Improper Data Collection Mechanisms, an Important Cause for Erroneous Corporate Metrics

Dr. Alexandru Petchesi[1]

[1] McMaster University

Received: 12 February 2010 Accepted: 2 March 2010 Published: 15 March 2010

Abstract

This paper intends to highlight one of the very important but most often overlooked aspects related to the challenges of the customization of information systems due to the lack of repeatability and reproducibility during data collection.

Index terms— Knowledge Management, Data Validation, Repeatability and reproducibility of data collection, Corporate metrics.

1 INTRODUCTION

any companies in the 21 st century are monitoring their regular activities through performance metrics. To calculate performance metrics data needs to be collected in a well defined way and stored for analysis purposes. To accomplish this goal many companies invest significant amounts of money into information systems such as Enterprise Resource Planning and Manufacturing Execution Systems to collect and report such data (Fig. **??**).

2 Fig 1: Corporate metrics: Pie chart of expenses

The strategy that many companies use to implement their information highway is through the acquisition of off-theshelf solutions which are then customized to the needs of the company. As currently there is no one software solution that can provide all information services needed by a company, the solution to build an information highway adopted by many companies is to purchase best of breed solutions and then integrate them. These integrations presented and will present quite a lot of challenges to companies due to the large variety of technologies used to implement them. This paper intends to highlight one of the aspects related to the challenges of the customization of information systems due to the lack of repeatability and reproducibility during data collection.

3 About-Alexandru

4 THE PROBLEM

What can go wrong during data collection process that can affect the quality and quantity of data we are collecting and therefore the reports we are generating from our information systems? I would like to present in this paper one of the major risk factors that has an impact on the data collected by an information system, the repeatability and reproducibility of the data collection process.The problem will be exemplified with a very simple case, for a shop floor control system. Imagine working in a manufacturing company that produces a certain product. One of the very important metrics related to manufacturing a product is the quality of the product which is measured in most companies through metrics such as first pass or rolled throughput yield. To calculate metrics such as the ones mentioned above, companies need to collect information on the products they manufacture such as the number of products with defects. An important characteristic of the data is related to its granularity , mainly related to the categories of defects that can be identified on a product. The data collection process of such data is done in many companies through operators which need to visually identify the cause of failure, then pick their data manually from a list of options offered to them by the software. Data collected in this manner lead to reports such as the Pareto chart presented below that gives decission makers within a company the information needed to identify the route causes of problems and take the necessary actions based on them. Therefore the accuracy of such a data collection process is very important as a report as the one presented below gives people in a company an image of the realities within the production process from within a company. If data is -distorte d? the image provided through the reporting mechanism is also distorted and does not reflect the realities from the factory.

5 M

The reports as the one presented in Fig. 2 provide companies images of the realities from within the factory and give them clues on the area of the process where they need to act upon to start improvement projects.People in information technology are very familiar with the -Ga rbage in Garbage Out? principle. To reduce or even completely eliminate this problem from a software application, the information technology community has developed defensive programming techniques. One of the best practices of data collection tells us that, in order to assure that the data we collect from the end users is right it is preferred to employ in a in the graphical user interface of a software application, pre-defined selection mechanisms such as combo boxes, selection lists etc. These allow the end-users to easily perform single or multiple selections of values from a well defined set.The data set for such a list is usually defined by the subject matter expert working on the business side with IT experts in charge with the customization of the tool. During the lifetime of the software product, employees using the software will use such a combo box to pick the proper values and submit them into the database for storage. When we are inputting data into an information system, the IT best practices are telling us, we need to make sure that we avoid the garbage in garbage out principle. The major effect of the pronciple above is that once the data collected is -con taminated? in our storage it will affect all the information systems from within our architecture that use this data source as the master. The result is that erroneous information will spread all across the company and this information can cost us significant amount of data due to spending the company might make as a result of the reports provided (Fig. **??**).

6 Fig 3. A combo box

What can go wrong during such a data collection mechanism that can affect the quality and quantity of data we are collecting? Everything seems to be properly set up from an IT prespective, but the employees of the company using the reports are sometimes complaining about discrepancies between the realities they are aware of and the data from the reports. Many of them become quickly frustrated and start loosing the trust on the reports provided to them many times by expensive software tools with a steep learning curve.The experiment presented below will identify an overlooked way of erroneous data entering an information system due to the lack of repeatability and reproducibility of the data collection process.

III.

