





A documentary analysis of Victorian Government health information assets' websites to identify availability of documentation for data sharing and reuse in Australia

Merilyn Riley, *GDipEpidBiostats*¹ ,
 Monique F. Kilkenny, *PhD*^{2,3} ,
 Kerin Robinson, *PhD*¹ ,
 Sandra G. Leggat, *PhD*^{1,4} 

Health Information Management Journal

1–9

© The Author(s) 2023



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/18333583231197756

journals.sagepub.com/home/himj



Abstract

Background: Health data sharing is important for monitoring diseases, policy and practice, and planning health services. If health data are used for secondary purposes, information needs to be provided to assist in reuse. **Objectives:** To review government health information asset websites to ascertain the extent of readily available, explanatory documentation for researcher sharing and reuse of these data. **Method:** Documentary analysis was undertaken on selected Victorian Government health information assets' websites in Australia. Data were obtained on nine information-categories: data custodian; data context; data dictionary; quality controls; data quality; limitations; access process; privacy/confidentiality/security and research requests/outputs. Information-categories were compared by dataset type (administrative or population-health) and by curating organisation (government or other agency). Descriptive statistics were used. **Results:** The majority of the 25 websites examined provided information on data custodian (96%) and data context (92%). Two-thirds reported access process (68%) and privacy/confidentiality/security information (64%). Compared with population-health websites, administrative dataset websites were more likely to provide access to a data dictionary (67% vs 50%) and information on quality controls (56% vs 44%), but less likely to provide information on the access process (56% vs 75%) and on research requests/outputs (0% vs 56%, $p = 0.024$). Compared with government-curated websites, other agency websites were more likely to provide information on research requests/outputs (80% vs 7%, $p < 0.001$). **Conclusion:** There is inconsistent explanatory documentation available for researchers for reuse of Victorian Government health datasets. Importantly, there is insufficient information on data quality or dataset limitations. Research-curated dataset websites are significantly more transparent in displaying research requests or outputs.

Keywords (MeSH)

routinely collected health data; data sharing; secondary data analysis; data curation; data accuracy; health information systems; health information management

Introduction

In an infodemic, where very large volumes of both accurate/reliable and inaccurate data and information are circulated, appropriate “infodemic management” is critical to minimise its potential adverse impacts, especially to public health (Wilhelm et al., 2023). Misinformation in this context can result in confusion and widespread mistrust of health leaders and scientific data. Despite the voluminous generation of health data arising from digital health and the risks of misinformation, there are driving forces towards enhancing data linkage to facilitate information sharing between, and across, health agencies (Australian

Government Department of Health and Aged Care, 2022). “The power of our data to solve key policy challenges grows exponentially as we make it more complete, more joined up and more available” (Australian Government Department of Health and Aged Care, 2022a: 2). The sharing and reuse of data does not occur in isolation. In Australia, health data are collected in specific contexts for

¹La Trobe University, Australia

²Monash University, Australia

³Florey Institute of Neuroscience and Mental Health, Australia

⁴James Cook University, Australia

Accepted for publication August 12, 2023.

Corresponding author:

Merilyn Riley, School of Psychology and Public Health, La Trobe University, Kingsbury Drive, Bundoora, VIC 3086, Australia.

Email: merilyn.riley@latrobe.edu.au

Correction (October 2023): Article updated to correct authors' order in the article. Please see (<https://doi.org/10.1177/18333583231208191>) for more details.

explicit purposes by governments, registries and health agencies at local, state and federal levels.

Much of the person-level health data that are provided to governments is a by-product of the clinician–patient information exchange. Transformation occurs as these data progress through multiple extraction, clinical coding and reporting processes, until they are finalised as an information asset at the relevant state and territory Department of Health. By this point in the journey, they have already become secondary data, consistent with Berg and Goorman's (1999) proposed "law of medical information." That is, the further data are removed from their original context and the more diverse purposes for which they may be used, the greater the task "to disentangle it from the context of its production" (Berg and Goorman, 1999: 51).

During the early stages of the development of data quality research, Lee and Strong (2003: 15) identified the importance of "knowing-why. . . behind routine data production activities" as it empowered data consumers to understand and question data quality issues and provide solutions. Awareness of the need for appropriate infrastructure (i.e. documentation, processes, technology) for sharing and reuse of data became even more apparent during the subsequent decade with the establishment of the Open Data Charter in 2015 (Open Data Charter [ODC], n.d.). Experts from universities, funding organisations, publishers and data scientists collaborated to develop the infrastructure requirements for effective sharing and reuse of data. This resulted in the development of the FAIR principles (findability, accessibility, interoperability and reusability) to promote good data management and reuse of data (Wilkinson et al., 2016).

