

DATA QUALITY ASSESSMENT

PROCEDURE

Table of Contents

[1. Purpose 4](#_Toc11066421)

[2. Background 4](#_Toc11066422)

[3. Approval and Effective Date 4](#_Toc11066423)

[4. Related Policy Instruments 4](#_Toc11066424)

[5. Scope or Application 5](#_Toc11066425)

[6. Definitions 5](#_Toc11066426)

[7. Procedures 6](#_Toc11066427)

[7.1 High-Level Process Diagram 6](#_Toc11066428)

[7.2 RASCI Matrix 7](#_Toc11066429)

[7.3 Detailed Procedures 8](#_Toc11066430)

[Stage 1 – Understanding the Data 8](#_Toc11066431)

[Stage 2 – Profiling and Identifying Critical Fields 9](#_Toc11066432)

[Stage 3 – Assessing for Fit for Purpose 10](#_Toc11066433)

[Stage 4 – Improving and Monitoring Data 11](#_Toc11066434)

[APPENDIX 1 – Definitions 13](#_Toc11066435)

**Document Control**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Revision & Schedule History   |  |  |  |  | | --- | --- | --- | --- | | Revision | Changed By | Last Reviewed | Review Scheduled | | 1.0 | H. Espinoza |  |  | | 2.0 | A. Satterthwaite |  |  | | 2A | L. Purves |  |  |   Change Summary   |  |  |  | | --- | --- | --- | | Date | Summary | Changes made by: | | April 2018 | Summarized DQA manual which was created and reviewed from the period June 2017 to April 2018 | H. Espinoza | | January 2019 | Re-formatted | H. Espinoza | | March 2019 | Reviewed and modified | 1. Satterthwaite | | June 2019 | Reviewed and modified | L. Purves | |  |

**Consulted**

This procedure was created in collaboration with:

|  |  |  |
| --- | --- | --- |
| **Department** | **Branch** | **Division** |
| Corporate Services | Data Analytics and Visualizations |  |
| Regional Clerk’s Office | Information Asset Management |
| Finance | Information Technology Services | Solutions Delivery |
| Strategy & Architecture |
| Procurement Office | Procurement Policy & Performance |
| Transportation Services | Business Planning & Technology | Transportation Technology & Data |
| Community & Health Services | Integrated Business Services | Business Intelligence & Solutions |
| Environmental Services | Business Planning & Operations Support | Technology Data and Information Management |
|  | Operations Maintenance & Monitoring | Process Control Systems and Supervisory Control and Data Acquisition |

# Purpose

This document establishes a systematic way to conduct a data quality assessment at York Region. It provides a consistent governed methodology for York Region that reflects DAMA’s Data Management Book of Knowledge (DM-BOK) approach to data quality. It outlines the tasks and expected output for each task with the goal of improving data quality. Information is an important asset for York Region and quality data that users can trust is an important determinant of how well the Region is able to plan, make decisions and provide services to the residents.

# Background

Data quality assessment is more than a one-time data clean-up. It is a systematic way of ensuring that quality of the data will improve over time, so that staff can trust the data they use. Defining a consistent way of assessing data quality for various datasets will not only be helpful to data quality (DQ) stewards, who monitor the quality of the dataset, but will also help other Regional employees to be more conscious in improving quality of the information they are maintaining.

This procedure should be used in conjunction with the Data Quality Assessment Manual (eDOCS #[7648254](PCDOCS://YORK/7648254/R)), which provides more guidance and details.

# Approval and Effective Date

This document was developed in collaboration with Data Analytics and Visualization Branch and approved by the Regional Clerk. This came into effect on…….

For further inquiry regarding the content of this document, please contact…..