7 THE EXPERIMENT

We are going to illustrate the repeatability and reproducibility issues of data collection through an example from the electronic manufacturing industry. The experiment was conducted a long time ago and the purpose of it in this paper is for exemplification only. The experiment will present the Gage R&R methodology from Six Sigma, an important statistical tool that can allow us to determine the repeatability and reproducibility of a data input process. Table **??** The list of issues for each location on the board (the standard)

A printed circuit board (Fig. 4) was used and marked with 30 locations, some locations marked had defects some did not have any defects, to identify the accuracy of the data collection mechanism. Three operators were selected randomly to determine how close their selection of data from a particular set was to the standard (Table **??** 4). The three operators selected were presented with a set of allowable values and they were asked to pick defects from a list of standard defects providedby in information system. This data selection mechanism was used by them already in their daily activities through an information system, using data provied by a combo box, where they needed to select one value from a list. Their answers were collected in a spreadsheet and in case their answer matched the standard defined by the expert a PASS was introduced in the Gage R&R tool and a FAIL was introduced in case their selection did not match the standard. (Fig. 5).

The experiment was repeated a week later with the same operators on the same printed circuit board without informing them about the fact that it was the same product.

The data collected from the second session was introduced in a similar way in the Gage R&R tool, as seen below. The spreadsheet then calculated for us the differences between what each operator's option and the standard defined by the expert providing us very valuable information on the data identification and selection mechanism.

Using the Gage R&R method we looked at the consistency of the data selection mechanism for each individual, between individuals and against the standard. The conclusion we drew were pretty interesting! The statistical analysis of the repeatability and reproducibility of the data selection process are shown in Fig. **??**.

8 IV. CONCLUSIONS

As seen in the results above (Fig. **??**) there are significant discrepancies for all 4 categories tested. The report tells us that the values picked from the list by the operators and entered in the information system and the realities as defined by the standard are significantly different. If this data would have been entered into an information system the reports generated from the data entered would have been very different from the realities from the factory and actions might have been taken in the wrong direction by the team using reports based on the data. Therefore, it is important that any data entry process which collects data introduced by human operators based on non-numeric criteria must be validated on regular basis for the repetability and reproductibility. Without this validation the money invested in information systems will not provide the value adds they for and can even produce significant financial losses to companies due to erronous reporting.

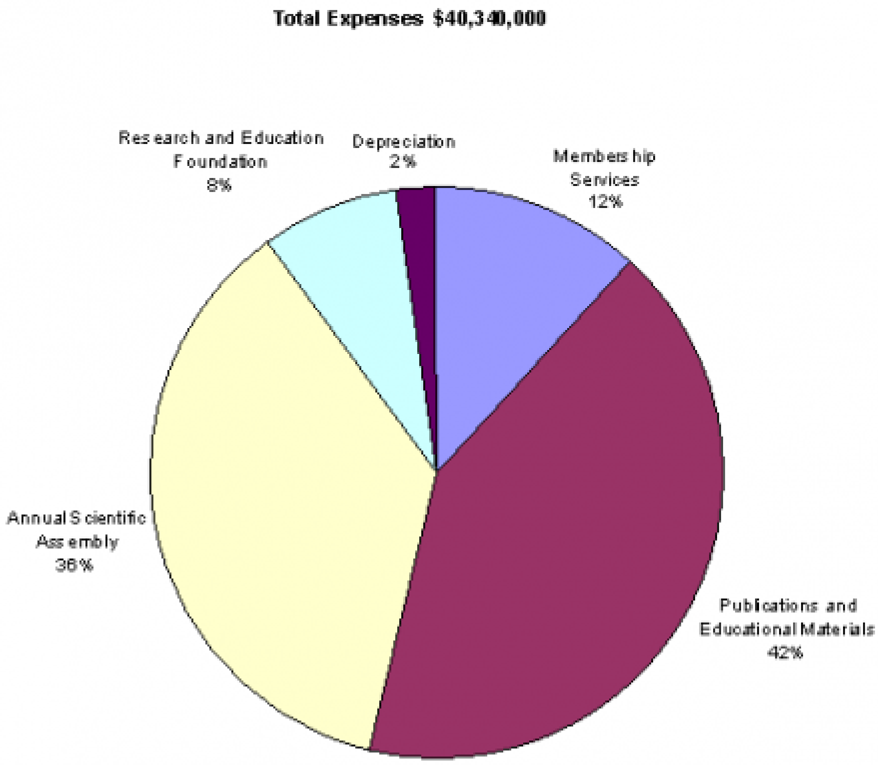


Figure 1:

