Successful Business Intelligence Systems: Improving Information Quality with the SAS® System

W. Droogendyk, Dofasco Inc., Hamiliton, Ontario, Canada
L. Harschnitz, Dofasco Inc., Hamilton, Ontario, Canada

Introduction

Dofasco Inc. manufactures flat rolled steel products, combining traditional Basic Oxygen Furnace, and Electric Arc Furnace technology at our Hamilton, Ontario plant with mini-mill technology at our joint venture plant in Gallatin, Kentucky. Over the past decade, wherever possible, Dofasco has been managing by data and information, rather than by intuition and experience. We have begun to treat data as a corporate asset, and realize that information can provide a competitive advantage. We have also become aware of how data moves through the business to become available as information. Figure 1 shows this pictorially.

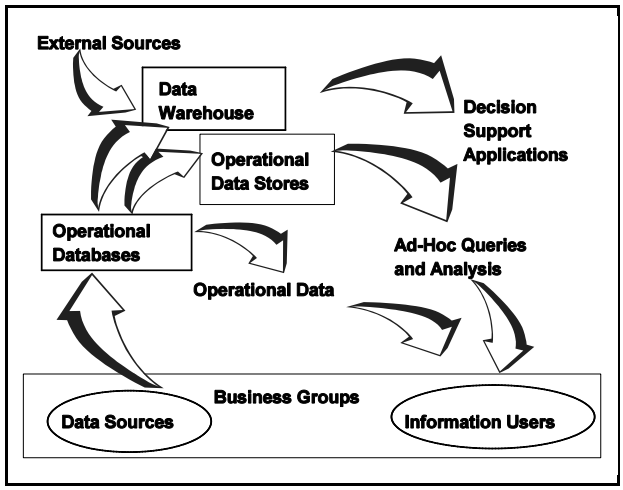

Figure 1 **:** Data/ Information Loop

We consider both our Operational Data Stores, and our Data Warehouse to be business intelligence systems. Both address the user need for connected, accessible, meaningful, and complete data. The Data Warehouse also has the characteristics of being static, disconnected from Operational Stores, and containing external data if required. From Figure 1 it can be seen that while information is extracted from the business intelligence systems, information quality must be built into the operational databases. Therefore, successful business intelligence systems begin with a rigorous Information Quality Program at the Operational level. Also, it can be seen that any information quality program becomes a joint endeavour between business groups and those responsible for the information systems.

This paper is an overview of our journey from recognizing and attacking data integrity problems to the development of an Information Quality Program, and some practical examples of leveraging the power of the SAS® System to get there and stay there.

2.0 Information Quality

The mission of an Information Quality Program is to maintain data, such that it can be combined with knowledge, to become business intelligence that provides a strategic or competitive advantage. This section of the paper discusses the basics of information quality. Understanding what information quality is, where problems may arise, how the lack of it affects business intelligence, and some of the solutions to information quality problems is the starting point of the journey to an Information Quality Program.

2.1 Understanding Information Quality

High information quality is achieved when high levels of data integrity and quality are combined with appropriate analysis and good business knowledge.

Data integrity can be described as how true the data values are to the current definitions and business rules. Data entered into the system is expected to be both individually and referentially accurate. Individually accurate in the sense that it is a valid value for that field, and referentially accurate in that it makes sense when combined with other fields in the same record. A daily high temperature of 91°F is individually accurate, but if it is attached to a record for New York City in January, it is not likely referentially accurate.

Data quality can be described as how well the data is structured to addresses business and analysis needs. A purchasing process may need the flexibility to either pay multiple invoices together, or parts of a single invoice separately. At the same time, the business may wish to be able to analyze costs by item. If this is the case, the data structure must be designed so that the cost of a specific item can be extracted from the appropriate payments for analysis.

2.2 Discovering Poor Information Quality

At best poor information is recognized immediately and excluded from the decision making process. At worst poor information is not recognized, and causes incorrect business decisions. In most cases, poor information is discovered at the end of the data/information loop when it reaches users as shown in Figure 1.

This type of information tends to be generated using data which has been stored for some time, and is difficult to repair. Modern operational systems, which rely on direct entry of disciplined data, are nearly impossible to correct after even a short period of time. Problems with data structures may require system changes before repair is possible.