More recently, the extensive and variable resources and requirements needed to transform data to an interoperable and reusable level have been identified (Huston et al., 2019; Khan et al., 2023; Tenopir et al., 2020,). Kim (2022) concluded that improved data quality and explanatory documentation surrounding datasets led to increased satisfaction in the reuse process for researchers. In their systematic review, Mc Grath-Lone et al. (2022) identified considerable inconsistency between researchers' understanding of what makes data ready for research. While not specifically focusing on reuse of data, Mc Grath-Lone et al. (2022: 1) identified five characteristics necessary for data to be defined as research-ready: "(a) available, (b) broad, (c) curated, (d) documented and (e) enhanced for research purposes." These authors found that documentation describing key characteristics of the data focused on the *availability* and *transparency* of information including context, purpose, creation and processing, coverage, quality and completeness, limitations, user guides, data governance and access, and use in research.

Previous studies have identified the difficulties in navigating the minefield of information to understand and access data for reuse purposes. For example, Williamson et al. (2022: 623) described the journey as clinician-academics in obtaining access to routine healthcare data as involving: "nine . . . stakeholders from four . . . organisations [who] took almost 3 years, including 15 initial or revised applications, assessments or agreements." Andrew et al. (2016) and Palamuthusingam et al. (2019) described

the difficulty in undertaking data linkage with multiple datasets, and across jurisdictions, due to inconsistencies in access policies, variable skill levels of both researchers and custodian/data owner staff, resourcing issues concerning time, staff, and money, data limitations, and restrictive and siloed policies and practices. In researching data reuse, York (2022: xvi) observed that "researchers lacked knowledge they desired about data [which] . . . frequently had a negative impact on their research."

There is a gap in Australian research on the provision of *available* and *transparent* documentation (i.e. trustworthy guidelines) for "research-ready" government health information assets for reuse purposes. Therefore, the aim of this study was to analyse selected government health information asset websites to ascertain the extent of explanatory documentation readily available for researcher access and reuse of these data.

Method

Study design

Documentary analysis (Bowen, 2009; Dalglish et al., 2020) in the form of an audit of website explanatory content was undertaken in March and April, 2023. This enabled the investigation of the type of documentation about a selected sample of government health information assets that was readily and publicly available to support the appropriate reuse of these datasets by researchers. The terms "information assets" and "datasets" have been used interchangeably in this article.

Sample and exclusion criteria

Riley et al. (2022) identified 28 datasets that were reported in the Victorian Department of Health information asset register in March 2019. These datasets, both administrative and population-health based, included person-level data and were associated with at least one peer-reviewed publication between 2008 and 2020. The websites of these information assets formed the basis of the current study (see Table A1). Any of the within-scope datasets that had subsequently become inactive or had been replaced by a new dataset were excluded.

Development of audit tool and data collection

An abstraction audit template was created based on recommendations from previous studies (Gilbert et al., 2018; Gordon et al., 2021; Mc Grath-Lone et al., 2022) and the Victorian Government (2019a) Data Access Policy. Sixteen information-categories recognised by these authors to be important for meaningful reuse of data were selected (see Supplemental Table S1). Audit information-categories included information on governance structures, funding source(s), purpose, scope, nature of collection (mandatory or voluntary), quality controls and data quality statements, participant privacy/confidentiality and security of data provision, meta-data, limitations, access process, access fee,

review of outputs before release, list of research requests or outputs publicly available, website address and comments. An audit data dictionary was compiled to assist with data collection (see Supplemental Table S2). One reviewer completed the abstraction. To validate the data abstraction and completion of the audit tool, a sample of five datasets (20%) was utilised to determine inter-rater agreement between the primary reviewer and a second reviewer.

This article focuses on the availability of documentation for 9 of the 16 categories for which data were abstracted; these relate to “research-readiness” (i.e. data custodian, context, data dictionary, quality controls, data quality statement, limitations, access process, privacy/confidentiality and security, and research requests or outputs). Each of the dataset websites was broadly categorised into administrative and population-based, as defined in Riley et al. (2022), for the analyses of information-categories. Datasets were also categorised into two groups based upon their data custodianship (i.e. “government-curated” and “other agency-curated”). Each dataset website was also anonymised and categorised into government, research and other for comparison of the individual information-categories.

