Data Quality

Michael McCarthy

15 August 2005

IMS5028 - CRM Systems

Introduction

- Data Quality Defined
- Characteristics of Data Quality
- Measuring Data Quality
- Data Profiling
- Causes of Data Quality Issues
- Consequences of Poor Data Quality
- Objectives of Data Quality Management
- Data Quality Principles
- Data Quality Improvement Process

15 August 2005

IMS5028 - CRM Systems

Data Quality Defined

"Information quality - Consistently meeting all knowledge worker and end-customer expectations in all the characteristics of the information products and services they deem important. The degree to which information consistently meets the requirements and expectations of all knowledge workers who require it to perform their processes," Larry English

15 August 2005

IMS5028 - CRM Systems
Characteristics of Data Quality

- Data needs to be fit for the purposes for which it is intended to be used.
- **WHY?** Because poor quality data results in:
  - Increased costs
  - Reduced revenue
  - Poor customer satisfaction
  - Poor staff satisfaction
  - Contravention of the Privacy Principles

- The following characteristics are used to provide metrics on the quality of data within systems. These metrics are used to establish baselines for the improvement or degradation of data quality within the system:

  - **Completeness**
    - The presence of data values within each of the data fields.
    - Every data field has a defined and explicit meaning.
    - There are no nulls, spaces or blanks in the data fields.
  
  - **Validity**
    - The degree to which the data conforms to both its domain values, and its business rules. For example:
      1. Valid values for Title are "Mr", "Mrs", "Miss", "Dr".
      2. Valid values for "Gender" are "M" (Male) and "F" (Female).
      3. Home Telephone Number is 10 digits, first digit is "0".
      4. Home Telephone Number is 10 digits, second digit is "2" and State is "NSW".

  - **Non-duplication**
    - The degree to which there is a one-to-one relationship between the data and the entity it represents. For example:
      1. There may be several customer records for the same customer, each having different spellings of their name - "John Smith", "J. Smith", "J Smith", "JOHN SMITH".
      2. Having a "single source of truth" for data.

  - **Accuracy**
    - Measure or degree of agreement between a data value or set of values and a source assumed to be correct. Also requires conformance to business rules. For example:
      1. Title is "Mr" and Gender is "M". Problem - what if Title is "Dr"?
      2. Home Telephone Number is valid, but is it my Home Telephone number? I may have moved two years ago.
Characteristics of Data Quality

- **Timeliness (Currency)**
  - The availability of the data within the required timeframe, so that its usability is not degraded in any way.
  - There’s little point in knowing that a customer has deposited a large sum of money in their account four weeks after the event.
  - ‘Business-time’: monthly, weekly, daily, real-time?

- **Usability**
  - The degree to which the information is efficiently suited and usable for its intended purpose.
  - That is, does the user have all the complete, valid, and accurate data they need to do their job? And did they receive it in time?

Measuring Data Quality

- **Q. How do we know what is good data quality?**
  - A very good question!!!
- **A. If you can’t be bothered capturing it correctly, why capture it?**
- **How much will it cost to get it right?**

Sources of Best Practice

- Addresses:
  - Australia Post Postal Address File (PAF)
  - Industry Codes (ANZSIC)
  - Australian Bureau of Statistics
- Formatting standards
  - Australian Standards Institute
- Telephone Numbers
  - Telstra, Sensis
- Competitor Knowledge
- Regulators
Data Profiling

- **Column Profile**
  - The attributes of a column of data such as data type, size, range of values, frequency (in the sample), distribution of values, cardinality, null, uniqueness characteristics.

- **Dependency Profile**
  - Analysis of the values in a column against the values in other columns. This step identifies candidates for primary keys of the data set as well as defining grey areas where dependencies are met some of the time but not all of the time.

- **Redundancy Profile**
  - Compares data between entities of the same data set or source to determine columns containing overlapping or duplicate values. This type of profile identifies attributes with the same information but with different names and attributes with the same name but different business meaning.
  - Additionally, identifies redundant columns and those required to connect different entities.

