Managing data quality in IT and communications systems A proactive approach

Dayanand Kathapurkar
University College London

Abstract: Poor quality data in IT systems affects business operations and may even result in financial losses or penalties. Communication systems also face similar challenges due to increase in customised and context services. Conventional data quality initiatives are based on assessment and improvement of quality after data creation and hence are largely reactive in nature. IT systems will have better management of data quality if they are built with data quality as focus.

1. What is data quality and why is it important?

The word ‘Data’ is widely used and yet remains difficult to define. It is a means to express, store and maintain ‘information’ and ‘knowledge’. Institutions and businesses need to maintain and process data for a number of reasons mainly to support day to day operations. Compliance is another major reason that is the driver for businesses for the focus on data quality. [1] Data quality is equally important in communications systems due to the increase in customised and context services.

Defining the quality of data is a challenging task. Quality is defined as ‘conformance to requirements’[2] and ‘fitness for use’ [3]. Paper *a homogeneous framework to measure data quality* [4] attempts to define different parameters of data helpful in assessing the quality. These parameters are completeness, relevance, reliability, consistency, correctness, timeliness, precision and conciseness. This paper defines framework for quantification of data quality. However the paper acknowledges that this measurement is static and reactive. Moreover, some of the parameters, such as relevance and conciseness, are not easy to measure while precision may only be applicable in case of a numeric data entities. The parameters defined in the paper *‘Internal data alignment: best practices – GDS implementation roadmap for retailers and manufacturers’* by Global Commerce Initiative (GCI)/Cap Gemini [5] are a subset of the above and are measurable.

a. Completeness – all required values are recorded
b. Standards– based or conformance
c. Consistency – no different attribute values in different datasets
d. Accuracy – right values
e. Time stamped – validity time frame

Methods and tools assessing and managing data quality are broadly based around these characteristics. The data quality methods need to be integrated with the application so as to ensure the quality upfront rather than relying on correcting the data. One important characteristic is missing in the above and that is ‘Integrity’. These are the rules such as ‘when x = a, y = b’.

The data and metadata definitions within the databases and applications focus on completeness and conformance by defining rules and constraints. Design guidelines are required to be defined and followed for ensuring consistency. For example, a unique key, generated by a master system, could be used across all systems instead of storing textual address. This is an extension of the principle of normalisation. This approach and its effectiveness need to be evaluated. Accuracy and timeliness need audit and synchronisation. For example, verifying customer contact details whenever a contact is made or at a regular frequency could ensure timeliness and accuracy.

2. Does data quality deteriorate?

Data changes due to business processes and real life events, for example, fulfilment of requests or customer moving house. Data items in IT systems need to be maintained in order to ensure quality. Data in IT systems goes through lifecycle phases of creation, usage, updates and deletion. Then, how does the data quality deteriorate? The loss of electronically stored data happens through events such as system crashes. However, the data quality deteriorates when the rules to record and update data are out of step with the processes that result in the changes; in other words, when the systems or the data models can't keep pace with the changes in the business environment. For example, if the address data is not updated when the house splits into flats then the data quality deteriorates. For ensuring data quality systems should be made to act on the triggers or event notifications; planning permissions in case of addresses. Typically, when automated process doesn't exist, such triggers are received by users who update the data.

3. Who is responsible for data quality?

The next important challenge is to identify who is responsible for data quality. For maintaining the quality of data the users should follow proper rules and the system should force users to follow these rules. Achieving a right balance between the two is very difficult. If IT system doesn't enforce the rules and give users freedom then there is no surety that the data quality could be ensured. On the other hand, if too much of enforcement of rules brings rigidity in the system which could render it unusable and make it prone to data errors as timely data updates may not take place.

The argument above points us to a fact that fixing the responsibility of data quality is not an easy task. In his lecture Ted Friedman [1] explains that data quality is not just an IT systems issue but is more related to the usage of the systems by business users.

4. Can IT systems help ensuring data quality?

Various initiatives that the companies undertake to ensure data quality involve identifying the data owners, setting up data integrity teams, data warehousing etc. The Auditing and control type of methodologies involve data profiling, standardisation, synchronisation and data enrichment. These approaches are reactive as they don't ensure the quality while data is created or updated. Six sigma and Total Information quality management [6] (TIQM) initiatives emphasise on customer focus, measurement and improvement of data quality. These are relatively new concepts in data quality and the success of them needs to be verified.

In this paper, I propose a proactive approach to define and manage data quality. It is based on following principles. An approach to implement the solution is also suggested.

Continuous Measurement:

Define the quantitative measures for data attributes of customer address based on completeness, conformance, consistency, accuracy, Integrity and time-stamping. Not all of these are applicable to every data attribute. For example, completeness is important for post codes but not necessarily applicable to street names or locality.

Collect the measurements throughout the data lifecycle. It means every activity that inputs, deletes or updates the data has to be trapped to measure the quality.

