

Data Quality Assessment Manual

TABLE OF CONTENTS

[INTRODUCTION 3](#_Toc11406905)

[SECTION 1 – UNDERSTANDING THE DATA 5](#_Toc11406906)

[Goal 6](#_Toc11406907)

[Sub-Processes 6](#_Toc11406908)

[1.a Identifying the project 6](#_Toc11406909)

[1.b Gathering existing documentation 7](#_Toc11406910)

[1.c Conducting Initial Dataset Assessment 8](#_Toc11406911)

[Business Engagement 10](#_Toc11406912)

[Output 10](#_Toc11406913)

[Gating 10](#_Toc11406914)

[SECTION 2 – PROFILING & IDENTIFYING CRITICAL FIELDS 11](#_Toc11406915)

[Goal 11](#_Toc11406916)

[Sub-Processes 12](#_Toc11406917)

[2.a Profiling 12](#_Toc11406918)

[2.b Identifying Critical Individual Data Elements 13](#_Toc11406919)

[2.c Understanding processing 14](#_Toc11406920)

[Business Engagement 14](#_Toc11406921)

[Output 15](#_Toc11406922)

[Gating 15](#_Toc11406923)

[SECTION 3 – ASSSESSING FIT FOR PURPOSE 16](#_Toc11406924)

[Goal 16](#_Toc11406925)

[Sub-Processes 17](#_Toc11406926)

[3.a Assessing Profile results 17](#_Toc11406927)

[3.b Assessing Fit for Purpose 17](#_Toc11406928)

[Business Engagement 19](#_Toc11406929)

[Output 20](#_Toc11406930)

[Gating 20](#_Toc11406931)

[SECTION 4 – IMPROVING & MONITORING THE DATA 21](#_Toc11406932)

[Goal 21](#_Toc11406933)

[Sub-Processes 22](#_Toc11406934)

[4.a Developing Data Improvement Recommendations 22](#_Toc11406935)

[4.b Documenting Data Quality Improvement Plan 23](#_Toc11406936)

[4.c/d Initiating Improvement Projects and Implementing Actions 24](#_Toc11406937)

[4.e On-going Monitoring 25](#_Toc11406938)

[Business Engagement 25](#_Toc11406939)

[Output 25](#_Toc11406940)

[Gating 26](#_Toc11406941)

[Appendix A: Roles and Responsibilities 27](#_Toc11406942)

[Appendix B: PROJECT PLANNING FOR DATA QUALITY ASSESSMENT 28](#_Toc11406943)

[Appendix C: METADATA CAPTURE 32](#_Toc11406944)

[Appendix D: MEASURES FOR MONITORING DATA QUALITY 38](#_Toc11406945)

[Appendix E: HINTS ON GETTING THROUGH STAGES MANAGEABLY 56](#_Toc11406946)

# INTRODUCTION

About the Manual

Data quality assessment (DQA) is the process of statistically evaluating data in order to determine whether it meets the quality required to support its intended use in business processes. This Manual details a consistent data governance methodology that has been adopted for York Region, which can be applied in all departments. The methodology reflects the concepts from DAMA’s Data Management Book of Knowledge (DM-BOK).

The primary audience for this Manual includes data quality (DQ) stewards and staff fulfilling other data roles (collectively the DQ team) who will conduct data quality assessments. The core of data quality assessment, and a big part of the role of DQ stewards, is communicating with data users to establish what level of quality is needed. If you are a manager or a data user interested in improving data quality, you can also get an understanding of the different stages of data quality assessment by reading the document.

Having trusted good quality information is becoming important determinant of how well the Region is able to plan, make decisions and provide services to residents. What we get from applying a consistent approach to data quality is higher trust in the data you are responsible for, fewer complaints, less time wasted doing periodic data clean-ups. There’s a lot of effort up front, but the on-going effort of maintaining data is reduced, as everyone understands what they have to do, and users spend a lot less time trying to corroborate the data. This Manual tells you what you have to do to get the right input from those users.

This Manual supports the Data Quality Assessment Procedure (eDOCS #8353186). The Manual is divided into 4 major sections to represent the 4 stages of data quality assessment. Each section discusses the goal, the tasks or sub-processes, the business engagement required the output and the gating or decision points in order to proceed to the next stage.

At this point, there is not an approved software package for data quality assessment being used within the Region. MS SQL Profiler, Alteryx and Excel are currently being used to clean and profile data, but are not optimal for data quality assessment and monitoring. If and when a better solution is available, this Manual will be updated to reflect best practices on how it should be used.

Appendices provide further details on the following:

* DQ roles and responsibilities, as a RASCI matrix (see Appendix A)
* Estimates of how long different DQ tasks might take, intended to help with project planning (see Appendix B)
* Different metadata elements generated at each stage of the process (see Appendix C)
* A typology of quality measures that can be built to monitor data quality (see Appendix D)
* Hints that a DQ project team will want to consider as they move through each stage of the data quality assessment process (see Appendix E)

Training

There are two in-class courses currently on Data Quality. They are DA0007 (Data Quality: Concepts) and DA0008 (Data Quality: Steps).

Employees can register for in-class training through the Learning Management System (LMS) in [ROSIE](http://aix-psap3.ykregion.ca:8080/hr/signon.html). For detailed instructions on how to register for York Region courses and programs, view the ["Registration" page (pg. #19)](https://s3.amazonaws.com/online.fliphtml5.com/aryl/rthb/index.html#p=18) in the Employee Development Corporate Catalogue.

Contact Information

Data Analytics and Visualization Branch, Corporate Services Department

Email data@york.ca or call 905.30.4444 ext 73282

# SECTION 1 – UNDERSTANDING THE DATA

 

**Figure 1.1**

## Goal

The first stage of data quality assessment is getting a good understanding of the data. Selecting a dataset for assessment is a prioritization exercise, based on the relative importance of the dataset to the corporation and the department. A DQ project team then gathers existing information about the dataset and develops an initial assessment of how the data is being used. Find out from the data users any issues with the dataset as they presently understand and use it. This stage also involves looking at the data lifecycle.

## Sub-Processes

### 1.a Identifying the project

The project team for a data quality assessment project will likely consist of a project lead, data quality (DQ) stewards for the dataset, subject matter experts, and other data roles who apply DQA tools to speed up the assessment process and application support staff (and ITS staff who may be required to give access to back-end data tables). Stakeholders include immediate users from the business unit that acquires or produces the data, downstream users from other business units that the team is aware of and even external agencies and the public.

Collectively the project team needs to agree on the specific objective of the assessment project. Each dataset at least requires an initial baseline assessment that will determine what level of quality is needed to effectively support the business functions. An assessment can also focus on maximizing use-value, by choosing to concentrate on how downstream and external users can be better served by the upstream processes. Or an assessment can focus on preparing for Open Data publication.

The DQ steward or other data roles may not be involved in prioritizing datasets that have corporate, cross-departmental or major departmental significance. This is often done by management. But once assigned to a priority dataset, it is good for the DQ steward to know what other priority datasets have been identified by management – data tend to connect with other data eventually.

Several DQ stewards may be assigned for a complex dataset; each may have a subset of the data to monitor. A database may have tables for work-orders, addresses, and contacts, and each of these might have separate DQ steward. Depending on how many business-functions the dataset records, it can be effective to constrain the scope to a few tables representing a discrete business process, rather than trying to assess the entire dataset, to keep the effort at a manageable scale.

In the initial stages of an assessment for complex dataset, it makes sense for the DQ stewards to work together. Each DQ steward should listen particularly for users’ comments on the data elements they are responsible for, and for its relationships with other data elements.

During scoping, management should decide if privacy, security and access concerns will be addressed or not. These are issues affecting quality that need to be addressed for every dataset, but they may have already been addressed in a separate project. The Corporate Access and Privacy Officer’s advice may be sought by the assessment team. It may be necessary to use the Privacy Impact Assessment and Cybersecurity processes if the scope warrants it.

The project lead should document the scope of the data quality assessment in a project charter. A charter ensures that there is no misunderstanding between the various stakeholders. Depending on the scope of the dataset and the assessment, this document can be shorter or more detailed.

### 1.b Gathering existing documentation

The task is to find any existing materials related to the dataset. Then within the system where metadata are stored, create fields that will link or associate the documentations. This can be done by any team member but is best done by the application support specialist or a DQ steward.

Some documentation may have been created by business analysts when the application was configured for entry, processing, or display of the data. Data models, business rules, process diagrams, and formulas can often be found as part of the use cases developed by business analysts as requirements for the application and database. Standards are sometimes published by the authorities that maintain them, such as regulatory agencies.

Where materials exist as hard-copy only, scan the materials and store them in an approved repository (e.g., eDOCS) so a link can be made from the metadata record to the scanned document. Where documentation is missing, the team should decide if there is value in creating it as part of the next task. Use an existing materials checklist to track which materials have been found or need to be created (See Figure 1.2).

**Figure 1.2**

|  |
| --- |
| Table: Existing Materials Checklist |
| Document Type | Found (Y/N) | Create (Y/N) |
| Business requirements |  |  |
| Conceptual data model |  |  |
| Logical data model |  |  |
| Physical data model |  |  |
| Business rules for data capture |  |  |
| Business process manuals |  |  |
| Standard operating procedures |  |  |
| Product support |  |  |
| User guides |  |  |
| Processing algorithms (e.g., ETL process) |  |  |
| SIPOC data flow diagram (Suppliers/Source-Input-Process-Output-Customers/Consume)  |  |  |
| Relevant standards |  |  |
| Regulations affecting capture, processing or reporting of data |  |  |
| Data lifecycle analysis |  |  |
| Privacy impact assessment |  |  |
| Security assessment |  |  |
| Application access procedures |  |  |

### 1.c Conducting Initial Dataset Assessment

The project team should have a good idea of what business processes the data support from the charter. During the initial dataset assessment, the project team develops its understanding of how the data flows between immediate, downstream and external users, and differentiates how the data is used by each user. Starting from existing documentation, and by asking questions of each user, the DQ steward should form a complete picture of how the data is created, processed and consumed. This may take a while if there was not much existing documentation of the business processes and the dataset.

As more information is gathered, it may become apparent that existing documentation is out-of-date. Update the documents as necessary to reflect the current business process and use of the dataset. As relevant business terms accumulate, document them in a glossary so that all the users share a common terminology. If there are differences between how immediate and downstream users define terms, these differences should be noted.

If there is not already a data model, create one. A profiling tool should be able to assist in creating a physical model, but a conceptual model is also helpful for users to understand the key relationships between different data elements (the physical model will contain application process tables and record ID’s that aren’t much interest to users).

If there is no existing data flow diagram, this should be created (Visio is a good tool for capturing this kind of information). The goal is to capture the flow of information through the Region, and what it is being used for. Track the flow of data from when it is received/ created, to its use as an input for downstream and external analyses. How is the data passed from one process to the next? This begins the process of understanding how and why the data is important to different users, and what are the determinants of data quality for these datasets.

Ask users about their business processes. Ask about their problems or concerns around the data they use and log these as issues. Potential users of the dataset should also be engaged, even if they are not currently using the dataset.

If there is not already a lifecycle analysis (most applications do not have this), this should be created. Lifecycle analysis helps understand when the data is at its most valuable to the business process, and helps avoid putting effort when it is not needed. It also can help control risk associated with keeping information longer than necessary, and can indicate the appropriate time to consolidate or archive data. All of this also helps compliance with the retention bylaw. The analysis documents when the data is created, how it is stored, how long will it kept and what is the process to dispose of the data. Users will determine how long data should persist, and whether it should be deleted or archived. Recognize that the disposition needs may vary between users. In some cases, archiving of aggregated data may be preferable to archiving of raw data. These determinations should be made with respect to the business processes the data supports. Include in the lifecycle analysis migration strategy if archiving a part of a whole of the data will move to another system for long-term preservation. The information management staff can assist in the analysis.