Causes of Data Quality Issues

1. **Differing definitions of data elements and business rules between business units using the same common data elements or systems.** For example:
   1. Differing currency conversion rates, different times.
   2. Multiple code sets, e.g. ANZSIC Code:
      - Defined by the Australian Bureau of Statistics
      - Different set maintained in the Front-line User Instruction Manual.
      - Different set maintained in the source systems.
      - Different set maintained in the ETL programs.
   3. Poor address quality by allowing customers to define address - vanity suburbs.

2. **The use of operational systems to support business processes rather than providing downstream information to the rest of the business.** Example:
   1. Using default (usually the first on a list) code for processing a transaction to bypass validation; "9999", "TBA".

3. **Lack of suitable validation rules at point of entry.** For example:
   1. Data fields not mandatory.
   2. Not checking for ten digits for Home Telephone Number.
   3. Not using an appropriate tool to validate input, e.g. address validation using the Postal Address File.
Causes of Data Quality Issues

- Changes in business processes that are not reflected in the operational system due to the cost of modification. For example:
  1. The omission of recording a critical piece of data such as postcode in a system.
  2. “Stuffing” the data into fields no longer used but still available, for example, “TBA” in the Home Telephone Number.
  3. Adding a new piece of data into an existing field already in use.
- The ability of data within a system to become stale (i.e. only accurate at the time of recording).
  1. A customer’s occupation being recorded only at the time of account opening with no defined mechanism for updating this information on a regular basis.

Causes of Data Quality Issues

- Operational staff either not aware of or not rewarded for data quality outputs – more usually volumes – results in workarounds or “get the data in quickly regardless of accuracy”.
- System conversions

Consequences of Poor Data Quality

- Poor Customer Satisfaction. For example:
  1. Repeated requests for the same information
  2. Sending statements to wrong address
- Increased costs. For example:
  1. Postage
  2. Return to sender
  3. Outbound calls
- Reduced Revenue Potential. For example:
  1. Lack of “Single Customer View” (duplicate customer records)
  2. Impaired marketing effectiveness (poor segmentation)
  3. Impaired contactability
Consequences of Poor Data Quality

- Potential for non-compliance
  - National Privacy Principles
  - Reserve Bank
  - APRA
  - Basel II
  - Sarbanes-Oxley
  - IFRS

Objectives of Data Quality Management

- Enhance business user confidence in the quality of information by providing status reports that are easily accessible, current and accurate.
- Enable data quality and assurance to be built into the process of sourcing data at the earliest possible stage in the information lifecycle by understanding the issues and characteristics of the data within source systems.
- Provide mechanisms to monitor the quality of data on an ongoing basis to allow remediation to take place proactively rather than reactively.
- Provide an environment that makes transparent the reconciliation process between sources of data and downstream information systems.

Data Quality Principles

- Joint Responsibility
  - Joint IT / Business responsibility to correct data quality issues using the most appropriate and cost-effective method(s) for the problem at hand. These methods include:
    - Business process changes
    - Source system validation enhancements
    - Database updates
    - Manual input (correction)
- Fix at Source
  - Ultimately, the point of data creation should be the target for resolving ongoing data quality problems.
  - Data quality fixes at any other point result in a cost of rework.
Data Quality Principles

- **Measure to Improve**
  - Where it is not possible to measure the quality of data it will not be possible to measure any improvements in data quality over time.
  - In these instances data quality initiatives are a waste of resource and should not be undertaken.
  - Data quality must be measured and baselines established, so that improvements or degradation may be recorded and actioned appropriately.

- **Cost Effectiveness**
  - It must be possible to tag the targeted data elements with a downstream value, for example, the cost of an outbound marketing call to an invalid telephone number, to ensure the prioritisation of data quality initiatives.
  - If the downstream value cannot be calculated or is unknown, the data has no clear value to the business.

Data Quality Principles

- **Regular Appraisal**
  - Once a data quality assessment has been undertaken, success criteria established and an action plan formulated, it is necessary to reappraise the data quality initiatives and metrics on a regular basis to ensure continual improvement.
  - This may take the form of incorporation of quality metrics into a team or individual performance measures (KPIs).
Measuring/Cleansing Data Quality

- Profiling Tools
  - Axio Evoke
  - Ascential
- Cleansing Tools
  - Ascential
  - Data Flux
  - Trillium
- Visit these web sites and others

Conclusion

- Thank you