# Related Policy Instruments

1. Regulatory

* Municipal Act
* Municipal Freedom of Information and Protection of Privacy Act (MFIPPA)
* Personal Health Information Protection Act (PHIPA)
* Bill 54 – Records Retention By-law 2014-52

B. Related Policy

* Information Management and Sharing Policy ([eDOCS #7766050](pcdocs://YORK/7766050/R))
* Data Quality Assessment Manual ([eDOCS #7648254](pcdocs://YORK/7648254/R))

# Scope or Application

This procedure applies to on-premise or cloud-based datasets that have been prioritized as corporately or departmentally significant. Elements of the procedure can be applied to data the Region acquires from other organizations to ascertain how good the data is in relation to York Region’s purposes.

It can be used by:

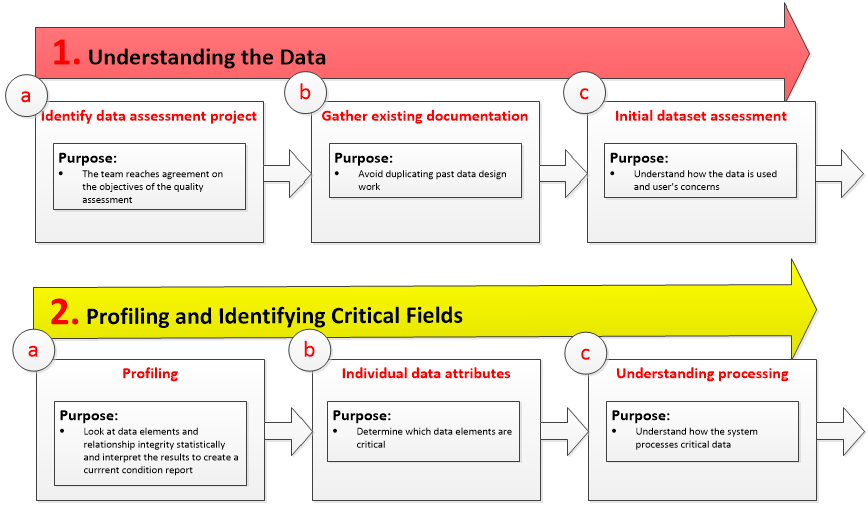
* DQ stewards and other staff who participate in data quality assessments
* Managers or data users interested in improving data quality

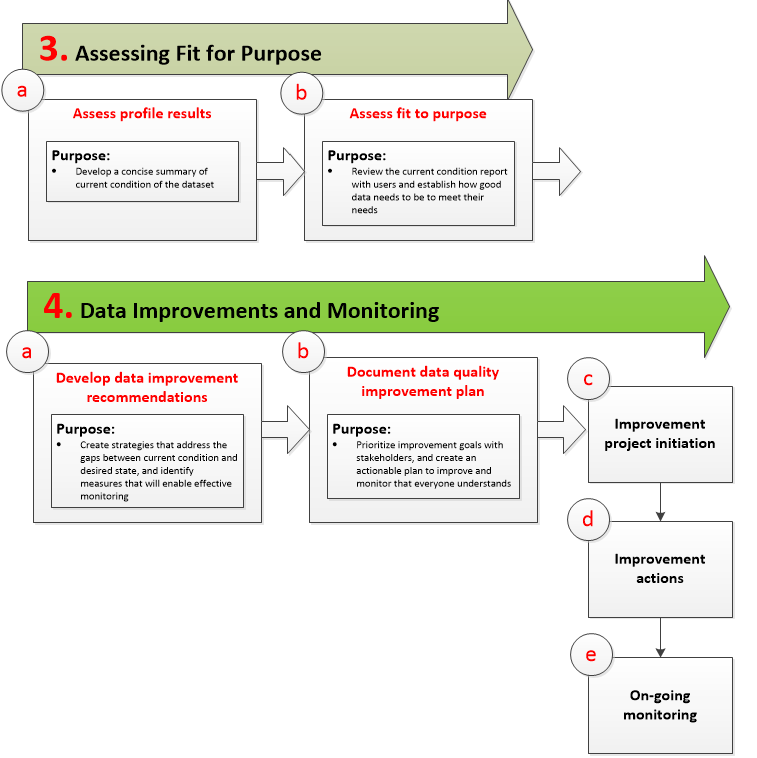
# Definitions

All terms included in this document can be found in Appendix 1.