In Dofasco's case, the need for increased information quality was recognized when an inaccurate trend in quality data

appeared over the course of several months. Fortunately, good business knowledge prevented this information from influencing decisions. What was lost, however, was the ability to use the analysis to target potential opportunities. The project to repair the data, and solve the root causes of the integrity problem consumed a portion of ten people's time, and six months to complete.

2.3 Sources of Poor Information Quality

Poor information quality occurs when the information available is inaccurate or misleading. It is caused by breakdowns in the processes and infrastructure that generate the information, and is rarely a people issue. Data creation processes, data structures, and the analysis techniques used to transform the data into information may all be sources of poor information quality.

Problems with data include incorrect or missing values, and referential integrity violations. They may stem from inadequate training, systemic problems within the database loaders, or a change in the business process that has not been reflected in the operational system.

Problems with data structure include vague or incorrect definitions, conflicts with business processes, unresolved many to many relationships, and data which is not retained. In most cases the data structures have not evolved as the business processes have, or are inconsistent because of stove pipe development.

Breakdowns in the analysis techniques used to transform data into information are usually related to such things as data used incorrectly, connected incorrectly, or analyzed in an invalid way. In many cases, the data structures contribute to these problems with inconsistent field names and definitions. Also, there may be little or no support for the knowledge worker in determining the correct way to approach the analysis.

2.4 Some of the solutions

In order to maintain good business information, a comprehensive information quality program is required which is based on strong data principles, architecture, stewardship, and data management processes.

To address data problems, a proactive data integrity program at the data sourcing point of the loop is essential. This allows for early detection of errors, and provides the maximum potential for correction of those errors. The identification and solution of systemic problems is key to continually improving data quality, thereby reducing the resources consumed by data monitoring and repair.

Data structure problems are addressed by the maintenance of data and business process models, and the storage of Metadata. It is critical that data and business process models be fully connected.

Knowledge workers require training and support in data and data structures so that they are adept in early recognition of poor information, and can assist in the correction process. The development and maintenance of a good metadata navigator is a useful way to provide partial support, as is a centralized query and data support group.

Dofasco has, over the last several years, been working in various areas to improve information quality. These activities are currently being consolidated and expanded into an Information Quality Program. Sections 3 and 4 of this paper describe the various activities.

3.0 Elements of an Information Quality Program

Once the need for an Information Quality Program has been identified, and its objectives and deliverables determined, the next step is to design the elements of the program. The following section describes the elements of an information quality program, and how they are being implemented at Dofasco.

3.1 Principles

Well defined, and supported data principles are critical to the success of an information quality program. These are the beliefs which govern all other actions and processes. Dofasco has the following data principles which drive the information quality program.

Data is a corporate asset. This expresses the understanding that data is owned by the whole corporation, and that its use can provide a strategic competitive advantage to the corporation. Data is valuable, and needs to be preserved and nurtured like any other corporate asset.

Data is shared and reusable. This acknowledges that data is dependent only on the business process which creates it. Once created, it can be shared and reused by many systems and people, given that users are responsible for the integrity of their analysis.

Data evolves with the business. This acknowledges that a business is an ever changing set of processes, responding to customer, shareholder, and employee needs; as well as market, and community changes. As business changes, some new data is required, some old data is obsolete, and some data changes in its relationship to other data.

Data has a single definition and a single source, which is as close to the point of creation as possible. This is the key to stopping data anarchy, and defining responsibility. If a data element has only one definition, every time that an instance of that data element is created or used, it is with regard to that static definition. Instances are created in a consistent way, and analysis of the data yields clean information. Further increasing the consistency of data is the fact that there is only one source, which is as close to the point of creation as possible. This expresses the belief that data is most accurate and complete at its point of creation, and that its creator is responsible for its accuracy. This ensures that the responsibility for creation consistency is well known, and can be accompanied by the authority to enforce data integrity. Also, this prevents the inadvertent corruption of data by competing sources, and the inefficient use of resources in redundant data capture.

Data will be gathered through the implementation of a single logical data model. This is to acknowledge that the best way to gather data which is shared and reusable is within a single

logical model which relates to the business process.