If deemed within the charter scope of the assessment, privacy, security and access considerations can be addressed. The Region has defined procedures and templates for Privacy Impact Assessment and Cyber-security assessment. Consider whether the different users have appropriate access to the dataset, consistent with their business purpose. Are there constraints on their use that could potentially be removed? Can the data be anonymized to make it fit for broader and even public consumption?

As much as possible the DQ stewards should be involved in these consultations and document preparation. The more knowledge of users’ concerns they gain will provide the basis for ongoing stewardship.

## Business Engagement

The DQ steward and the project team must engage managers and subject matter experts to identify appropriate data assessment projects. Prioritizing which datasets should be quality controlled is fundamentally a management activity, as it relates to business objectives and resources. The business unit must supply the DQ team with documentation about the business process, the dataset, and internal processing rules. The business unit managers must ascertain if staff has a current understanding of the data collection process. The DQ team will also need to engage business subject matter experts (both in the business unit creating the data and in business units using the data) to determine what business processes the data supports, and if there are compliance issues.

## Output

* Project Charter - documents the dataset for assessment, the objective of the assessment, and the team doing the assessment and their roles
* Business process documentation for the dataset – such as business rules, links in metadata, business glossary, data model, data flow diagram (showing how the dataset is used by different users) and a data lifecycle analysis
* Issue log for the dataset, listing user’s data gap and concerns

## Gating

This stage is about understanding the data. When everyone on the project team feels they have assembled enough information on how the data is being created and used, it is time to move onto the next stage. It is important that the DQ stewards, especially if recently assigned, are aware of the responsibilities of DQ stewardship, as they will be taking a significant role as the process moves forward and will most likely lead subsequent data quality assessments projects.

Refer to the introduction section of this Manual for training resources and to the Roles and Responsibilities around DQ Stewardship (eDOCS #7807003).

# SECTION 2 – PROFILING & IDENTIFYING CRITICAL FIELDS



**Figure 2.1**

## Goal

At this stage, the goal is togenerate statistical profile of fields and relationships within the dataset. This can be automated with a suitable profiling tool. The DQ team looks deeper at the structure of the dataset, collecting information about each field in the dataset. Data users are asked to identify which of the fields are critical for their business purpose. They should also be asked to provide definitions of these critical fields, to ensure that there are not different understandings between data users.

Basic information about each field is collected. The metadata associated with a dataset should provide a lot of information to users about whether they should use it, including when it was last updated. If the metadata is missing, users have little capacity to assess whether the data is appropriate for their purpose.

The DQ team also will get an understanding of how fields are processed. Formulas or computer programs can add systemic errors to a dataset, so it makes sense to understand how they work. Remember that what is seen on the screen is not necessarily how it is stored in the database.

## Sub-Processes

###

### 2.a Profiling

Profiling a dataset involves statistical evaluation of each field for a variety of measures. What proportion of the entries in the field is null? What is the average value in the field? What is the distribution of values in the field? What is the cardinality of the field? Depending on the type of field and the nature of the values in the field being assessed, different profile results can be generated.

Getting several profiles for each field helps to understand what is going on in the database. Profiling should take into account quality dimensions (completeness, timeliness, consistency, validity, integrity) that are likely to be significant for users. Other data roles, such as a data scientist, can assist the DQ stewards in selecting profile measures that are appropriate to the field.

A profiling tool will also look at the relationships between fields, and can detect when expected relationships are missing; for example, where a child record is missing a relationship to a parent record.

Profile measurements can be standardized for comparison between different time periods. Taking a ratio measurement, for example, of missing values, using the formula:

 1 minus (missing values/total values)

The formula allows for a comparison from the initial baseline profile measurement date to a profile taken after improvements were implemented in data collection, regardless of how many new data points were collected in the meantime. It is important to establish a repository to store this timeline based information for later comparisons.

Depending on the profiling tool used, business rules for each field may be suggested by the software. Rules suggested by the profiling tool should be checked for veracity by the project team before being used or added to metadata.

Profiling will show a number of errors in the data. It can be tempting to fix them straight away but resist the temptation. More meaningful and permanent interventions will be developed in subsequent steps of the DQA process, with consideration of how the errors were generated in the first place.

For the initial assessment, it is important to be as comprehensive as possible, looking at all fields. This provides a baseline to compare later improvement projects and re-assessments against. Attention to primary and foreign keys and the relationships they embody is important in the initial assessment, as problems in these fields can have serious impact on how intelligible the dataset is to users.

A profiling tool can be used to measure metadata completeness. Of all the potential fields, how many are missing, and how many are populated? This can be expressed as a formula for comparative purposes. The result can be entered into a completeness field for metadata (metadata about metadata), providing a quality measure that has meanings for users. As the data quality assessment process proceeds, this measure can be recalculated to show progress within the assessment project. It should also be compared over time to ensure that the metadata does not degrade. This task can be completed by a profiler, but the DQ steward should be aware of how complete the metadata is, as this impact the users’ capacity to evaluate a dataset before they use it. There is also a qualitative aspect to metadata assessment in that the entries should be intelligible to metadata users. This requires a more subjective evaluation as to whether the metadata reads well.

The formula for metadata completeness can be adjusted as the data quality assessment program matures, but an initial score should include quantitative and qualitative measures. Focusing on fields that have been deemed critical is a practical way of avoiding having the same weighting criteria for fields that have low meaning for users. The ratio of how many field-level attributes have completed the description, business rules and formula filled in provides a simple quantitative measure. The qualitative measure is a reader assessing how readable and meaningful the field descriptions for the critical fields are.

A simple metadata completeness score thus might be $\left(^{b+c+d}/\_{a}\right)X e $ , where

1. number of critical fields
2. # of non-null descriptions for critical fields
3. # of non-null business rules for critical fields
4. # of non-null formulas for critical fields
5. A subjective score of meaningfulness on a scale of 1-3, where:

1 - Not clear what the fields are for

2 - Somewhat clear

3 - Very clear, no ambiguity

This provides a metadata completeness score out of 100.

Again, it is important to establish a repository to store this timeline-based information for later comparisons.

### 2.b Identifying Critical Individual Data Elements

The DQ stewards pay close attention to the data elements that have assigned responsibility for. They need to ask detailed questions about the fields and relationships which are critical for users, and then understand how those fields are maintained. They need to ask the people who provide or input data what they do. They also ask the staff or service providers who built the system how the fields get processed, modified and displayed by the application system on the data model, and how table and field level metadata are created for the dataset. A profiling tool should be able to accomplish much of this, but business definitions that the various users of the data understand and can agree with will need to be entered manually.

Profiling tools will provide-field level metadata for every field in a database, but commercial application databases often have field names that do not mean much to users and may also have numerous tables and fields that are of limited interest to users. To make automated profiling intelligible, it may be necessary to provide aliases for fields that correspond more closely to the labels in the user sees in an application, and mask system tables and fields, so they do not add clutter can make the results less intelligible.

Ask immediate, downstream and external users which fields are critical for their business purposes. Sometimes a non-critical field for the immediate user is critical to a downstream user. Primary and foreign keys are critical if they relate to other critical fields.

###

### 2.c Understanding processing

For the fields that the users have identified as critical, document how they are maintained. Questions include: How many people are responsible for entering data into this field? Do they process data as they are entering it (with the potential for introducing errors)? Are there controls on what they can enter, either as another staff person checking for correct data entry, or a system-controlled process? Are there complex Extract Transform Load (ETL) processes for this field? If so, who maintains the ETL process? Developing an understanding of business processing of data is an activity for the DQ steward and the system administrator. Advanced profiling tools can help surface the formulas and/or business logic used to process data by reverse engineering.

Capture formulas and business logic in metadata, and then characterize fields as containing raw data, lightly or highly processed data. Raw data is data as entered or as received. Lightly processed data is that subject to a simple arithmetic formula based on two or three inputs, where it is easy to understand what is happening to the data. Highly processed data is that subject to complex formulas based on four or more terms or containing if/then logic, or ETL-type processes, where most people would have difficulty understanding what is happening to the data. Highly processed data can be a high-risk area for quality because processing can hide data entry errors, and formulas can get corrupted.

## Business Engagement

The DQ team must work with business subject fields through the various quality dimensions (completeness, timeliness, consistency validity, integrity)) to determine which fields are most critical to the various business processes. They must also work together to determine which quality measures are significant to users.

## Output

* Field definitions, including which fields users see as critical to their business processes
* Series of statistical measures about different fields and the relationships between them. This output can be stored as metadata about tables and individual data elements, fields, or domains. Depending on the capabilities of the profiling tool, this can include suggested business rules the profiling tool detects for each field
* Documentation of the processing or manipulation of critical fields

## Gating

The profiling stage is about getting a statistical review of the current condition of the data. The stage is complete when all the fields have been profiled, and users have identified which of these fields are critical to their business needs. The DQ team should have a good understanding of whether each critical field should be characterized as raw, lightly or highly processed before moving on to the next stage.

# SECTION 3 – ASSSESSING FIT FOR PURPOSE

 

**Figure 3.1**

## Goal

In this stage the DQ team interprets the profile results to determine the current condition of the data. Then they’ll work with the data users to determine how good it needs to be in order to reliably support their business processes. This stage is where data quality targets are defined based on user requirements. The current condition of the data is compared or assessed against the purposes it is put to by immediate, downstream and external users.

## Sub-Processes

###

### 3.a Assessing Profile results

The DQ team reviews each profile result to determine whether there are any problems with the field and the relationships it participates with. Pay attention to primary and foreign key relationships for orphaned records. Where the nature of the problems is not completely clear from the profile result alone, examine the data directly where this is possible to gain an understanding of the issue, or query the data further to see if understanding emerges. The emphasis should be on detecting problems, not deciding what to do about them. Do the profile results confirm or contradict the perceived problems users brought forward in during the initial assessment? Document issues from the review so that they can be discussed with the data users later.

If the profile tool used was able to generate suggested rules, check whether these rules correspond to the rules that were captured in existing documentation and from users. Where there is a discrepancy, determine which rule is most appropriate for the business process and document it. Any changes this necessitates to the embedded business logic (and this should be a very rare occurrence) should follow the appropriate software development life cycle and change approval process before production deployment.

###

### 3.b Assessing Fit for Purpose

This is a highly consultative part of the data quality assessment process. The profile results are shared with users. The DQ steward will elicit feedback from the various data users (and their managers) about whether the condition of the data is adequate for their needs, at the level of specific fields, and in relation to specific problems identified in the profile results assessment.

To conduct this review, bring forward the existing documentation collected during the first stage (Understanding the Data) and all the metadata collected in subsequent steps (Profiling) as resources for the participants. Start by reviewing the use cases and perceived problems collected during the initial consultation with data users. Tie the issues users reported in during that consultation to the field level metadata collected during profiling. Then bring in the profile assessment results. While most of the attention should be on the fields that users identified as critical, other fields should be in mind.

The review can also be done with immediate, downstream and external users altogether, or separately. There can be value in the different groups of users hearing from each other, but it may be more practical to meet with different groups separately. While there is a tendency to value the needs of the immediate users more highly (their business process is typically most closely aligned with the dataset, and their business unit usually bears the budget impact of collecting or acquiring the data), it is important to realize that data quality is defined collectively by all the users, including those that are downstream users. The overall value of the data is also tied to all the use cases it is put to, not just the use cases defined by the immediate users. The DQ steward should be involved in all these reviews so that the perspectives of downstream and external users are clearly known to someone in the immediate users’ group.

Together with the users, go through what they are using the data for, and how closely it conforms to their needs, at the level of fields and the condition of the data. What are the specific questions they are hoping to answer with the data? Is the current condition of the data adequate, or would their questions be better served by improved data? Do the users create reports that use specific fields? Are the reports used to help inform decisions, or are decisions entirely based on the reports? What are the implications if the data is wrong? What if the data is invalid, incomplete or simply not available for the questions and reports in a given time period? How timely is data delivery for each purpose?

