Data Quality Guidelines

Document No: CDO-001
Updated: July 22, 2024
Issued by: Chief Data Officer

1.0 Purpose

The purpose of the Data Quality Guidelines is to establish common guidelines for data quality management across State of Hawaii agencies. Through effective data quality management, state agencies can promote trust, improve operational efficiency, and enable better service to the citizens of Hawaii with accurate and timely information.

2.0 Authority

Hawaii Revised Statutes (HRS)[1] Section §27-44, established within the Office of Enterprise Technology Services, in the Department of Accounting and General Services, and under the authority of the Chief Information officer, the Chief Data Officer to develop, implement, and manage statewide data policies, procedures, standards, and guidelines. HRS §27-44 also established a Data Task Force to assist the Chief Data Officer in developing the State's data policies, procedures, and standards.

3.0 Scope

3.1. State Agencies

The Data Quality Guidelines apply to all state agencies.

The Data Quality Guidelines provide high level guidelines on data quality. Each agency shall develop additional policies and guidelines as necessary according to relevant federal and state laws and regulations, both at the data set level and at the individual field level, to ensure compliance in its operations. When a conflict exists between the Data Quality guidelines and an agency's policy, the more restrictive policy will take precedence.

3.2 Definitions

[1] HRS §27-44. https://www.capitol.hawaii.gov/hrscurrent/Vol01_Ch0001-0042F/HRS0027/HRS_0027-0044.htm

As developed by the Federal Committee on Statistical Methodology (FCSM) and informed by the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) and other sources, data quality is the degree to which data capture the desired information using appropriate methodology in a manner that sustains public trust.

Data Quality Management (DQM) refers to the set of practices, processes, and tools used to ensure that an organization's data is accurate, complete, reliable, and consistent over time. DQM includes continuous monitoring, data validation, issue detection, data correction, and improvement efforts to maintain high-quality data throughout its life cycle.

3.3 Covered Use

The Data Quality Guidelines apply to handling of data in all data sets managed by state agencies. This includes, but is not limited to systems in the cloud, on premises, and/or on local drives.

The Data Quality Guidelines shall be applied to the entire data life cycle from data creation, data collection, data cleansing and transformation, data storage and modeling, data science and analytics, data visualization, impact tracking, to data retention. The Data Quality Guidelines shall also be applied to all data applications including Machine Learning[2] and Artificial Intelligence.[3]

The Data Quality Guidelines are created with reference to the following international and national data quality frameworks and guidelines:

- ISO/IEC 25012 data quality standards[4]
- Office of Management and Budget guidelines[5]
- Data Quality Assessment Framework of International Monetary Fund[6]
- DAMA International common practice[7]
- Relevant state and federal standards
- Research on key data quality dimensions

4.0 Information Statement

[2] Refer to 7.0 Definitions of Key Terms.

[3] Refer to 7.0 Definitions of Key Terms.

[4] ISO25012. https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

[5] OMB Section 515 Information Quality Guidelines. https://www.govinfo.gov/content/pkg/FR-2002-02-22/pdf/R2-59.pdf

[6] UN Data Quality Assessment Framework. https://unstats.un.org/unsd/methodology/dataquality/

[7] Data Management Body of Knowledge (DAMA-DMBoK). https://www.dama.org/cpages/home

4.1 General Data Quality Guidelines

- **Accuracy:**

Data Accuracy refers to the correctness, truthfulness, and reliability of data. Data values shall be correct and free from errors, especially those that occur due to incorrect data entry or faulty processes. Data shall correctly represent the real-world scenario or event it is supposed to depict.[8, 9]

- **Completeness:**

Data Completeness refers to the extent that all required data is present in a data set with no missing values and accessible for meaningful analysis. The data shall include all information necessary for the intended analysis or operation.[10,11]

- **Uniqueness:**

Data Uniqueness refers to the principle that each record in a data set shall be distinct and not duplicated. No two records shall be identical in all their fields, particularly in key fields that uniquely identify each record.[12]