Analysis

Manifest content analysis, as described by Kleinheksel et al. (2020), was used to summarise the main findings of the document audit. This approach involved the systematic investigation of large volumes of easily observable textual data, often incorporating “surface-level analysis [which] assumes there is objective truth in the data that can be revealed with very little interpretation” (Kleinheksel et al., 2020: 128). Fisher’s exact test, $\alpha=0.05$, was calculated in OpenEpi Version 3.01 (Dean et al. 2014), to identify associations for categorical variables where cell numbers were <5 . Cohen’s Kappa was calculated to determine inter-rater reliability between the sample extraction for both reviewers (Cohen, 1960). The Landis and Koch (1977) interpretation of Cohen’s Kappa was used to describe agreement levels: 0–0.2 (*slight agreement*), 0.21–0.40 (*fair agreement*), 0.41–0.60 (*moderate agreement*), 0.61–0.80 (*substantial agreement*) and 0.81–1.0 (*almost perfect to perfect agreement*).

Ethics

No ethical approval was required for this study as all documents utilised were available in the public domain.

Results

Of the 28 within-scope information assets, 25 datasets were included in the audit (see Table A1). Criteria for exclusion included datasets that had been replaced with new service providers and/or data collection processes ($n=2$), and one dataset that no longer actively collected data at the time of website extraction.

Based on a sample of five datasets, there was fair to perfect agreement between two reviewers, on the availability of each information-category on dataset websites. Of the

Table 1. Comparison of selected information-categories available on websites, by type of dataset.

Information-categories	Type of dataset	
	Administrative N=9	Population N=16
	n (%)	n (%)
Data custodian	8 (89)	16 (100)
Contextual information	9 (100)	14 (88)
Data dictionary (meta-data)	6 (67)	8 (50)
Quality controls	5 (56)	7 (44)
Data quality statement/information	0 (0)	1 (6)
Limitations	1 (11)	1 (6)
Access process	5 (56)	12 (75)
Privacy/confidentiality & security	6 (67)	10 (63)
Research requests/output publicly available	0 (0)	9 (56)*

Note. This table must be interpreted with caution due to the small numbers.

*Statistically significant $p < 0.05$.

nine information-categories, six had perfect agreement (data custodian, contextual information, quality controls, limitations, privacy/confidentiality/security, requests/outputs publicly available), two had moderate to substantial agreement (data quality information and data dictionary respectively) and one information-category (access process) had fair agreement (Landis and Koch, 1977).

Table 1 highlights the variability in the number of information-categories available on websites for both the administrative and population-health datasets. Overall, the websites for most datasets included information on the data custodian, context, and privacy/confidentiality and security. Proportionally, more administrative (than population-based) dataset websites contained information on data dictionary and quality controls, while the websites for the population-health datasets contained more information on the access process. Both administrative and population-health based datasets contained limited information on data quality statements (0% and 6% respectively) or dataset limitations (11% and 6% respectively). Compared with administrative datasets, population-health datasets provided significantly more information on research requests or outputs (0% vs 56%, $p=0.024$). Table 2 provides a comparison of selected information-categories available on data custodian websites by government-held status. Based on both data custodianship types, the majority of websites for both government- and other agency-curated datasets provided information on data custodian (93% and 100% respectively) and contextual information (93% and 90% respectively). Compared with government, other agency-curated datasets were more likely to hold publicly available information on research requests or outputs (7% vs 80%; $p < 0.001$).

There was variability in the availability of specific information-categories provided by curating organisation (Table 3). Only 11 (44%) of the website datasets provided information on six or more of the categories. The highest number of available information-categories (i.e. eight) was

Table 2. Comparison of selected information-categories available on data custodian websites by government-held status.

Information-categories	Government-curated N= 15	Other agency-curated* N= 10
	n (%)	n (%)
Data custodian	14 (93)	10 (100)
Contextual information	14 (93)	9 (90)
Data dictionary (meta-data)	7 (47)	7 (70)
Quality controls	7 (47)	5 (50)
Data quality statement/information	1 (7)	1 (10)
Limitations	1 (7)	0 (0)
Access process	9 (60)	8 (80)
Privacy/confidentiality and security	9 (60)	7 (70)
Research requests/output publicly available	1 (7)	8 (80) [†]

*"Other" refers to research centres, screening services and industry associations. [†]Statistically significant $p < 0.05$.

provided on websites related to datasets curated by other agencies. No dataset websites included information on all categories. The websites with the lowest number of available information-categories were all government-curated datasets. There was considerable variation in the number of information-categories available for each of the government-curated administrative datasets (Gov1-admin to Gov9-admin). Of these nine datasets, four (44%) included 6/9 of the information-categories (67%) compared to only one (17%) of the government-curated population-health datasets (Gov10-pop to Gov15-pop), which included 6/9 of the information-categories. The information-categories that were well reported (by more than half of the within-scope datasets) included data custodian, contextual information, data dictionary, access process and measures for ensuring privacy/confidentiality and security.