Within specific fields, are there anomalies that create issues or inefficiencies for users? Are addresses, for example, formatted appropriately? Is there a service or a lookup available that could validate entries in a field? What are the perceived limitations of the data? From the perspective of various users, are there any fields that could be defined differently, or that are not captured as part of the current data collection process, that would help the dataset meet the needs of the users?

Asking these questions is to ascertain what good quality data would look like in terms of the purposes users are trying to achieve, along with the dimensions of completeness, timeliness, validity, consistency, and integrity. The user’s expectations can be classified later if necessary. Relevance should also be considered – perhaps this is not the appropriate dataset for the purpose. All responses should be captured in the metadata as expectations relative to specific business processes. An assessment of how much of a gap there is between the expectation and the current condition of the data should also be made, again with respect to the specific business process.

An evaluation grid (See Figure 3.2) can be used to help quantify the gap, and should be completed for each business process the dataset it is used for. The grid uses a 1-5 Likert scale that many people are familiar with and prompts for the subjective evaluation of each quality dimension.

**Figure 3.2**

|  |
| --- |
| Table: Subjective Evaluation Criteria for Dataset Fit for Purpose |
| Subjective evaluation criteria  | **Complete-ness****(data expected is there)** | **Timeliness****(delivered when needed)** | **Validity****(reflects reality accurately)** | **Consistency****(over time)** | **Integrity****(no orphaned data)** |
| 1 | Serves business process well | Clearly informs decisions | No perceived risk |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |
| 3 | Business process able to function | Minor uncertainty around decisions | Small risk |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |  |
| 5 | Compromises business process | Compromises decision certainty | Creates risk |  |  |  |  |  |

Where the dataset is published as Open Data, the project team should consider writing a guidance document for external users that can be posted as part of the metadata. Providing a means of evaluating data quality can significantly increase the value of the data to external users, who otherwise have to make many assumptions about data quality. The guidance document would require updating whenever improvements are made to the dataset.

At the conclusion of this step, the project status of the dataset should be changed to “baseline assessment complete”, and the overall status changed to “governed”.

## Business Engagement

The DQ team, including the DQ steward, must work with data users (and their managers) to determine how good the data needs to be to support the business processes. Typically there is an educational component to this engagement. Collectively the DQ team works with data users to determine whether there is a gap between the current state of the data and the desired state. This needs to be done critically.

## Output

* Summary of the profile results for each data field, identifying any obvious problems or anomalies with the data, as well as the results of any further investigations. This can also include any discrepancies between documented and suggested business rules. The report needs to be in a format that can be readily shared with the project team, immediate, downstream and external data users.
* Documentation of the desired quality for the data, in relation to the current condition of the data, and an understanding of what the dataset should and should not be reasonably used for. The documentation should reflect the dimensions of quality, with respect to completeness, timeliness, validity, consistency, and integrity.
* A guidance document for external users that can be posted as part of the metadata.

## Gating

This stage (Assessing Fit for Purpose) is complete when the profiling assessment had been documented so that the current condition of the dataset is readily intelligible to other members of the project team, immediate, downstream and external users, and all the major users of the dataset have been consulted about their needs relative to the current condition of the dataset, and establishing the desired state.

# SECTION 4 – IMPROVING & MONITORING THE DATA



**Figure 4.1**

## Goal

This stage is focused on developing potential improvements to the dataset that can close the gap between the current condition and the desired quality targets. This requires an analysis of where the errors originate, whether with the people, the process or the technology. Each source of error will require a different form of remediation. Once potential data improvements have been defined, the DQ team works with users and managers to create a prioritized Data Quality Improvement Plan that can be acted on over time to make the dataset fit for its user’s various purposes.

This stage also involves developing measures specific to each quality dimension that monitor whether each improvement is sustained over time.

Data quality improvements and monitoring are not strictly speaking part of assessing data quality, but once a data quality improvement plan has been approved, actions to address the issues and implement the recommendations should commence, assuming resource availability. These quality improvement projects act on one or more recommendations, and result in improvements being made to specific fields in the data. These improvements are sustained because each improvement implementation includes deploying quality measurement queries that will be monitored by the DQ steward on a regular basis, enabling early detection of quality degradation.

## Sub-Processes

###

### 4.a Developing Data Improvement Recommendations

Based on an understanding of the gap between user expectations and the current condition of the data with respect to different dimensions of quality, the DQ team will formulate potential improvements that will reduce the gap.

For each critical field where there is a gap between expectation and current condition for one or more quality dimensions, develop a plan to improve the data. Some critical fields may have a similar gap profile – these can potentially be treated similarly. Conduct root-cause analysis and review documented data collection practices, data input, database maintenance activity, delivery practices, or even data disposition that may cause data quality degradation. Ask if the gap is a result of misaligned data concepts, data definitions, or missing schema elements? Analyze where the errors originate, whether with the people, the process or the technology. What action would improve quality from the perspective of immediate, downstream or external users, keeping in mind how they use the data?

It is important that any improvements are not just quick fixes; unless the improvements tackle the causes, either in the input, maintenance or delivery practices, the gap will reappear after a short time. Most effective improvements will have a change management or training component to them, whether to modify practices of input staff and users, or to inform them of changes made at the system level. To ensure that the improvement goals are being met, and to detect relapses in practice, appropriate measures of data quality will be developed to monitor whether the improvement is sustained over time.

Quality monitoring measurements can be in-line or periodic. In-line measurement occurs when the data is being processed and provides immediate feedback on whether the condition of the data meets expectations. Periodic measurement is monitoring every so often to ensure that the condition of the data has not changed in some significant way. Measures are specific to the various dimensions of data quality: completeness, timeliness, validity, consistency, and integrity.

Choose measures appropriate to the recommended improvement from the list in **Appendix D**. Often the measures will need to be modified or adapted to match the data being measured. The measure chosen should be normalized, so those quality measurements taken at different times can be compared, even though more records have been added to the dataset over time. Determine whether the measure will be applied to the whole dataset, or, if the dataset is very large, to a part of the dataset. Also define acceptable levels of deviation that represent an allowable percentage of errors appropriate to the business process and the desired quality target, that will trigger action for the DQ steward monitoring the dataset. Determining measures and monnitoring triggers is a technical task that will require the assistance of a data scientist, data quality analyst or similar role. The measures should be carefully explained the DQ steward, so they know exactly what is being measured and how.

Define a response protocol for staff, led by the DQ steward, to follow in the event greater than acceptable levels of deviation are discovered. This is a key part of sustaining quality – defining the actions to take when quality declines, based on the root cause analysis as to why the quality issues emerge in the first place.

### 4.b Documenting Data Quality Improvement Plan

The previous step will lead to a comprehensive list of potential data improvements that addresses all the short-comings of critical fields in the dataset, and a set of measures that could be used to monitor changes. In many cases, implementing the whole list will exceed the capacity of immediately available resources. So it makes sense to prioritize which improvement actions should occur first, taking into account the positive impact the actions will have on data quality and the cost and ability to implement them.

One aspect of data improvement recommendations that has costs associated with it is creating the metadata repository that will capture the output from the measures so that data quality can be tracked over time. Ideally, this metadata can be integrated with the dataset it is measuring, but with proprietary applications, this may not be possible. Figuring out the cost of implementing measures and a metadata repository will likely require the assistance of technical staff beyond the DQ team.

The project team, including the DQ steward, should meet with immediate, downstream and external users to review and rank the potential data improvements. Collectively the users should prioritize the data improvements, keeping in mind where the funds to make the improvements are coming from. Typically the originating branch that manages the dataset will be paying for most of the improvements, so that will lend weight to the prioritization exercise. But managers from that group should also be aware that the overall quality of the dataset is determined by all the users, and without addressing at least some of the concerns of downstream or external users, the dataset will not become a trusted source. In some cases, downstream users may be able to contribute budget towards data improvements that benefit their purposes. Budget inputs are less likely from external users, but there is the ongoing reputational risk that at some point should be reflected in the prioritization.

Following the prioritization with users, a member of the project team should draft a project plan that indicates which improvement recommendations will be implemented and in what order, with timelines, resource requirements, and budget. This project plan should be reviewed with users to ensure that the user community fully comprehends what is being proposed. This is important as a trust-building measure.

Where the budget and resources are small, and within the scale of resources usually associated with the dataset, the DQ team should be empowered to implement them. Where the budget and resources are significant, the DQ team should review the project plan with management for authorization to proceed. Ultimately approval of the prioritized project plan will lie with management in the group that manages the dataset.

### 4.c/d Initiating Improvement Projects and Implementing Actions

Data quality improvement projects implement recommendations of the data quality improvement plan. Several discrete projects may be required to implement all of the recommendations in the Plan, reflecting the priorities given to different recommendations and the availability of resources over time.

The data quality improvement team may be different than the data quality assessment team, as it may have staff with more technical database skills.

The data quality improvement projects have to take account of the metadata repository that will capture the output from the measures so that data quality can be tracked over time. The first data quality improvement project for a dataset should always consider the metadata requirements of additional recommendations that may be implemented in subsequent improvement projects so that the metadata repository does not have to be rebuilt.

Data quality improvement projects should include change management and communications components that will help sustain the improvement. Staff involved in data capture need to understand the objectives of the improvement, and have suitable reinforcement for any behavior changes. Training for all users on data definitions, business process changes and improvement objectives and realized outcomes will also help build trust.

A good number of data improvement projects will involve data cleansing, fixing errors and missing values in the dataset. Removing bad data, if not done properly, can have a serious impact on data in other table, especially when arbitrarily assigned GUIDs are used. Whenever possible data fixes should be piloted in a test environment before they are applied to the production database.

### 4.e On-going Monitoring

It is the DQ steward’s on-going responsibility to monitor data quality. If quality degrades below the agreed upon targets, the DQ steward will co-ordinate any interventions necessary to bring it back up again.

Following the data quality improvement projects, either through data collection methods, processing formula changes or system changes (e.g., implementing a drop-down selection box on a previously unstructured field), a periodic reassessment of the affected fields should be done to ensure that the changes had the desired impact. Reassessment using data profiling tools should be done to determine what impact the improvement has. The reassessment should target just the fields that were the improvement targets.

Once data improvements have been implemented, the DQ steward should regularly run the quality measurement queries to ensure that quality targets are being met. The response protocol if the targets are not met should be clear for all data entry staff. If the business processes change, the requirements against the data should be reassessed. DQ stewards should initiate another quality assessment. However, this may be smaller than the initial baseline assessment, with more specific objectives and may look at a subset of the data fields the initial baseline assessment examined.

## Business Engagement

Business subject matter experts must be involved in brainstorming practical steps to improve the data. They should take into account the impacts of the proposed activities on the business process. Business managers must be engaged in determining resources available to improve the data as well as identifying which proposed improvements would most benefit the business process and the user community.

## Output

* Set of prioritized recommendations for improving data quality and corresponding measure specifications for ensuring that data quality is sustained over time
* Approved project plan for improving data quality
* Dataset improvements with corresponding measures implemented for monitoring

## Gating

This stage is never complete, insofar as the DQ steward has an on-going responsibility to monitor data quality improvements. However, the DQ team can consider the stage complete when the data users agree to the prioritized set of recommendations that can then be implemented as separate data quality improvement projects. For each recommendation that addresses a gap between the current condition and its purpose, a quality monitoring measure type should be identified, and the formula for the measure should be specified.

#

# Appendix A: Roles and Responsibilities

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.