3.2 Architecture

The data architecture is designed to support the beliefs expressed in the data principles. It provides a logical framework for the creation and use of data. Dofasco's data architecture begins with its choice of a Unix® based, client/server infrastructure, which uses an Oracle® RDBMS and HP® servers. This choice of a flexible, open infrastructure was made to position Dofasco to be able to take advantage of distributed processing and purchased applications.

Data elements are created in a standard data modelling process, which ties data to the appropriate business process, and are represented in a normalized form A standard set of metadata is captured for each data element which includes, but is not limited to, the name, definition, allowed values, and data stewards. Data element names are based on a set of standards and abbreviation rules, and, along with the data models, are reviewed prior to implementation. Operational databases tend to be implemented in normalized form, but decision support databases may be denormalized. The data warehouse is multi-tiered, and subject oriented. Databases are designed first, forcing applications to be data driven.

Data cannot always be directly shared from a single database. Data confidentiality, or the number of system users may lead to a requirement for data replication. When this occurs Dofasco uses a replication strategy that requires replication to be fed from the source of the data, with a predetermined synchronization method.

Information is extracted from the databases using a standard set of user query tools. Microsoft's Access®, and Excel® , Platinum's Forest and Trees®, and the SAS System provide a wide range of tools for data access and analysis.

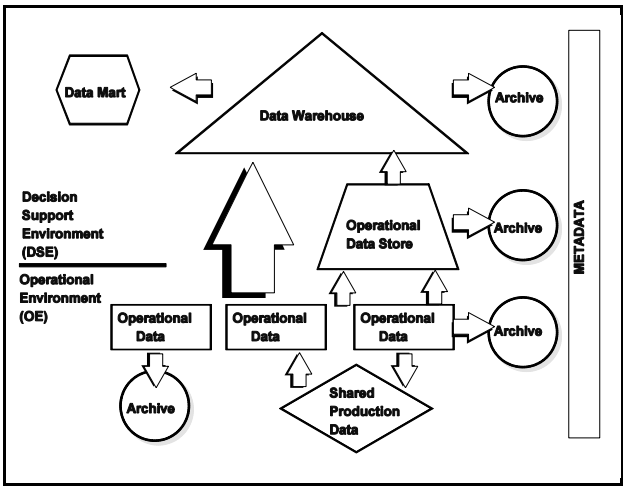

Figure 2: Data Architecture

3.3 Stewardship

Data stewardship is an excellent way to approach the management and maintenance of data as a corporate asset. At Dofasco the data stewardship program began in 1994, and is expanding to cover all data elements.

Dofasco splits the stewardship responsibilities into four categories. Strategic data stewards are responsible for definitions, and the business rules that govern allowed values and the creation of data. Operational data stewards are responsible for the creation of instances of data which comply with the data definitions and the business rules. Knowledge stewards/workers are responsible for the correct use of data and analysis techniques in converting data to information. Data experts support all three types of stewards by having an in depth understanding of data and systems in a particular business area.

3.4 Management System

Data principles, data architecture, and data stewardship are all implemented through a series of data management processes. These processes are the joint responsibility of data stewards and the Data Services group within Information Systems.

Processes exist for identifying and defining data elements through the connection of business processes to data, and for capturing and storing metadata. These processes are the responsibility of the strategic stewards, with Data Services facilitating consensus discussion and acting as metadata custodians.

Processes for data creation, data auditing, and data repair are the responsibility of the operational steward. They are assisted by the data experts, who are also heavily involved in driving the data quality improvement process. The assessment of the cost of non-quality data is a critical measurement for this process.

Information delivery processes are tied to the standard query and analysis tools. Data Services provides data training, and support for the standard tools , to increase the understanding of knowledge stewards.

4.0 Leveraging the SAS System to improve Information Quality

The SAS System is a strong, comprehensive, and flexible tool that is used to support the processes within our Information Quality program. The following section gives examples of how Dofasco uses SAS to do this.

4.1 Screening, Monitoring

The capabilities of the data step allow for complex screening and monitoring of data for integrity errors. The purpose of screening and monitoring is to quickly identify data problems so that the data can be repaired and the processes corrected.

