

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Data dictionaries: essential tools for the ethical and transparent use of integrated data

Rebecca S. Pepe^{1,*} and Kristen Coe²

Submission History

Submitted:	05/12/2024
Accepted:	30/06/2025
Published:	13/10/2025

¹University of Pennsylvania,
School of Social Policy & Practice,
Philadelphia, 19104

²City of Philadelphia, Office of
Integrated Data for Evidence and
Action, Philadelphia 19102

Abstract

Data transparency lays the groundwork for the ethical use of administrative data. This is particularly true for linked administrative data within integrated data systems (IDS). Data dictionaries, resources that maintain the metadata of the information housed in an IDS, offer a tool to ensure transparency throughout the data life cycle. The FAIR Principles, which assert that data be Findable, Accessible, Interoperable, and Reusable provide a useful framework by which to measure the effectiveness of data dictionaries in the IDS context. This paper uses the FAIR Principles to discuss the ways in which data dictionaries serve as tools in the ethical and transparent use of integrated data as well as the challenges that remain. Linked administrative data is a valuable source of information for programmatic and academic research. Data dictionaries facilitate the ethical handling of this sensitive information and maintain a commitment to transparency in data inquiry and research.

Keywords

administrative data linkage; data integration; metadata; data dictionary; data transparency; data ethics; FAIR principles

*Corresponding Author:

Email Address: peper@upenn.edu (Rebecca S. Pepe)



Introduction

Administrative data are the information collected when government agencies perform routine provision of services and programming [1, 2]. Integrated data systems (IDS) link this individual-level data by a unique identifier or combination of identifiers, to produce a composite record that illustrates how people navigate social services over time [1, 3, 4]. This information can be repurposed to facilitate research and evaluation [1, 2, 5], but at a time when trust in the government is near historic lows [6], particular care must be taken to ensure this work is done ethically and for the good of the individuals represented in the data [5, 7].

Transparency in data collection and use is foundational to the ethical handling of integrated data [8, 9]. Data dictionaries maintain metadata on the information housed in an IDS and hold promise as agents of transparency and ethical data use [1, 10, 11]. Specifically, the FAIR Data Principles, which advocate that data be Findable, Accessible, Interoperable, and Reusable, provide a useful framework to understand how data dictionaries serve the goal of data transparency and the obstacles in realising their potential [12]. These principles set guideposts for good data management and stewardship which is a pre-condition to advancing innovation in social service research and programming [12, 13]. The current paper seeks to describe the benefits and challenges of data dictionaries as a tool to advance transparency in administrative data collection, integration, and use.

Benefits

The creation and maintenance of IDS data dictionaries foster transparency in data collection and use. Embodying the first FAIR principle of Findability [12], data dictionaries serve as a menu of potential data elements from which government staff and external researchers, with the proper permissions and authorisations, can make selections for use in analytic inquiries. Like the sections of a restaurant menu, findability is improved with a well-organised dictionary. Even while the data elements remain secure and subject to request procedures, publishing the resource on a website or open data portal further advances findability. Clear, richly defined metadata that is broadly accessible democratises who can request data for analysis. The tool levels the playing field, ensuring that all parties have equitable access to information held by the IDS.

Furthering the FAIR principle of Accessibility [12], data dictionaries foster the ability to discern exactly what information is available and facilitate specificity in data requests. Knowing which data elements exist in a file or database and the identifiers of those data elements allows the researcher to request just those data elements. Data from governmental agencies are often stored in several linked files. While a natural tendency may be for researchers to request the entire file (e.g., the demographics file), the provision of a data dictionary allows the researcher to specify exactly which components are required (e.g., race, gender), preventing the unnecessary disclosure of additional information (e.g., date of birth, country of origin).

For systems that house data covered by the Health Insurance Portability and Accountability Act (HIPAA),

narrowing requests to the exact elements necessary for a specific inquiry allows the IDS to maintain HIPAA's minimum necessary Privacy Rule which requires a covered entity to "make reasonable efforts to limit use, disclosure of, and requests for protected health information to the minimum necessary to accomplish the intended purpose" [14]. This guideline offers a useful rule of thumb for integrated administrative data in general, regardless of whether its data is governed by HIPAA as it maximises protections for sensitive, identifiable data [15, 16].

Metadata can prevent unnecessary data requests. For instance, location data may be available in three forms: address, XY coordinates, and census tract. Providing researchers with this information at the outset allows them to identify which level of specificity is necessary for their analysis. If the more anonymised census tract is sufficient, sharing the exact address or XY coordinates can be avoided. When government and research oversight bodies are confident that researchers are only requesting the data elements essential for their authorised purpose, it becomes easier to approve data requests. With improved specificity and resultant efficiency, more data requests can be reviewed and authorised.