Discussion

It is well documented that successful reuse of data requires the availability of appropriate infrastructure and documentation to provide data reusers with sufficient knowledge to manage and analyse the data (Khan et al., 2023; Wang et al., 2023). Under the *Victorian Protective Data Security Standards V2.0*, government organisations are required to maintain an information asset register (Office of the Victorian Information Commissioner [OVIC], 2019a), with the aim to ensure "consistent identification . . . for public sector information across its lifecycle" (OVIC, 2021: 11). A recommended information asset register template is provided by the OVIC (2019b), which contains provision for a number of the information-categories that we have identified as essential for appropriate documentation for reuse of government health information assets. Despite the provision of such a template, our findings demonstrated there was considerable inconsistency in the documentation available on government health information asset websites.

Purpose of datasets and curating organisations

Most of the within-scope datasets were curated by three categories of organisations (specifically, research centres, government, industry associations). Each reflected a different approach to database management dependent upon the

organisational curation processes. Jahnke and Asher (2012: 1) identified this lack of cross-organisational conformity as "one of the major challenges facing data curation today." This was reflected in our audit of information-categories available on dataset websites. The most obvious outcome, the lack of standardisation of the available documentation, supported Mc Grath-Lone et al.'s (2022) findings on the inconsistency of information available on specific information assets. Grouping these datasets for analysis and categorising them as either administrative or population-health, or by curator (government or other agency), proved unhelpful. There was only one statistically significant outcome between these groupings (i.e. datasets curated by "other agency" were more likely to provide information on research requests or research outputs). It proved more valuable to analyse the information-categories for each dataset separately. Some dataset websites demonstrated the availability of most of the specified information-categories to assist researcher knowledge in the use of their datasets, whereas others provided very little information.

The lack of some information-categories on the websites of administrative datasets is not surprising. Despite governments' calls for increased open data sharing across services, many administrative datasets are not set up for reuse for research (Mc Grath-Lone et al., 2022; Nikiforova & McBride, 2021). Our analysis of the government administrative datasets identified four that provided six information-categories on their website; these possibly represented administrative datasets that are well utilised. Riley et al. (2022) previously identified that only 4 of the 28 Department of Health (DoH) information assets under study were associated with over half of the resultant 756 publications, thereby supporting advice from the *DataVic Access Policy Guidelines* that "high-value datasets should be prioritised" (Victorian Government, 2019a). Not all government information assets are provided with the same curation resources, highlighting that some datasets will be more "research-ready" than others (Tenopir et al., 2020).

Data access, privacy and confidentiality and security

The ability to locate and easily access a dataset is a fundamental contributor to a researcher's decision on whether or

Table 3. Availability of selected information-categories on dataset websites, April 2023 (n = 25).

ID no.	Data custodian	Contextual information	Data Dictionary (meta-data)	Quality controls	Data quality statement	Strengths and limitations	Access process	Privacy/confidentiality and security	Research requests/publicly available	Total
Gov1-admin	-	Yes	-	-	-	-	-	-	-	1
Gov2-admin	Yes	Yes	Yes	Yes	-	-	Yes	Yes	-	6
Gov3-admin	Yes	Yes	Yes	Yes	-	-	-	Yes	-	5
Gov4-admin	Yes	Yes	Yes	Yes	-	-	Yes	Yes	-	6
Gov5-admin	Yes	Yes	Yes	Yes	-	-	Yes	Yes	-	6
Gov6-admin	Yes	Yes	Limited, fields only	Limited	-	-	Yes	Yes	-	4
Gov7-admin	Yes	Yes	Yes	Yes	Yes	-	Yes	-	-	6
Gov8-admin	Yes	Yes	-	-	-	-	-	-	-	2
Gov9-admin	Yes	Yes	Yes	-	Limited	Yes	?	Yes	-	5
Gov10-pop	Yes	Yes	-	-	-	-	Yes	-	-	3
Gov11-pop	Yes	Yes	-	Yes	-	-	Yes	Yes	-	5
Gov12-pop	Yes	-	-	-	-	-	Yes	Yes	-	3
Gov13-pop	Yes	Yes	-	-	-	-	-	-	-	2
Gov14-pop	Yes	Yes	Yes	Yes	-	-	Yes, in annual report	Yes	-	6
Gov15-pop	Yes	Yes	-	-	-	-	-	-	Yes	3
Res1-pop	Yes	Yes	Yes	Yes	-	-	Yes	Yes	Yes	7
Res2-pop	Yes	Yes	Yes	Yes	-	-	Yes	Yes	Yes	7
Res3-pop	Yes	Yes	Yes	Yes	Yes, in annual reports	Limited, in annual report	Yes	Yes	Yes	8
Res4-pop	Yes	Yes	Yes, upon request	Yes	Limited	-	Yes	Yes	Yes	7
Res5-pop	Yes	Yes	Yes, upon request	Yes	-	-	Yes	Yes	Yes	7
Res6-pop	Yes	Yes	Yes	-	Limited, in annual report	Limited, in annual report	Yes	-	Yes	5
Other1-pop	Yes	Yes	Yes	-	-	-	Yes	Yes	Yes	6
Other2-pop	Yes	Yes	-	-	-	-	-	-	-	2
Other3-pop	Yes	Yes	Limited, fields only	-	-	-	Yes	Yes	Yes	5
Other4-pop	Yes	-	-	-	-	-	-	-	-	1
Total (n = 25)	24	23	14	12	2	2	17	16	9	

