THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT

Defining Data Quality Dimensions



Abstract This paper has been produced by the DAMA UK Working Group on "Data Quality Dimensions". It details the six key 'dimensions' recommended to be used when assessing or describing data quality.



DEFINING DATA QUALITY DIMENSIONS

BACKGROUND

The term *data quality dimension* has been widely used for a number of years to describe the measure of the quality of data. However, even amongst data quality professionals the key data quality dimensions are not universally agreed.

This state of affairs has led to much confusion within the data quality community and is even more bewildering for those who are new to the discipline and more importantly to business stakeholders.

Socrates said, “The beginning of wisdom is the definition of terms”. Hence, the goal of this whitepaper is to define the key data quality dimensions and provide context so there can be a common understanding for industry professionals and business stakeholders alike.

Sir Karl R. Popper built on this saying "I do not say that definitions may not have a role to play in connection with certain problems, but I do say it is for most problems quite irrelevant whether a term can be defined (or not). All that is necessary is that we make ourselves understood." This certainly reinforces the idea that dimensions are indicators helping us to measure and communicate the quality of data, as opposed to defining what the data itself means or represents.

In May 2012, DAMA UK asked for volunteers to join a working group to consider the issue and produce some best practice advice. The response was overwhelming and demonstrated the need for such a piece of work.

Other data management professional organizations have also been keen to support this initiative and as such we were very pleased to welcome Julian Schwarzenbach, Chair of the BCS Data Management Specialist Group and Gary Palmer, charter member of IAIDQ to join the working group.

Contents

Defining Data Quality Dimensions 1
BACKGROUND 1
WHAT IS A DATA QUALITY DIMENSION? 3
CONTEXT 3
APPLICATION 4
HOW TO USE DATA QUALITY DIMENSIONS 5
SIX CORE DATA QUALITY DIMENSIONS 6
COMPLETENESS 7
UNIQUENESS 9
TIMELINESS 10
VALIDITY 11
ACCURACY 12
CONSISTENCY 13
OTHER DATA QUALITY CONSIDERATIONS 14
GLOSSARY 15
AUTHORS 16
EXTERNAL SOURCES OF REFERENCE 16

WHAT IS A DATA QUALITY DIMENSION?

A *Data Quality (DQ) Dimension* is a recognized term used by data management professionals to describe a feature* of data that can be measured or assessed against defined standards in order to determine the quality of data.

For example:

- A test data set is measured as 93% complete
- The result of an accuracy assessment for a data item in a test data set was 84%

A *DQ Dimension* is different to, and should not be confused with other dimension terminologies such as those used in:

- other aspects of data management e.g. a data warehouse dimension or a data cube dimension
- physics, where a dimension refers to the structure of space or how material objects are located in time

Note: For the purpose of this paper the term data quality dimension is taken to mean:

something (data item, record, data set or database) that can either be measured or assessed in order to understand the quality of data

** Characteristic, attribute or facet*

CONTEXT

The best practice laid out in this document is designed to assist data quality practitioners when looking to assess and describe the quality of the data in their organizations.

This document defines the six best practice definitions as generic *data quality dimensions*. This will help to reduce uncertainty and confusion that may arise when considering data quality. It is suggested that these dimensions and definitions should be adopted by data quality practitioners as the standard method for assessing and describing the quality of data. However, in some situations one or more dimension may not be relevant.

The intention is for organizations to use these dimensions to measure the impact of the poor data quality in terms of cost, reputation and regulatory compliance, etc.

APPLICATION

This paper provides a checklist of dimensions that users can choose to adopt when looking to assess the quality of the data in their organization. It is not a prescriptive list and use of the dimensions will vary depending on the business requirements and industry involved.

To aid the use and application of these dimensions, each dimension is illustrated by an example in a fictitious school scenario. This has been deliberately chosen as something that everyone can relate to, regardless of the industry in which they work.

