Data Provenance Visualization in Brazilian Public Health Dashboards

Johne M. Jarske[*] Universidade de São Paulo
Jorge R. de Almeida[†] Universidade de São Paulo
Lucia V. L. Filgueiras[‡] Universidade de São Paulo
Leandro M. R. Velloso[§] Universidade de São Paulo
Tania L. Santos[¶] Faculdade de Tecnologia de Cotia

Abstract

The use of dashboards to disseminate information and facilitate decision making is an increasingly common practice on companies and governments. However, there are many open questions on how to make reliable dashboards. This paper aims to address dashboard's reliability issues through the visualization of a set of metadata that describes the origin and changes undergone by the data over time, that supports data confidence and validity, known as data provenance. By investigating dashboards published by the Brazilian public health agencies and from abroad, and searching for best practices in the literature, our research team proposes the adoption of a three-layer provenance model, which we expect to address the presented problem. Enabling usable data provenance visualizations, we expect to enhance trust on data driven communication of health-related issues to public managers, press, and the lay public.

Index Terms: data provenance visualizations—provenance—visualization—; user experience—visualization design—design—; unified health system—health—provenance—

1 Introduction

The use of dashboards to disseminate information and enable decision making are ubiquitous. Nearly all industries, non-profit, private or public organizations built and employed dashboards to support data-driven decision making [28]. Despite the dissemination of dashboards, many suffer from a lack of information related to data provenance.

Data provenance consists on a set of metadata that describes the origin and changes undergone by the data over time, that supports data confidence, quality, ensure reproducibility, and reinforce trust in the data [19]. Quality information is strongly linked to provenance and is essential to evaluate data reliability. However, providing user-friendly provenance data visualizations and is still a pressing challenge [15]. The need to address these challenges can be observed in dashboards developed by the Brazilian Ministry of Health to disseminate health-related information to public-health managers, press, and the lay public.

Brazil has a public health system denominated Unified Health System (SUS). SUS is a decentralized and hierarchical public system which management is shared among its three federative levels: federal, state, and municipal. The Brazilian Ministry of Health (federal level) is responsible for formulating, defining standards, inspecting, monitoring, evaluating SUS policies and actions, and disseminating information [7].

The geographic and population Brazilian dimensions allow us to infer the logistical and management issues to maintain the SUS structure, not only related to the flow of supplies and equipment, facilities, and people management, but also in the management of the information flow, which is produced and shared among its three administrative levels and shared with other society actors.

Between 2020 and 2022, our research team carried out the project entitled Infovis for Public Health, funded by PAHO (Pan American Health Organization), meeting the demand of DEMAS/SE-MS (Departamento de Monitoramento e Avaliação do SUS, Secretaria Executiva do Ministério da Saúde – Monitoring and Evaluation Department of SUS, Ministry of Health Executive Secretary). The project was conducted in two stages. The objective of the first stage was to improve the user experience (UX) of public health managers, based on the understanding on how they perceive the usefulness and usability of the information presented on SAGE (Sala de Apoio a Gestão Estratégica - Strategic Management Support Room) dashboards, aiming to propose improvements in the information visualization. The second stage of the project aimed to implement a design system portfolio for dashboards with a complete set of visualization guidelines, a content architecture model, a new dashboard development process that includes a set of activities focused on understanding the users' needs and desires in terms of information and an user-centered dashboard evaluation model.

Among the main issues raised by the users in the series of interviews carried out by the project team were those related to the information quality displayed on the panels. Among the issues, was the lack of information about the data provenance, to the point of compromising users' trust in the information displayed.