Note. See Supplemental Table S3 for an audit of all information-categories (i.e. n = 16).

*Other screening services, industry associations+.

?Indicates that unclear information is available so it is not possible to make a determination.

Govt: government; admin: administrative; pop: population-health; Res: research organisation.

not to reuse it (Gregory et al., 2020). Data access issues have historically been one of the major barriers in the reuse of government data (Crusoe and Melin, 2018; Mc Grath-Lone et al., 2022). Almost 70% of datasets included in this study provided information on the access process on their websites and almost all contained details of the data custodian. There was one government administrative dataset that contained neither data custodian nor access process. Despite the absence of this information on the specific website of interest, Riley et al. (2022) had previously identified that this information asset had been utilised in more than 50 publications between 2008 and 2020, demonstrating that information on websites is not the only avenue through which researchers can access information. In their data discovery research, Liu et al. (2022: 3) identified that “social networks and word of mouth are the most used sources of discovering/collecting data by researchers.”

Our audit identified that almost 70% of the datasets provided information about privacy/confidentiality and/or security issues for release of data, regardless of whether the dataset in question was administrative or population-health focused. Most organisations indicated their data were covered under a range of ethical (e.g. National Statement on Ethical Conduct in Human Research) or legal (e.g. Privacy Principles, Health Records Act, Privacy and Data Protection Act) requirements, guidelines and principles.

Identification of participants/patients in provision of government health data to researchers is a major concern for governments in their role as data custodian (Office of the Australian Information Commissioner [OAIC], 2018). Many of the information asset websites provided assurance that only aggregated or de-identified data would be released to researchers, unless there were exceptions outlined in Australian Privacy Principle 6, “Use or disclosure of personal information” (OAIC, n.d.).

Knowledge and use of information asset documentation

Liu et al. (2022: 3) identified that researcher data discovery behaviour often involved exploration of data attributes (e.g. “measurement, level of granularity, quantity and coverage. . . suitability of data formats, and quality of metadata”) prior to accessing data. Mc Grath-Lone et al. (2022) identified the importance of *available* and *transparent* documentation for dataset “research-readiness.” Results reported from our study focused on contextual information, meta-data dictionary, data quality statements (information) and limitations documentation.

Contextual information. Understanding context is an essential requirement for data reuse as insight into the purpose for which something is collected impacts upon how it is interpreted (Pasquetto et al., 2019; Wang et al, 2021). This understanding is confirmed from our documentation audit where all but one of the datasets provided contextual information on the relevant websites.

Meta-data dictionary (meta-data). One of the underlying pillars of the FAIR principles (Wilkinson et al., 2016) is the

interoperability of data. Effective data reuse necessitates a standardised approach to data structure and definitions (Sinaci et al., 2013); this is the role of meta-data (Zhou et al., 2021). The document audit identified that almost 60% of datasets included a data dictionary on their website. While this demonstrates a significant proportion of agencies that identify the importance of meta-data in the promotion of data use, there is still some way to go before we reach a higher level of interoperability.

Data quality statement or information. Liu et al. (2022: 3) concluded that “data quality was the most critical data attribute [identified] by researchers in their data discovery efforts.” Our audit of within-scope information assets identified a large proportion (83%) of datasets that did not provide data quality information. From an analysis of the peer-reviewed publications of 28 government population-health information assets, Riley et al. (2022) discovered that 11 had published studies on the data quality of their information. The websites of these datasets were included in this document audit; however, none of these dataset websites referred to these data quality publications or provided links to the studies. Researcher knowledge of information asset data quality would be enhanced by the provision of links on the respective websites to these publications or by the provision of data quality statements/information, as recommended by the Victorian Public Service Information Management Framework (Victorian Government, 2020).