Before attempting to use data quality dimensions, an organization needs to agree the quality rules against which the data needs to be assessed against. These rules should be developed based upon the six data quality dimensions, organizational requirements for data and the impact on an organization of data not complying with these rules. Examples of organizational impacts could include:

- incorrect or missing email addresses would have a significant impact on any marketing campaigns
- inaccurate personal details may lead to missed sales opportunities or a rise in customer complaints
- goods can get shipped to the wrong locations
- incorrect product measurements can lead to significant transportation issues i.e. the product will not fit into a lorry, alternatively too many lorries may have been ordered for the size of the actual load

Data generally only has value when it supports a business process or organizational decision making. The agreed data quality rules should take account of the value that data can provide to an organization. If it is identified that data has a very high value in a certain context, then this may indicate that more rigorous data quality rules are required in this context.

HOW TO USE DATA QUALITY DIMENSIONS

Organizations select the data quality dimensions and associated dimension thresholds based on their business context, requirements, levels of risk etc. Note that each dimension is likely to have a different weighting and in order to obtain an accurate measure of the quality of data, the organization will need to determine how much each dimension contributes to the data quality as a whole.

A typical Data Quality Assessment approach might be:

1. Identify which data items need to be assessed for data quality, typically this will be data items deemed as critical to business operations and associated management reporting
2. Assess which data quality dimensions to use and their associated weighting
3. For each data quality dimension, define values or ranges representing good and bad quality data. Please note, that as a data set may support multiple requirements, a number of different data quality assessments may need to be performed
4. Apply the assessment criteria to the data items
5. Review the results and determine if data quality is acceptable or not
6. Where appropriate take corrective actions e.g. clean the data and improve data handling processes to prevent future recurrences
7. Repeat the above on a periodic basis to monitor trends in Data Quality

The outputs of different data quality checks may be required in order to determine how well the data supports a particular business need. Data quality checks will not provide an effective assessment of fitness for purpose if a particular business need is not adequately reflected in data quality rules. Similarly, when undertaking repeat data quality assessments, you should check to determine whether business data requirements have changed since the last assessment.

Whilst most data quality dimensions can be assessed by analyzing the data itself, assessing accuracy of data can only be achieved by either:

- Assessing the data against the actual thing it represents, for example, when an employee visits a property; or
- Assessing the data against an authoritative reference data set, for example, checking customer details against the official list of voters

SIX CORE DATA QUALITY DIMENSIONS

The six core dimensions of data quality are:

1. **Completeness**
2. **Uniqueness**
3. **Timeliness**
4. **Validity**
5. **Accuracy**
6. **Consistency**