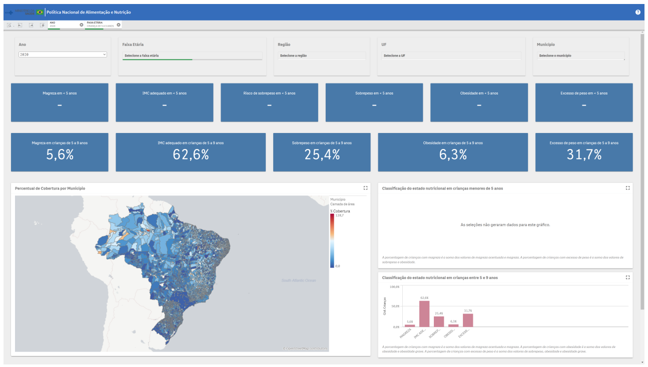
When the information is used for decision making, provenance can be fundamental to improve accuracy, decision time and confidence [14]. For example: i) display the data collection date could help identify whether population health information was obtained based on current data or projections based on the 2010 Census, what is crucial to understand whether the information is subject to distortion and inaccuracy; ii) display the data sources could help to identify objectives, methods of collection, measurement and calculation; iii) identify those responsible for the data could help to communicate data issues and ask for support; iv) display the original data address can make possible analysis reproduction and to carry out different techniques.

Since 2011, the Brazilian Access to Information Law has guided the implementation and promotion of open data, what includes provenance metadata [21]. However, still there is no official policy concerning the use of data provenance in information delivered by the government. For example, Figure 1 shows a SAGE dashboard that does not show any information about its data provenance. It does not offer any annotation or metadata to help the data interpretation. This is not, unfortunately, an isolated exception.

The incorporation of provenance metadata into health databases is a complex task, that needs to be carried out in the long-term including the adoption of database architecture such as HL7- Provenance (Health Level Seven) [6]. However, experiments become necessary to show that it is possible and relevant to implement a provenance structure that allows dashboard users to visualize those involved in the production, manipulation, and dissemination of health data.

The current work is a continuity of Infovis For Public Health

[*] e-mail: johne.jarske@usp.br
[†] e-mail: jorgerady@usp.br
[‡] e-mail: lfilguei@usp.br
[§] e-mail: leandrovelloso@usp.br
[¶] e-mail: tania.ls@outlook.com.br


Política Nacional de Alimentação e Nutrição
ANO
2020
FAIXA ETÁRIA
CRIANÇA DE 5 A 9 ANOS
Ano
2020
Faixa Etária
Selecione a faixa etária
Região
Selecione a região
UF
Selecione a UF
Município
Selecione o município
Magreza em < 5 anos
–
IMC adequado em < 5 anos
–
Risco de sobrepeso em < 5 anos
–
Sobrepeso em < 5 anos
–
Obesidade em < 5 anos
–
Excesso de peso em < 5 anos
–
Magreza em crianças de 5 a 9 anos
5,6%
IMC adequado em crianças de 5 a 9 anos
62,6%
Sobrepeso em crianças de 5 a 9 anos
25,4%
Obesidade em crianças de 5 a 9 anos
6,3%
Excesso de peso em crianças de 5 a 9 anos
31,7%
Percentual de Cobertura por Município
Município
Camada de área
% Cobertura
128,7
0,0
South Atlantic Ocean
Classificação do estado nutricional em crianças menores de 5 anos
Classificação do estado nutricional em crianças entre 5 e 9 anos
100,0%
50,0%
0,0%
Qtd. Crianças
5,6%
62,6%
25,4%
6,3%
31,7%
MAGREZA
IMC ADE...
SOBREP...
OBESID...
EXCESS...

As seleções não geraram dados para este gráfico.

A porcentagem de crianças com magreza é a soma dos valores de magreza acentuada e magreza. A porcentagem de crianças com excesso de peso é a soma dos valores de sobrepeso e obesidade.

A porcentagem de crianças com magreza é a soma dos valores de magreza acentuada e magreza. A porcentagem de crianças com obesidade é a soma dos valores de obesidade e obesidade grave. A porcentagem de crianças com excesso de peso é a soma dos valores de sobrepeso, obesidade e obesidade grave.

Figure 1: SAGE Dashboard about infant nutrition and feeding [9]. The dashboard does not show any information about its data provenance and also does not offer any annotation or metadata to help the data interpretation.

project, addressing open issues related with data provenance pointed by the first two project stages. The aim of this paper is to provide a framework that can enable the dissemination of usable data provenance and guidance for future visualization techniques. Enabling usable data provenance visualizations, we expect to enhance trust on data driven communication of health-related issues to public managers, press, and the lay public.