This type of transparency also serves to ensure data quality and assurance. The development and maintenance of IDS data dictionaries necessitate cross-discipline discussion of what data are being collected and how they are being translated into the IDS. Their updating and maintenance offer a mechanism by which agencies can revisit decisions about data sharing and access. Data elements of low quality, limited capture, or high sensitivity may be considered for editing or removal. This allows for the investigation of why data capture is suboptimal and creates opportunities to update coding schemas about programs and services that may change over time. For example, a social service agency may have a legacy intake form that asks clients a question that most staff find inappropriate and outdated. As a result, most staff encourage clients to skip the question. The data dictionary might then note a 10% completion rate for that data element. By examining the metadata, the agency could become aware of the issue and consider revising the legacy form to address staff concerns about the inappropriate question. Researchers will also know not to request data elements like these because of their insurmountable missing data problems.

Having a data dictionary improves the efficiency and accuracy of data requests and analytic projects. Because they offer research teams a guide to what information is available, data dictionaries facilitate the early discussion of construct operationalisation, use of proxies, and data quality prior to project initiation. They provide standardised definitions of data elements that support well conceptualised inquiries with a high probability of successful execution. In addition to documenting the format of raw data (i.e., data as collected), some data dictionaries also include frequently requested transformations of data elements. For example, an infant's expected due date and date of birth might be transformed into a prematurity indicator. By documenting the method of transformation in the data dictionary and offering the resulting transformed data to all researchers, the IDS can promote scientific replicability and the Reusability tenet of the FAIR Principles. Reusing data extraction code instead of creating custom variations of similar

indicators for each researcher has the added benefit of creating efficiency in governmental analytic offices.

Challenges

This level of transparency is not without challenges. One challenge in the development of data dictionaries is the reliance on collaboration among cross-disciplinary staff. It requires program staff who are familiar with the services offered by municipalities and can explain how and why data are captured at the front line. Technical staff on the other hand, are tasked with determining how data are loaded, stored, and maintained in the IDS. Then, analytic staff must navigate which data elements to use to conceptualise and answer research inquiries. While each of these roles are essential to the success of IDS and associated data projects, they each have unique training, points of view, and terminology. This is further nuanced by the variation in their subject matter expertise (e.g., health, child welfare, justice). Collaborators often occupy different levels and sit in different areas of the organisation, adding complexity to the FAIR goal of Interoperability [12]. For data to be Interoperable, as defined by the FAIR Principles, it is critical that the metadata of data dictionaries utilise shared, accessible language. One technique that can attenuate this challenge is to include multiple data element names in the resource. Specifically, data dictionaries should include the names of data elements that are intuitive to program staff as well as technical staff. The inclusion of a crosswalk of both what the data element is called in the IDS as well as how it would be known to a data requester, ensures that all parties have a shared understanding of what information is being referenced.

A second challenge to ensuring interoperability in the development of data dictionaries is defining who its audience will be. A resource created for program staff seeking answers to research questions about the services their organisation offers will differ from one made for technical staff to inform the building and maintenance of the IDS. Similarly, there is a tension between a desire to create a resource that is comprehensive in its scope versus one that prioritises readability and ease of use. There is a natural tendency to want to include every detail about each data element, but this often comes at the cost of a dynamic resource that will be actively utilised. The FAIR principle of Reusability can be helpful here [12]. When determining whether to maintain additional information about an IDS, one should ask whether it is necessary to replicate previous empirical work. One may consider whether it is critical to capture how a data element changes over time. For example, gender may have historically been captured as a binary indicator but is now a categorical variable with several possible responses. For researchers to measure gender trends over time, it is critical to capture these modifications as they occur. However, the challenge remains that this information is often revealed iteratively through the process of resource development and updating. It can be difficult to anticipate what information will be critical to maintain and to develop a comprehensive system of documentation at the outset. One helpful approach to this challenge is to maintain supplementary lists of details to be documented or decisions to be revisited when reconciling data and finalising the resource.

A large body of work advocates the integral role of community voice in data collection and use [10, 17, 18]. Those individuals who are represented in the data should have a say in how their data is captured, shared, and used [10, 18]. While there is public support of government data used to identify services and supports for clients [10], there is little evidence of support for the sharing of municipal data across agencies which is required for these improved processes to take place. This challenging duality underscores the need to engage with communities to lead decision-making efforts in how data is collected by source agencies and used through the IDS. Data dictionaries offer a tool to inform this conversation and facilitate data sovereignty.

Examples

In their 2020 work “Data Feminism,” Catherine D’Ignazio and Lauren F. Klein underscore the importance of metadata, explaining that the data in a government dictionary “do not look very technically complicated. The complicated part is figuring out how the business process behind them works” [19]. They explain that the consequence of insufficient metadata is analysis void of context and thus prone to misinterpretation.