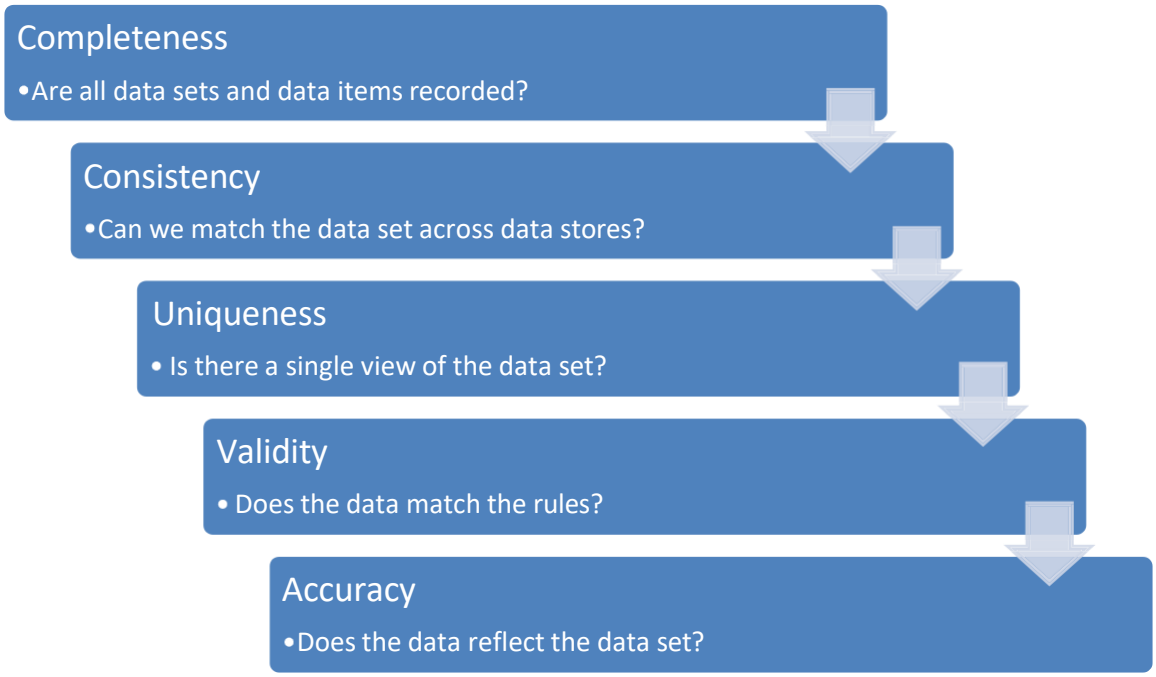


Figure 1 - Example of the application of different data quality dimensions to a data set

COMPLETENESS

Title	**Completeness**
Definition	The proportion of stored data against the potential of "100% complete"
Reference	Business rules which define what "100% complete" represents.
Measure	A measure of the absence of blank (null or empty string) values or the presence of non-blank values.
Scope	0-100% of critical data to be measured in any data item, record, data set or database
Unit of Measure	Percentage
Type of Measure: • **Assessment** • **Continuous** • **Discrete**	Assessment only
Related dimension	Validity and Accuracy
Optionality	If a data item is mandatory, 100% completeness will be achieved, however validity and accuracy checks would need to be performed to determine if the data item has been completed correctly

Example(s)	Parents of new students at school are requested to complete a Data Collection Sheet which includes medical conditions and emergency contact details as well as confirming the name, address and date of birth of the student. Scenario: At the end of the first week of the Autumn term, data analysis was performed on the 'First Emergency Contact Telephone Number' data item in the Contact table. There are 300 students in the school and 294 out of a potential 300 records were populated, therefore 294/300 x 100 = 98% completeness has been achieved for this data item in the Contact table.
Pseudo code	Count 'First Emergency Contact Telephone Number' where not blank in the Contact table/ count all current students in the Contact table.

Footnote:

1. Measure critical data for completeness first; incompleteness in non-critical data may not matter to the business.

UNIQUENESS

Title	Uniqueness
Definition	No thing will be recorded more than once based upon how that thing is identified.
Reference	Data item measured against itself or its counterpart in another data set or database.
Measure	Analysis of the number of things as assessed in the 'real world' compared to the number of records of things in the data set. The real world number of things could be either determined from a different and perhaps more reliable data set or a relevant external comparator.
Scope	Measured against all records within a single data set
Unit of Measure	Percentage
Type of Measure: • **Assessment** • **Continuous** • **Discrete**	Discrete
Related dimension	Consistency
Optionality	Dependent on circumstances
Example(s)	A school has 120 current students and 380 former students (i.e. 500 in total) however; the Student database shows 520 different student records. This could include Fred Smith and Freddy Smith as separate records, despite there only being one student at the school named Fred Smith. This indicates a uniqueness of 500/520 x 100 = 96.2%
Pseudo code	(Number of things in real world)/(Number of records describing different things)
External Validation	IAM Asset Information Quality Handbook Principles of Data Management, Keith Gordon

Footnote:

1. Uniqueness is the inverse of an assessment of the level of duplication

TIMELINESS

Title	Timeliness
Definition	The degree to which data represent reality from the required point in time.
Reference	The time the real world event being recorded occurred.
Measure	Time difference
Scope	Any data item, record, data set or database.
Unit of Measure	Time
Type of Measure: • **Assessment** • **Continuous** • **Discrete**	Assessment and Continuous
Related dimension	Accuracy because it inevitably decays with time.
Optionality	Optional dependent upon the needs of the business.
Example(s)	Tina Jones provides details of an updated emergency contact number on 1st June 2013 which is then entered into the Student database by the admin team on 4th June 2013. This indicates a delay of 3 days. This delay breaches the timeliness constraint as the service level agreement for changes is 2 days.
Pseudo code	Date emergency contact number entered in the Student database (4th June 2013) minus the date provided (1st June 2013) = a 3 Day delay.

Footnote:

1. Each data set will have a different proportion of volatile and non-volatile data as time acts differently on static and dynamic records.

