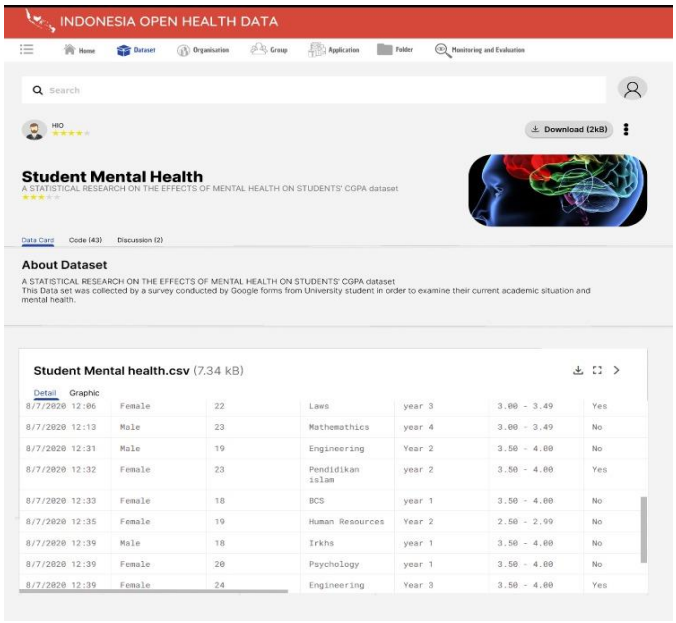

Fig. 8. Data visualization of Open Health Data (1)

 ISSN: 2579-7298

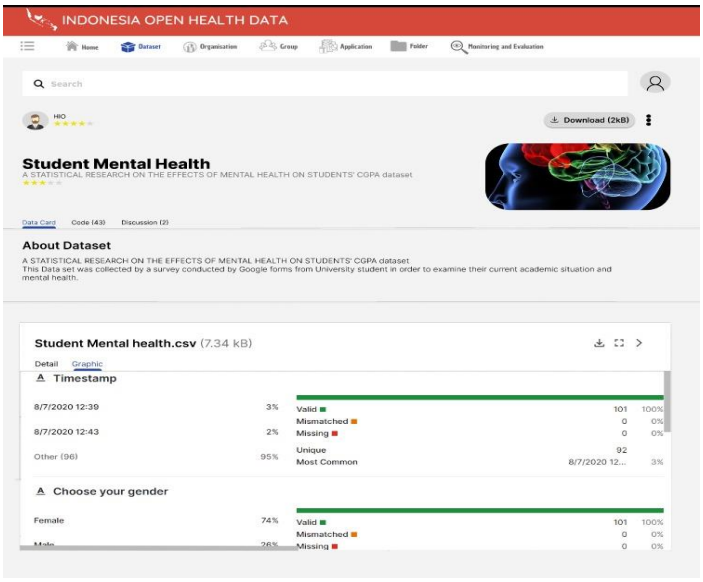

Fig. 9.Data visualization of Open Health Data (2)

In the graphical form, users can see data visualization in graphs or diagrams. The chart types are lines and points, points, bars, and columns. The displayed graph is adjusted to the column parameters contained in the data.

Table 1 presents the usability testing result using a Likert scale 1 to 5, which 1 is strongly disagree and 5 is strongly agree. The main component we used as questions is focused more on the technical aspect of Open data. There are nine questions given to same participants who involved in data collection stage, i.e.: (1) the necessity of the development; (2) the quality of data; (3) the usability of data; (4) the accessibility of data; (5) the readability of data in terms of font and color; (6) the readability in terms of screen composition; (7) the readability of functions button; (8) the navigation; and (9) response time to access the datasets.

Table 1. Usability testing result

Items		Tests	Mean scores
Development goals	1	Is development necessary?	4.86
	2	Good/high data quality	3.57
	3	Good/high data usability	3.86
	4	Good data accessibility	4.14
Readability	5	Good font and color	4.57
	6	Good compositions of screen	4.43
	7	Good function buttons	4.00
Efficiency	8	Good navigation	4.57
	9	Fast response time to access datasets	4.00
Grand mean			4.22

IV. Conclusion

In this paper, we have proposed an Open Health Data System. The usability testing result shows that our proposed Open Health Data has a grand mean of 4.22 (of scale 5). This shows that the performance of our proposed Open health Data is excellent. However, there are certain factors should be considered carefully, such as:

a. Data collection problem. The data is collected from various sources and organizations created by different stakeholders for different purposes without consideration of integration.
b. Confidentiality of personal data and patient data transmission issues. Although Indonesian Health Law No. 36 of 2009 states that health data can be shared based on public interest or in response to an emergency, there remain some concerns that data transmission to other institutions violates patient data protection.

c. Resources issues. Data sets we found often need to be completed and real-time, and there are discrepancies between the system and the manual reports. This is shown from the usability testing result, which is data quality score (3.57 of scale 5) is the lowest followed by data usability (3.86 of scale 5).