4.1.1 Product Serial Number Integrity

Each steel coil which we produce is identified using a serial number, which is used to collect processing information as the coil is finished. The serial number consists of a single alpha character, from a restricted list, followed by a number between 10000 and 99999. The serial numbers are to be used consecutively and uniquely. Gaps in serial number ranges in Dofasco's historical data were discovered

accidentally and a program was written to detect and report on these gaps. Corrective actions to the data transfer between the operating and historical systems were implemented, and monitoring continues on a daily basis. Initial errors measured in the order of 10% and have dropped to a new level of 60 ppm. These are manually corrected as detected. The statements given below are part of the program used to identify these errors.

```
*****COMPARE SERIAL NUMBERS OF ADJACENT RECORDS
AND OUTPUT SKIPS, DUPLICATES, INVALID SERIALS ETC;

data missing;
length error $20;

do i=1 to last by 1;
  set sers point=i nobs=last;
  this_one=substr(ser_no,2,5);
  last_ser=ser_no;
  j=i+1;
  set sers point=j nobs=last;
  next_one=substr(ser_no,2,5);
  next_ser=ser_no;
  missing=(input(next_one,6.0))  - (input(this_one,6.0)+1);

  if missing gt 0 then do;
     error='MISSING SERIAL(S)';
     output;
  end;
  else if ' ' le next_one lt '10000' then do;
     error='INVALID SERIAL';
     output;
  end;
  else if missing = -1 then do;
     error='DUPLICATE SERIAL';
     output;
  end;
*****CORRECT GAP FOR NEW SERIAL PREFIX IS -89999;
  else if missing lt -1 then do;
     missing=89999+missing;
        if missing=0 then error='NEW PREFIX OK';
          else error='LOOK FOR CAUSE';
     output;
  end;
if j=last then stop;
end;
stop;
run;
```

4.1.2 Unique Code Integrity

Dofasco's customer and processing requirements sometimes result in a single coil being split into several parts. When this occurs a part number is added to the original serial number. For traceability, we replicate the entire coil history for each part. To accommodate certain types of analysis, such as yield calculations, the unique code field is set to 'N' for any replicates. Other entries which are not replicates have null values in this code. This code is determined through a complex set of programs which occasionally do not work properly, and cause errors in corporate measures. The recursive read program below collects violations and reports their occurrence in a simple list form for issuance to the parties concerned.

```
select distinct
     a.ser_no, a.ser_part_no, b.prev_split_op_no, a.pce_wt "a_pce_wt",
    c.pce_wt "c_pce_wt", a.unique_cd "a_uniq", c.unique_cd "c_uniq"
from act_oper a,act_oper c,ser b
where a.ser_no = c.ser_no
  and a.ser_no = b.ser_no
  and a.ser_part_no < c.ser_part_no
  and a.ser_part_no = b.ser_part_no
  and a.ser_part_no > '0'
  and a.unique_cd is null
  and c.unique_cd is null
  and a.ser_no between 'A10000' and 'A10001'
  and a.pce_wt=c.pce_wt
  and a.op_seq_no=c.op_seq_no;
```

4.1.3 Serial Part Number Integrity

Should a coil be split, it retains its original serial number but the default part number of "0" is changed to 1 to 9 as applicable. A serial with a part number of "0" can have no other part numbers. The program below collects records for all part number = "0" occurrences and non "0" occurrences and compares the two lists. Serials in both lists are identified for correction.

```
create view zero as
   select * from connection to oracle
     (  select distinct a.ser_no
          from  act_oper  a
            where
              a.pce_process_date >= '10-mar-96'
              and a.ser_part_no='0');

 create view not_zero as
    select * from connection to oracle
      ( select distinct a.ser_no
          from  act_oper  a
            where
              a.pce_process_date >= '10-mar-96'
              and a.ser_part_no > '0');

proc sql;
create view both as
select distinct a.*
   from zero a ,not_zero b
      where a.ser_no=b.ser_no;
```