VALIDITY

Title	Validity
Definition	Data are valid if it conforms to the syntax (format, type, range) of its definition.
Reference	Database, metadata or documentation rules as to the allowable types (string, integer, floating point etc.), the format (length, number of digits etc.) and range (minimum, maximum or contained within a set of allowable values).
Measure	Comparison between the data and the metadata or documentation for the data item.
Scope	All data can typically be measured for Validity. Validity applies at the data item level and record level (for combinations of valid values).
Unit of Measure	Percentage of data items deemed Valid to Invalid.
Type of Measure: • **Assessment** • **Continuous** • **Discrete**	Assessment, Continuous and Discrete
Related dimension	Accuracy, Completeness, Consistency and Uniqueness
Optionality	Mandatory
Applicability	
Example(s)	Each class in a UK secondary school is allocated a class identifier; this consists of the 3 initials of the teacher plus a two digit year group number of the class. It is declared as AAA99 (3 Alpha characters and two numeric characters). Scenario 1: A new year 9 teacher, Sally Hearn (without a middle name) is appointed therefore there are only two initials. A decision must be made as to how to represent two initials or the rule will fail and the database will reject the class identifier of "SH09". It is decided that an additional character "Z" will be added to pad the letters to 3: "SZH09", however this could break the accuracy rule. A better solution would be to amend the database to accept 2 or 3 initials and 1 or 2 numbers. Scenario 2: The age at entry to a UK primary & junior school is captured on the form for school applications. This is entered into a database and checked that it is between 4 and 11. If it were captured on the form as 14 or N/A it would be rejected as invalid.
Pseudo code	Scenario 1: Evaluate that the Class Identifier is 2 or 3 letters a-z followed by 1 or 2 numbers 7 – 11. Scenario 2: Evaluate that the age is numeric and that it is greater than or equal to 4 and less than or equal to 11.

ACCURACY

Title	Accuracy
Definition	The degree to which data correctly describes the "real world" object or event being described.
Reference	Ideally the "real world" truth is established through primary research. However, as this is often not practical, it is common to use 3rd party reference data from sources which are deemed trustworthy and of the same chronology.
Measure	The degree to which the data mirrors the characteristics of the real world object or objects it represents.
Scope	Any "real world" object or objects that may be characterized or described by data, held as data item, record, data set or database.
Unit of Measure	The percentage of data entries that pass the data accuracy rules.
Type of Measure: • **Assessment** • **Continuous** • **Discrete**	Assessment, e.g. primary research or reference against trusted data. Continuous Measurement, e.g. age of students derived from the relationship between the students' dates of birth and the current date. Discrete Measurement, e.g. date of birth recorded.
Related Dimension	Validity is a related dimension because, in order to be accurate, values must be valid, the right value and in the correct representation.
Optionality	Mandatory because - when inaccurate - data may not be fit for use.
Applicability	
Example(s)	A European school is receiving applications for its annual September intake and requires students to be aged 5 before the 31st August of the intake year. In this scenario, the parent, a US Citizen, applying to a European school completes the Date of Birth (D.O.B) on the application form in the US date format, MM/DD/YYYY rather than the European DD/MM/YYYY format, causing the representation of days and months to be reversed. As a result, 09/08/YYYY really meant 08/09/YYYY causing the student to be accepted as the age of 5 on the 31st August in YYYY. The representation of the student's D.O.B.–whilst valid in its US context–means that in Europe the age was not derived correctly and the value recorded was consequently not accurate.
Pseudo code	((Count of accurate objects)/ (Count of accurate objects + Counts of inaccurate objects)) x 100 Example: (Count of children who applied aged 5 before August/YYYY)/ (Count of children who applied aged 5 before August 31st YYYY+ Count of children who applied aged 5 after August /YYYY and before December 31st/YYYY) x 100

CONSISTENCY

Title	Consistency
Definition	The absence of difference, when comparing two or more representations of a thing against a definition.
Reference	Data item measured against itself or its counterpart in another data set or database.
Measure	Analysis of pattern and/or value frequency.
Scope	Assessment of things across multiple data sets and/or assessment of values or formats across data items, records, data sets and databases. Processes including: people based, automated, electronic or paper.
Unit of Measure	Percentage.
Type of Measure: • **Assessment** • **Continuous** • **Discrete**	Assessment and Discrete.
Related Dimension(s)	Validity, Accuracy and Uniqueness
Optionality	It is possible to have consistency without validity or accuracy.
Example(s)	School admin: a student's date of birth has the same value and format in the school register as that stored within the Student database.
Pseudo code	Select count distinct on 'Date of Birth'

OTHER DATA QUALITY CONSIDERATIONS

It is crucial to understand and manage the six core dimensions. However, there are additional factors which can have an impact on the effective use of data. Even when all six dimensions are deemed to be satisfactory, the data can still fail to achieve the objective.