This risk intensifies when adding the complexity of integrated data from numerous distinct sources. If available at all, it can be common to receive a data dictionary like the sample in Figure 1. While this can appear sufficient at face value, integrating the FAIR principles produces a data dictionary more like Figure 2. From this view we understand what data is available for extraction, where the data is coming from, how it is defined, and how the definitions have changed over time. The same data element (e.g., date of birth) could come from any number of sources. The source used by a researcher can have implications for data quality and accuracy. Further, an *integrated* data system will typically offer a version of that data element created from all sources. This form is preferred for most applications because it prioritises the most common or reliable sources when available (see row 7 of Figure 2). By understanding the different versions available for a specific data element, data requestors and data stewards are better equipped to discuss the best element for a particular application.

Figure 1: Sample data dictionary with limited documentation of data elements.

Data Element	Values
Gender	M: Male
	F: Female
	N: Nonbinary
	O: Other
Country of Origin	Free Text
Date of Outreach	DDMMYYYY
Date of Death	DDMMYYYY
Cause of Death	ICD-9 and ICD-10 codes
Date of Birth	DDMMYYYY
Age at Death	Integer

Figure 2: Sample data dictionary with comprehensive documentation of IDS data elements.

	Data Element	IDS Name	Source Agency	Source Name	Notes	Values
1	Gender	gender_ohs	OHS	sex	The gender the client self-identifies at intake. "Nonbinary" and "Other" options were introduced on 01/01/2019.	M: Male F: Female N: Nonbinary O: Other
2	Country of Origin	country_ohs	OHS	imm_origin	Not available to external researchers	Free Text
3	Date of Outreach	outreach_date_ohs	OHS	service_date	The date service was provided.	DDMMYYYY
4	Date of Death	death_date	MEO	dod_ddmmyy	Note there is a 30+ day lag between the time of death occurs and receipt of data on death record	DDMMYYYY
5	Cause of Death	death_cause	MEO	cod	ICD-9 is used for underlying cause of death for deaths in years 1979-2015; ICD-10 is used for deaths in years 2015-present. Missing in >10% of incidence.	See pdf entitled "Underlying and Multiple Causes of Death Codes"
6	Date of Birth	dob_meo	MEO	dob_ddmmyy	Date of birth as recorded on the death record. Missing in 5% of cases.	DDMMYYYY
7	Date of Birth	dob_ids	IDS	--	The most frequently found date of birth for an individual across all data systems within the IDS. Preferred for many applications to reduce missing data and data entry errors. See technical handbook for details.	DDMMYYYY
8	Age at Death	death_age	IDS	--	The IDS calculates age at death by finding the difference between birthdate and date of death both of which come from the Medical Examiner's Office (MEO) data file. If the date of birth is not available in the MEO file but is available in the IDS, the IDS date of birth will be used. To protect privacy, Age at Death is the preferred field to request if date of birth is not otherwise required.	Integer
Abbreviations: IDS: Integrated Data System OHS: Office of Housing Services MEO: Medical Examiner's Office						

In practice, this information can be displayed in a variety of ways. Technical manuals and tables are two common formats. Some jurisdictions have developed more dynamic and interactive ways to share similar information. For example, in the Allegheny County Data Warehouse's publicly available QuickCount data, the Department of Human Services has created a series of metadata pop-ups that appear when the user hovers their cursor over data elements [20]. Individuals can click on these pop-ups to navigate to a "Program Definitions" page with even more detail. The Program Definitions page includes sections such as "explanation of results" and "what else should I know about this data," which detail data availability, reporting lags, and other relevant information [21]. The clear, intuitive, and up-to-date information provided in systems like these prevent the data from "idling" on their portals, awaiting users to undertake the intensive work of deciphering the bureaucratic arcana that obscures their significance" [19].

Conclusion

Data dictionaries are critical tools that ensure transparency in data collection and use. It is through these resources that IDS can promote equity, quality, and good data governance via FAIR Principles. As seen throughout the field of integrated population data, the challenges that persist demonstrate that the work is not only technical but also relational. Data

dictionaries offer just one tool to support these conversations and efforts to build public trust in the use of administrative data to better serve clients and advance the field of social welfare research.

Conflict of interests statement

None declared.

Ethics statement

This work did not require ethics approval because it did not involve human participants or the use of personal data.

Data availability statement

No datasets were used in the preparation of this manuscript.