- **Timeliness:**

Timeliness of Data refers to how up-to-date and available the data is when it is needed for decision-making, analysis, and operational processes. Timely data is crucial for making informed decisions based on the most current and relevant information.[12]

- **Consistency:**

Data Consistency refers to the uniformity of data across a system or among different systems. It ensures that the same piece of information is represented identically wherever it appears, maintaining integrity and reliability. It includes methods such as utilizing standardized formats, definitions, and coding structures.[10 ,11]

- **Validity:**

Data Validity refers to the extent that data confirms to the expected formats, rules, and constraints. It ensures that data is usable and meaningful within the intended context. Departments shall verify that data values fall within expected ranges and

[8] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, "Methodologies for data quality assessment and improvement," ACM Computing Surveys (CSUR), vol. 41, p. 16, 2009.

[9] D. McGilvray, Executing data quality projects: Ten steps to quality data and trusted information: Morgan Kaufmann, 2008

[10] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," Communications of the ACM, vol. 39, pp. 86-95, 1996.

[11] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," Journal of management information systems, vol. 12, pp. 5-33, 1996

[12] D. McGilvray, Executing data quality projects: Ten steps to quality data and trusted information: Morgan Kaufmann, 2008.

comply with defined formats to maintain validity.[9,13]

Each dimension has its associated risks and risk remediation plans, with interconnections with other data guidelines such as privacy, equity, open data, classification, and cataloging (under Polices of https://data.hawaii.gov/).

Table 1: Data Quality Dimensions, Risks, Remediation Plans, Interdependencies, and Measurements

Dimension	Risks	Remediation Plans	Interdependencies	Measurement to consider
Accuracy	Flawed analyses, misinformed decisions, damaged credibility	- Validation processes (verification, cross-referencing) - Data quality audits & reviews	Accurate data minimizes privacy breaches (protects personal details), avoids biased outcomes (ensures fair assessment for all), and allows for informed public participation (transparency). It also streamlines classification (accurate categorization).	- Error rate: percentage of incorrect data entries - Discrepancy rate: frequency of data mismatches identified during audits
Completeness	Skewed analyses, inaccurate reporting, missed opportunities	- Capture all data elements (validation checks) - Regular data audits & supplement missing data	Protects privacy (reduces risk of exposing partial information in breaches), Equity (ensures fair assessment for all), Open Data (offers transparent resources for public analysis). Complete data sets allow for a more holistic view, reducing privacy risks, and enable better classification (comprehensive	- Percentage of missing data elements - Rate of completion: proportion of records with all required fields populated

[13] L. L. Pipino, Y.W. Lee, R.Y. Wang, "Data quality assessment," Communications of the ACM, vol. 45, pp. 211-218, 2002.

			picture for accurate categorization) and cataloging (all information available for retrieval).	
Uniqueness	Inflated metrics, inaccurate reporting, wasted resources	- Data cleansing strategies (deduplication) - Preventive measures to minimize duplicates	Managing duplicates ensures fair representation in open data (avoids skewed results) and facilitates efficient classification (reduces redundancy) and cataloging (eliminates unnecessary entries).	- Duplicate rate: ratio of duplicate records to total records - Deduplication effectiveness: percentage reduction in duplicates after cleansing
Timeliness	Missed opportunities, inaccurate analyses, obsolete insights	- Regular update schedules & efficient procedures - Automated notifications & prioritize timely updates	Timely data updates are crucial for maintaining privacy (avoids outdated information exposure), promoting equity (ensures everyone has access to the latest information for fair assessment), and supporting open data initiatives (provides users with current data for analysis). It also supports classification (uses most recent data for categorization) and cataloging (ensures retrieval of the latest information).	- Data latency: average time lag between data collection and availability - Update frequency: rate at which data is refreshed according to the schedule
Consistency	Integration challenges, misinterpretation, errors	- Standardize formats, definitions,	Consistent data guidelines safeguard privacy (reduces misinterpretation	- Inconsistency rate: number of inconsistencies detected