Note that job titles may differ from the roles used in the table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step | **Understanding the Data** | **Profiling Critical Fields** | **Assessing Fit for Purpose** | **Improvements & Monitoring** |
| Stage | 1a | 1b | 1c | 2a | 2b | 2c | 3a | 3b | 4a | 4b | 4c | 4d | 4e |
| 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 plan | Ongoing monitoring |
| **Business process 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** |  |  |  |  |  |
| **Data steward** | **C** | **R** | **A** | **I** | **C** | **C** | **I** | **R** | **R** | **C** | **S** | **A,R** | **A,R** |
| **Quality analyst** | **S** | **S** | **R** | **A,R** | **A,R** | **A,R** | **A,R** | **A** | **A** | **R** | **S** | **S** |  |
| **Data scientist** |  |  |  | **S** |  |  | **S** |  |  |  | **S** | **S** |  |

# Appendix B: PROJECT PLANNING FOR DATA QUALITY ASSESSMENT

A data quality assessment project takes time and planning. Each stage requires different resources and skills, and the level of effort for the DQ steward and other data roles will vary. Depending on the data quality tools available, some of the tasks can be automated. The following estimates of time required relate to a targeted DQA project (not a whole system!) and are provided to assist project planning and identify what resources the DQ steward and other data roles should draw upon.

**Stage 1 - Understanding the data**

This step should not take long. Depending on the size of the dataset identified and the scope of the user community, the project charter can be more or less formal, and the document can be shorter or more detailed (various departments currently use different project charter templates). While this stage has a consultation component, in getting agreement on the objective of the assessment project, this can often be done over the phone and/or through email. The project lead has responsibility for most of the work in this step.

**Figure B.1.1**

|  |
| --- |
| Table: Estimated Effort for Identifying Data Quality Assessment Project |
| **Task** | **Responsible** | **Effort** |
| Identify target dataset | Project lead  | 0.25 hour |
| Identify team | Project lead | 0.50 hour |
| Identify DQ steward | Project lead | 0.25 hour |
| Determine objective | All | 2 hours |
| Write charter | Project lead | 1 hour |
| Enter metadata | Project lead with DQA support staff | 0.25 hour |
| **Total** |  | 4.25 hours  |

**Figure B.1.2**

|  |
| --- |
| Table: Estimated Effort for Gathering Existing Documentation |
| **Task** | **Responsible** | **Effort** |
| Gather documents | DQ steward/ any project member  | 5 – 14 hours |
| Create links to documents | DQ steward/ any project member |  1 hour |
| **Total** |  | 6-15 hours  |

**Figure B.1.3**

|  |
| --- |
| Table: Estimated Effort for Initial Assessment |
| **Task** | **Responsible** | **Effort** |
| Create data model | Profiler |  2 hour |
| User consultation(s) | DQ steward/ any project member  |  4 -8 hours |
| Logging issues in metadata | DQ steward/ any project member |  1 hour |
| Data flow diagram | DQ steward/ any project member |  4 hours |
| Dataset lifecycle analysis | DQ steward/ any project member |  4 hours |
| Business terms to glossary | DQ steward/ any project member |  2 hours |
| **Total** |  | 21-25 hours  |

**Stage 2 – Profiling and Identifying Critical Fields**

The effort for initial profiling will be dependent on how many fields are in the dataset, and how many relationships it contains. It will also depend on the type of data within the dataset – it is often easier to determine which profile measures to use against numerical data than against non-numerical fields.

**Figure B.2.1**

|  |
| --- |
| Table: Estimated Effort for Profiling |
| **Task** | **Responsible** | **Effort** |
| Determining which profile measures to use against each field | Profiler | 1 hour |
| Running profile measures | Profiler | 2 hours |
| Adding profile results to metadata if not automated | Profiler | 1 hour |
| **Total** |  | 4 hours |

**Figure B.2.2**

|  |
| --- |
| Table: Estimated Effort for Metadata Completeness Profiling |
| **Task** | **Responsible** | **Effort** |
| Running metadata completeness routine | Profiler | 0.25 hour |
| Entering metadata completeness field | DQ steward/ profiler | 0.25 hour |
| **Total** |  | 0.50 hour |

**Figure B.2.3**

|  |
| --- |
| Table: Estimated Effort for Individual Data Elements |
| **Task** | **Responsible** | **Effort** |
| Field level metadata creation | Profiler  |  2 hours |
| Adding business description field | DQ steward |  2 hours |
| Consultation with users | DQ steward/ any project member |  4 hours |
| Adding criticality field | DQ steward/ any project member |  2 hours |
| **Total** |  | 10 hours  |

While understanding processing can be complex, it is focused on a few critical fields and involves relatively limited consultation.

**Figure B.2.4**

|  |
| --- |
| Table: Estimated Effort for Understanding Processing |
| **Task** | **Responsible** | **Effort** |
| Describing processing and capturing formulas | DQ steward and system administrator |  4 hours |
| Adding processing rules field to metadata | DQ steward |  2 hours |
| Adding level of processing characterization field to metadata | DQ steward |  0.5 hour |
| **Total** |  |  6.5 hours  |

**Stage 3 – Assessing Fit for Purpose**

**Figure B.3.1**

|  |
| --- |
| Table: Estimated Effort for Assess Profile Results |
| **Task** | **Responsible** | **Effort** |
| Analyzing profile results | Profiler and DQ steward | 2 hours |
| Documenting issues | Profiler and DQ steward | 1 hours |
| Adding issues/problems to metadata | Profiler | 1 hours |
| Putting analysis into a format that can be communicated to project team and users | DQ steward | 4 hours |
| **Total** |  | 8 hours |

This is the most intensive part of the data quality assessment process, involving the whole project team, and any additional immediate, downstream and external users.

**Figure B.3.2**

|  |
| --- |
| Table: Estimated Effort for Assess Fit for Purpose |
| **Task** | **Responsible** | **Effort** |
| Organize review meetings with users | Project team |  1 hour |
| Review profile results with users | DQ steward and project team | 12 hours |
| Documenting issues | DQ steward and project team |  1 hour |
| Adding expectations/problems to metadata | DQ steward and project team |  1 hour |
| Guidance document for open data  | DQ steward and open data coordinator |  1 hour |
| **Total** |  | 17 hours |

**Stage 4 – Data Improvements and Monitoring**

**Figure B.4.1**

|  |
| --- |
| Table: Estimated Effort for Document Future Measurement Recommendations |
| **Task** | **Responsible** | **Effort** |
| Review gaps and develop improvement recommendations | DQ steward and other data roles | 2 hour |
| Chose appropriate measures | DQ steward and other data roles | 1 hour |
| Modify appropriate measures to fit dataset | DQ steward and other data roles | 1 hour |
| Write up recommendations in metadata | DQ steward | 1 hour |
| **Total** |  | 5 hours |

**Figure B.4.2**

|  |
| --- |
| Table: Estimated Effort for Document Future Measurement Recommendations |
| **Task** | **Responsible** | **Effort** |
| Develop costing and resource requirements for improvements | Project team, DQ steward and other data roles | 7 hr. |
| Determine suitable measure metadata repository | ITS/DAVS  | 4 hr. |
| Review and prioritize improvements with users | DQ steward and other data roles | 4 hr. |
| Write up project plan | Project team | 3 hr. |
| Review project plan with users and get approval | DQ steward | 2 hr. |
| **Total** |  | 20 hrs. |

#

# Appendix C: METADATA CAPTURE

With an advanced data quality assessment application suite, which typically combines data profiling, metadata management, and data cleansing tools, a lot of metadata capture can be automated (these tools sometimes also include lineage and usage tracking, capturing who is using the database). Even in the absence of such tools, though, metadata capture at each stage of the data quality process is important – metadata provides the capacity for users and DQ stewards to see what is going on with the data over time.

The following tables indicate what kinds of metadata are captured at each stage on the process. Depending on the tools available, this capture will be manual or automated.

**Stage 1 – Understanding the Data**

Ideally metadata should be captured when a data quality assessment process starts. This ultimately helps subsequent data quality teams understand what has already been done to the dataset. This also includes determining the purpose of the project (domain), when the project will start (when the project charter was agreed upon), and identifying the DQ stewards (for continuity between projects addressing the same dataset).

This metadata can be captured in a metadata management tool. This would begin the automation of the DQA project. It should also be reflected in that data catalogue that all users have access to.

**Figure C.1.1**

|  |
| --- |
| Table: Metadata for Identifying Data Quality Assessment Project |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Project Name | Descriptive name so as not wholly reliant on York Region Data Catalogue ID | No |
| Project Type | Drop-down selection of Baseline assessment, specific improvement, periodic reassessment, open data | Yes |
| Date Project Started | So history of DQA activity is apparent | Yes |
| DQA Project Status | Enter Started when the project charter is agreed on. This field will be updated as the project progresses.  | Yes |
| DQ steward | Who is the DQ steward responsible for quality? Helps continuity, users know who DQ steward during this DQA activity was | Yes |
| DQ Project Team Members | Helps the next DQA activity | No |

Metadata to capture when gathering existing documentation is the link to the documents and the type of document.

**Figure C.1.2**

|  |
| --- |
| Table: Metadata for Gather Existing Documentation |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Document Type | The type of document, from a controlled domain of document types (see Existing Materials Checklist) | Yes |
| Link path to document | Hyperlink or path to where the document is stored | Yes |
| Date document prepared/revised | When was the document prepared/revised? This field can provide insight as to whether the document reflects changes made to the business process | No |

Entries are made to the issues log based on consultations with data users. Links to any additional documents created through this step are added to the metadata (see Step 2 for metadata field details).

**Figure C.1.3**

|  |
| --- |
| Table: Metadata for Initial Assessment |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| User Process | The immediate, downstream or external business process using the dataset | Yes |
| User name | A named user associated with the user process field above – can make subsequent follow-up easier | No |
| Quality Expectation | Expected levels of completeness, reliability (accuracy), timeliness, access | Optional |
| Issue | Perceived shortcomings in relation to expectations | Optional |
|  |  |  |

**Stage 2 – Profiling and Identifying Critical Fields**

Metadata captured during initial profiling forms the baseline from which improvements can be measured. This should be organized so that profiles from different dates can be compared.

**Figure C.2.1**

|  |
| --- |
| Table: Metadata for Profiling |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Field ID | The ID of the field being assessed | Yes |
| Profile date | Date the profile tool was run against the dataset | Yes |
| Profile measure 1 | Profile measure description, such as # of non-null values, % of null values, average value, % of missing relationships, cardinality | Yes |
| Profile result 1 | Result returned by a profile measurement as a standardized measure, e.g., ratio | Yes |
| Profile measure 2 | ,, | Yes |
| Profile result 2 | ,, | Yes |
| Profile measure x | ,, | Yes |
| Profile result x | ,, | Yes |

Metadata completeness measures should also be reflected in the data catalogue

**Figure C.2.2**

|  |
| --- |
| Table: Metadata for Metadata Assessment |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Metadata completeness score | Calculated measure out of 100, of how many entries have been completed and how meaningful the metadata is, providing data user with a sense of how reliable the metadata is | Yes |

Metadata for critical fields should also be reflected in the data dictionary. Some entries should surface in a business terms glossary.

**Figure C.2.3**

|  |
| --- |
| Table: Metadata for Individual Data Elements |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Table Name | An intelligible table name and machine alias | No |
| Table Purpose | What kind of data is held in this table? The field above – can make subsequent follow-up easier | No |
| Field Name | An intelligible field name and machine alias | Yes |
| Field Business Description /Definition | What real-world phenomena does the data in this field represent?  | Optional |
| Business Rules | What are the business rules for populating this field? Is it a required field, does it have restricted cardinality, etc. | No |
| Criticality | How critical is this field for users?  | Optional |

Internal processing rules, especially where not obvious to users, should be captured. Highly processed data is a potential source of error, so care should be taken to ensure the processing rules are captured correctly in the metadata. This will aid future investigations of data issues.

**Figure C.2.4**

|  |
| --- |
| Table: Metadata for Understanding Processing |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Field ID | The ID of the field being assessed | Yes |
| Processing rules | Description and formulas of how data in the field is processed | Yes |
| Level of processing | Does this field contain raw data, lightly or highly processed data? | Yes |

**Stage 3 – Assessing Fit for Purpose**

When systematic issues are found when assessing the profile results, these should be logged so they can be dealt with later.