Design rules within application using these measure to ensure the acceptable level of data quality

Incentivise: Provide incentives to users of the system for maintaining the quality of records, which they are responsible for, above acceptable level and with efficiency. Setting the acceptable levels is important as Acceptable level may not mean acceptable quality. Also they may be different for frontoffice and backoffice teams.

Triggers: Trigger the correction process for the team of experienced and better equipped staff to improve quality. The levels of incentives for the team could be different as they may be provided with longer time and more tools.

Penalise: Penalties for underachievement and deterioration of quality. Scores for penalties will depend on the acceptable levels.

Audit and synchronisation: Physical audit, verification with the customer and synchronisation of data with Master sources are the ways to ensure data quality in an application

Illustration:

Let's take an example of customer addresses and how the framework could be designed. The scores can be designed as

If criteria are not met – 0 point
If criteria are partially met – 1 point
If criteria are fully met – 2 points

The example of the scoring system below is superficial and simplistic and may not be directly used. Detailed analysis and design work, supported by profiling and data requirements, is needed to define acceptable framework.

Field	Description	Data Quality characteristic	Scores
Post Code	• Max length but not fixed • Format restrictions • Verification with master possible	• Completeness • Conformance • Accuracy	Max – (2+2+2)=6 Min – (1+2+2) = 5 Acceptable–(1+2+2)=5
Town/City, County and Country	• Specific values • Combination of Town/City and postcode are unique	• Accuracy • Integrity	Max – (2+2)=4 Min – (2+2)=4 Acceptable – (2+2)=4
Street name and locality	• No fixed format • Alphabetical • Aliases exist • Fuzzy logic required for verifying accuracy	• Consistency • Accuracy	Max – (2+2)=4 Min – 0 Acceptable – (2+0) = 2
House number or building/business name	• Different formats • Alphanumeric	• Consistency • Accuracy	Max – (2+2)=4 Min – 0 Acceptable – (2+0) = 2
Flat	• Not mandatory • Variants possible	• Consistency (if applicable)	Max – 2 Min – 0 Acceptable – 0
Total Score			Max – 20 Min – 9 Acceptable – 13

Table 4.1: Illustration of Data Quality scoring system

Address data can be verified with the Postal Address File (PAF). Due to the amount of churn PAF undergoes frequent updates. We can design a points scoring system that will trap every input and update to these fields and award points based on criteria as above. A number of scenarios are possible when a customer approaches.

A complete match is found and customer agrees full points are scored for that entry.

It is possible that a customer address may not be verified against PAF, even partially. The entry will score the points accordingly. It still has to comply with the rules as above. The acceptability of the entry will be based on minimum points to be scored.

Bulk changes are made to the addresses. When they are fed into the system, the scores for affected customer addresses will go down due to mismatches. These are the triggers for the updates. The scores are restored once the updates are successful.

Customer may request changes to their address entry such as new house name or business name. These will alter the score and trigger a feedback to PAF. Once PAF is updated suitably the score for that entry is altered again.

Using the scores: The scores can be designed and used as below

Initially, the benchmarking will be required for setting maximum, minimum and acceptable values of the score.

Targets can be defined for the frontoffice and backoffice teams however Incentives and penalties will need to be agreed as there may be people issues.

Individual Scores are used for defining field level quality measures and analysis where as consolidated scores could be used for various cause and effect analyses e.g. Overall quality of address data, quality achievable when for new customers etc.

Advantages and way forward:

The possible advantages of this approach can be summarised as:

Quantification of data quality at any instance is expected to make the cause – effect analyses easy. It will also help using statistical methods.

The responsibilities of the IT system and the user groups will become clear.

Defining the criteria will depend on the clarity and understanding during the design stage. Like any application the framework may evolve as understanding improves.

There may be performance implications of capturing the scores instantaneously and maintenance issues with the framework. To mitigate the risk of causing performance bottlenecks, initially the measures could be defined at higher lever i.e. at the full address level than at every field level.

This is an intrusive method hence creating a framework for existing application is also a challenge.

I am about to start a proof of concept trial for this approach which will be followed by experimental implementation. Until any success is achieved this approach remains theoretical.

5. References:

[1] http://www.gartner.com/it/products/podcasting/asset_145611_2575.jsp Lecture (podcast) by Ted Friedman, VP Gartner, on ‘Cost of poor data quality’ on 1st March 2006

[2] Crosby PB (1979) ‘*Quality is Free*’ New York: McGraw Hill.

[3] Joseph M. Juran. 1988. ‘*Juran on Planning for Quality*’. New York: Free Press.

[4] Monica Bobrowski, Martina Marre, and Daniel Yankelevich. *A homogeneous framework to measure data quality*. In Proceedings of the International Conference on Information Quality (IQ), pages 115-124, Cambridge, MA, 1999.

[5] http://www.gartner.com/it/products/podcasting/asset_145611_2575.jsp Lecture (podcast) by Ted Friedman, VP Gartner, on ‘Cost of poor data quality’ on 1st March 2006

[6] http://www.dmreview.com/article_sub.cfm?articleId=1011016 by Larry English, DM review October 2004.