

This paper may not be reproduced or published without the explicit permission of SBSL. Contact info@sbsworld.com for further details

Common Data Defects and Data-cleansing

Overview

" Business is built on enduring Relationships."

" More business from existing customers is ten times more profitable than seeking new customer's business."

Typical data stored in a Data warehouse will be summary of transactions carried out with the customers of the business. So the data already provides a way to look at a customer and see his relationship with the business. Unfortunately this way of looking at a customer is not good enough – you do not get to see the customer the way he sees you. Valuable business insights are missed if the data is used without carrying out the steps of De-Duplication and House Holding. Also there will be avoidable wastages if data which is not De-Duplicated and House-Holded is used for mailing lists.

On the assumption that Data Cleansing has been carried out on the data the House-Holding De-duplication is then done on this data to find out the number of unique individuals. "House-holding" can also be done subsequently to finalise the number of unique households from the de-duped data.

This data can now be used as the foundation for creating a data repository/ warehouse, which contains clean, value-added and consolidated data.

What is Data Cleansing

Data capture systems are often designed keeping in view a Department's perspective or particular business activities perspectives. But a company has more than one Department/ Center or a line of business. Data consolidation across the Company is not possible without Data Cleansing. More particularly this problem becomes acute when there are mergers and takeovers or when business expands without the Data Capture systems being upgraded/ updated simultaneously. When data is captured for operational / legacy systems from input forms, quite often it is not in a usable format for any off-line usage and analysis. This is largely due to inadequate care taken in designing the manual forms and the actual data entry quality. Since these errors hardly affect the operational systems, many times the data remains in its original form for years. This problem becomes more serious when the data from two or more operational systems is brought together for different kinds of analyses. Analysis using such erroneous data can be seriously flawed. Hence, there is a critical need for Data Cleansing, as the first step towards creating a consolidated view of the data.

Data Cleansing involves different steps such as:

- a. Standardize formats – bring data from different systems/ platforms onto a common format
- b. Standardize Name and Address data content – includes transposition of names, repositioning of address elements, standardizing abbreviations etc

- c. Value Additions – such as adding titles to names, gender fixing, Pin code fixing, telephone mapping etc.

Data Defects

Some Examples of Data Defects are given below. Consider an example of a database consisting of fields such as *Name* (in a single field), *Address* (in multiple fields), *Telephone No.*, *Fax No.*, *Sex*, *Age* and *Date of Birth*; and pulled together from three different operational systems:

- a. **Inconsistent Data Values**
Names entered in different sequences
Such as Surname–First Name–Middle Name or First Name–Middle Name–Surname in the same *Name* field
- b. **Domain Schizophrenia –**
Fields being used for different purposes
Such as Fax No. stored in a field for *Telephone No.*
- c. **Missing Data Values**
Data not entered for all records
Such as missing PIN codes, Sex codes
- d. **Incorrect Data Values –**
Caused by transposition of key strokes while entering data
Such as PIN Code entered as 411 000 (which does not exist)
- e. **Domain Value Redundancy**
Non-standardised data values in which two or more values mean the same thing
Such as one operational system gives Age in Years and the other gives Date of Birth
- f. **Non-Atomic Data Values**
Multiple facts entered in the same field
Such as fax and telephone numbers entered in the same field in one of the operational system