**Figure C.3.1**

|  |
| --- |
| Table: Metadata for Assess Profile Results |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Field ID | The ID of the field being assessed | Yes |
| Problem 1 | Description of an issue detected within the field, or of a relationship the field participates in | Yes |
| Problem 2 | ,, |  |
| Problem x | ,, |  |

For each data quality target document the reason for the target relative to a business process, and an assessment of the difference between current condition and expected quality. This will help later prioritization. When an assessment has been completed, the data catalogue status can be changed to “governed”.

**Figure C.3.2**

|  |
| --- |
| Table: Metadata for Assess Fit for Purpose |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Field ID | The ID of the field being assessed | Optional |
| Expectation 1 | Detailed description of the expected quality of data within the field, or of a relationship the field participates in, relative to a specified business process | Yes |
| Difference Severity 1 | Assessment of difference between the current condition and expected quality | Yes |
| Expectation 2 | ,, |  |
| Difference Severity 2 | ,, |  |
| Expectation x | ,, |  |
| Difference Severity x | ,, |  |
| DQA Project Status | Change status to Baseline Completed when Assess Fit for Purpose step and all preceding steps have been completed.  | Yes |
| Governed | Add Governed as a status when the initial assessment is complete (there must also be a DQ steward) | Yes |

**Stage 4 – Data Improvements and Monitoring**

Each improvement suggestion should be captured as metadata, together with a measurement specification, and action threshold.

**Figure C.4.1**

|  |
| --- |
| Table: Metadata for Document Future Measurement Recommendations |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Field ID | The ID of the field being assessed | Optional |
| Recommended improvement | A specific action to be taken to improve data quality in one or more specified fields of the database | Yes |
| Measurement type | The monitoring tool that will be used to detect changes in the condition of the data after the improvement is made (see Appendix D for measurement types) | Yes |
| Measurement specification | Formula for calculating the measure, specifically tailored to the database and field being monitored | Yes |
| Acceptable level of deviation | Percentage of permissible errors  | Yes |
| Response protocol | Directions to staff on how to proceed if the measure results show greater than acceptable levels of deviation | Yes |

This step captures priorities against different improvement recommendations.

**Figure C.4.2**

|  |
| --- |
| Table: Metadata for Document Data Improvement Plan |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Field ID | The ID of the field being assessed | Optional |
| Recommended improvement | A specific action to be taken to improve data quality in one or more specified fields of the database | Yes |
| Priority  | A ranking relative to other recommended improvements | Optional |
| Implementation date | Future date the improvement will be implemented (modified when the improvement is completed)  | Optional |

# Appendix D: MEASURES FOR MONITORING DATA QUALITY

**What are measures for?**

Despite our best efforts to engineer automated data processing, things still go wrong. We put the effort into making a good dataset, spending the time to cleanse data of errors and retrain staff on data entry, and then all of a sudden it all goes bad, and it looks like it will take three months to fix it again. And we do not know why it went wrong.

Measures help us detect trends in data quality, and alert us to respond before data quality is completely compromised, hopefully avoiding a large-scale intervention to restore the expected level of quality. Implementing measurement controls and monitoring is a good way to avoid the risks and costs associated with bad data.

Defining measures to ensure data improvements are sustained over time is both an art and a science. The art is properly categorizing the issue that creates the problem and the solution to it. This helps select an appropriate measure or set of measures that can be layered together. The science is creating the precise query or process control that provides intelligible information about quality trends in the dataset.

**Figure D.0.1**

|  |
| --- |
| Table: 50 Different Measures for Monitoring Data Quality (derived from Laura Sebastian Coleman’s typology) |
| **Quality Dimension** | **Sustainment control or query** | **In-line/periodic** |
| Completeness | Metadata completeness | Periodic |
|  | Availability for processing | Inline |
|  | Record counts to control records | Inline |
|  | Record total to control totals | Inline |
|  | Record size compared to past sizes | Inline |
|  | Record expected length | Inline |
|  | Non-nullable fields | Inline |
|  | Date criteria | Inline |
|  | Reasonability based on date criteria | Inline/Periodic |
|  | Missing fields critical to processing | Inline |
|  | Reasons for rejecting records | Inline/Periodic |
|  | Balance record counts though a process | Inline |
|  | Reasonable balance of inputs and outputs | Inline |
|  | Balance amounts through a process | Inline |
|  | Ratio of summed amount fields | Inline |
|  | Defaults from derivation | Inline/Periodic |
|  | Overall sufficiency for defined purpose | Periodic |
|  | Overall sufficiency of dataset measures and controls | Periodic |
| Timeliness | Is data delivered on time? | Periodic |
|  | Timely availability of data for access | Periodic |
|  | Data processing duration | Inline |
| Consistency | Consistent formatting in one field | Periodic |
|  | Consistent formatting cross table | Inline/Periodic |
|  | Defaults | Periodic |
|  | Cross table defaults | Periodic |
|  | Field content defaults | Periodic |
|  | Consistent column profile | Periodic |
|  | Consistent multicolumn profile | Periodic |
|  | Consistent cross table multicolumn profile | Periodic |
|  | Chronology consistent with business rules | Periodic |
|  | Consistent time elapsed | Periodic |
|  | Chronology consistent with business rules across tables | Periodic |
|  | Distinct count of represented entities | Inline/Periodic |
|  | Ratio of distinct count to two represented entities | Inline/Periodic |
|  | Consistent amount field calculation across secondary fields | Inline |
|  | Consistent record counts by aggregated date | Periodic |
|  | Consistent amount field by aggregated date | Periodic |
|  | Consistent cross table amount field calculation | Inline |
|  | Consistent cross table amount field by aggregated date | Periodic |
|  | Consistency compared to external benchmarks | Periodic |
| Validity | Single field detailed validity check | Periodic |
|  | Single field aggregate validity check | Periodic |
|  | Multiple field validity check | Periodic |
|  | Cross table detailed validity check | Periodic |
| Integrity | De-duplication | Inline |
|  | De duplication reasonability | Inline/Periodic |
|  | Parent child referential integrity | Periodic |
|  | Child parent referential integrity | Periodic |

There are a lot of concepts behind measurement. There is the object of measurement (the data model, data is received, dataset content, data processing, cross-table content), the purpose of measurement (assessment, process control, monitoring), periodic versus in-line measurement, taking and processing the measurement, response protocols for when the result is not good.

Measures offer different results - some measures provide an absolute score of errors, others look at the reasonableness of data inputs. Some measures detect changes in the dataset over time, revealing patterns that would otherwise not be apparent. These measures require that operational metadata for measure results is available for calculating and comparing against historical means. Some sophisticated measures require control records be passed with source data that allow comparisons to ensure all the data expected is processed by the database system.

While it would be nice to reconfigure applications to incorporate process controls (largely the inline checks), the Region has many proprietary systems that did not incorporate data quality control mechanisms, and that do not enable easy changes to interfaces. We also receive data from sources that we have little control over. But we can build queries that read the data after it has been recorded in the dataset (largely the periodic checks).

The measures listed in this appendix have been adapted from different sources. The main source is Laura Sebastian-Coleman’s Measuring Data Quality for Ongoing Improvement. (2013, Morgan Kaufman). Although other authors identify dimensions of quality differently, this source provided the best objective measures for monitoring. Leo Pipino *et al*’s Data Quality Assessment (Communications of the ACM, April 2002) provided the model for some of the formulas used to illustrate how to implement the measures.

**D.1. Completeness measures**

**D.1.1 Metadata completeness (1)**

An initial formula for metadata completeness was developed for assessing critical field metadata.

A simple metadata completeness score thus might be $\left(^{b+c+d}/\_{a}\right)X e $ , where:

a. number of critical fields

b. # of non-null descriptions for critical fields

c. # of non-null business rules for critical fields

d. # of non-null formulas for critical fields

e. A subjective score of meaningfulness on a scale of 1-3, where

1 - Not clear what the fields are for

2 - Somewhat clear

3 - Very clear, no ambiguity

This provides a metadata completeness score out of 100. The measure provides a basic sense that attention is being paid to database management. This formula does not give credit for metadata before it gets into assessment and formal governance. That metadata could be called un-scored.

A more mature metadata completeness score could use a richer formula for metadata completeness that includes critical field metadata as one element weighed among others:

* Basic Dublin Core equivalent metadata (current Data Catalogue standard) 20%
* Critical field metadata 50%
* Usage and Limitations 20%
* Data dictionary 10%

This more mature measure of dataset completeness would give users a better measure of metadata utility. But it is also harder to maintain.

**D.1.2 Availability for processing (7)**

This is a process control that measures receipt of data inputs and is useful where there are several data sources that have to be amalgamated (e.g., inputs from nine local municipalities). If not all the data inputs are available for processing, users may not be aware that the resulting amalgamated dataset is incomplete. To set up this measure, the different dataset inputs need to be named, so they can be flagged as missing.

**Data quality threshold and response protocol**

In the event that an input is missing, add a flag to the dataset metadata to indicate that it is incomplete. For some processes, this may result in delaying further processing and report generation until the missing input is available. Reference – LSC pp.232-233.

**D.1.3 Record counts to control records (8)**

This in-line measure checks that all the records expected were delivered from one system to another. It is always possible that an input sent from another system will not contain all the information it should. Much data is sent through queries. Receiving a second query result that is simply the count of records that were sent enables a check on the completeness of the input before the input is processed into the receiving system.

**Data quality threshold and response protocol**

In the event of a discrepancy between the record count and the number of records received in the data delivery, staff should contact the data input source to request a re-delivery of the data.

 Reference – LSC pp.233-234.

**D.1.4 Record total to control totals (9)**

Similar to record counts to record controls, a second control file can be sent from the source system that queries a critical field total, such as the dollar total. If the dollar amount in the records received, does not match the amount in the control record, staff should contact the data input source to request a re-delivery of the data.

 Reference – LSC pp.234-235.

**D.1.5 Record size compared to past sizes (10)**

Many processes are regular enough that they contribute a similar number of records from period to period, give or take a few. This in-line measure provides a reasonableness test that will often be more applicable in the York Region context than getting control files from local municipal systems. Unlike the control file based measures, it requires operational metadata to be maintained for the number of records received in each period. To detect anomalies, compare the count of records received this period to the median count of records received in previous periods.

**Data quality threshold and response protocol**

Where the count of records received in this period deviates more than two or more standard deviations from the median of records received in previous periods, examine the data. Look for truncations, missing records, or if the deviation is on the higher side, look for duplicate records. Contact the data input source to determine if the anomaly was expected at their end (e.g., “the records show much less/more activity than usual. Was that actually the case?”). If it was not, request a re-delivery of the data.

 Reference – LSC pp.235-237

**D.1.6 Record expected length (11)**

Much as a data delivery can be received that is missing records, it can also be missing columns in a table. This in-line process measure is intended to detect missing fields. Before adding a received record into a regional database, the record would be checked programmatically, depending on its format, that it contained the right number of fields, the right number of commas, or if all the field definitions were present in an XML file. These kinds of checks can be incorporated into load sequences, or into ETL scripts, that provide notice when fields are missing.

**Data quality threshold and response protocol**

The scripted process needs to notify staff to investigate anomalies, and potentially request a re-delivery or investigation of the data from the source system. In some cases, the script would stop the data load process.

 Reference – LSC pp.237-238

**D.1.7 Non-nullable fields (12)**

Some fields allow null values, for other fields only non-null values are permissible. This in-line control measure reduces the risk of incomplete data by stopping processing of nulls in mandatory fields. Ideally, this is a process control. Where controls cannot be added to a proprietary system, the measure can be used as a frequently run query.

For mandatory fields, establish an allowable level of null values. Create a query that measures

$$1-\frac{Count of nulls in period}{Count of records in period}$$

**Data quality threshold and response protocol**

When the threshold (e.g., 95%, 99% or 99.5%) for an acceptable level of nulls is not met, the staff is notified to investigate causes and make appropriate corrections.