2

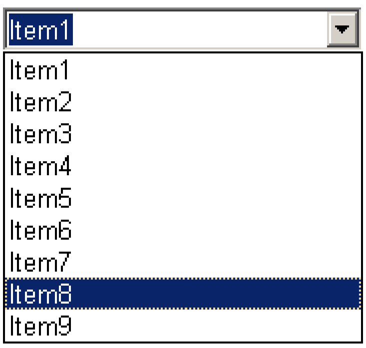


Figure 2: Fig. 2

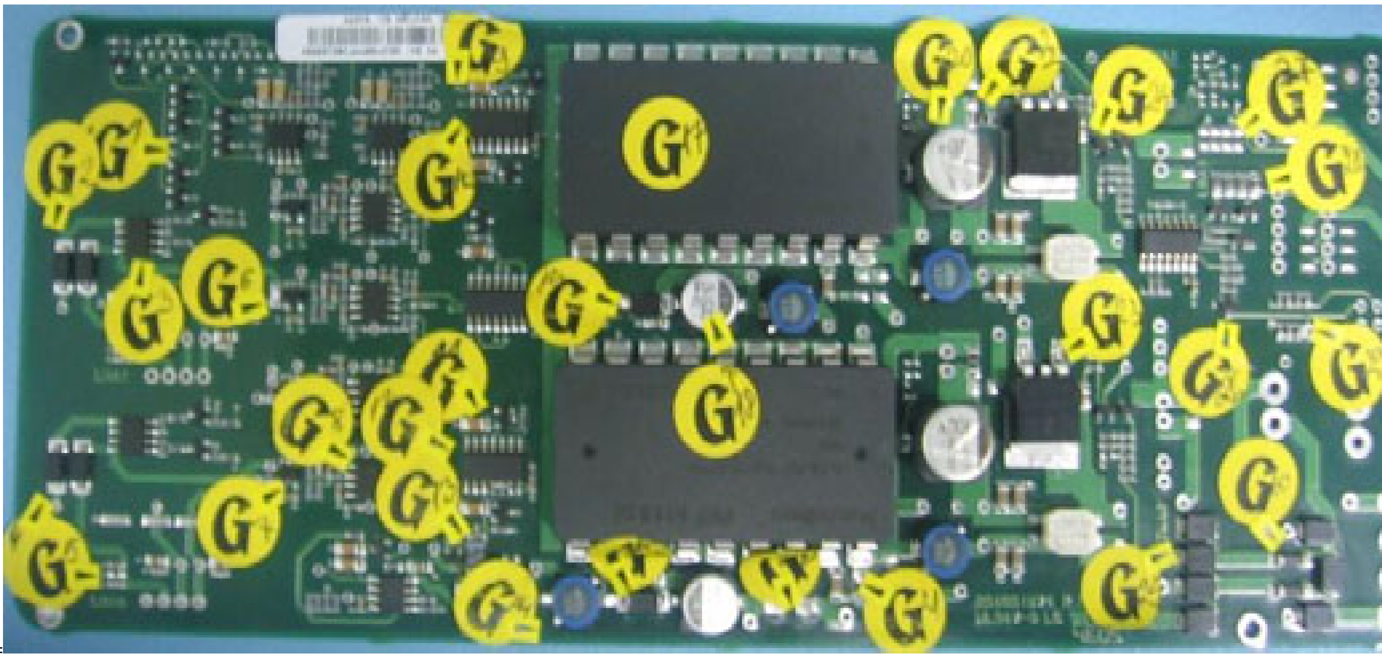
4 

Figure 3: Fig 4 .

SCORING REPORT

Attribute Legend[a] (used in computations)

1 PASS

2 FAIL

DATE: 26.09.2006

NAME: Expert

PRODUCT: Defect identification

BUSINESS: VISUAL INSPECTION

Known Population		Operator 1		Operator 2		Operator 3		All operators agree within and between each Other	All Operators agree with standard
								Y/N	Y/N
Sample #	Attribute	Try #1	Try #2	Try #1	Try #2	Try #1	Try #2	Agree	Agree
1	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
2	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
3	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
4	PASS	FAIL	FAIL	PASS	PASS	FAIL	FAIL	N	N
5	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
6	PASS	FAIL	FAIL	PASS	PASS	FAIL	FAIL	N	N
7	PASS	FAIL	PASS	PASS	PASS	PASS	FAIL	N	N
8	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
9	PASS	FAIL	PASS	PASS	PASS	FAIL	FAIL	N	N
10	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
11	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
12	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
13	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
14	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
15	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
16	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
17	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
18	PASS	PASS	PASS	FAIL	FAIL	FAIL	FAIL	N	N
19	PASS	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	Y	N
20	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	N	N
21	PASS	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	Y	N
22	PASS	PASS	FAIL	PASS	PASS	FAIL	PASS	N	N
23	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	N	N
24	PASS	PASS	PASS	PASS	FAIL	FAIL	PASS	N	N
25	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	N	N
26	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	N	N
27	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y
28	PASS	PASS	PASS	PASS	PASS	PASS	FAIL	N	N
29	PASS	PASS	PASS	PASS	PASS	FAIL	PASS	N	N
30	PASS	PASS	PASS	PASS	PASS	PASS	PASS	Y	Y

5

Figure 4: Fig 5 .