2 The Concept of Data Provenance

The practice of collecting information about the origin of physical or digital objects is known as provenance. Moreau and Groth [23] define provenance as a set of metadata that describes the people, institutions, entities, and activities involved in influencing, producing, or delivering any type of physical or digital object. The concept of provenance was imported from other areas where the origin of the object is relevant to the perception of value, as in the case of works of art [29]. The importance of provenance metadata lies in ensuring product quality, reproducibility, and reliability, as the user can access the origin of product information whenever necessary [23]

According to Yazici [31], the record of provenance information is fundamental to provide reusable, reproducible, optimized and fault-tolerant solutions as it facilitates the location and exploitation of the original, intermediates and final data sources to ensure verifiability and reliability.

The W3C consortium defined a model for data provenance, the PROV model. It has three main components: entities, agents, and activities. Entities consist of physical, digital, conceptual objects for which provenance metadata are recorded. Activities represent the work, actions, or processes required to create an entity. Agents can be an organization, a person, a software, or an inanimate object, which plays a role in an activity, so they are assigned a degree of responsibility [18].

According to Gil and Miles [18], the basic relationships between the PROV components are:

- **wasDerivedFrom**: an entity is derived from another entity.
- **wasGeneratedBy**: an entity was generated through an activity.
- **used**: an activity used an entity to generate a new entity.
- **wasAssociatedWith**: an entity is associated with an agent.
- **wasAttributedTo**: An activity is assigned to an agent.

Figure 2 shows an overview of the structure of PROV Model with its three main components and their basic relationships.

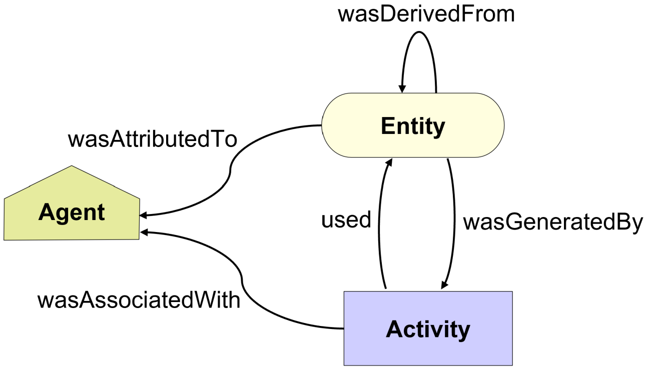


Figure 2: Overview of the structure of PROV records. Elaborated by the authors based on PROV Model Primer [18].

The W3C PROV-O - PROV Ontology [22] provides a complete description of PROV metadata.

An example of provenance relationship can be seen on Figure 3. The entity :map was derived from :mapdataset, was generated by the :designActivity using :mapDataset. The :designActivity was attributed to :SAGE on behalf of :CGGIE. The :mapDataset was derived from :samuData and :geoData, was generated by the :agregationActivity using :samuData and :geoData. The :agregationActivity was associated with:SAGE on behalf of :CGGIE. :geoData was associated with :IBGE and :samuData with :DEMAS.

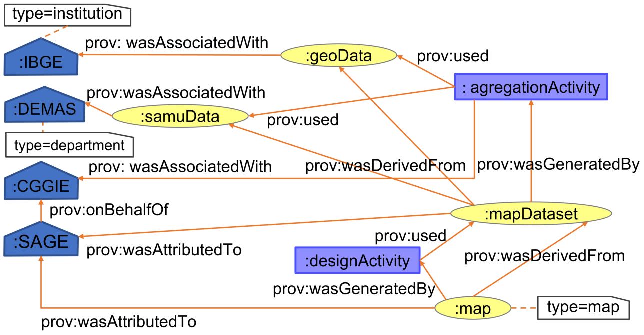


Figure 3: Sample of provenance diagram, elaborated by the authors, based on PROV-O: The PROV Ontology [22].