# Procedures

## **High-Level Process Diagram**





## **RASCI Matrix**

A RASCI chart indicating Responsibilities (R), Accountabilities (A), Supporting (S), Consulted (C),and Informed (I) roles for each step and stage of the data assessment process follows.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Stage | **Understanding the Data** | | | **Profiling Critical Fields** | | | **Assessing Fit for Purpose** | | **Improvements & Monitoring** | | | | |
| Step | **a** | **b** | **c** | **a** | **b** | **d** | **a** | **b** | **a** | **b** | **c** | **d** | **e** |
| Role | Identify Assessment Project | Gather documentation | Initial data assessment | Profiling | Critical data attributes | Understanding processing | Assess profile results | Assess fit for purpose | Develop recommendations | Document improvement plan | Initiate improvement project | Implement improvement project | Ongoing monitoring |
| **Business unit manager** | **A,R** | **A** |  |  |  |  |  | **I** |  | **A** | **A,R** |  |  |
| **Business subject matter expert** |  | **C** | **C** |  | **C** | **C** |  | **C** | **C** |  |  |  |  |
| **Immediate data user** |  |  | **C** |  | **C** |  |  | **C** | **C** |  |  |  |  |
| **Downstream data user** | **C** |  | **C** |  | **C** |  |  | **C** |  |  |  |  |  |
| **External data user** |  |  |  |  |  |  |  | **C,I** |  |  |  |  |  |
| **DQ steward** | **C** | **R** | **A** | **I** | **C** | **C** | **I** | **R** | **R** | **C** | **S** | **A,R** | **A,R** |
| **DQ analyst** | **S** | **S** | **R** | **A,R** | **A,R** | **A,R** | **A,R** | **A** | **A** | **R** | **S** | **S** |  |
| **Other data roles** |  |  |  | **S** |  |  | **S** |  |  |  | **S** | **S** |  |

## **Detailed Procedures**

| **Step No.** | | **Action** | **Performed By** | |
| --- | --- | --- | --- | --- |
| Stage 1 – Understanding the Data | | | | |
| 1.a | | Identify the data quality assessment project:   * Prioritize datasets that are of enterprise and/or departmental importance * Decide on the target datasets for data quality assessment * Determine project scope, objectives, and team members, DQ stewards or other stakeholders who will be involved with the project   Output of this step is an approved Project Charter | Business unit manager | |
| 1.b | | Gather existing documentation:   * Business requirements document and relevant standards * Data models and existing metadata * Business rules and data flow diagram   The goal of this step is to avoid duplication of documentation that may already exists | DQ steward | |
| 1.c | | Conduct initial assessment of the target dataset:   * Document how the data is used, who the data producers and users are, the business processes that the data supports, the systems that provide or consume data to/from this target dataset, how the dataset is being maintained or updated, how the data is consumed to generate reports or analytics * Consult with data producers and users on known quality issues * Review or develop data process flow diagram and defining relevant business terms * Define the information lifecycle for the dataset   Output of this step can include: updated documentation, Supplier-Input-Process-Output-Consumer (SIPOC) diagram, business glossary terms, lifecycle analysis, an initial understanding of the quality expectations for the dataset | DQ analyst | |
| Step No. | Action | | Performed By |
| Stage 2 – Profiling and Identifying Critical Fields | | | |
| 2.a | Perform data profiling:   * Assess field-level data columns (in the dataset) statistically * Assess structure rules between columns in the dataset   Output of this step includes field-level data analysis, and machine-detected business rules suggestions | | DQ analyst |
| 2.b | Identify which data fields are critical:   * Determine which data fields are most critical for immediate, downstream and external users   Output of this step is a list of the critical data fields in the dataset | | DQ analyst |
| 2.c | Understand processes behind the critical data fields:   * Understand whether any critical data fields are processed fields * Assess risk of processes being corrupted   Output of this step is information about how critical data fields are processed | | DQ analyst |