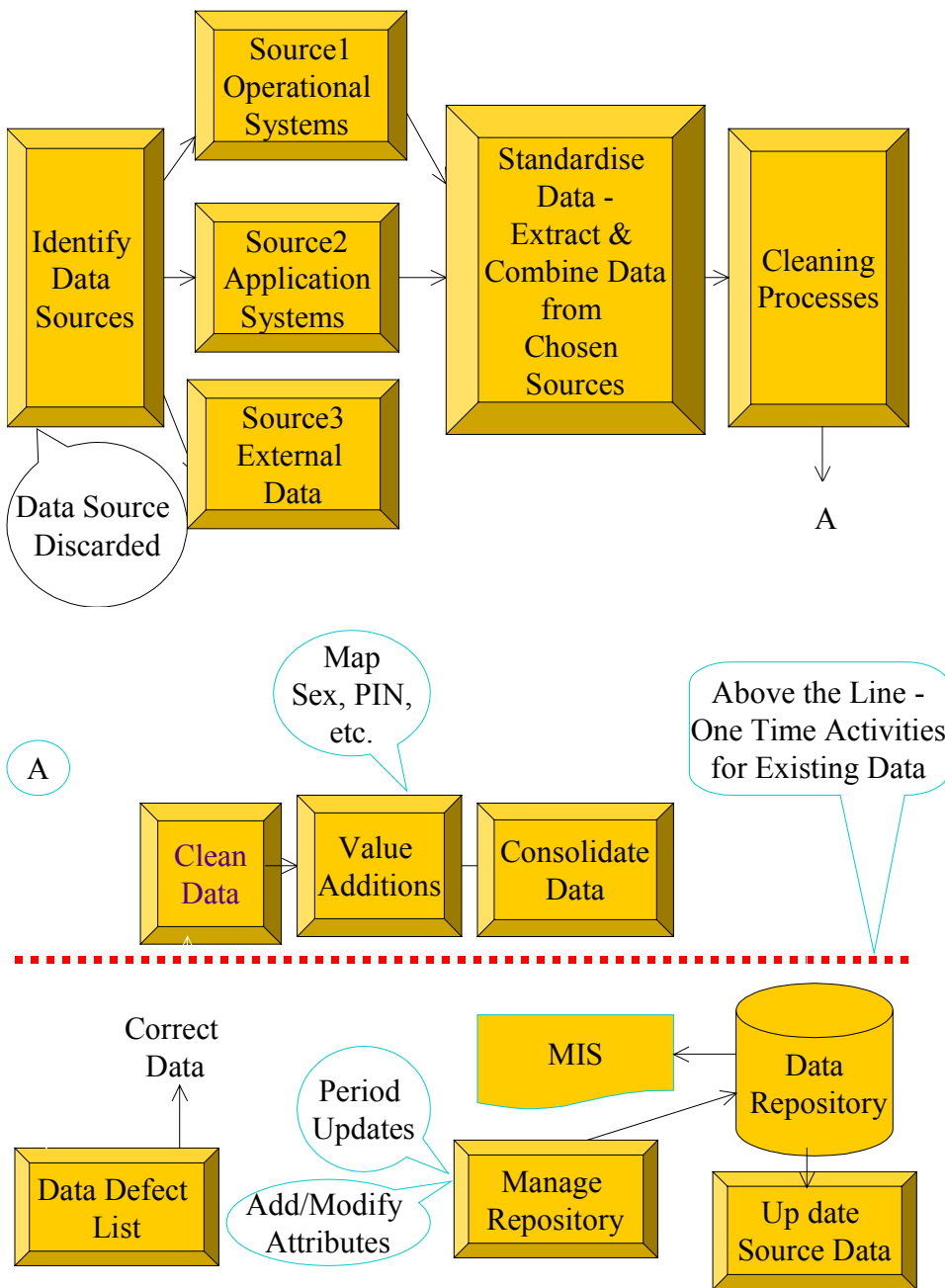
How Cleansing Operations Are Done

The most logical way to remove all such defects would be to go back to the manual data entry forms and either do the proof reading for each record or re-enter the data again. However, these methods are laborious and time-consuming.

Using certain tools and data masters, it is now possible to pass the raw data through a series of computer programmes to get cleaned data. The exception lists and defects lists that are simultaneously generated can then be subjected to manual cleaning using the source documents.

The entire operation can be pictorially represented as follows –

Data Cleansing - Stages



Data Mining

With the data warehouse / repository in place, we can now embark upon Data Mining exercises that would yield information leading to a better understanding of the Customer, along parameters such as:

- Geographic and Demographic segmentation
- Relationship History
- Pattern Identification
- Customer Profiling

Ultimately all these and many other related techniques enable an organisation to communicate better with its own Customers and to build better relationship with them because they see them better.

Why Spectrum

Experience - We have very significant experience in Data Capture and Data Cleaning. Over the last 10 years, we have been engaged in the formation and marketing of information products. Hence, we have created our own databases of 6+ Gigabytes of data, growing currently at the rate of 45-50 MB per month. We have 200+ in-house and out-sourced manpower that is sensitised to these tasks.

Existing Clientele - We have created Data Repositories for entities in the Finance and Telecom Industry

Data Quality Assurance – We have our own in-house systems in place for this purpose, including the conduct of a regular Data audit by an external firm of Chartered Accountants. The Institute of Chartered Accountants of India has given us the responsibility of publishing the ICAI's Professional Pronouncements. Also, we would be happy to adhere to and demonstrate compliance with Client-specified methods and standards.

Confidentiality and Data Security – All our clients need these to be paramount. We have well-oiled systems for this purpose.

Software Capabilities – We have developed our own Full Text Database engine, which is used as the backbone for String-matching operations that yield excellent results in the process of de-duplication and house holding. We have developed automation techniques, software and large, detailed data masters that are especially amenable to data cleansing of names databases in India. We have developed Data Mining applications that are already at work in a number of client locations in the Finance Industry in India.

There can be three business drivers for undertaking Data Cleansing activities. Firstly cleansed data **prevents wastages**, which take place when a Company communicates with its existing customers using defective databases. Secondly, cleansed data lets a Company **see the Customer in its entirety as he sees the Company**. Thirdly Data Mining activities can be carried out to **discover hidden patterns for more business opportunities. The process of Data Cleansing is partially amenable to software solutions but it must be supplemented by manual efforts.**

Copyright © Spectrum Business Support Ltd. All Rights reserved