In the Big Data era, the amount of data created by has grown each year exponentially. As the amount of data grows, data provenance also grows. The requirements for the analysis and visualization of provenance are ongoing issues in many different scenarios [31]. According to Yazici at. al 31, effective data visualization is an important tool to make data provenance more accessible.

The most basic view of metadata is in the form of structured text. The W3C-PROV proposes, however, a diagramming model [18] (see Figure 3). Hoekstra and Groth [20] propose the use of the Sankey diagram, and other authors suggest the use of node links [2, 3, 5]. Yazici et al. [31] proposed a graph visualization with summarization methodology to generate short summaries for massive graph and Yazici et al. [30] a graph visualization associated with an graphical comparison approach between provenance files. Most of these approaches were designed to display a large amount of data and were made for the use of provenance researchers and specialists.

As data provenance visualization is not a easy task from a lay public and requires some provenance visualization literacy. As stated by Yazici at. al 31, users can become overwhelmed trying to understand and explore the meaning of data provenance due to its sheer volume and complexity. Data provenance visualization should be integrated seamlessly in the dashboard to enable a satisfactory data comprehension.

3 Methods

To address the issue of providing a framework that can enable the dissemination of usable data provenance and guidance for future visualization techniques ainming to enhance trust on data driven communication of health-related issues to public managers, press, and the lay public,this study had, so far, the following studies: i) exploratory search in COVID-19 dashboards from different countries, with the aim of identifying if, and how, these panels deal with the provenance issues; ii) address the findings of the "Infovis for Public Health" project regarding data provenance: the experience perception of health managers from municipalities, states and federation, obtained through interviews and usability inspections, in the use of panels made available by SAGE; the understanding of information and provenance flow through the SUS systems; definition of the design system framework that points out the need for specifying data provenance.

4 Results

The results of the research so far are presented in this section.

4.1 Provenance in COVID-19 dashboards

To identify how different countries deal with provenance metadata in their dashboards. The COVID-19 dashboards were selected because they are a source of common concern for all countries researched. We researched 12 institutional COVID-19 dashboards from different countries in the 5 continents. The results are displayed in Table 1.

Table 1: Samples of COVID Dashboards

Country	Source	Provenance-related metadata
Brazil	[8]	Data source, last update, link to the data source.
Argentina	[10]	Data source, last update and update frequency, annotations.
Chile	[11]	Data source, data period, last update. It contains a tab called Notas Técnicas (Technical Notes) with annotations about the data.
Colombia	[12]	Last update, link to the open data. It contains a tab called Notas (Notes) with annotations about the data.
United States	[16]	Data source, period, last update, for each graph: start date, postdate, link to the data source, annotations and link to metadata and definitions.
Canada	[24]	Data source, last update, ownership of the data, link to the data source, lots of annotations, archived reports.
Germany	[17]	Data source, last update, annotations, link to the data source.
United Kingdom	[1]	Data source, last update, last update for each graph, annotations, metadata, link to data source.
France	[13]	Data source, link to the open data.
India	[27]	Last update, annotations.
Australia	[26]	Data source, last update, annotations.
South Africa	[25]	Data Source, last update, annotations.

Provenance visualization was not a main concern. In all of 12 researched dashboards provenance were not an explicit subject. Instead of looking for provenance, we looked for metadata commonly related to provenance such as data source, last update, update frequency, start date, data period, link to the data source, link to archived reports, and annotations. The most common metadata found were the data source and last update.

United States and Canada documents have the most well instrumented dashboards with metadata, including provenance related metadata and specially annotations.

4.2 User needs

The subject of the data provenance is still unknown by most users, but its need was identified from the way the interviewed users characterized the information available at the SAGE dashboards: fragmented; outdated; divergent from one system to another; not reflecting reality; not reflecting the information entered by the municipalities; not informing the origin; available without known need; unreliable. Provenance metadata is necessary so that the above issues are properly understood or addressed to those responsible for disseminating the information on the dashboards.