 Reference – LSC pp.238-239

**D.1.8 Date Criteria (16)**

Date fields are important for data processing, as well as for their content. This in-line process control measurement compares the content of the date fields with criteria established in the metadata that provides minimum and maximum expected dates (e.g., data received monthly can be expected to have data from the previous month - you do not expect to see a record dated two years prior, or two months into the future). Dates meeting the criteria are valid for loading to the target dataset. This type of “date” control can be incorporated into load sequences, or into ETL scripts, that provide notice when dates are outside of the expected timeframe.

**D.1.9 Reasonability based on Date Criteria (17)**

This measure compares the number of “date” field records that do not meet expectations with the historical sequence of past measurements.

To detect anomalies, compare the count of records outside the expected date ranges received this period to the median count of records outside the expected date ranges received in previous periods.

**Data quality threshold and response protocol**

If the current count of records outside the expected date ranges received this period exceeds three or more standard deviations of the median count of records outside the expected date ranges received in previous periods, investigate the causes.

**D.1.10 Missing Fields Critical to Processing (18)**

Data from a source system can be imported to a target system with an allowable number of errors, or records can be inspected before they are processed, and those records with errors can be rejected before they are processed. Pre-inspection avoids having errors in critical fields that might create problems for downstream processing and reports, albeit at a cost of having some records missing. Where records are rejected, it is worth tracking how many records are rejected over time so the impact of their absence can be gauged, and to detect changes in the data sources.

Create a query that that provides a ratio of the number of rejected records for each period to the number of accepted records for the period. Normalize it as

$$1-\frac{Count of rejected records in period}{Count of accepted records in period}$$

Compare this with the median of past ratios from the operational metadata.

**Data quality threshold and response protocol**

If the current ratio of rejected to accepted records exceeds three or more standard deviations of the median ratio of rejected to accepted records received in previous periods, or an alternate threshold determine by subject matter experts, investigate the causes. This measure can be an effective way of showing improvements in source systems, as the index ratio improves over time. Whenever the index ratio declines, the causes should be investigated.

 Reference – LSC p.247

**D.1.11 Reasons for Rejecting Records (20)**

Data can be rejected in initial inspection, and can also be rejected as it is being processed. If data is tagged for the reason it is rejected, this information can be used to analyze patterns of data rejection. These patterns can be used to inform data improvements in source systems.

 Reference – LSC pp.249-250

**D.1.12 Balance Record Counts through a Process (19)**

Record balancing is a process control measure that reconciles data process inputs with data process outputs, to detect if anything went amiss during processing that resulted in records being deleted or duplicated.

Implementing this in-line measurement requires a routine to count the records at the beginning of a process, and minimally at the end of the process. Counts can also be taken at intermediate steps in data processing if there is a need to isolate where problems occur.

**Data quality threshold and response protocol**

Once implemented, this process control can run without support. Unless there is a mismatch, processing can be assumed to be running smoothly. If a mismatch is reported either during processing or at the conclusion, the processing routines will require fixing and testing.

 Reference – LSC pp.248-249.

**D.1.13 Reasonable Balance of Inputs and Outputs (21)**

Where records are being combined or split during processing, it is not possible to use definitive balance record processes to ensure that nothing went amiss during processing. Given stable systems, however, the expectation that the ratio of inputs to outputs will be consistent is reasonable. This measure compares the current ratio of input records to output records with the historical medians ratios at different points in the process.

**Data quality threshold and response protocol**

If the current ratio of input to output records exceeds three or more standard deviations of the median ratio of input to output records in previous periods, or an alternate threshold determine by subject matter experts, investigate the causes.

 Reference – LSC pp.250-251

**D.1.14 Balance Amounts through a Process (22)**

Data is sometimes unexpectedly dropped during complex data processing. Like balance record counting, balancing amounts is a process control measure that reconciles data process inputs with data process outputs, to detect if anything went amiss during processing. But it detects changes during processing to numeric fields and allows a focus on critical information.

Implementing this in-line measurement requires a routine to sum across one field at the beginning of a process, and minimally at the end of the process. Sum totals can also be taken at intermediate steps in data processing if there is a need to isolate where problems occur. Rejected records can be part of the balancing, reconciling three amounts instead of two.

**Data quality threshold and response protocol**

Once implemented, this process control can run without support. Unless there is a mismatch, processing can be assumed to be running smoothly. If a mismatch is reported either during processing or at the conclusion, the processing routines will require fixing and testing.

 Reference – LSC pp.251-253.

**D.1.15 Ratio of Summed Amount Fields (23)**

Where records are being combined or split during processing, it is not possible to use definitive balance amount processes to ensure that nothing went amiss during processing. Given stable systems, however, the expectation that the ratio of inputs to outputs will be consistent is reasonable. This measure compares the current ratio of input amounts to output amounts with the historical medians ratios at different points in the process.

**Data quality threshold and response protocol**

If the current ratio of input to output amounts exceeds three or more standard deviations of the median ratio of input to output amounts in previous periods, or an alternate threshold determine by subject matter experts, investigate the causes. Reference – LSC pp.253-255.

**D.1.16 Defaults from Derivation (24)**

This measure tracks derived default values to see how frequently they are being used. It monitors risk in data processing stability. Derived data will be internally generated according to a set of defined rules that typically rely on inputs from multiple sources. At the end of the set of defined rules, there is usually a clause that allows for the possibility that the data meets none of the stipulated conditions. Given stable systems, however, the expectation that the incidence of derived defaults should be consistent is reasonable. If there is a deviation from the expected incidence, it may reflect changes in the business process for the inputs or unintended changes in the processing environment.

Create a query that counts the frequency of derived default values. Normalize this as

$1-\frac{count of defaults in a period }{count of records in a period}$ .

Compare this measure with the historical median of all previous results from this measure.

**Data quality threshold and response protocol**

Data quality thresholds for derived default value frequency are set as two or more standard deviations from the historic average. This sets a level of reasonableness, allowing for some variation around the mean. Changes in the frequency of derived default value usage are not necessarily bad and may reflect an underlying change to the business process. If so, investigate, and document the business process changes in the metadata. However if there has been no change in the business process, investigate the processing environment carefully to determine if the deviation is the unintended consequence of system changes. There may be other unintended consequences elsewhere in the data.

 Reference – LSC pp.255-256.

**D.1.17 Overall Sufficiency for Defined Purpose (47)**

This is a periodic review of the entire dataset to determine if it still fits business needs. Do the data fields still address the business purposes. Ideally, data users will identify any short-comings in relation to specific business needs changing, but a formal periodic review ensures that their broader changing business requirements are not overlooked. A dataset profile, including the number of records present in the dataset, time periods covered, etc. should be conducted to inform the review.

**Data quality threshold and response protocol**

If shortcomings in the dataset relative to changing business needs are detected in consultation with dataset users, any gaps in the dataset should be addressed, by adding appropriate fields for the business purpose. This may entail altering data collection methods, protocols, and tools.

 Reference – LSC pp.296-297.

**D.1.18 Overall Sufficiency of Dataset Measures and Controls (48)**

This is a periodic review of the quality measurement regime. Unlike other measures, this is directed not at the data itself, but at the metadata generated by other measures. It entails a periodic review of the adequacy of each of the measures chosen to monitor data quality. Are they comprehensive and acting as effective quality controls? If they are not, the measures themselves should be altered. This may be due to changing business processes, critical fields were not targeted for measurement, or because the measures initially were chosen do not appear to be providing sufficient insight into the quality of the dataset.

**Data quality threshold and response protocol**

Based on a review of the how each measure is constructed, and the metadata each measure produces, a determination of adequacy is made, in relation to the effectiveness of the measure. Is it producing the insight necessary for ensuring data completeness, timeliness, consistency, validity or integrity at the agreed-upon quality threshold? When a measure does not, the measure should be re-formulated and tested. Reference – LSC pp.297-298.

**D.2. Timeliness measures**

**D.2.1 Is Data Delivered on Time? (6)**

If data is expected on a schedule (e.g., monthly delivery) the delivery date can be compared with the scheduled date as an in-line measure to monitor data delivery. This is a useful measure when downstream processes are impacted by late delivery.

In the metadata, define the expected delivery period. Define an acceptable level of lateness. Each time data is received (or delivered), record the date of actual delivery. The measure is the median of past differences between expected delivery and actual delivery.

**Data quality threshold and response protocol**

If the current data delivery date exceeds three or more standard deviations of the previous median of differences between expected and actual delivery dates, investigate to uncover the causes, and remediate to ensure that this will be less likely to occur in the future.

Reference LSC pp.229-232

**D.2.2 Timely Availability of Data for Access (26)**

This measure compares the time data is posted to a store where it is accessible to data users at the time it was scheduled to be available. It can be applied to the whole database, or to specific tables that are of importance to data users. In the metadata, define expected dates data that will be available to users. Define an acceptable level of lateness. Each time data is posted for user access, record the date of posting. The measure is the median of past differences between expected posting date and actual posting date.

**Data quality threshold and response protocol**

If the current data delivery date exceeds three or more standard deviations of the previous median of differences between expected and actual delivery dates, investigate to uncover the causes, and remediate to ensure that this will be less likely to occur in the future. The measure can be used to trigger a message to users when data will be posted late.

Reference: LSC pp.259-260.

**D.2.3 Data Processing Duration (25)**

Data processing takes time. Differences in how long it takes to run similar data processing tasks may indicate problems with the data, or a problem with data processing task itself. This measure can be complicated by other processes running in the same environment and requires a detailed understanding of the processing environment to be fully leveraged.

To implement this measure, determine the programmatic beginning and end points in the process and from these calculate task duration. Compare this period’s task duration against the average of previous period’s task duration.

**Data quality threshold and response protocol**

If the current data processing task duration exceeds three or more standard deviations from the previous average of task duration, investigate to uncover the causes, and remediate.

 Reference: LSC pp.257-259.

**D.3. Validity measures**

**D3.1 Single Field Detailed Validity Check (27)**

If a domain of valid values is contained in a reference table, values in the database can be compared against this table to ensure they are valid. The measure returns the frequency of records that match each value in the domain reference table, as well as the frequency of invalid values. The measure can then be used to detect change over time in the level of validity (for fields where validity is extremely critical, use this measure as an intake check before the values are accepted into the database). Create a query that returns the frequency usage of each valid value contained in the reference table, and the frequency of each value not contained in the reference table. Note that the query logic is similar to that for the Consistent Column Profile Check, so this measure can be readily combined with a consistency check. Compare each value frequency with the historical results from this measure to detect changes in validity.

**Data quality threshold and response protocol**

If the value frequency for any valid domain value exceeds three or more standard deviations of the value frequency for the same valid domain received in previous periods, investigate the causes. Changes in the frequency of valid domain values aren’t necessarily bad and may reflect an underlying change to the business process. If so, investigate, and document the business process changes in the metadata. However if there has been no change in the business process, investigate to understand where the change is coming from, and take appropriate action. Where a larger number of invalid values are being detected, the domain of valid values may need to be reconsidered and adjusted.

 Reference – LSC pp.261-264.

**D.3.2 Single Field Aggregate Validity Check (28)**

This measure summarized the results of the Single Field Detailed Validity Check as a percentage of valid and invalid values. Building on the query from the Single Field Detailed Validity Check, calculate the percentages of valid and invalid values. Compare the current results with the historical results from this measure to detect changes in validity.

**Data quality threshold and response protocol**

The project team should decide on a validity compliance target with an acceptable level of deviation (e.g.., 95%, 99%, 99.5%). If the percentage of valid domain values is below the target, investigate the causes. The investigation should start with the detailed results from the Single Field Detailed Validity Check, to understand the pattern of invalid values, and the trend of values generally. Changes in the distribution of valid and invalid values may reflect an underlying change to the business process; if so, document the business process changes in the metadata, and adjust the domain of valid values as necessary. However if there has been no change in the business process, investigate to understand where the change is coming from, and take appropriate action.

 Reference – LSC pp.264-266.

