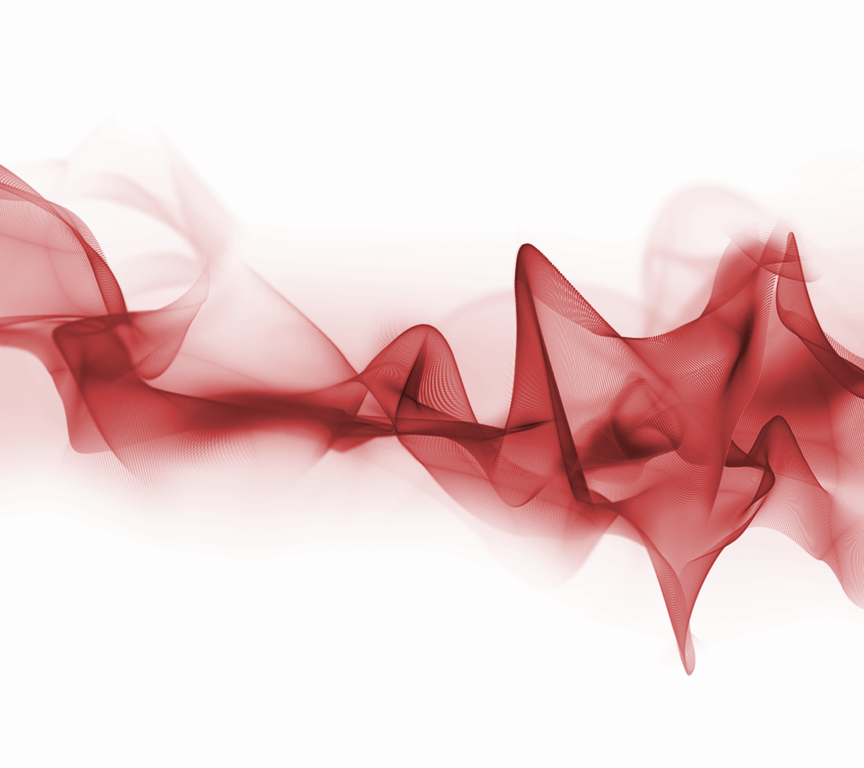
Data Quality Training

Pre-Cursory Materials to Develop a DQ Analyst Induction Package

Basic Skills to Review Data Quality



© Perspicacity Ltd 2018

|  |  |
| --- | --- |
| Document Control | |
| Version | V2.0 |
| Created | 2018.04.30 |
| Updated | 2019.12.20 |
| Original Author | Matthew Bishop, Perspicacity Ltd |
| Contact | [matthew.bishop@perspicacityltd.co.uk](mailto:matthew.bishop@perspicacityltd.co.uk) / 07545 878906 |

© Perspicacity Ltd 2018

This work product is licensed under the creative commons attribution 2.0 license. This allows worldwide, royalty-free, non-exclusive rights to share, distribute and create derivative work products (including for commercial gain) as long as the terms of the license are adhered to. Full details of the terms can be found at:

<https://creativecommons.org/licenses/by/2.0/uk/legalcode>

# Table of Contents

[Table of Contents 3](#_Toc27743804)

[Data Quality Strategy - Introduction 4](#_Toc27743805)

[Data Quality Strategy Series 4](#_Toc27743806)

[A little bit about Perspicacity Ltd 4](#_Toc27743807)

[About this document 5](#_Toc27743808)

[Methods to Review Data Quality 5](#_Toc27743809)

[Audit 5](#_Toc27743810)

[What do we mean by Audit? 5](#_Toc27743811)

[The downsides of Audit as an approach to Data Quality 5](#_Toc27743812)

[Validation 6](#_Toc27743813)

[What do we mean by Validation? 6](#_Toc27743814)

[The downsides of Validation as an approach to Data Quality 6](#_Toc27743815)

[Measurement 6](#_Toc27743816)

[What do we mean by Measurement? 6](#_Toc27743817)

[The downsides of Measurement as an approach to Data Quality 6](#_Toc27743818)

[Approaches to Reviewing Data Quality 6](#_Toc27743819)

[Approach to systematically review data quality 6](#_Toc27743820)

# Data Quality Strategy - Introduction

## Data Quality Strategy Series

This document is part of an open source series of documents to facilitate the establishment of a Data Quality strategy and function within an organisation. The full open source library can be found at <https://github.com/perspicacity-ltd/DataQualityReporting>