4.3 SUS – Information and Provenance Flow

The SUS is a very complex public health system based on a decentralized administration shared among its three federative levels: federal, state, and municipal. The municipal level is responsible to carry out the health services to the population; the state level to coordinate regional initiatives and planning; and the federal level to the general coordination, planning, resources distribution, and monitoring and control.

Through the applications provided by DATASUS (informatics department of the SUS), the state and municipality levels, among other information, submit all data related to the services provided to the population.

Based on the information supplied by the states and municipalities, the federal level through its different departments can manage, monitor, and control the established health services and programs. Among SUS departments, DEMAS (Monitoring and Evaluation Department), is responsible for articulating, developing, monitoring, and evaluating actions in the public health system and guaranteeing the dissemination of strategic information, providing support for decision making through the SAGE portal.

The SAGE portal provides strategic public health information to public health professionals, public health managers at the three administrative levels, press, researchers and the interested public. Figure 4 illustrates this flow of information and provenance.

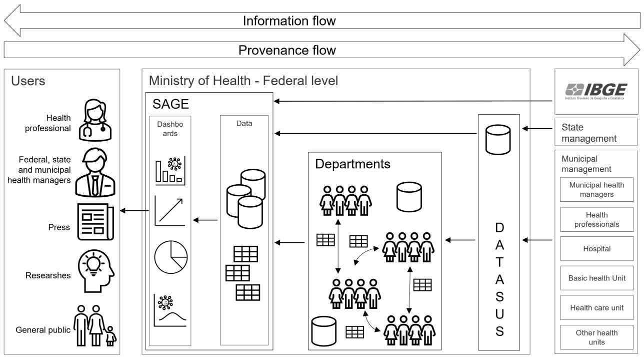


Figure 4: Information and provenance flow through SUS administrative levels.

4.4 Provenance and Design System

The design system for dashboards proposed by the research project Infovis for Public Health was created with the main objective of supporting a set of guidelines for the production of data visualization in public health dashboards. The Design System provides the basis for standardizing visual elements for data visualization projects, presenting good practices and requirements in the use of health charts and maps.

The design system also recommends that each dashboard or graph should contain functions that enable the access to information about the graph interpretation, explanatory notes, data description and data provenance.

5 THREE LAYER PROVENANCE VISUALIZATION

In 1997, Tim Berners Lee, amid the huge boost of the WEB, invented by himself, 8 years before, in 1989, suggested that any respectable WEB application should have an ”Oh, yeah?” button: “At the toolbar (menu, whatever) associated with a document there is a button marked ”Oh, yeah?”. You press it when you loses that feeling of trust. It says to the Web, ”so how do I know I can trust this information?”. The software then goes directly or indirectly back to meta-information about the document, which suggests a number of reasons.” [4].

To meet the users' needs discussed in session 4.2, a provenance visualization should follow the idea of the ”Oh, yeah?” designed by Tim Berners. Every time the user loses the feeling of trust, he can use provenance to track the transformation of the data.

For the most dashboard users, it is not necessary to follow the data provenance until the origin, but until the nearest data owner, responsible for the data. In the context of SUS, the CGGIE is the department responsible to maintain SAGE dashboards. Each dashboard attends a demand of an internal department responsible for a health program and consequently, for its related data. These departments are, in general, the owner of the data, responsible for its analysis, interpretation, maintenance and accuracy.

Tracking the data provenance, the user can identify issues in the visualization and the data itself, download the data at each transformation step, and can identify the agent responsible for each activity (data transformation). Adding a communication support (e-mail or chat, for example) in the provenance visualization, the user can send messages to the agent responsible for each activity to ask questions, to make suggestions or to point errors.

In this way, it is proposed a three-layer provenance covering the user interface; the data used by the user interface; and the data sources from which the datasets used by the user interface were extracted (generally supplied by the data owner). Figure 5 points out the proposed three-layer architecture.

