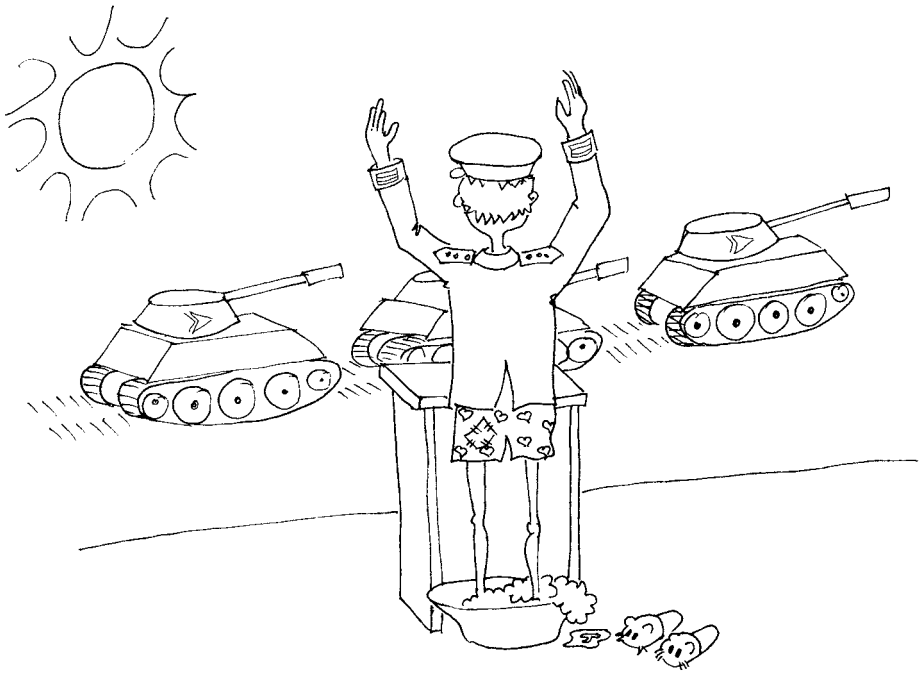
9. The Implications of Data Quality

"In each system the data quality is such that it merely permits the system to operate."



9.1. The Business Implications of Data Quality

One of the basic principles of information technology is the understanding of 'garbage in, garbage out'. This saying draws attention to the fact that computers will produce nonsensical output if they are given nonsensical input. An example: A person received a letter from their bank recommending one of their credit card products. This recommendation surprised the customer as they had already owned this very product for over half a year. The letter confused them. Had the bank forgotten about their existing card? Could they acquire a new, second credit card? Would the bank like to raise the customer's credit limit? The customer reasoned that the bank was probably aware of the existence of their existing card so most likely they were entitled to receive an additional card. After some consideration the customer went into one of the local branches of the bank to start the application procedure. Unfortunately, the clerks very soon found out that the customer was not, in fact, eligible to receive an additional card, and the customer was disappointed. Could they continue to trust the bank to keep their money if it made such simple mistakes?

Although this mistake did not cause any financial damage, it is a good example of how a seemingly minor administrative mistake can cause a significant loss of trust (in the example above, what happened was that the incorrect recording of customer data caused the credit card and account management systems to fail to cooperate properly and to send out numerous mails to customers with no regard to their product histories).

Cases like the above occur on a daily basis in everyday life. However, the number of such mistakes may quickly reach a level which is unacceptable, can damage brand value and increase the daily work of employees who handle data and customers. This raises concerns within the company about how data should be handled and how data quality should be improved. As mistakes such as the one described above are commonly due to a high number of errors and are of a wide variety of types, they are best handled using data mining techniques.

Our years of experience in the field suggest that data quality is a rather relative concept. In any system the data quality is such that it merely permits the system to operate at an 'acceptable' level. If a system operates without considerable problems then the data quality is generally considered to be adequate. Problems mostly arise when changes are made to systems and what is required of them, as data quality changes in parallel. Data quality mostly deteriorates in these cases as it no longer meets increasing quality requirements. In the example above, a new credit card product was introduced but data quality issues meant that the credit card and account management systems did not work together as planned. Another example of this is when a phone number database is used by administrators to call customers. In this

case, the formats of the numbers in the database may permissibly be different. However, if the same database is used for automated sms texting then both the form (e.g. dash and slash characters) and content (e.g. country and dialling codes) should be unified to the format required by the texting system.

|204|

9.2. Data Quality Issues

9.2.1. Data Quality Requirements

The errors that are easiest to spot are field level data errors. For each field a domain can be defined that restricts the field values. If a value is outside of the domain then an error is flagged up. A 'negative' wage or an age of 150 are good examples of such errors. Blank fields with no values count as errors too. Errors at a record level may violate certain dependencies. Records of gynaecological interventions carried out on male patients would hopefully be caused by problems with data!

These flaws in data can be handled by creating a list of requirements, defined by business domain experts. This list can be considered to be a part of the system and regular data checks can be run using it. To minimize type errors, a meta database can be created which describes all the specified criteria (e.g. we assign a domain descriptor to each field) along with checking routines. Special cases may require custom check routines based on their specification. Ambiguities which arise during system development must be handled with the cooperation of domain experts. All the assumptions made about the data/system must also be documented.

Errors in data may arise from the problematic operation of a system caused by the existence of large amounts of data (for example, large numbers of returned mails or large amounts of duplicates of internal customer records). Other errors may occur when the functionality of an operating system is expanded (as with the example of the phone database above). Data quality problems can be categorised into two main groups:

- Value errors. These types of errors can easily be spotted as:
  - they deviate from a formal restriction (e.g. Hungarian VAT numbers are restricted to 8 digits followed by a dash then a digit then a dash and finally 2 digits);
  - their domains are out of the expected range (e.g. the name given for the city name field is not contained in a list of possible cities); or
  - the data is atypical and outside the normal constraints of the expected parameters (e.g. the date of birth of a customer is earlier than 1900.01.01 or their weight is greater than 300 kg).

- Contextual data flaws. Data with these types of errors may appear to be correct. However, it can cause problems relative to a reference point;
  - Contextual flaws within a record (e.g. a specified city does not match its zip code, a customer's first name is inconsistent with their gender);
  - Contextual errors across records:
    - Technical (items violating the requirement for uniqueness, binding and obligatory filling in requirements);
    - Duplications: when an item (e.g. a customer) appears in the system more often than is needed
    - 'Dummy' values: values that carry no real information but are entered only because of an obligation to enter something (e.g. an obligatory VAT number field is filled in with the number "11111111").

Data consolidation is handled in a similar way to duplicates. Typical cases of consolidation include:

- the master data of a firm needs enriching with data from one or more external data sources (e.g. we would like to merge income data from the Central Statistical Office with our corporate master data);
- during a data integration process the master data from systems which are managed side by side are consolidated; or
- a list of the target customers of a marketing campaign should be filtered to exclude existing customers.

The most complex errors are system level errors and occur when an error is implied with a whole set of records. How can these errors be diagnosed? If we were to document each and every requirement for data and its diagnosis we would end up with a document many times larger than the system specification itself. This is why reference points are used to serve as a compass.

9.2.2. Reference Points

Reference points are criteria used for making data comparisons and to identify deviations and errors.

A data warehouse is often designed to replace certain systems in order to supply the very same information more efficiently. In this case we still have the original systems as references. Source systems are capable of producing some useful reports.

It helps if we ask the opinions of domain experts while validating test results. There are always some products, customers or indicators that people have had experience with. This is why the domain experts along with the developers should specify some