References

1. Hawn Nelson A, Jenkins D, Zanti S, Katz M, Burnett T, Culhane D, et al. Introduction to Data Sharing and Integration. Actionable Intelligence for Social Policy: University of Pennsylvania; 2020

- May. Available from: <https://aisp.upenn.edu/wp-content/uploads/2020/06/AISP-Intro-.pdf>.
2. Jutte DP, Roos LL, Brownell MD. Administrative Record Linkage as a Tool for Public Health Research. *Annual Review of Public Health*. 2011;32:91–108. Available from: <https://doi.org/10.1146/annurev-publhealth-031210-100700>.
3. Dunn HL. Record Linkage. *Am J Public Health Nations Health*. 1946;36:1412–6.
4. Fantuzzo J, Henderson C, Coe K, Culhane D. The Integrated Data System Approach: 2017; Available from: <https://repository.upenn.edu/handle/20.500.14332/59953>.
5. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. *Big Data & Society*. 2017;4:2. Available from: <https://doi.org/10.1177/2053951717745678>.
6. Bell P. Public Trust in Government: 1958–2024 [Internet]. Pew Research Center. 2024 [cited 2024 Nov 14]. Available from: <https://www.pewresearch.org/politics/2024/06/24/public-trust-in-government-1958-2024/>.
7. Hawn Nelson A, Zanti S. Four Questions to Guide Decision-Making for Data Sharing and Integration. *International Journal of Population Data Science* [Internet]. 2023 [cited 2024 Nov 14];8. Available from: <https://doi.org/10.23889/ijpds.v8i4.2159>.
8. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. *Journal of Biomedical Informatics*. 2012;45:165–72. Available from: <https://doi.org/10.1016/j.jbi.2011.10.006>.
9. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang L-C, Smith P, et al. GUILD: GUIDance for Information about Linking Data sets†. *Journal of Public Health*. 2018;40:191–8. Available from: <https://doi.org/10.1093/pubmed/idx037>.
10. Hawn Nelson AL, Zanti S. A framework for centering racial equity throughout the administrative data life cycle. *Int J Popul Data Sci*. 2020;3:1367. Available from: <https://doi.org/10.23889/ijpds.v5i3.1367>.
11. Lawrence R, Barker K. Integrating Data Sources Using a Standardized Global Dictionary. In: Abramowicz W, Zurada J, editors. *Knowledge Discovery for Business Information Systems* [Internet]. Boston, MA: Springer US; 2002 [cited 2024 Nov 29]. p. 153–72. Available from: https://doi.org/10.1007/0-306-46991-X_7.
12. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3. Available from: <https://doi.org/10.1038/sdata.2016.18>.
13. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: The roles of common data elements and harmonization. *Journal of Biomedical Informatics*. 2020;107. Available from: <https://doi.org/10.1016/j.jbi.2020.103421>.
14. US Department of Health and Human Services. Minimum Necessary FAQs [Internet]. Health Information Privacy. [cited 2024 Nov 17]. Available from: <https://www.hhs.gov/hipaa/for-professionals/faq/minimum-necessary/index.html>.
15. Agris JL. Extending the Minimum Necessary Standard to Uses and Disclosures for Treatment: Currents in Contemporary Bioethics. *Journal of Law, Medicine & Ethics*. 2014;42:263–7. Available from: <https://doi.org/10.1111/jlme.12140>.
16. Deitch J. Protecting Unprotected Data in Mhealth. *Northwestern Journal of Technology and Intellectual Property*. 2020;18:107–28. Available from: <https://scholarlycommons.law.northwestern.edu/njtip/vol18/iss1/4>.
17. Bensenor IM, Goulart AC, Thomas GN, Lip GYH, Management on behalf of the NHRG on AF. Patient and Public Involvement and Engagement (PPIE): first steps in the process of the engagement in research projects in Brazil. *Brazilian Journal of Medical and Biological Research*. 2022;55:1–3. Available from: <https://doi.org/10.1590/1414-431X2022e12369>.
18. Lewis T, Gangadharan SP, Saba M, Petty T. Digital Defense Playbook: Community Power Tools for Reclaiming Data [Internet]. *Our Data Bodies*; 2018 [cited 2023 Oct 16]. Available from: https://www.odbproject.org/wp-content/uploads/2019/03/ODB_DDP_HighRes_Single.pdf.
19. D'Ignazio C, Klein LF. *Data Feminism*. Cambridge: The MIT Press; 2020. Available from: <https://doi.org/10.7551/mitpress/11805.001.0001>.
20. Allegheny County Department of Human Services. Allegheny County Data Warehouse. 2024. Available from: <https://alleghenycountyanalytics.us/2024/02/07/allegheny-county-data-warehouse/>.
21. QuickCount Data Tool [Internet]. Allegheny Analytics. [cited 2025 Jun 16]. Available from: <https://www.alleghenycountyanalytics.us/quickcount-data-tool/>.

Abbreviations

IDS:	Integrated Data System
FAIR:	Findable, Accessible, Interoperable, Reusable
HIPAA:	Health Insurance Portability and Accountability Act