Treasury Board of Canada Secretariat
Secrétariat du Conseil du Trésor du Canada
Canada

Guidance on Data Quality

Published: 2024-01-01

Published by Treasury Board of Canada, Secretariat
90 Elgin, Ottawa, Ontario, K1A 0R5, Canada

Catalogue Number: BT48-42/2024E-PDF
ISBN: 978-0-660-69835-9

This document is available on the Government of Canada website at www.canada.ca

This document is available in alternative formats upon request.

Aussi offert en français sous le titre : Orientation sur la qualité des données

Guidance on Data Quality

Date of publication: January 10, 2024

On this page

Preamble

This guidance provides departmental officials with a common vocabulary for understanding data quality and advice on how to approach it in practice in support of the following:

- sections 4.3.1.1 and 4.3.2.1 of the *Policy on Service and Digital*
- section 4.3.1.3 of the *Directive on Service and Digital*
- sections 6.3.1, 6.3.3 and 6.3.4 of the *Directive on Automated Decision-Making*

1. Key concepts

▶ In this section

1.1 What is data quality?

Data quality is a characteristic of data defined by nine dimensions: access, accuracy, coherence, completeness, consistency, interpretability, relevance, reliability and timeliness.

Why is data quality important?

The quality of data affects whether and how easily users can find, share and use it when they need it. High data quality supports evidence-based decision-making and the use of automated decision systems, and can enhance the design and delivery of policies, programs and services across government.

Data quality can also assist departmental officials in confirming whether data meets the needs and goals of particular users, that is, if it is fit for purpose. However, data can be considered fit for purpose without meeting all data quality dimensions to the same extent. While departmental officials are encouraged to consider data quality in terms of the purpose for which the data is being used, not all dimensions of data quality will be equally relevant in every context. Similarly, data quality should also be considered throughout the life cycle of the data in question, as the relevance of each dimension may vary depending on the stage at which data's quality is considered.

2. Data quality dimensions

The nine dimensions of data quality provide departmental officials with a common vocabulary to define and assess the quality of data. The dimensions can help departmental officials identify and articulate various data quality issues, and address them to ensure that data is usable and relevant to their needs. Compliance with the dimensions can strengthen government-wide data governance, sharing and reuse.

The dimensions may overlap, and the importance of each might vary depending on the specific needs of the user. In some cases, it might also be necessary to find a balance between them. Departmental officials can provide more details to existing dimensions or consider additional ones to suit the specific type of data with which they are dealing. Departmental officials are encouraged to consult experts (for example, data stewards, data custodians, data providers, subject matter experts) who have the right knowledge to address enquiries related to each dimension of data quality.

Access

Access describes how easy it is to discover, retrieve, process and use data.

Access refers to whether users know about the data and have the right permission to access it. Even with access, users may not have the ability to process or manipulate the data to suit their needs. This could be, for example, because of technical limitations, not having enough resources, not having the necessary information, or certain policies and legislation having restrictions on how they can use the data.

Accuracy

Accuracy describes the degree to which data describes the real-world phenomena it is intended to represent.

Data is accurate when it represents a phenomenon adequately. Assessing accuracy depends on the context, methodology and the validity of underlying hypotheses or assumptions. In public sector organizations, keeping data accurate can mean making sure that the data collected when providing services matches the information clients shared. To ensure accurate data for policy and program initiatives, users often have to check the data against trusted sources and evaluate how the data was collected and processed in the first place.

Coherence

Coherence describes how easily a user can compare and link data from one or more sources.

A coherent dataset follows standard structures and classifications that are widely accepted. Users can improve data coherence by using organizational, federal, national or international standards, and specifically those prescribed as a Government of Canada (GC) enterprise data reference standard. When data is more coherent, it can be easily reused and combined with other data, allowing users to integrate and compare it.

Completeness

Completeness describes the degree to which data values are sufficiently populated.

Data can be considered complete when it has the entries that users need to use it appropriately. This includes both the dataset and additional information that helps users understand the dataset in their specific businesses.

Consistency

Consistency describes the degree to which data is internally non-contradictory.

Consistency makes sure that all the connections and relationships between the different parts of the dataset make sense and are logically correct.

Interpretability

Interpretability describes how much data can be understood in its appropriate context.

A dataset is interpretable if a user, human or machine can understand its entries, determine why and how it was collected or created, and determine whether it is relevant to a policy, program, service or other government initiative.

Relevance

Relevance describes how well the data supports a specific goal or objective.

The relevance of data depends on whether it provides useful information or insights to achieve what the user wants to do. Assessing whether data is relevant depends on the situation and the user’s needs. The same data could be important for one task but not another.

Reliability

Reliability describes how well differences in data can be explained.

Reliability is about data meeting user expectations over time. A dataset is reliable when users can explain how it evolves or changes over time. Reliability also involves making sure the data remains intact and unaltered or is altered only in a documented way through data integrity checks.