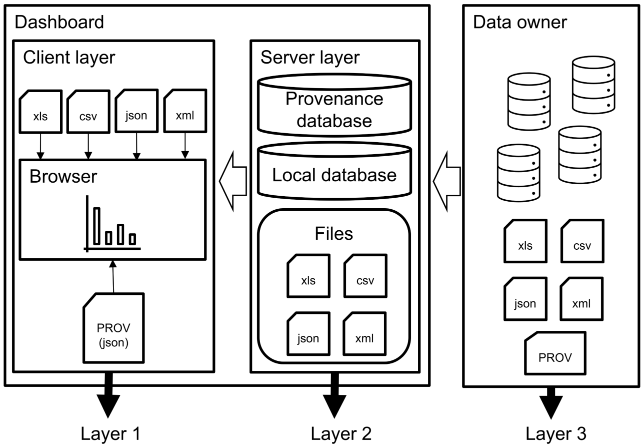


Figure 5: The three-layer provenance model.

For each layer, the following provenance metadata must be captured: i) artifact name (panel, graphic, data file, database); ii) date of creation; iii) start date (of use); iv) end date (when applicable); v) version; vi) previous version; vii) agent responsible; viii) requester; ix) resources used; x) and address for data access. To complement the provenance metadata, annotations or other metadata can be inserted to support the understanding of the resources involved.

Additionally, since it does not exist yet a systematic program for automated capture of provenance metadata and its incorporation in the Ministry of Health's databases, some limitations are imposed. If, on the one hand, it is not possible to map all data transformations and data interchange, on the other hand, it is possible to identify the main stages and their respective entities (data), activities (interchange and transformations undergone) and agents (responsible for the entities and activities) involved. A three-layer provenance metadata can be easily captured by automatic tools, or, at least, manually. However, manually collecting and maintaining provenance information can be cumbersome and error prone.

Another important aspect of a three-layer provenance model can allow a more user-friendly provenance visualization. According to Yazici [31], an effective data visualization is a key step to enable a better understanding of a complicated data such as provenance.

Yazici [31], argues that data provenance can easily become large, messy, and difficult to interpret. Hence, there is a need for visualization approaches such as summarization or, as our proposal, to refine the scope of the provenance visualization to the most important parts for the users.

The three-layer approach will enable users to explore, beyond what has already been proposed in the literature, a set of more user-friendly visualizations, including infographics.

6 CONCLUSION

This paper is part of an ongoing project. The research goes in the direction of better understanding requirements of a provenance visualization model and considering that this is a project that intends to bring the provenance visualization to the users of dashboards, it is necessary to adopt appropriate user experience evaluation tools.

So far, we have identified through an exploratory review some of the most relevant approaches to data provenance visualization; we explored COVID-19 dashboards maintained by the governments of 12 countries (a common concern to all of them) and observed how provenance-related metadata was addressed; identified user needs that can be addressed with data provenance visualization; identified the information flow through the three administrative levels of the SUS and how to trace data provenance; add to our dashboard design system the need of an ”Oh, yeah?” button like solution to display information about the graph interpretation, explanatory notes, data description and provenance and, finally; we proposed a three-layer data provenance approach that will enable a set of user-friendly provenance visualizations, including infographics.

The next necessary step to our research is to execute a systematic review on data provenance visualization to identify the state-of-art on this subject; mature our three-layer provenance model and experiment a new set of provenance visualization approaches.

ACKNOWLEDGMENTS

The authors wish to thank PAHO (Pan American Health Organization) for the for the financial support, FUSP (Fundação de Apoio Universidade de São Paulo) for the project execution and the Brazilian Ministry of Health, especially the DEMAS team for all the information provided and the partnership.

REFERENCES

[1] U. H. S. Agency. Coronavirus (covid-19) in the uk, 7 2022.

[2] M. K. Anand, S. Bowers, and B. Ludäscher. Provenance browser: Displaying and querying scientific workflow provenance graphs. pp. 1201–1204, 2010.

[3] L. Bavoil, S. P. Callahan, P. J. Crossno, J. Freire, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Vistrails: Enabling interactive multiple-view visualizations. pp. 135–142, 2005.