4.1.4 Missing Disposition Reason Codes

Whenever product processing deviates from its planned routing, for reworking, repairs, scrapping or reapplication (divert), we require that operational personnel post the reason for this event. These reasons are essential to Dofasco's corporate Quality Improvement Projects, both for initiation and tracking. Dofasco's Cost of Quality system relies extensively on the integrity of these reason codes. The program below is an example of searching for and reporting missing codes. The resultant listing is made available to the operational areas for their use to correct records which are incomplete or incorrect.

```
SELECT OPER_CD1 LABEL='OPERATION PASS',
       OPER_YMD LABEL='OPERATION DATE',
       COIL_WHO  LABEL='SERIAL',
       COIL_PAR LABEL='PART',
       OPER_WT LABEL='WEIGHT',
       MILL_PRO LABEL='PRODUCT',
       DISP_CD LABEL 'DISPOSITION',
       OPER_CD2 LABEL='OPERATION',
       CUSTOM_C LABEL='CUSTOM CODE',
       DEFECT_1, DEFECT_2
  FROM ADR.ACTUALOP AS A
   WHERE OPER_YMD GE "9612111" AND UNIQUE_C NE 'N'
```

```
 AND (((DISP_CD ge '0'
  OR CUSTOM_C IN ('1','5','6','8','9','D','E','G','J','K','L','P','S','T'))
 AND DEFECT_1 LT 'A00' AND DEFECT_2 LT 'A00'
  OR (DEFECT_1 LT 'A00' AND DEFECT_2 GE 'A00'))
 OR (DISP_CD LT '1' AND DEFECT_1 GE 'A00' AND DEFECT_2
GE 'A00'));
```

4.1.5 Customer Service Call Reports

These call reports are used to communicate and resolve difficulties between Dofasco and our customers. Often, manufacturing responses are required and the reports need to be filled out accurately. Various date fields are used to determine the speed of response for various activities within Dofasco's customer service system. These activities are measured and reported. The program below looks for reports which are incomplete or have incorrect data. The printouts are forwarded to the account representatives for follow up.

```
select      a.rpt_year_no      "RPT_YEAR",a.rpt_servc_repr_cd
"RPT_SERV",
     a.rpt_no "RPT_NO", a.complaint_allow_flg "JUSTFIED"
  from  rpt_cust a
    where  a.contact_date > (sysdate - 180)
           and a.claim_close_date is null
      order by rpt_year_no, rpt_servc_repr_cd, rpt_no;

select      a.rpt_year_no      "RPT_YEAR",a.rpt_servc_repr_cd
"RPT_SERV",
           a.rpt_no "RPT_NO",sum(a.disp_pce_wt) "DISP_WT"
  from  rpt_disp a
    where  a.rpt_year_no=96
       group by rpt_year_no, rpt_servc_repr_cd, rpt_no
         order by rpt_year_no, rpt_servc_repr_cd, rpt_no;

select rpt_year_no "RPT_YEAR",rpt_servc_repr_cd "RPT_SERV",
     rpt_no "RPT_NO",contact_date "CONTDATE",
     contact_name "CONTACT",req_resp_date "REQ_RESP",
     resp_pers_name "RESPNAME",
     action_cd "ACTIONCD",act_resp_date "ACT_RESP",
     rpt_comp_date "RPT_COMP",complaint_allow_flg "JUSTFIED"
  from   rpt_cust
    where  contact_date > '01-jan-95';
```

Various data step statements are used to check data validity and make comparisons, similar to the ones listed below.

```
rpt_time=rptcomp-cont_dat;
response=actresp-(dateprt(req_resp));
if justfied  = 'Y' and disp_wt <= 0 then output;
else if justfied = 'N' and disp_wt > 0 then output;
if actioncd='Y' and actresp in (., 0) then do;
        overdue=today()-(datepart(req_resp));
        if sign(overdue) lt 0 then overdue=.;
end;
if rpt_time lt 0 then output;
if 0 lt reqresp lt cont_dat then output;
if 0 lt actresp lt cont_dat then output;
if rptcomp= . or cont_dat= . then output;
if justfied in (' ','I') and rptcomp ge '01jan96'd then output;
```

4.2 Reporting

SAS is also used to produced listings and graphs which can be used to facilitate correction, or to quantify problems. The two examples which follow show the code required to produce simple reports and graphs.