Data may be perfectly complete, unique, timely, valid, accurate and timely. However if data items are in English and the users don't understand English then it will be useless.

It may be useful to ask these additional questions about your data.

Usability of the data - Is it understandable, simple, relevant, accessible, maintainable and at the right level of precision?

Timing issues with the data (beyond timeliness itself) - Is it stable yet responsive to legitimate change requests?

Flexibility of the data - Is it comparable and compatible with other data, does it have useful groupings and classifications? Can it be repurposed, and is it easy to manipulate?

Confidence in the data - Are Data Governance, Data Protection and Data Security in place?
What is the reputation of the data, and is it verified or verifiable?

Value of the data - Is there a good cost/benefit case for the data? Is it being optimally used?
Does it endanger people's safety or privacy or the legal responsibilities of the enterprise?

Does it support or contradict the corporate image or the corporate message?

GLOSSARY

Term	Definition
Dimension	Generally a characteristic by which you can regard or summaries something
Assessment	Checked by testing with an algorithm, reference to trusted data or secondary research. i.e., max length, pattern sequence, lookup to a trusted source like postal data or manual reference real world things.
Continuous Measurement	Items checked periodically e.g. age, weight or height, which change over time.
Discrete Measurement	Checking whether something which is an absolute measure is true or false e.g. gender, date of birth, birth place.
Data Item	An individual field in a data record referred to as a column in a relational database.
Record	A record is a set of related values.
Data Set	A dataset (or data set) is a collection of data, usually presented in tabular form.
Database	An organized collection of data.
Measure	The process of establishing the extent to which the dimension conforms to its unit of measure.
Scope	The uses to which the dimension is applicable.
Unit of Measure	The method of measurement.
Related dimension	Defines any associated dimensions.
Optionality	Optional, not required or mandatory.
Thing	Data item, record, data set or database.
Reference	The rules against which a dimension is being compared.
Pseudo Code	Using simple English to represent complex programming language or a method of describing complex computer programming code

AUTHORS

- Nicola Askham - The Data Governance Coach; DAMA UK Committee Member
- Denise Cook - Senior Manager, Data Governance, Security & Quality, Lloyds Banking Group, Fellow of the BCS
- Martin Doyle - CEO, DQ Global
- Helen Fereday - Data Management Consultant, Aviva UK Health
- Mike Gibson - Data Management Specialist, Aston Martin
- Ulrich Landbeck - Data Management Architect, Microsoft Corporation
- Rob Lee - Group Head of Information Architecture, Lloyds Banking Group
- Chris Maynard - Director, Transforming Information Ltd
- Gary Palmer - Chief Alchemist, Information Alchemy; Charter Member IAIDQ
- Julian Schwarzenbach - Director, Data and Process Advantage; Chair, BCS Data Management Specialist Group

EXTERNAL SOURCES OF REFERENCE

- DAMA Body of Knowledge – First Edition
- DAMA Dictionary of Data Management DAMA Body of Knowledge – First Edition
- DAMA Dictionary of Data Management - 2nd Edition
- IAIDQ Glossary
- Institute of Direct Marketing Award in Data Management
- Institute of Asset Management Asset Information Quality Handbook
- Siemens Industry Online Support
- Wikipedia Data Consistency Entry
- Execution - MiH
- The Practitioner's Guide to Data Quality Improvement - David Loshin
- The TIQM Quality System for Total Information Quality
- Management – Larry English (MIT Information Quality Industry Symposium, July 15-17, 2009)
- Data Quality, The Accuracy Dimension - Jack E Olson
- Improving Data Warehouse and Business Information Quality - Larry English

This paper represents the views of DAMA UK and the Data Quality Dimensions Working Group and not necessarily the viewpoint of the organizations which the authors work for.