**D.3.3 Multiple Field Validity Check (29)**

Relationships between fields can be examined for validity. For example, fields for postal code and province may only contain valid values. But the relationship between a given postal code and the corresponding value in the province field may be invalid. Complex queries may be required to measure across multiple fields to monitor multiple field validity within a table.

**Data quality threshold and response protocol**

The project team should decide on a validity compliance target with an acceptable level of deviation (e.g.., 95%, 99%, 99.5%). If the percentage of valid domain values is below the target, investigate the causes. The query may also be written to flag records where validity is in question.

 Reference – LSC pp.266-267.

**D.3.4 Cross Table Detailed Validity Check (41)**

Relationships between fields in related tables can be examined for validity. Cross table validity checks are used for complex relationships reflecting complex business rules. Complex queries with table join may be required to measure across multiple fields to monitor multiple field validity within a table.

**Data quality threshold and response protocol**

The project team should decide on a validity compliance target with an acceptable level of deviation (e.g.., 95%, 99%, 99.5%). If the percentage of valid domain values is below the target, investigate the causes. The query may also be written to flag records where validity is in question.

 Reference – LSC pp.291-292.

**D.4. Consistency measures**

**D.4.1 Consistent formatting in one field (2)**

This measure is useful when data imports are coming from several sources that may vary over time, or when several people are entering data (e.g., getting data from 9 local municipalities or staff changing).

Define rules for formatting the field, such as data type, and level of precision (e.g., thousands, one or two decimal places, specific abbreviations to use). Add this definition to the metadata. Profiling tools can recognize inconsistent formatting. Or write a SQL query that will detect deviation from the defined format. Normalize the measure as

 $1- \frac{count of deviations}{count of all records}$.

Run the query periodically, matching the frequency of inputs. Record the results in metadata to enable trend detection.

**Figure D.4.1.1**

|  |
| --- |
| Table: Metadata for Consistent Formatting in One Field Measure |
| **Metadata**  | **Description and rationale** | **Mandatory field(s)** |
| Data Catalogue ID | The ID of the dataset being assessed | Yes |
| Field ID | The ID of the field being assessed | Yes |
| Measure | The measure being taken | Yes |
| Date | Date the measure was taken | Yes |
| Result | Normalized result of the measure | Yes |

**Data quality threshold and response protocol**

The project team should decide on a compliance target with an acceptable level of deviation (e.g.., 95%, 99%, 99.5%). This will partly be dependent on whether formatting influences how the field is processed after data entry. If deviations larger than the acceptable level are detected after a data improvement, it may reflect a change in business processes from one of the data sources. Investigate and take appropriate corrective action.

 Reference - LSC pp.225-227.

**D.4.2 Consistent formatting cross table (3)**

Fields of the same type within a dataset should have the same formatting rules. Where inputs from multiple sources may vary over time, there is a chance that one or more sources may change formats unannounced on one or more fields. This measure looks across fields. For fields that should have the same formatting, develop a query to check format as in Consistent Formatting in one Field above. Run the query against each field that should share the format. This will provide a series of normalized measures. Take the average of the measures to index consistent formatting cross table.

This measure is generally part of an initial assessment. After an improvement, it is best to set up an inline measure against received data as it is processed. When this is not feasible, use this measure periodically. Reference – LSC p.227

**D.4.3 Defaults (4)**

Checking whether default values are being applied correctly where data is missing or unavailable helps maintain database consistency. Permissible default values should be recorded in the metadata field business rules. Use a profiling tool or develop a query that isolates non-permissible default values (e.g., search for “0” when “–“ has been defined as the permissible default) from permissible values. Normalize the measure as

$1- \frac{count of non-permissible default values}{count of all default values}$.

**Data quality threshold and response protocol**

Problems in defaults are typically caused by inconsistent data entry. Where the level of non-permissible default values climb past the threshold, determine which data entry staff are creating the errors, and train them to use permissible default values only.

 Reference – LSC p.227

**D.4.4 Cross table defaults (5)**

Fields with similar formatting should have similar default values to avoid confusing data users. If this is an issue for the dataset, it should be dealt with during a data improvement project. For simplicity, run a query as developed in Default measure above against different fields and compare them.

 Reference – LSC pp.228-229.

**D.4.5 Field content defaults (15)**

This measure tracks valid default values to see how frequently they are being used. This is an important trend measure for critical fields that allow default values, as default values are valid, and other tests will not detect subtle changes that default values can introduce to a dataset. Changes over time may indicate business process changes or changes in the phenomena that is being captured. Create a query that counts the number of times permissible defaults are used. Normalize this as

$1-\frac{count of defaults in a period }{count of records in a period}$ .

Compare this measure with the historical mean of all previous results from this measure.

**Data quality threshold and response protocol**

Data quality thresholds for default value frequency are set as two or more standard deviations from the historic average. This sets a level of reasonableness, allowing for some variation around the mean. Changes in the frequency of default value usage are not necessarily bad and may reflect an underlying change to the business process. If so, investigate, and document the business process changes in the metadata. However if there has been no change in the business process, investigate to understand where the change is coming from, and take appropriate action.

 Reference – LSC pp.241-243.

**D.4.6 Consistent Column Profile (30)**

For stable processes, a consistent value distribution is expected for any field over time. This measure examines the value frequency distribution and compares it with previous value frequency distributions to detect abnormal deviations. This measure uses the same query logic as the Single Field Detailed Validity Check, and so it can be readily combined with a validity check. The measure can then be used to detect change over time in the value frequency distribution. Create a query that returns the frequency usage of each value in the field. Compare each value frequency with the historical results from this measure to detect changes in validity.

**Data quality threshold and response protocol**

If the frequency of any value exceeds three or more standard deviations of the frequency for the same value received in previous periods, investigate the causes. Changes in the frequency of values are not necessarily bad and may reflect an underlying change to the business process. If so, investigate, and document the business process changes in the metadata. However if there has been no change in the business process, investigate to understand where the change is coming from, and take appropriate action.

 Reference – LSC pp.267-270.

**D.4.7 Consistent Multicolumn Profile (33)**

Where there are business rules that define relationships between fields within a table, it is possible to measure the frequency distribution of related sets of values between the fields. This frequency distribution can be compared with previous frequency distributions to detect drift over time. Because this measure is complex and can result in a lot of results to analyze, filters can be applied to look at specific sets of values. Create a query that looks at the relationship frequency between values in two or more fields, resulting in a frequency distribution for a set of values across the fields. Compare the results with the frequency distribution in previous periods.

**Data quality threshold and response protocol**

If the frequency for any set of values exceeds three or more standard deviations of the frequency for the same set of values received in previous periods, investigate the causes. Changes in the frequency of sets of values aren’t necessarily bad, and may reflect an underlying change to the business process. If so, investigate, and document the business process changes in the metadata. However if there has been no change in the business process, investigate to understand where the change is coming from, and take appropriate action. Where this measure is built against critical fields as part of the acceptance testing of data inputs, exceeding three standard deviations may stop the data load until the causes have been investigated.

 Reference – LSC pp.274-277.

**D.4.8 Consistent Cross Table Multicolumn Profile (42)**

This measure is similar to the Consistent Multicolumn Profile, but targets business rules that create relationships between fields in different tables. It measures the frequency distribution of sets of values across tables. Create a query with table joins that returns the frequency distribution of sets of values across fields in different tables. Compare the results with those for previous periods. Filters can be used to reduce the number of returns for easier analysis.

**Data quality threshold and response protocol**

If the frequency for any set of values exceeds three or more standard deviations of the frequency for the same set of values received in previous periods, investigate the causes. Changes in the frequency of sets of values aren’t necessarily bad, and may reflect an underlying change to the business process. If so, investigate, and document the business process changes in the metadata. However if there has been no change in the business process, investigate to understand where the change is coming from, and take appropriate action.

 Reference – LSC pp.292-293.

**D.4.9 Chronology Consistent with Business Rules (34)**

Business processes have chronological constraints that imply business rules – one action cannot occur before a previous action. Where different date fields are used to record when different sequential actions occurred, the consistency of chronology can be measured. Create a query that measures the ratio of dates being in the expected order. This provides a measure of the level of illogical dates that can be compared with the ratios from previous periods for consistency.

**Data quality threshold and response protocol**

Where explicit or implied business rules for process chronology are violated (e.g., an application cannot be approved in less than two days after it has been received), investigate the causes for erroneous data entry.

 Reference – LSC pp.277-278.

**D.4.10 Consistent Time Elapsed (35)**

This measure calculates a difference between two date fields, in terms of years, months, days, hours, minutes or seconds, depending on the nature of the business process. Within a stable business process, there should be some consistency, for example, on the time elapsed between a work order being created and the work being completed. Create a query that measures the frequency distribution of time elapsed between sequential dates in a business process. The results can be compared with the frequency distribution from previous periods for consistency over time.

**Data quality threshold and response protocol**

For business processes where timelines are regular and expected, thresholds can be established. Differences in relation to the thresholds may indicate problems in data entry. For business processes where the timelines are less regular, comparing the frequency distribution with previous periods may indicate a change in the conditions the business process operates in, or a data entry issue, both warranting investigation.

 Reference – LSC pp.279-281.

**D.4.11 Chronology Consistent with Business Rules Across Tables (43)**

This measure tests for the same issues as the Chronology Consistent with Business Rules measure, but does so across tables, requiring a table join as part of the query.

**Data quality threshold and response protocol**

Where explicit or implied business rules for process chronology are violated (e.g., an application cannot be approved in less than two days after it has been received), investigate the causes for erroneous data entry.

 Reference – LSC pp.293.

**D.4.12 Distinct Count of Represented Entities (31)**

This measure compares the distinct count (the number of unique values) to a threshold, historical counts or total records. It provides a measure of reasonability for transactional data before it is loaded into a database, on the basis that a stable business process will give similar results in different periods, allowing for the current period to be compared with previous periods. Using this measure requires a definition of what a distinct entity within the database is. Create a query that counts the distinct number of entities present in the dataset (e.g., unique customers). The query results are compared with results from previous periods, comparing either raw numbers with previous averages, percentage of median, or a ratio of the distinct entities to the total number of records (e.g., ratio of unique customers to total number of orders).

**Data quality threshold and response protocol**

Thresholds can be defined by subject matter experts or as three standard deviations from the median. If the thresholds are exceeded, the data load is paused until an investigation shows the causes.

 Reference – LSC pp.270-272.

**D.4.13 Ratio of Distinct Count of Two Represented Entities (32)**

Consistency in relation to past inputs can be applied to the relationship between two entities. For example, the ratio of traffic tickets issued to the number of officers issuing traffic tickets might be expected to be consistent over time. This reasonability measure can be used to check incoming data against previous periods before it is loaded into the data store. Define the distinct entities. Create a query that provides a ratio of the number of distinct entities in one field with the number of distinct entities in a second related field. Compare the ratio with the average ratio from previous periods.

**Data quality threshold and response protocol**

If the ratio exceeds three standard deviations from the average ratio from previous periods, investigate the causes before loading the new data to the data store.

 Reference – LSC pp.272-274.

**D.4.14 Consistent Amount Field Calculations across Secondary Fields (36)**

This in-line measure looks at consistency over time across an amount field and one or more distinct entities in other fields. For example, the dollar value (expressed either as a total or as a percentage) against different types of construction can be compared against previous periods. Define the distinct entities in the secondary field. Create a query that sums the amount field (for the denominator), and also tabulates the amounts against the distinct entities. Normalize these results as percentages of the total amount, and/or calculate the average amount per distinct entity. Compare against average percentage mix and/or average amount per entity from previous periods.

**Data quality threshold and response protocol**

A threshold of greater than three standard deviations from the mean of percentages or amounts from previous periods can be used. If the threshold is exceeded, investigate the causes of the deviation and make any necessary corrective actions before loading to the data store. Reference – LSC pp.281-284.