[4] T. Berners-Lee. Cleaning up the user interface, section-the” oh, yeah?”-button, 1997.

[5] A. Chebotko, S. Lu, S. Chang, F. Fotouhi, and P. Yang. Secure abstraction views for scientific workflow provenance querying. *IEEE Transactions on Services Computing*, 3:322–337, 2010.

[6] H. Community. Hl7 fhir specification: Resource provenance - content, 2011.

[7] A. N. Constituinte. Constituição da república federativa do brasil, 1988.

[8] M. da Saúde Brasil. Painel coronavírus, 7 2022.

[9] M. da Saúde Brasil. Política nacional de alimentação e nutrição, 7 2022.

[10] M. de Salud Argentina. Monitor público de vacunación, 7 2022.

[11] M. de Salud Chile. Covid-19 en chile, 7 2022.

[12] I. N. de Salud Colombia. Coronavirus (covid-19) en colombia, 7 2022.

[13] M. des Solidarités et de la Santé et Santé publique France. Covid-19 - france, 7 2022.

[14] B. Dy, I. Nazim, A. Poorthuis, and S. C. Joyce. Improving visualisation design for effective multi-objective decision making. *IEEE Transactions on Visualization and Computer Graphics*, 2626:1–12, 2021. doi: 10.1109/TVCG.2021.3065126

[15] H. Figgemeier, C. Henzen, and A. Rümmler. A geo-dashboard concept for the interactively linked visualization of provenance and data quality for geospatial datasets. *AGILE: GIScience Series*, 2:25, 2021. doi: 10.5194/agile-giss-2-25-2021

[16] C. for Disease Control and Prevention. Covid data tracker, 7 2022.

[17] R. K.-I. Germany. Robert koch-institut: Covid-19-dashboard, 7 2022.

[18] Y. Gil and S. Miles. Prov-primer: a primer for the prov data model, 2013.

[19] M. Herschel, R. Diestelkämper, and H. B. Lahmar. A survey on provenance: What for? what form? what from? *VLDB Journal*, 26:881–906, 12 2017. doi: 10.1007/s00778-017-0486-1

[20] R. Hoekstra and P. Groth. Prov-o-viz - understanding the role of activities in provenance. *Provenance and Annotation of Data and Processes. IPAW 2014. Lecture Notes in Computer Science, vol 8628.*, 1:2015–2019, 2015. doi: 10.1007/978-3-319-16462-5 18

[21] J. M. Jardim. A lei de acesso à informação pública. *Tendências da pesquisa brasileira em ciência da informação*, 5, 2012.

[22] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. *PROV-O: The PROV Ontology*. World Wide Web Consortium, 2013.

[23] L. Moreau and P. Groth. *PROV-OVERVIEW - An Overview of the PROV Family of Documents*, vol. 3. 2013. doi: 10.2200/S00528ED1V01Y201308WBE007

[24] G. of Canada. Covid-19: Outbreak update, 7 2022.

[25] D. of Health. Covid-19 - online resource news portal, 7 2022.

[26] D. of Health and A. Care. Coronavirus (covid-19) case numbers and statistics, 7 2022.

[27] M. of Health and F. Welfare. Covid-19 facilities in states union territories, 7 2022.

[28] A. Sarikaya, M. Correll, L. Bartram, M. Tory, and D. Fisher. What do we talk about when we talk about dashboards? *IEEE Transactions on Visualization and Computer Graphics*, 25:682–692, 2019. doi: 10.1109/TVCG.2018.2864903

[29] L. C. teacher) Moreau and I. Foster. *Provenance and annotation of data : International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006 : revised selected papers*. Springer, 2006.

[30] I. M. Yazici and M. S. Aktas. A usability study on data provenance visualization approaches. Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/UYMS54260.2021.9659779

[31] I. M. Yazici, E. Karabulut, and M. S. Aktas. A data provenance visualization approach. *Proceedings - 2018 14th International Conference on Semantics, Knowledge and Grids, SKG 2018*, pp. 84–91, 7 2018. doi: 10.1109/SKG.2018.00019