2 Organisational Requirements

Tell me: how many DQ criteria are we actually talking about?

2.1 DQ Criteria (7+2)

Some time ago, we encountered 45 DQ criteria. To us it seemed important to determine how they might be checked and we found the following six essential checking methods:

- Methods which should be covered by expert's approval (e.g. interpretability, granularity, necessity)
- Surveys (e.g. objectivity, relevance, intelligibility trustworthiness, added value)
- Methods covered by IT security /business monitoring (e.g. access opportunity, access security, temporary availability)
- Automatic measuring methods
- Methods requiring a visual inspection or a document check (traceability, normative consistency, origin)
- Methods comprising audits / follow-up examinations (e.g. applicability, applicability of audits)

We think it doesn't make sense to engage the DQ team in tasks that are already professionally handled by other teams in the company. Therefore the DQ criteria may in principle be confined to 7 automatically measurable criteria and two documentary criteria (7+2).

G. Morbey, *Data Quality for Decision Makers,*
DOI 10.1007/978-3-658-01823-8_2, © Springer Fachmedien Wiesbaden 2013

Table 1: DQ Criteria (7+2)

	DQ criteria[4]	Description
Automatically measurable	(1) Completeness per row (horizontal completeness)	Is there any missing or defective data in a record? All data is entered according to business needs.
	(2) Syntactical correctness (conformity)	Is there data in a non-standardised format? The data fits into the specific format.
	(3) Absence of contradictions (consistency)	Which data values are contradictory? The data do not contradict integrity specifications (business rules, empirical values) or defined ranges of values (within the data pool, in comparison with other data pools, in time elapsed).
	(4) Accuracy incl. currency	Which data is wrong or expired? Correct and up to date (timeliness) notation of existing names, addresses, products etc.
	(5) Absence of repetitions (free of duplicates)	Which data records or contents of columns are being repeated? No duplicates (search for synonyms and similarities), no homonyms, no overlapping (continuity), everything is precisely identifiable (uniqueness).
	(6) Business referential integrity (integrity)	Which reference data or relations are missing? There will not be any clients without a contract, products will be listed, …
	(7) Completeness (Cross check sums, vertical completeness)	Is there data consistency over all systems? For instance: at an appointed date the number of contracts in the data source is exactly the same as the number of contracts in the DWH.

[4] DQ Criteria 1 through 7 correspond to the 7 Basic DQ Rule Types

Table 1: (continuation)

	DQ criteria	Description
Documentary	(8) Availability of documentation (findability)	Can the data be found easily and quickly (e.g. using common "search"-functions) Are the data tagged?
	(9) Normative consistency	It has to be assured that the naming and meaning of certain data is the same over all systems, processes and departments of the organisation.

Actually, in the course of an internal survey, I already came across 19 dimensions of DQ, but I like 7+2 a good deal more. Sometimes less is better, isn't it? However, from my staff I heard before that in matters of Solvency II just 3 DQ criteria are sufficient.

Indeed, in the context of Solvency II, only three DQ performance indicators are requested: appropriateness, accuracy and completeness of data. But when it came to the question of how these criteria should be checked, we had to fall back on the 7+2 criteria introduced above. In the meantime we are convinced that each DQ criterion comprising automatic and/or documentary components can be described with our 7+2 criteria. Let's take a look at the next figure.

Thank you so far. Thinking it over seriously, criterion no. 9 - "normative consistency" - is rather tricky. We will have to turn our company upside down if that actually needs to be achieved.

You've got it! ☺ No! Let's stay realistic; even for criterion 8, "availability of documentation", nobody being even a little familiar with actually running IT systems would ever expect that each data field is documented and traceable! However, in regard to economic key figures, only a single documentation – and not various – should exist, using the same labels across all sources. For example, if sales figures

(field/column "fee") are aggregated from various business departments, it simply must not occur that data from department A includes insurance tax, while data from department B doesn't.

For the case that a central documentation of the most important data definitions of the company is aimed for, it would furthermore be reasonable that the system supports the departments in the course of harmonizing business terms (definitions). It is important to keep everything in a reasonable frame.

What about „technical correctness" as a DQ criterion?

If you are referring to concepts like "correctness in regard to contents" or "absence of professional errors", I unfortunately have to disappoint you. This is one of the problems computers cannot solve[5]. As an alternative, you might try to define evaluable sub-aspects of "technical correctness" and test these sub-aspects, like actuality, completeness, valid syntax, absence of contradiction etc. On the level of single cases the sum of these sub-aspects might be sufficient to cover what is meant by "technical correctness". However, the isolated use of terms like "correctness", "correctness with regard to contents" and "technical correctness" in a DQ environment may be revealed as wishful thinking, or security strategies which DQ just cannot provide.