		coding structures - Data governance policies & training	and potential misuse), promote equity (ensures everyone uses data consistently for fair assessment), and strengthen open data (improves data clarity for public analysis). Consistent formats also facilitate data integration across systems for classification and cataloging, enabling efficient organization and retrieval.	- Adherence rate: compliance with predefined data guidelines and formats
Validity	Erroneous analyses, incorrect conclusions, compromised decision-making	- Robust validation checks (formats, data types, rules) - Regular data quality assessments & rectify issues	Safeguards against unauthorized access/misuse (Privacy). Valid data upholds the integrity of open data initiatives and ensures data adheres to established guidelines, minimizing privacy risks. It also supports effective classification (avoids errors in categorization) and cataloging (ensures data is retrievable based on valid formats and types).	- Invalid entry rate: percentage of data not meeting format or range specifications - Validation success rate: rate of successful validation checks

4.2 Additional Geospatial Data Quality Elements

In addition to the general data quality principles outlined in Section 3.1, geospatial data quality encompasses specific considerations unique to its location-based nature. Here are some key aspects to consider for geospatial data quality:

- **Positional Accuracy**: Positional Accuracy refers to how closely the geospatial data represents the actual location on the Earth's surface. Factors like data collection methods, coordinate systems used, and resolution all influence positional accuracy.[14]
- **Reference System Consistency**: Consistency in the reference system ensures that all data adheres to a well-defined and uniform standard. This provides a common frame of reference for all geospatial elements, enabling accurate integration and analysis across data sets.[15]
- **Geospatial Completeness**: Geospatial completeness refers to the extent to which geospatial data is comprehensive and thorough. While completeness is generally important for all data types, in geospatial data, it extends beyond just attribute information. It also encompasses the completeness of geographic features themselves, ensuring that all relevant features are present without any gaps or missing sections.[16]
- **Topological Consistency**: Topological consistency refers to the principle that ensures the spatial relationships between features are logically correct.[17]
- **Lineage**: Lineage refers to tracking the origin, processing steps, and transformations undergone by geospatial data. Understanding lineage allows users to assess the pedigree and potential biases introduced during data creation.[18]
- **Scale:** Scale refers to the level of detail and extent of coverage represented by the data. The scale of the data (e.g., 1:24,000) shall be clearly documented and appropriate for the intended use.[19]

5.0 Compliance

The Data Quality Guidelines shall take effect upon publication. The Chief Data Office may amend at any time; compliance with guidelines is strongly recommended.

6.0 Contact Information

[14] Federal Geographic Data Committee (FGDC). (1998). Content standard for digital geospatial data. https://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf

[15] Open Geospatial Consortium (OGC). (2019). Reference systems for OGC geospatial standards. https://www.ogc.org/

[16] Goodchild, M. F. (2011). Geographic information science and systems (4th ed.). Springer. ISBN 978-3-642-16771-1. Chapter 3.

[17] National Center for Geographic Information & Analysis (NCGIA). (2012). Core curriculum for geographic information science. https://umaine.edu/scis/ncgia/

[18] International Standards Organization (ISO). (2015). Geographic information - Lineage (ISO 19115:2003).

[19] Muehrcke, P. C., Muehrcke, J., & Sh Muehrcke, D. (2004). Map use, reading, analysis, and interpretation (6th ed.). JP Publications. ISBN 978-0-7668-2779-7. Chapter 2.

Submit all inquiries and requests for future enhancements to the Chief Data Officer in the Office of Enterprise Technology Services, Department of Accounting and General Services, at data@hawaii.gov.

Additional data related policies and guidelines documents can be found at data.hawaii.gov.

7.0 Definitions of Key Terms

All terms shall have the meanings found in the Data & AI Glossary (under Glossaries on https://data.hawaii.gov/).