Timeliness

Timeliness describes the amount of time between the end of the period to which data pertains and the time at which that data is available to meet user needs.

Timeliness measures the delay between two time points: the moment the data covers and the moment users can actually use that data. Timeliness describes the degree to which users have access to the data when they need it.

3. Good practices for data quality

This guidance pairs the nine dimensions with corresponding good practices to provide a common approach for assessing and managing data quality. The list of recommended practices is not meant to be exhaustive but rather to enable departmental officials to interpret and apply the nine dimensions consistently. They can also be used to inform approaches to assessing, maintaining or improving data quality. The dimensions and good practices can be used for all types of data, and departmental officials can adjust them to fit their specific needs. Each practice may not be relevant in all cases or carry the same significance from one context to the next; it is left to the discretion of departmental officials to decide if, when, and how to apply each one.

Access

- Develop an inventory or catalogue of datasets to support policy, programs or services.

- Apply metadata reference standards (that is, prescribed enterprise-wide or departmental metadata reference standards) to describe concepts, variables or classifications in your data assets, in accordance with the *Standard for Managing Metadata* and the *Standard on Geospatial Data*.
- Establish processes for how your organization documents, retains, publishes, archives and disposes of the data it collects or creates.
- Assign security categorization to data assets, as required under the *Directive on Security Management*.
- Define access rights, privileges and restrictions, in compliance with the *Directive on Security Management* and the *Directive on Identity Management*.
- Confirm processes and procedures to support the production of data in response to requests for information under the *Access to Information Act* and the *Privacy Act*.
- Use plain language (see the *Canada.ca Content Style Guide*) and machine-readable formats (for example, CSV (comma-separated values), XML (Extensible Markup Language), JSON (JavaScript Object Notation)) to make data easier to share, process, use, publish and archive, including in the required metadata.
- Give users many ways to access and extract data, such as making data available in multiple formats and through accessible application programming interfaces (APIs) developed in accordance with the Government of Canada Standards on APIs.
- Invest in data infrastructures to provide easy and secure access to data, in accordance with the "cloud-smart" approach established in the *Directive on Service and Digital*. Sensitive data (Protected B, Protected C or Classified) should be held in systems within the geographic boundaries of Canada or within GC organizations abroad (see the *Direction on the Secure Use of Commercial Cloud Services: Security Policy Implementation Notice (SPIN)* and *Government of Canada Security Control Profile for Cloud-Based GC Services*).

- Work in the open by default and publish data to the Open Government Portal in accordance with the *Directive on Open Government* and as permitted by applicable federal privacy, security and intellectual property frameworks.
- Conduct surveys to identify barriers to finding, accessing and using data in your organization.
- Report any unauthorized access or use of data to designated security officers and, where personal information is involved, to departmental privacy officials. This may lead to notification to the Treasury Board of Canada Secretariat and the Office of the Privacy Commissioner of Canada, as required under the *Directive on Privacy Practices*.

Accuracy

- Check with trusted data sources to verify content and understand the data's context. If data errors are identified, review with trusted sources and determine how to address the errors.
- Ensure that data is described according to prescribed metadata reference standards (that is, enterprise or departmental level) so that users can determine its accuracy. Relevant metadata could include information about the source, purpose and method of collection, processing, revisions, coverage, and data model and related assumptions.
- Ensure that data adequately represents any domains (for example, geographic areas, populations) contained within it.
- Adhere to valid value ranges, where applicable. Explanations for outliers should be provided to data users.
- Develop rules to check data for errors, including duplication within a dataset. Apply the rules throughout the data's life cycle, particularly during collection and sharing.
- Ensure that methods used throughout the data life cycle minimize biases and statistical errors (such as sampling errors) (refer to the total

survey error framework and Gender-Based Analysis Plus (GBA Plus).

- Ensure that an authoritative source exists for data, where possible.
- Develop processes that allow for personal information to be corrected or updated if requested (see the *Directive on Privacy Practices*).
- Work with subject matter experts to check the concepts and assumptions used or how closely the data matches what the user wants to capture.
- Provide information about the measure of error or uncertainty of the data (for example, standard error, confidence intervals), where applicable.
- Ensure that outputs of artificial intelligence (AI) systems (for example, generative AI used or deployed by a department) are assessed for accuracy, including through bias testing.

Coherence

- Where applicable, adopt GC enterprise data reference standards, including for encoding and formatting the data.
- In the absence of an applicable GC enterprise data reference standard, adopt or adapt existing departmental, national and/or international data standards, and document differences in practices, particularly when sharing data with other organizations or publishing data to the Open Government Portal. Relevant standards could be domain-specific, designed for specific types of data (for example, statistical, geospatial).
- Consistently record data reference standards used in a data inventory or catalogue or in data-sharing agreements. If new data reference standards are developed, document reasons for not using existing and applicable GC enterprise or departmental data reference standards.
- Define, classify and represent data elements based on common data architectures, in accordance with the Government of Canada Enterprise Architecture Framework.