Two examples from day-to-day business:

(1) After bonuses for new customers are paid, some old customers are generated anew in the system with small modifications – for instance in name or address. DQ then detects that 75% of the data

[5] A topic in „theoretic computer science", the so-called "halting problem" of the Turing machine, applied to correctness of data.

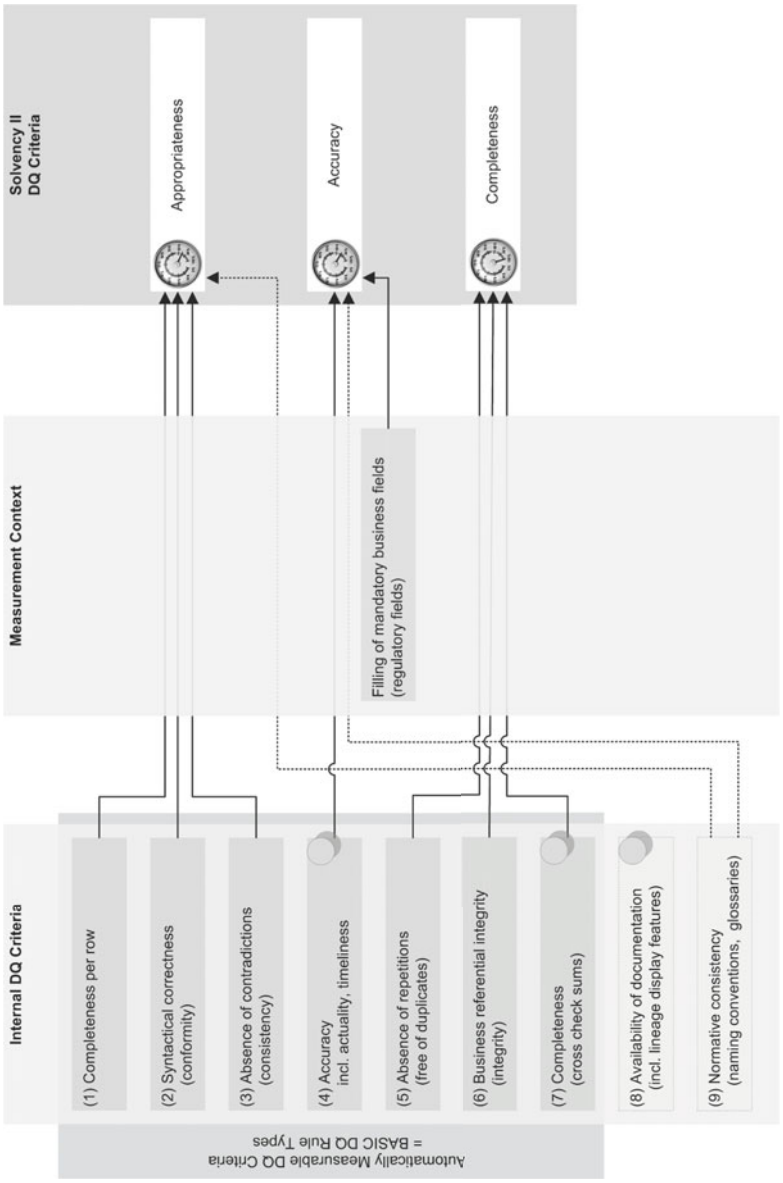

Figure 1: Mapping Internal DQ Criteria to SII DQ Criteria

of two customers are identical. The employee confirms that these actually are two customers and that everything is all right. Is that "correctness with regard to contents"?

(2) By the end of the year, in some companies the contract applications regularly increase and at the end of the first following quarter they are cancelled without any interaction with the client. Is that "professional correctness"?

I hope these examples illustrate the limitations of DQ. It is not an objective of DQ to serve this kind of investigative purpose. It's up to the particular employees responsible for the data (data owner) and other services to provide professional and content correctness, meaning that the data gives an account of the real conditions.

Examples for DQ Rules

Please give me some examples for automatically measurable DQ criteria, so things become clearer! Shall we order some more wine?

Sure, no problem. But you should keep in mind, even if the arbitrary naming of DQ rules might have a trivial touch, they are only definable after a thorough data analysis (data profiling), which has to be conducted together with the respective data users/ data owner. This guarantees that later on DQ defects in the defined data areas can be automatically detected, quantified and if necessary corrected.

A possible relief: If the data profiling repeatedly reveals no critical abnormalities, there is no need to define DQ rules.

Example (1) Completeness per row (horizontal completeness)

Rule „customer record for automobile insurance“: Here it is verified that all fields which are minimally needed for creating a customer record for the automobile insurance are actually filled in. If more than 5% of the data do not reach the criterion, the responsible department is informed.