| **Step No.** | **Action** | **Performed By** |
| --- | --- | --- |
| Stage 3 – Assessing for Fit for Purpose | | |
| 3.a | Assess profile results. This task includes:   * Reviewing profile results and summarizing for presentation to data users * Understanding where data quality issues are, and investigating apparent anomalies * Determining if discovered rules match the prescribed rules   Output of this step is the profile results analysis | DQ analyst |
| 3.b | Assess fit for purpose:   * Review profile results analysis with data users * Understand implications of having wrong, invalid, incomplete, inconstant or unavailable data in the dataset * Establish data quality targets with users that will meet their business needs   Outputs of this step are data quality targets for critical data fields, and a baseline quality assessment | DQ steward |

| **Step No.** | **Action** | **Performed By** |
| --- | --- | --- |
| Stage 4 – Improving and Monitoring Data | | |
| 4.a | Develop data improvement recommendations that address quality gaps:   * Create strategies that address gaps between current and desired state of data * Identify appropriate data quality assessment framework measures that will enable monitoring * Develop either in-line or periodic specifications for each measure   Outputs of this step are recommendations on how to improve quality with measurement specifications | DQ steward |
| 4.b | Document data quality improvement plan.   * Prioritize improvement goals and actions with stakeholders * Prioritize actions relative to other datasets through stewardship governance   Output of this step is a data quality improvement plan | DQ analyst |
| 4.c  4.d | Initiate and implement improvement projects and actions. This may include several discrete projects:   * Plan and implement specific data improvement recommendations * Construct queries that measure data quality * Determine the metadata repository that will capture the output from the measures and will provide the capability to track over data quality over time * Educate staff involved in data capture   Outputs of this step are tangible improvements to data and queries to measure aspects of data quality | Business unit manager  with  DQ steward |

| **Step No.** | **Action** | **Performed By** |
| --- | --- | --- |
| 4.e | Monitor continuously.   * Regularly run the quality measurement queries to ensure that quality targets are being met. * Intervene when quality falls below established targets   Outputs of this step are periodic status reports on data quality, which may include lists of anomalies or discrepancies | DQ steward |

# APPENDIX 1 – Definitions

**Data quality** – a measure of how well data fits the business purposes it is being used for. Data quality can be measured in relation to are commonly understood dimensions of data quality: completeness, timeliness, validity, consistency, integrity, and relevance. The dimensions can be broken out into measureable variables which can be used to assess the change in quality.

**Data quality (DQ) analyst**– this could be the DQ steward or a staff with statistical skills to analyze the data. Also responsible for profiling and identifying critical fields that will benefit from data quality assessment process as well as documenting the data improvement plan.

**Data quality (DQ) steward** – refers to the staff responsible for monitoring and ensuring quality of assigned dataset. The person has the authority to make corrections on the dataset to improve quality. A DQ steward may be responsible for all the data in an application, a group of tables representing one business function in the application, a table or even just a group of fields in a table.

**Data users** – people who consume information, often times, integrating them from different sources to create a something with business value. The following are the classes of users:

* Immediate users – staff from the business unit that acquires or creates the data – they are directly responsible for the business process that the data represents
* Downstream users – staff working for another business unit who take outputs from immediate users. They will usually have different expectations of the data, reflecting the business process they are engaged with.
* External users - other public agency, business entity, media organizations, general public or anyone outside York Region, who use data for a variety of purposes – such as ensuring regulatory compliance; coordinating development activity, service delivery and other business processes; informing their own processes; assessing York Region’s performance relative to other jurisdictions; or to exploring innovation initiatives.

**Data consumers -** downstream application processes that intake output from the source system.

**Data producer -** staff that capture or enter data into a source system, capturing information from the real world and representing it digitally. They can be immediate users; in many business areas, the function of producing (input) and consuming (analysis) the data may be done by different people.

**Source system** - system in which data is originally captured or created as raw unprocessed data – it could be applications hosted by York Region (for example, Citiworks), or hosted by other organizations (for example, Ontario Province Social Assistance Management System), or entirely external to York Region (for example, Municipal Property Assessment Corporation - MPAC database).