For our subsequent work, we should ensure that the data has high quality. In addition, the data should be usable, accessible, and approved legally and ethically by doing some testing and evaluations for health experts and policymakers.

References

[1] JKR. Nayek, "Evaluation of Open Data Government Sites: A Comparative Study," Library Philosophy and Practice (e-journal). 1781. 2018. https://digitalcommons.unl.edu/libphilprac/1781

[2] MRA. Nurrahma, "Transparansi Data Pembangunan Kesehatan Perspektif Open Government Data Transparency of Health Development Data from Open Government Data Perspective", Journal of Governance and Administrative Reform 1:1. June 2020.. https://e-journal.unair.ac.id/JGAR/index

[3] P. Kostkova, H. Brewer, S. de Lusignan, E. Fottrell, B. Goldacre, G. Hart, P. Koczan, P. Knight, C. Marsolier, RA. McKendry, E. Ross, A. Sasse, R. Sullivan, S. Chaytor, O. Stevenson, R. Velho, J. Tooke, "Who Owns the Data? Open Data for Healthcare", Front. Public Health 4:7. 2016.

[4] C. Culnane, BIP. Rubinstein, V. Teague. "Health Data in an Open World", December 2017. https://www.researchgate.net/publication/321873477_Health_Data_in_an_Open_World

[5] KP. Seastedt, P. Schwab, Z. O'Brien, E. Wakida, K. Herrera, PGF. Marcelo, "Global healthcare fairness: We should be sharing more, not less, data", PLOS Digit Health 1(10): e0000102. 2022. https://doi.org/10.1371/journal. pdig.0000102

[6] R. Hejlan, J. Crompvoets, "Open health data: Mapping the ecosystem", Digital Health Volume 7: 1–16. 2021.

[7] J. Piasecki, PY. Cheah, "Ownership of individual-level health data, data sharing, and data governance", BMC Medical Ethics 23:104. 2022. https://doi.org/10.1186/s12910-022-00848-y

[8] L. Hurbean, D. Danaiata, F. Militaru, AM. Dodea, AM. Negovan, "Open Data Based Machine Learning Applications in Smart Cities : A Systematic Literature Review". Electronics 10, 2997. 2021. https://doi.org/10.3390/electronics10232997

[9] Q. Chen, W. Wang, F. Wu, S. De, R. Wang, B. Zhang, X. Huang, "A Survey on an Emerging Area: Deep Learning for Smart City Data", IEEE Transactions on Emerging Topics in Computational Intelligence, 3, 392-410. 2019.

[10] ODI 2017. https://opendatabarometer.org.

[11] ODIN 2022. https://odin.opendatawatch.com › countryProfile › IDN.

[12] MJ. Islami, "Implementasi Satu Data Indonesia: Tantangan dan Critical Success Factors (CSFs)", Jurnal Komunika 10:1. Juni 2021. DOI: 10.31504/komunika.v9i1.3750.

[13] B. Wicaksono, D.S. Rusdianto, AH. Brata, "Pengembangan Sistem Portal Satu Data Indonesia Pada Kantor Staf Presiden Menggunakan Comprehensive Kerbal Archieve Network (CKAN)", Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 2:8, 2882–2888. 2018.

[14] M. Gryseels, N. Manuel, L. Salazar, P. Wibowo, "Ten ideas to maximize the socioeconomic impact of ICT in Indonesia", (March), 1–20. 2015.

[15] https://opendatabarometer.org/4thedition/?_year=2016&indicator=ODB

[16] Open Data Unit, "Open Data Strategy 2017 -2022". 2017. Dublin.

[17] DS. Sayogo, S. Budi, C. Yuli, "Critical Success Factors of Open Government and Open Data at Local Government Level in Indonesia", International Journal of Electronic Government Research, 14:2, 28–43. 2018. https://doi.org/10.4018/IJEGR.2018040103

[18] I. Susilowati, S. Wisnaningsih, D. Silviawati, "Perlindungan Hukum terhadap Hak Privasi dan Data Medis Pasien di Rumah Sakit X Surabaya", Jurnal Wiyata. 2018.

[19] J. Kucera, D. Chlapek, J. Klímek, M. Nečaský, "Methodologies and Best Practices for Open Data Publication", Databases, Texts, Specifications, Objects. 2015.

[20] Krotova, A. Mertens, M. Scheufen, "Open data and data sharing: An economic analysis", Series/Report no.: IW-Policy Paper No. 21/2020. 2020.

[21] European Data Portal, "Creating Value Through Open Data", Report study on the Impact of Re-use of Public Data Resources. November 2015.

[22] The World Bank 2022. https://opendatatoolkit.worldbank.org/en/technology.html.