**D.4.15 Consistent Record Counts by Aggregated Date (37)**

This periodic measure counts records by date ranges to reflect the business process. For example, the count of ambulance call-outs dealt with in a month. This count can then be compared with previous periods, or, if there is seasonality to the business process, to the same period in previous years. The date range of aggregation should be defined meaningfully in relation to the business process so that short term volatility is not a factor in the comparison. Create a query that aggregates record counts by date. Compare with the mean from previous periods, taking seasonality into account if necessary.

**Data quality threshold and response protocol**

Significant changes may indicate missing or duplicate records. Thresholds of three or more standard deviations from the mean of previous date aggregation periods should be investigated.

 Reference – LSC pp.284-286.

**D.4.16 Consistent Amount Field by Aggregated Date (38)**

This periodic measure totals amounts by date ranges to reflect the business process. For example, the total fees collected during a four week period. This amount can then be compared with previous periods, or, if there is seasonality to the business process, to the same period in previous years. The date range of aggregation should be defined meaningfully in relation to the business process so that short term volatility is not a factor in the comparison. Create a query that aggregates an amount field by date. Compare with the mean from previous periods, taking seasonality into account if necessary.

**Data quality threshold and response protocol**

Significant changes may indicate missing or duplicate records. Thresholds of three or more standard deviations from the mean of previous date aggregation periods should be investigated.

 Reference – LSC pp.286-288.

**D.4.17 Consistent Cross Table Amount Field Calculations (44)**

This in-line measure looks at consistency over time across one or more amount fields and one or more distinct entities in other fields, where the data is in related tables. For example, the dollar value (expressed either as a total or as a percentage) against different types of construction can be compared against previous periods. Define the distinct entities in the secondary field. Create a query or set of queries with table joins that sums the amount field (for the denominator), and also tabulates the amounts against the distinct entities. Normalize these results as percentages of the total amount, and/or calculate the average amount per distinct entity. Compare against average percentage mix and/or average amount per entity from previous periods.

**Data quality threshold and response protocol**

A threshold of greater than three standard deviations from the mean of percentages or amounts from previous periods can be used. If the threshold is exceeded, investigate the causes of the deviation.

 Reference – LSC pp.293-294.

**D.4.18 Consistent Cross Table Amount Field by Aggregated Date (45)**

This periodic measure totals amounts by date ranges to reflect the business process, where the data are contained in related tables or systems. For example, the total fees collected for two related processes during a four week period. This amount can then be compared with previous periods, or, if there is seasonality to the business process, to the same period in previous years. The date range of aggregation should be defined meaningfully in relation to the business process so that short term volatility is not a factor in the comparison. Create a query with table joins that aggregates one or more amount fields by date. Compare with the mean from previous periods, taking seasonality into account if necessary.

**Data quality threshold and response protocol**

Significant changes may indicate missing or duplicate records. Thresholds of three or more standard deviations from the mean of previous date aggregation periods should be investigated.

 Reference – LSC pp.294-295.

**D.4.19 Consistency Compared to External Benchmarks (46)**

External benchmarks from standards organizations or municipal peers can be helpful in assessing data quality as well as the efficiency of business processes. Column or multicolumn profiles are typical measures for comparison with external benchmarks. While benchmarks from standards organizations should be based on validated data, caution should be taken those benchmarks from peer organizations as these have not necessarily been validated consistently.

**Data quality threshold and response protocol**

Significant deviations from benchmarks should be investigated and the causes understood. Where the causes relate to data quality, corrective actions should be taken.

 Reference – LSC pp.295-296.

**D.5. Integrity measures**

**D.5.1 De-duplication (13)**

This inline measure is a process control that detects and removes duplicate records from an input source (usually a transactional system that allows multiple instances of a record) before the records are brought into a database. This requires programming to detect unique records, and choose the most appropriate record to input to the database, often the record with the most recent timestamp. These kinds of checks can be incorporated into load sequences, or into ETL scripts.

 Reference – LSC pp.239-240

**D.5.2 De-duplication reasonability (14)**

Where there is an expectation of consistency in the number of duplicate records from a transactional input source that contains multiple instances of the same record, duplication can be monitored over time to detect anomalous changes in the source data. If the process control developed for de-duplication is configured to count in each period how many duplicate records are rejected, a historical comparison can be created.

**Data quality threshold and response protocol**

If the current number of duplicate records rejected exceeds three or more standard deviations of the previous median of duplicates rejected, contact the source system

 Reference – LSC pp.240-241

**D.5.3 Parent Child Referential Integrity (39)**

This measure detects orphan records. Referential integrity is the degree to which data in two or more tables related through foreign keys are complete. Such relationships may be between fact tables (e.g., an order without a customer), or between a fact table and a reference table (e.g., allowable domain values). Validity checks will take account of relationships between fact tables and reference tables, but relationships between fact tables may have anomalies caused by timing issues in data entry.

Create queries to detect orphan records as a percentage of total parent records for critical relationships.

**Data quality threshold and response protocol**

Run the queries periodically, and compare them with historic values for previous periods. If there is drift away from tolerable levels established by the data users, investigate to determine the cause.

 Reference – LSC pp.288-290.

**D.5.4 Child Parent Referential Integrity (40)**

Where there is an expectation that a parent record will have at least one child record, and the timing of data entry is not simultaneous, periodic integrity checks can be used to assess the degree of completeness. Create queries to detect parent records not having child records as a percentage of total parent records for critical relationships.

**Data quality threshold and response protocol**

Periodically run the queries, and compare the results with historic values for previous periods. If there is a drift away from tolerable levels established by the data users, investigate to determine the cause.

 Reference – LSC pp.290.

# Appendix E: HINTS ON GETTING THROUGH STAGES MANAGEABLY

**Stage 1 – Understanding the data**

**E.1.1 Identify Data Assessment Project**

* A data quality assessment project should be defined by at least one business process: the business drives data quality requirements (significant data is often related to several business processes). In a few cases, it may not be immediately apparent how data connects to a business process; if so, hold off spending time on quality assessment until the relationship is apparent.
* The project may look at all data in an application, a subset of it, or a collection of related datasets: be careful not to confuse an application with a business process – while they are often consistent, sometimes applications are tracking more than one business process.
* Remember that data quality is all about expectations, so relate to the business process at the branch, department, and enterprise and community levels: do not be tempted to confine data quality assessment just to the branch that captures the data.
* Start with the obvious stakeholders - others may emerge as the assessment progresses: While immediate users of data are usually obvious because they work within the branch creating the data and have access to the source system, downstream users are not always as obvious. Engage the downstream users you know when determining the scope and type of assessment, but do not forget there are possibly enterprise and external users whose perspectives are also important – tracing the information flows will help find them.

**E.1.2 Gather Existing Documentation**

Any missing documents will be created during the assessment process: keep the time needed for this step short by focusing on finding existing documentation. Do not attempt to fill the gaps by creating missing documents at this point – it will be more efficient to do it later when more of the user community is engaged during the initial dataset assessment and profiling tasks.

**E.1.3 Initial Dataset Assessment**

* The downstream audience is not always considered in application design, so downstream data flow and user input are important to capture: procurement tends to focus on applications rather than datasets, with the needs of the business unit paying for the application foremost. It can be tough to figure out who the downstream users are, especially if they are using data through a data export rather than through the application, but it is users collectively who define what quality is in relation to a dataset.
* For context link data to the application and the business: at this step try to keep users focused on the dataset as a whole rather than on particular data elements when they are answering questions. Individual data elements are looked at more rigorously in later steps.
* The Region’s glossary of terms helps define the business purpose of data and helps make data unambiguous: building the business glossary helps understand the relationships with other datasets, which becomes important in later steps when improvements to the data are being considered.

**Stage 2 – Profiling and Identifying Critical Fields**

**E.2.1 Profiling**

* This stage can be done concurrently with initial data assessment and can provide some details to assist with the field-level data attributes metadata: running a profile tool against data early on can help speed the collection of metadata, and helps automate the assessment process.
* Initial assessment should be as comprehensive as possible. For periodic reassessments and improvement projects you should only profile critical data fields related to the task: the goal of an initial assessment is to learn about the dataset and inventory the problems with it, so as many fields as possible should be profiled. These problems will then be prioritized, and the most important ones dealt with. It does not make sense to profile all the fields after an improvement project that focuses only on a few fields.
* Profile primary and foreign keys as these affect dataset validity. Missing relationships between parent and child records are among the most common problems impacting data quality. It makes sense to pay attention to primary and foreign keys when profiling, even though they do not contain meaningful attributes.
* Most organizations have more data quality issues than they will ever be able to address: by being able to compare the profiles between different fields, we can quickly identify which data quality issues are having the most impact.
* After an improvement project, compare the field profile with an initial profile to see if there is a difference: by dating when profile measures were captured, ~~longitudinal~~ tracking of data quality over time becomes possible, and the series of profiles becomes a time-series measure of quality, enabling detection of new errors.
* Metadata completeness may be reconsidered after the profile has been assessed: good metadata is integral to data quality. Additional metadata added during profiling counts as data improvement. As it takes little time to run a metadata completeness assessment, this is an easy way to show quick progress on one aspect of data quality.

**E.2.2 Individual Data Attributes**

* Metadata is required for all targeted field-level attributes: system tables and fields used by the application for processing and report generation do not need to be documented as field level metadata. Their content where relevant is captured in the next step.
* Assessing criticality for different users minimizes the work in later stages and focuses DQA effort where it counts. As it is impossible to correct all the data errors that can occur, determining which fields are important to users allows data improvement effort to concentrate on data elements where it will make the most difference to users.

**E.2.3 Understanding Processing**

* Complex processing can be a high-risk area for quality: processing can hide data entry errors, and formulas can get corrupted without users noticing. Characterizing the processing as raw, lightly or highly processed helps later determination of fields to measure for unexpected differences over time that may reflect corruption of processing routines for whatever reason.

**Stage 3 – Assessing Fit for Purpose**

**E.3.1 Assess Profile Results**

* Profiling does not tell if data meets expectations: while this step can reveal a lot of different problems with a dataset through statistical analysis of each field, it is the expectations of users that ultimately define quality, so profiling analysis is not enough in itself to direct improvements – the analysis has to be reviewed by the users.

**E.3.2 Assess Fit for Purpose**

* Remember that data quality is all about expectations; as many users as possible should be consulted to help define data quality, paying attention to the gap between the current condition of the data and the desired quality relative to the purposes users put the data to.
* Expectations should be expressed in terms of quality dimensions: completeness, validity, consistency, timeliness, integrity (and relevance): the more closely user expectations can be expressed in these terms, the easier it is to develop improvement strategies that can be monitored in the next step.
* Absence of expectations may indicate that data is not critical to that user, dig deeper to find out: sometimes downstream users will attempt to use data for purposes that it is ill-suited to meet, being unaware of more appropriate data sources, or use data to tangentially inform or corroborate decisions that are primarily informed by other sources.
* For Open Data publishing the DQ team may decide to develop data quality documentation for users: external and even anonymous users can benefit greatly from seeing quality assessments.
* Avoid deep discussion of the source of errors until the next stage, otherwise it may get in the way of users being able to agree on data quality targets as they start blaming others in the data supply chain.
* Indicate in metadata any data limitations, what it should not be used for: this guidance helps immediate and downstream users determine how much to rely on this data source.

**Stage 4 – Data Improvements and Monitoring**

**E.4.1 Develop data improvement recommendations**

* Measures are constructed to detect unexpected conditions or changes in data content: measures are what DQ stewards use to maintain data quality in the long run with minimal effort, so it is worthwhile to put the effort in to develop them carefully. In many cases constructing appropriate measures will require the involvement of other data role that is familiar with specifying query structures that reflect quality dimensions.

**E.4.2 Document data quality improvement plan**

* Good maintenance practices help sustain data quality: even though the project plan deals with implementing specific technical improvements, the plan should account for how these improvements will be sustained over time through an education component that develops good maintenance practices amongst staff involved in data acquisition.