READING

5

Data Integrity
(an excerpt)

by Marcus Boudville and Adam Taylor

Marcus Boudville is at Aviva Investors (United Kingdom). Adam Taylor (United Kingdom).

DIMENSIONS OF DATA QUALITY

2

The scope of what data is can be varied and wide ranging. Dictionary definitions vary too. One such definition describes data as "facts and statistics collected together for reference or analysis" (this definition comes from Oxford University Press). Ultimately, data is all information in all forms. In the context of investment performance measurement, it is easier to consider data as being in one of two distinct forms. The first form is structured in the sense that data sits within a system or application, such as an accounting or a portfolio performance system. It is extractable and organized in a way that makes it useful and suitable for the application's use. But data can also be less structured and can be found in Microsoft Excel spreadsheets, Microsoft Word documents, or paper documents. We will describe the dimensions of data quality, and by "dimensions," we mean the different ways in which we can think about data quality, as follows:

Accuracy: Data are valid and correct.

Completeness: Data are complete given the intended purpose.

Conformity: Data conform to standards and rules.

Consistency: There is a consistent application of rules across different systems.

Timeliness: Data must be available in time for use and regularly updated.

Lineage: There is a record and knowledge of where the data come from.

Such descriptions and examples are shown in Exhibit 1. These are characteristics or aspects of data that need to be considered when thinking about data quality. We consider both structured and unstructured data in the discussion of the dimensions. An overarching principle is that data should be "fit for purpose."

© 2020 CFA Institute. All rights reserved.

Exhibit 1 Dimensions of Data Quality

Dimension	Description	Examples and Scenarios
Accuracy	Data are valid and correct. There is an appropriate level of accuracy of the data for its intended purpose. This outcome can be achieved through validation to an authoritative source of data or by comparing multiple sources of the same data.	Example of arrangements where one could apply different accuracy rules: "Report any instances where the NAV has moved over 25 bps over a set period of time when being used for regulatory reports." "Report any instances where the NAV has moved over 40 bps over a set period of time when being used to prepare client reports."
Completeness	Achieved when the dataset is complete given the purpose for which it was intended. It is worth noting that some data are mandatory and some are optional; this will all be part of the expected data collection.	An operations analyst may have a list of all open positions in the Investment Book of Records (the internal firm register of open positions). The analyst expects this information to be held by various brokers and counterparties. As a result, he or she will compare the independent multiple broker and counterparty position records with the internal Investment Book of Records view.
Conformity	Ensuring that the data conforms to internally defined standards and rules. For example: • character length (maximum number of characters allowed) • character type (alphanumeric, integer, numeric, Boolean, float, string, array, etc.)	When transferring a file from System A into another system or application, a check is run to ensure that the character type of the attribute in the file being sent matches what is expected to be received from the data warehouse (the in-house data store). A breach is reported if a mismatch is found.
Consistency	Ensuring that data quality rules are applied consistently across one or many systems or applications for the same attribute or dataset.	Net asset value (NAV) information is held in systems A, B, and C. If there is a rule to "report any instance where the NAV has moved over 5 bps a day" in System A but in Systems B and C the number is set at 10 bps and 15 bps, respectively, then the data consumers will have differing levels of confidence within the NAV data attribute (within different systems).

Exhibit 1 (Continued)

Dimension	Description	Examples and Scenarios
Timeliness	With regard to data quality processes, timeliness would ensure that the data are being reviewed to ensure that they are not stale. The rule(s) will test whether the attribute has updated per the expected frequency. Actual frequency is compared with the expected frequency. Generally, timeliness refers to getting information at the right time and quickly enough to meet reporting and analytical requirements.	The NAV of a daily priced fund is expected to be updated on a daily basis. If the date of the NAV attribute value is not the expected most recent date, allowing for the time the attribute is expected to be updated and public holidays, then it is stale and not appropriate to use.
Lineage	Understanding where the data have come from (the origin), what happens to the data, and where the data move over time. It allows errors to be traced and helps ensure the data quality for all its intended purposes.	A particular data attribute is created and maintained in a particular system (say, ABC) and is distributed to a different system (DEF). For example, a performance system receives inputs from multiple source systems—where a particular source data item is missing or incorrect, it may be necessary to fix the problem at the source, and therefore, knowledge of the source of each data item is required.

Note: The above descriptions of different dimensions of data quality are broadly based on contributions by Brian Buzzelli (Acadian Asset Management).

EXAMPLE 1

Dimensions of Data Quality

The data governance officer, responsible for developing an environment that ensures transparency of all current data governance–related policies at a fund management firm, is meeting with an investment consultant to describe the firm's data governance arrangements. She describes how the firm ensures that when data move from one system to another, there are no issues or errors arising from a mismatch of data attributes.

What dimension of data quality is the data governance officer *most likely* describing?

A Accuracy

B Conformity

C Consistency

Solution

B is correct. The character type and attribute in the file being sent must match what is expected to be received from the data warehouse (the in-house data store). A breach is reported if a mismatch is found. A is incorrect because accuracy requires data to be valid and correct. C is incorrect because consistency is ensuring that data quality rules are applied consistently across one or many systems or applications for the same attribute or dataset.