- Ensure that concepts, definitions and classifications are compatible within and across datasets so that data can be compared and combined, both internally and across the GC and external organizations.
- Use concordance tables to show discrepancies and transitions between standards used across data sources.
- Reduce data duplication across datasets to improve the integrity of the data and ensure that data is unique.

Completeness

- Ensure that no entries, columns or rows essential to the dataset are missing or incomplete.
- Keep values, concepts, definitions, classifications and methodologies up to date.
- Assign mandatory and optional labels to columns or rows in a dataset to make it easy to determine its completeness.
- Add the appropriate metadata describing the context and purpose of the data's acquisition. Metadata could also flag privacy, confidentiality or accuracy considerations that have an effect on completeness.

Consistency

- Develop rules to validate logical relationships encoded in a dataset. This could include rules formalizing the relationship between two interrelated variables.
- Regularly validate the consistency of datasets. Validation processes should be standardized and automated to support efficiency.
- Maintain a record of consistency issues identified in data validation procedures, and periodically review validation rules to keep them adequate and effective.
- Get the appropriate metadata from the data provider to learn about a dataset's entity classes, the values they are intended to permit, and the relations that hold among them.

Interpretability

- Adopt, adapt or develop controlled vocabularies so that key concepts are named and defined consistently in a dataset. Refer to prescribed metadata or data reference standards. Follow prescribed data reference standards governing permissible values for elements in a dataset (reference data, master data).
- As set out in the *Standard for Managing Metadata*, apply prescribed metadata reference standards for definitions and procedures to clarify why and how the data was collected and its security categorization, considering the needs of target audiences.
- Document the information needed to meaningfully interpret the data, including the original intent for the data collection and calculation methods, and maintain links between this documentation and the data throughout its life cycle.
- Inform users of the limitations of the data.

Relevance

- Establish processes to consult stakeholders on their data needs. This could be complemented by examining data inventories or catalogues to identify what already exists and minimize redundant data collection (see the *Guideline on Service and Digital* for guidance on information and data collection).
- Identify data requirements and sources based on business objectives and user needs.
- Assess and document how data assets meet data requirements in order to gauge their relevance. This could involve tracking how data assets are used and reused.
- Use the results of relevance assessments to inform future data acquisition and related life-cycle management and governance activities.

- Establish criteria to balance business needs and privacy and security risks when collecting data (see Statistics Canada's Necessity and Proportionality Framework).
- Ensure that the institution has the legislative authority to collect or create data about an identifiable individual and that the collection is directly related to an operating program or activity within the institution.
- Preserve data and associated metadata that have historical or archival value in accordance with the *Library and Archives of Canada Act* and supporting policy instruments.

Reliability

- Clearly document how data is collected and analyzed to make it easier for third parties to check and maintain the integrity of the data production process.
- Identify and document sources that can directly or indirectly change a dataset. Sources of change could include what the data represents, how the data was collected, data capture and storage technologies, data processing platforms, legislative or regulatory measures, policy requirements, and cyberattacks.
- Test data collection or creation instruments before using them, and document calibrations and account for differences in results.
- Record changes to your data assets so that users can determine where they came from and how they have evolved (that is, document through metadata).
- Identify and document dependencies among data assets in a data architecture or when analyzing data.
- Make concepts, definitions and classifications compatible over time. Specify and explain discrepancies in how these elements are maintained over time.

- Protect data assets from fraudulent or unauthorized activities that could undermine their credibility. This includes defining, implementing and maintaining security controls to meet information technology (IT) security requirements, in accordance with the *Directive on Security Management* and the *Directive on Privacy Practices*.
- Employ digital preservation approaches to monitor and guard against the deterioration of data assets over their life cycle. Conduct regular data integrity checks (through the use of hashing or checksums) and document any evidence of deterioration in accordance with the *Library and Archives of Canada Act* and supporting policy instruments.
- Report tampering or unauthorized destruction of data assets to designated security officers.
- Ensure that the data has an authoritative source, where possible.

Timeliness

- Identify users' current and future data needs, including considerations of time (reference periods, legislative or policy requirements, service standards).
- Consult with data providers about whether data needs can be met without delay, and inform data users of any expected issues, including the data provider's ability to meet timelines established in data-sharing agreements.
- Ensure that data providers have a data release schedule that documents the stages of the data production process and accounts for discrepancies and delays (such as through contingency planning).
- Publish preliminary data to the Open Government Portal where appropriate and in accordance with the *Directive on Open Government*.

ISBN: 978-0-660-69835-9

Date modified:

2024-02-01