Rule „not empty for regulator authorities“: It is checked that previously defined data fields are not empty before being further processed. All data sets failing this criterion are immediately handed over to the responsible employee in an adequate way.

Example (2) Syntactic correctness (conformity)

Rule „format gender“: In a selected data pool the following labels are accepted: m, f, male, female, masculine, feminine, homen, mulher, hermaphrodite, fm: formerly man, fm: formerly woman. Empty or differently completed data sets have to be marked for an immediate correction.

Rule „format date“: Data in the fields A, C, E are permitted only in one of the following patterns dd.mm.yy, ddmmyy and mm/dd/yyyy. Any other pattern will cause problems in further computations. Incorrect data sets are rejected and handed over to the responsible employee.

Example (3) Absence of contradictions (consistency)

Rule „gender and title“: Despite all implemented plausibility checks, sometimes the data is partly contradictory. In such a case it is checked whether title and gender are contradictory.

Rule „premium smaller than amount insured”: A violation of

this rule sometimes occurs when a new product has been introduced and if there are reasons to not implement such a plausibility check in the near future.

Example (4) Accuracy including currency

Rule „actuality of the business report“: The business report compares actual data with data dating back up to four years. A report is up to date if at least 20 to 30 % of the data sets have a time signature from the past twelve months.

Rule „actual portfolio“: The register does neither include only formerly distributed nor future products („valid until“ <today, „valid from”>today).

Example (5) Absence of repetitions (free of duplicates)

Rule „unique annual premium“: The combination of the fields “number of contract”, “year” and “amount” may only appear once in table x.

Rule „unique customer“: There may not be two or more data records in which the following data fields show similarities of more than 95 %.

Example (6) Business referential integrity (integrity)

Rule „only known processing indicators (KPI)“: No data will be transferred if their KPI is missing in the table of “List of workable KPI in the DWH”.

Rule „no fakes“: There will be no insurance application with a product name, which is not listed, in the “actual portfolio”.

Example (7) Completeness (cross check sums)

> Rule „completeness of reserves“: The difference between sum x and number y is always bigger than 0 and the difference for the instance should not exceed 0,01 % in a report period.
>
> Rule „sales figures“: The difference between the sum of the sales figures and the number Y from DWH may not exceed 0,3%. If so, department XYZ has to be informed immediately.

That’s it!

Thanks for the material. However, I can imagine that most of your examples are already taken account of in our company.

As I conceded initially, I believe that the lion’s share of the cases described is already taken care of. But from active DQ projects I learned that there is hardly ever a stringent system or a protocol on the results of DQ tests. We frequently encounter confusions with protocols from job control and ETL. They usually inform us only about aborts in the processing, why they happened (DQ deficits maybe a trigger) and how much data has been processed.

How can you be sure that DQ does not serve as a disguise for the new development of lots of applications just for testing purposes if other programs are running correctly?

Yes, and in a little while even more new programmes will appear in the market to check whether everything was correct in past DQ tests. And over all checks and crosschecks we will forget that the company has more substantive things to accomplish and louse up business in spite of the best possible DQ.

Sorry for being sarcastic, but somehow, you are right. There actually

is this hazard. Besides all efforts to do things right, the objective of DQ has to be kept in mind! Which is to detect errors in the data and taken as a whole to make data more reliable. DQ cannot serve as test- or acceptance-authority for already implemented or new programmes, since that would really mean doing the same thing twice over. But we can utilize DQ in a highly meaningful way for measurements before and after migrations with consistent criteria. However, if defects are discovered, it is not always possible to attribute them to bugs in the new programme or to handling problems.

Okay, we had better leave the further discussion to the experts. But may I ask for a short explanation of the differences between data profiling and DQ rules!

Of course! Let's try to do that with the help of the previously introduced analogy of water quality. We consider data profiling as a kind of "global analysis of water quality", where analytical chemists do not yet know what they are looking for. They apply some standard procedures, for example the search for the number of germs (in the case of data we would speak about column analysis). Because of earlier incidences the chemists already have reasonable suspicions. As soon as the disturbing contaminations are identified, the critical spots will be equipped with a simple monitoring device. For instance a small box with pH value monitor, contrast agent and scale. Depending on the water source or position, there are several individual options to keep the pH value within a certain range. Swinging back to DQ: DQ rules serve as a simple monitoring mechanism and if necessary, DQ deals with certain defects.

I think we have had enough DQ for today. Let's drop the subject and close with the following thought:

http://www.springer.com/978-3-658-01822-1

Data Quality for Decision Makers
A dialog between a board member and a DQ expert
Morbey, G.
2013, XI, 78 p. 7 illus., Softcover
ISBN: 978-3-658-01822-1