- **Data Quality:** Data Quality refers to the degree to which data capture the desired information using appropriate methodology in a manner that sustains public trust.[20]
- **Data Classification**: Data classification refers to the process of categorizing data based on its sensitivity.[21]
- **Risks**: Risks refer to the extent to which an entity is threatened by a potential circumstance or event, and typically a function of: (i) the adverse impacts that would arise if the circumstance or event occurs; and (ii) the likelihood of occurrence[22].
- **Remediation**: Remediation refers to the neutralization or elimination of a vulnerability or the likelihood of its exploitation.[21]
- **Machine Learning (ML):** Machine learning (ML) is a field within artificial intelligence. ML focuses on the ability of computers to learn from provided data without being explicitly programmed for a particular task.[23]
- **Artificial Intelligence (AI):** A branch of computer science devotes to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement.[24]

8.0 Revision History

Date	Description of Change
July 22, 2024	Published

9.0 Related Documents

[20] Information Quality Act. https://www.govinfo.gov/content/pkg/PLAW-106publ554/html/PLAW-106publ554.htm

[21] Information Systems Audit and Control Association (ISACA) Glossary. https://www.isaca.org/resources/glossary

[22] National Institute of Standards and Technology Glossary. https://csrc.nist.gov/glossary

[23] National Institute of Standards and Technology. https://www.nccoe.nist.gov/ai/adversarial-machine-learning

[24] U.S. Department of State. https://www.state.gov/artificial-intelligence/#:~:text=Artificial%20Intelligence%20and%20Society&text=%E2%80%9CThe%20term%20'artificial%20intelligence',influencing%20real%20or%20virtual%20environments.%E2%80%9D

[1] CDC-Early Hearing Detection and Intervention (EHDI). https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf

[2] Federal Committee on Statistical Methodology (FCCM). https://nces.ed.gov/FCSM/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf

[3] Department of Housing & Urban Development (HUD). https://www.hud.gov/sites/dfiles/CFO/documents/FY23_HUDDATAActDQP_7.10.2023-signed.pdf

[4] Institute of Education Sciences(IES). https://ies.ed.gov/ncee/edlabs/regions/central/resources/pemtoolkit/pdf/module-5/CE5.3.2-Data-Quality-Dimensions.pdf

[5] United States Agency for International Development (USAID). https://pdf.usaid.gov/pdf_docs/Pnadw112.pdf

[6] United States Department of Agriculture (USDA). https://www.ers.usda.gov/about-ers/policies-and-standards/data-product-quality/ers-data-product-quality-standards/

[7] Arkansas Department of Human Services (DHS). https://humanservices.arkansas.gov/wp-content/uploads/748-Exhibit-26-DCFS-Data-Quality-Plan_V3.3.pdf

[8] California Department of Health Care Services (DHCS). https://www.dhcs.ca.gov/Documents/CSD_YV/Family%20Services/PPSDS-Pv-Data-Quality-Standards-Jan-2021.pdf

[9] New Jersey Department of Environmental Protection (DEP). https://www.nj.gov/dep/srp/guidance/srra/data_qual_assess_guidance.pdf

[10] New Mexico Interagency Data Governance Council (DGC). https://webapp.hsd.state.nm.us/Procurement/docs/Data%20Services%20Info/Data%20Quality%20Plan/Data%20Quality%20Planv1.1.pdf

[11] Utah Homeless Management Information System (UHMIS). https://www.utah.gov/pmn/files/519855.pdf

[12] Ardagna, Danilo, Cinzia Cappiello, Walter Samá, and Monica Vitali. 2018. “Context-Aware Data Quality Assessment For Big Data.” Future Generation Computer Systems 89 (December): 548–62.

[13] Kugler, Zs., Gy. Szabó, H. M. Abdulmuttalib, C. Batini, H. Shen, A. Barsi, and G. Huang. 2018. “Time-Related Quality Dimensions of Urban Remotely Sensed Big Data.” International Archives

of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives 42 (4): 383–87.

[14] Abdellaoui, Sabrina, Ladjel Bellatreche, and Fahima Nader. 2016. “A Quality-Driven Approach for Building Heterogeneous Distributed Databases: The Case of Data Warehouses.” In Proceedings - 2016 16th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2016, 631–38. CCGRID ’16. Cartagena, Columbia: IEEE.