Data limitation documentation. Only one dataset provided information on the limitations of its data. The provision of this information is central to understanding how the data can be meaningfully reused for research (Mc Grath-Lone et al., 2022). Data provided for reuse purposes are significantly different from primary data, which are collected by a researcher for their own use (Sexton et al., 2017). Pasquetto et al. (2019: 2) posited the “data creators” advantage, specifically that those who create data have “intimate and tacit knowledge” that data reusers do not have. Either the data creator needs to collaborate with the data reuser to provide this knowledge, or appropriate documentation that outlines the strengths and limitations of the dataset needs to be provided so the reuser has an accurate understanding of the research possibilities in reuse of the dataset (Mc Grath-Lone et al., 2022).

Research transparency

Mc Grath-Lone et al (2022: 8) identified that “it is . . . important that there is a transparency in how the administrative data has been used in research.” Providing examples on how various datasets have been utilised in research not only provides guidance for researchers seeking to potentially use the data, but it may also prevent duplication and increase efficiency. If information is readily available on outputs that have been obtained from research requests, then, other researchers will not need to address the same issues or may directly connect with those who have already gathered the information. Other agency-curated datasets in our audit were transparent in providing their research

outputs. The government-curated datasets need to improve significantly in this area to match the other organisations.

Limitations

This study was limited to a selected number of information assets in the Australian state of Victoria. Between-state or between-country variations in outcomes may exist. The selection of pre-defined information-categories used to assess the level of detail and availability of information on dataset websites was limited to a subjective sample of topic areas that had been identified from the existing literature (Gilbert et al., 2018; Gordon et al., 2021; Mc Grath-Lone et al., 2022; Victorian Government, 2019a). Reviewer bias may exist in the extraction of data for information-categories from dataset websites; however, the kappa statistics demonstrated strong inter-rater agreement on the availability of eight of the nine information-categories. The presence of information-categories on a website does not imply evaluation of the usefulness of its content. This needs to be considered in any further evaluations of documentation provided for data reuse.

It is recommended that:

1. Data quality information is made readily available for high-value datasets to aid ease of reuse, including links to any published data quality articles.
2. Government-curated information assets should provide a publicly accessible list of research requests or outputs obtained from reuse of their data; and
3. Regular review of information asset website content be implemented to ensure the availability of up-to-date and meaningful documentation (i.e. trustworthy guidelines) to assist accessing and reusing data.

Conclusion

This study adopted an evidence-based approach to determine the extent of documentation (i.e. trustworthy guidelines) available on selected DoH information asset websites to support reuse of these data for research purposes. The findings have demonstrated inconsistency in the available website information, despite the provision of a recommended OVIC information asset register template. While contextual information, data custodian, access process, and privacy/confidentiality and security measures were found to be well to reasonably well reported, there was an overwhelming lack of information provided on the data quality or the limitations of the datasets for most websites and a lack of information on research requests or outputs for government-curated websites. Given that the *DataVic Access Policy Guidelines* recommend the use of a data quality statement (Victorian Government, 2019b), and have provided a data quality statement template, the lack of this information points to widespread omissions in appropriate data quality documentation provided for dataset reuse. There was also a lack of readily available information on the limitations of information assets to assist researchers in making correctly informed analyses of the data they use.

Acknowledgements

Thank you to Alan Riley for piloting the audit template and navigating the dataset website minefield.


Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: MFK received Future Leader Fellowship (105737) support from the National Heart Foundation of Australia. All other author (s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Merilyn Riley, GDipEpidBiostats  <https://orcid.org/0000-0003-4230-7062>

Monique F. Kilkenny, PhD  <https://orcid.org/0000-0002-3375-287X>

Kerin Robinson, PhD  <https://orcid.org/0000-0002-9037-6022>

Sandra G. Leggat, PhD  <https://orcid.org/0000-0002-2252-4302>

Supplemental material

Supplemental material for this article is available online.

References

- Andrew NE, Sundararajan V, Thrift AG, et al. (2016) Addressing the challenges of cross-jurisdictional data linkage between a national clinical quality registry and government-held health data. *Australian and New Zealand Journal of Public Health* 40(5): 436–442.
- Australian Government Department of Health and Aged Care (2022) Data strategy 2022–2025: Harnessing the power of data for better health, aged care and wellbeing. Available at: <https://www.health.gov.au/resources/publications/department-of-health-and-aged-care-data-strategy-2022-25?language=en> (accessed 19 April 2023).
- Berg M and Goorman E (1999) The contextual nature of medical information. *International Journal of Medical Informatics* 56(1–3):51–60.
- Bowen GA (2009) Document analysis as a qualitative research method. *Qualitative Research Journal* 9(2): 27–40.
- Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.
- Crusoe J and Melin U (2018) Investigating open government data barriers: A literature review and conceptualisation. In: *17th international conference on electronic government (EGOV)*, Krems, Austria, September 2019, pp. 169–183.
- DalGLISH SL, Khalid H and McMahon SA (2020) Document analysis in health policy research: The READ approach. *Health Policy and Planning* 35(10): 1424–1431.
- Dean AG, Sullivan KM and Soe MM (2014) OpenEpi Version 3.0.1: Open source epidemiologic statistics for public health. Available at: https://www.openepi.com/Menu/OE_Menu.htm (accessed April 2023).
- Gilbert R, Lafferty R, Hagger-Johnson G, et al. (2018) GUILD: GUIDance for Information about Linking Data sets. *Journal of Public Health* 40(1): 191–198.