* The series presents a set of documents as a starter for ten
* It can be used by organisations starting on their data quality journey and those who already have a data quality function
* It contains the following components of a data quality strategy:
  + Strategy & Exec Summary
  + Policy & Standard Operating Procedures
  + Technical Specification (including link to reporting suite at <https://github.com/perspicacity-ltd/DataQualityReporting>)
  + Highlight Report Template
  + Training Materials
  + DQ KiteMark Images

## A little bit about Perspicacity Ltd

Perspicacity provides decision support consultancy, coaching, & development to the NHS. They are passionate about reducing the cost of software development to the NHS and aspire to create an active community of NHS and commercial organisations. They all share a common goal of improving the quality and efficiency of patient care through better, and more informed, decision making.

Open source helps the healthcare community to do this by sharing software development, learning from each other, and help software meet the needs of every organisation without being constrained to a single solution or paying for the same piece of work over and again across different organisations.

Although these Data Quality open source products are suitable for any organisation, healthcare or not, they are here as a result of wanting to freely share Perspicacity's collective products and ideas across the NHS and to widen the benefit of good, but usually locally isolated, projects further.

Perspicacity's open source offerings can be found at <https://github.com/perspicacity-ltd>

If you'd like to find out more, please contact Matthew Bishop on 07545 878906 or [matthew.bishop@perspicacityltd.co.uk](mailto:matthew.bishop@perspicacityltd.co.uk)

# About this document

This document contains some basic elements that would be good to include in a basic DQ analyst induction package.

# Methods to Review Data Quality

There are 3 basic methods to reviewing data quality:

* + Audit
  + Validation
  + Measurement.

Each method has its advantages and disadvantages. A critical part of successful Data Quality review is knowing which to use in each Data Quality scenario.

## Audit

### What do we mean by Audit?

Audit takes a random sample from a specified cohort of data and looks for errors within the sample. This is a useful approach to identify previously unidentified errors and quantify the prevalence of errors for the purpose of prioritising correction resources. Regularly scheduled audit of cohorts of data, especially where risk has previously occurred, is recommended (e.g. auditing the RTT clock stops from last month to check whether any clock stops were inadvertent errors) to identify new errors and assess where correction efforts should be focussed.

Data can be corrected as audit is being undertaken.

### The downsides of Audit as an approach to Data Quality

* + The main downside to audit as an approach to data quality is that the random nature of the sample used in the audit can significantly skew the results. This can result in issues not being identified, issues being under-represented or issues being over-represented within the sample.
  + Depending upon the frequency of the audit and the amount of change within the cohort being sampled, the results can lose relevance (known as “going stale”) over time.
  + Whilst suitable to determine the prevalence of the data quality issue (also known as the correction “backlog”), it is usually difficult to determine the incidence (also known as the “rate of error”).

Whilst these downsides don’t eliminate the value of audit, it is best avoided as a method to quantify the amount of resource required to correct and maintain all except the smallest of data quality issues.

## Validation

### What do we mean by Validation?

Validation takes a specified cohort of data and looks for errors within the entire cohort. This is much more comprehensive than audit and

Data can be corrected as the validation is being undertaken.

### The downsides of Validation as an approach to Data Quality

Downsides – often delivered inconsistently, costly if not targeted, may lose relevance over time (or never achieve relevance in the first place)

* + Make your validation cohorts as small as possible by using DQ measures

## Measurement

### What do we mean by Measurement?

Measurement uses existing data to count & show potential errors in the data. This produces a conclusive list of records for correction when measuring timeliness and completeness.

Effective measurement should use the DQ measurement principles in the DQ Policy in order to maximise the success from measuring and correcting data quality.

### The downsides of Measurement as an approach to Data Quality

False positives, false negatives, insufficient data to measure properly, correction unlikely as part of the process (unless absolutely sure data can be corrected algorithmically)

# Approaches to Reviewing Data Quality

## Approach to systematically review data quality

Use the methods described below, in the following order:

1. audit a cohort of data where there is a perceived risk (the “risk cohort”), to find any unidentified data quality errors within the cohort
2. build a measure for each data quality error to flag records with that error
3. audit records flagged by each data quality measure (to find any false positives i.e. false flags within your data quality measures)
4. audit records not flagged by each data quality measure (to find any false negatives i.e. missing flags within your data quality measures)
5. rebuild each data quality measure, where necessary and possible, to eliminate false positives and false negatives
6. start validation and correction of each data quality measure within existing resource
7. estimate the audit, validation and correction resource requirement