[15] Li, Zhe, Sai Wu, Hongwei Zhou, Sheng Zou, and Tingting Dong. 2019. “An Overview on Quality Evaluation Constitution in Context of Big Data Application.” Journal of Physics: Conference Series 1302 (2): 1–6.

[16] Kulkarni, Anuja. 2016. “A Study on Metadata Management and Quality Evaluation in Big Data Management.” International Journal for Research in Applied Science & Engineering Technology (IJRASET) 4 (VII): 455–59.

[17] Onyeabor, Grace Amina, and Azman Ta’a. 2019. “A Model for Addressing Quality Issues in Big Data.” In Recent Trends in Data Science and Soft Computing, edited by Faisal Saeed, Nadhmi Gazem, Fathey Mohammed, and Abdelsalam Busalim, 843:65–73. Kuala Lumpur, Malaysia: Springer.

[18] Radhakrishnan, Asha, and Sarasij Das. 2018. “Quality Assessment of Smart Grid Data.” In 2018 20th National Power Systems Conference (NPSC), 1–6. Tiruchirappalli, India: IEEE.

[19] Ya, Li, Song Heliang, and Xu Yingcheng. 2020. “Studies on Data Quality Evaluation Index System for Internet Plus Government Services in Big Data Era.” Journal of Physics: Conference Series 1584 (1): 012014.

[20] Arolfo, Franco, and Alejandro Vaisman. 2018. “Data Quality in a Big Data Context.” In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), edited by András Benczúr, Bernhard Thalheim, and Tomáš Horváth, 11019 LNCS:159–72. Lecture Notes in Computer Science. Cham: Springer International Publishing.

[21] Ehrlinger, Lisa, and Wolfram Wöß. 2017. “Automated Data Quality Monitoring.” In 22nd MIT International Conference on Information Quality, 1–9. Little Rock, Arkansas, USA.

[22] Taleb, Ikbal, Hadeel T.El E Kassabi, Mohamed Adel Serhani, Rachida Dssouli, and Chafik Bouhaddioui. 2017. “Big Data Quality: A Quality Dimensions Evaluation.” Proceedings - 13th IEEE International Conference on Ubiquitous Intelligence and Computing, 13th IEEE International Conference on Advanced and Trusted Computing, 16th IEEE International Conference on Scalable Computing and Communications, IEEE International, no. February 2018: 759–65.

[23] Talha, Mohamed, Nabil Elmarzouqi, and Anas Abou El Kalam. 2020. “Towards A Powerful Solution for Data Accuracy Assessment in The Big Data Context.” International Journal of Advanced Computer Science and Applications 11 (2): 419–29.

[24] Zhou, Ningning, Guofang Huang, and Suyang Zhong. 2018. “Big Data Validity Evaluation Based on MMTD.” Mathematical Problems in Engineering 2018 (June): 1–6.

[25] Federal Geographic Data Committee (FGDC). (1998). Content standard for digital geospatial data. https://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf

[26] Open Geospatial Consortium (OGC). (2019). Reference systems for OGC geospatial standards. https://www.ogc.org/

[27] Goodchild, M. F. (2011). Geographic information science and systems (4th ed.). Springer. ISBN 978-3-642-16771-1. Chapter 3.

[28] National Center for Geographic Information & Analysis (NCGIA). (2012). Core curriculum for geographic information science. https://umaine.edu/scis/ncgia/

[29] International Standards Organization (ISO). (2015). Geographic information - Lineage (ISO 19115:2003).

[30] Muehrcke, P. C., Muehrcke, J., & Sh Muehrcke, D. (2004). Map use, reading, analysis, and interpretation (6th ed.). JP Publications. ISBN 978-0-7668-2779-7, Chapter 2.

[31] Information Systems Audit and Control Association (ISACA) Glossary. https://www.isaca.org/resources/glossary

[32] National Institute of Standards and Technology Glossary. https://csrc.nist.gov/glossary