- Gordon B, Barrett J, Fennessy C, et al. (2021) Development of a data utility framework to support effective health data curation. *BMJ Health & Care* 28: e100303.
- Gregory K, Groth P, Scharnhorst A, et al. (2020) Lost of Found? Discovering data needed for research. *Harvard Data Science Review* 2(2): 2–32.
- Huston P, Edge VL and Bernier E (2019) Reaping the benefits of Open Data in public health. *Canada Communicable Disease Report* 45(11): 252–256.
- Jahnke L and Asher A (2012) *The Problem of Data: Data Management and Curation Practices Among University Researchers*. Washington, DC: Council on Library and Information Resources. Available at: <https://www.clir.org/pubs/reports/pub154/> (accessed 28 April, 2023).
- Khan N, Thelwall M and Kousha K (2023) Data sharing and reuse practices: Disciplinary differences and improvements needed. *Online Information Review*. Epub ahead of print 7 February 2023. DOI: 10.1108/OIR-08-2021-0423.
- Kim Y (2022) A sequential route of data and document qualities, satisfaction and motivations on researchers' data reuse intentions. *Journal of Documentation* 78(3): 709–727.
- Kleinheksel AJ, Rockich-Winston N, Tawfik H, et al. (2020) Demystifying content analysis. *American Journal of Pharmaceutical Education* 84(1): 7113.
- Landis JR and Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174.
- Lee YW and Strong DM (2003) Knowing-why about data processes and data quality. *Journal of Management Information Systems* 20(3): 13–39.
- Liu YH, Wu M, Power M, et al. (2022) *Elicitation of data discovery contexts: An interview study* (1.0). Geneva, Switzerland: Zenodo.
- Mc Grath-Lone L, Jay MA, Blackburn R, et al. (2022) What makes administrative data “research-ready”? A systematic review and thematic analysis of published literature. *International Journal of Population Data Science* 7(1): 1718.
- Nikiforova A and McBride K (2021) Open government data portal usability: A user-centred usability analysis of 41 open government data portals. *Telematics and Informatics* 58: 101539.
- Office of the Australian Information Commissioner (OAIC) (2018) *De-identification and the privacy act*. Available at: <https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/handling-personal-information/de-identification-and-the-privacy-act> (accessed 12 April 2023).
- Office of the Australian Information Commissioner (OAIC) (n.d.) *Australian Privacy Principles quick reference. Chapter 6: Use or disclosure of personal information*. Available at: <https://www.oaic.gov.au/privacy/australian-privacy-principles/australian-privacy-principles-guidelines/chapter-6-app-6-use-or-disclosure-of-personal-information> (accessed 25 May 2023).
- Office of the Victorian Information Commissioner (2019a) *Victorian Protective Data Security Standards V2.0*. Available at: <https://ovic.vic.gov.au/information-security/standards/> (accessed 4 August 2023).
- Office of the Victorian Information Commissioner (2019b) *Practitioner Guide: Identifying and Managing Information Assets. Version 2.0*. Available at: <https://ovic.vic.gov.au/information-security/practitioner-guide-identifying-and-managing-information-assets/> (accessed 4 August 2023).
- Office of the Victorian Information Commissioner (2021) *Victorian Protective Data Security Standards. Version 2.0. Implementation Guidance V2.1*. Available at: <https://ovic.vic.gov.au/information-security/standards/> (accessed 4 August 2023).
- Open Data Charter (ODC) (n.d.) Our history. Available at: <https://opendatacharter.net/our-history/> (accessed 19 April 2023).
- Palamuthusingam D, Johnson DW, Hawley C, et al. (2019) Health data linkage research in Australia remains challenging. *International Medical Journal* 49(4): 539–544. Erratum in: *International Medical Journal* 2019; 49(9): 1195.
- Pasquetto IV, Borgman CL and Wofford MF (2019) Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review* 1(2): 2–32.
- Riley M, Robinson K, Kilkenny MF, et al. (2022) The suitability of government health information assets for secondary use in research: A fit-for-purpose analysis. *Health Information Management Journal*. Epub ahead of print 26 April 2022. DOI: 10.1177/18333583221078377.
- Sexton A, Shepherd E, Duke-Williams O, et al. (2017) A balance of trust in the use of government administrative data. *Archival Science* 17: 305–330.
- Sinaci AA and Laleci Erturkmen GB (2013) A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. *Journal of Biomedical Informatics* 46(5): 794–794.
- Tenopir C, Rice NM, Allard S, et al. (2020) Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLoS One* 15(3): e0229003.
- Victorian Government (2019a) Making data available. Available at: <https://www.data.vic.gov.au/datavic-access-policy-guidelines/making-data-available> (accessed 21 April 2023).
- Victorian Government (2019b) Preparing datasets before making them available. Available at: <https://www.data.vic.gov.au/datavic-access-policy-guidelines/preparing-datasets-making-them-available> (accessed 17 April 2023).
- Victorian Government (2020) Data quality guideline. Information management framework. Available at: <https://www.vic.gov.au/data-policies-and-standards#data-quality-guidelines> (accessed May 2023).
- Wang X, Duan Q and Liang M (2021) Understanding the process of data reuse: An extensive review. *Journal of the Association for Information Science and Technology* 72(9): 1161–1182.
- Wang P, Xe Y, L X, et al. (2023) Allocation of attention to metadata and retrieval functions: Implications for perceived value and open data discovery and reuse. *Journal of Librarianship and Information Science*. Epub ahead of print 24 August 2023. DOI: 10.1177/09610006231154529.
- Wilhelm E, Ballalai I, Belanger M, et al. (2023) Measuring the burden of infodemics: Summary of the methods and results of the Fifth WHO Infodemic Management Conference. *Journal of Medical Internet Research Infodemiology* 3: e44207.
- Wilkinson M, Dumontier M, Aalsberg I, et al. (2016) The FAIR principles for scientific data management and stewardship. *Scientific Data* 3: 160018
- Williamson K, Nimegeer A and Lean M (2022) Navigating data governance approvals to use routine health and social care data to evidence the hidden population with severe obesity: A case study from a clinical academic's perspective. *Journal of Research in Nursing* 27(7): 623–636.
- York J (2022) Seeking equilibrium in data reuse: A study of knowledge satisficing. PhD Thesis, University of Michigan, Michigan. Available at: <https://deepblue.lib.umich.edu/handle/2027.42/174439> (accessed 19 April 2023).
- Zhou H, Demartini G, Indulska M, et al. (2021) Evaluating the quality of repurposed data: The role of metadata. In: *ACIS 2021 Proceedings*, Sydney, Australia, p. 50.

Appendix

Table AI. Within-scope Victorian Department of Health information assets.

Datasets	Name in full
ACIR/AIR	Australian Childhood Immunisation Register/Australian Immunisation Register
ANZICS APD	Australian and New Zealand Intensive Care Society Adult Patient Database
ANZPIC	Australian and New Zealand Paediatric Intensive Care Register
AuSCR	Australian Stroke Clinical Registry
BSV	BreastScreen Victoria
CCOPMM	Consultative Council on Obstetric and Paediatric Mortality and Morbidity database
CSR	Cardiac Surgery Registry
ESIS	Elective Surgery Information System
HACC MDS	Home and Community Care Minimum Dataset
Life!	Life! Program Dataset
VAED	Victorian Admitted Episodes Dataset
VASM	Victorian Audit of Surgical Mortality
VCAR; VBDR;	Victorian Congenital Anomalies Register/Victorian Birth Defects Register
VCCAMM	Consultative Council on Anaesthetic Mortality and Morbidity Database. Replaced by Victorian Peri-Operative Consultative Council in July 2019
VCDC	Victorian Cost Data Collection
VCCR*	Victorian Cervical Cytology Registry Data Collection
VCOR	Victorian Cardiac Outcomes Registry
VCR	Victorian Cancer Register
VDI	Victorian Death Index
VEMD	Victorian Emergency Minimum Dataset
VES MDS	Victorian Eyecare Services Program Collection – Annual Returns
VHM*	Victorian Health Monitor
VPCR*	Victorian Psychiatric Register
VPDC	Victorian Perinatal Data Collection
VPHS	Victorian Population Health Survey
VRMDS	Victorian Radiotherapy Minimum Dataset
VSTR/VSTOR	Victorian State Trauma (Outcomes) Registry
VTP	Tuberculosis Undertakings Data Collection

*Excluded from audit.