4.2.1 Exception Listing

The following code produces a listing similar to the one shown as Table 1. This is typical of the listings which we use for ongoing data repair.

```
title1 'RECENT SERIALS MISSING FROM LOGDSE.PIECE
RESULT TABLE as of';
proc print data=missing label noobs;
label hr_est='Hot Roll Date'
     hr_shift='Hot Roll Shift Code'
     last_ser='Last Serial on File'
     next_ser='Next Serial on File'
     missing='Number of Serials Missing'
     error='Error Message';
var hr_est hr_shift last_ser next_ser missing error;
run;
```

Table 1: Typical Data Repair Listing

RECENT SERIALS ERRORS FROM LOGDSE.PIECE RESULT TABLE
as of 08:56 Friday, December 13, 1996

Hot Roll Date	Hot Roll Shift Code	Last Serial on File	Next Serial on File	Number of Serials Missing	Error Message
17SEP96	1	C40347	C40347	-1	DUPLICATE SERIAL
03NOV96	2	C63586	C63586	-1	DUPLICATE SERIAL
26NOV96	1	C73026	C73028	1	MISSING SERIAL(S)

4.2.2 Typical Graph

When we begin to audit data to see if a data quality problem exists, we typically produce graphs showing exception levels over time, such as the one shown below. Most of these are simple vbar charts which are generated with two or three lines of code such as the ones which follow.

```
proc gchart data=server.excep;
  vbar date/discrete sumvar=count;
run;
```

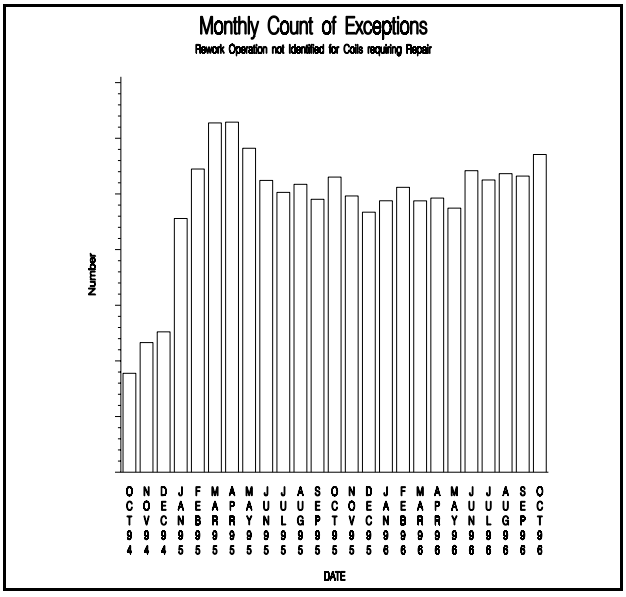

Figure 3: Typical Exception Graph

4.3 Improving

SAS is also used at Dofasco to assist with the primary objective of an information quality program, which is to improve. Run charts are used during periods of focussed improvement, and then gains are retained using control charts.

4.3.2 Disposition Reason Code not Reported

Example 4.1.4 showed how missing disposition reason codes are tracked and reported for correction. The following graph, shown as Figure 4, is used to monitor the ongoing improvement as a result of this tracking. This graph shows the dramatic improvement possible through an information quality effort.

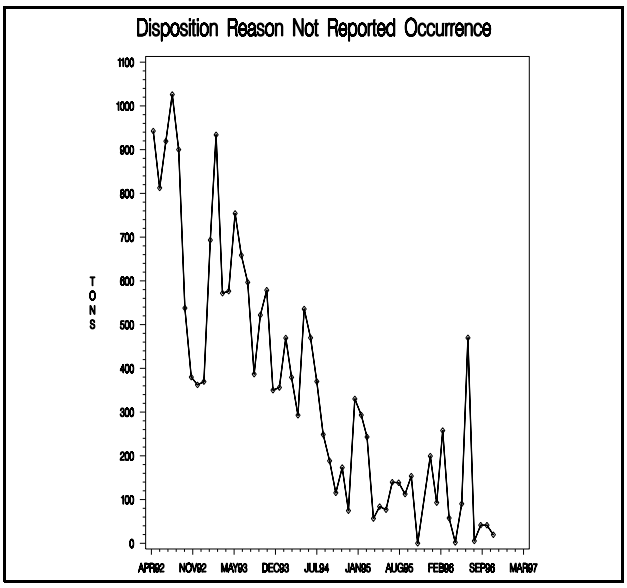

Figure 4 : Data Quality Improvement

4.3.2 Product Charge & Finish Weights

As the product moves through various operations, weight losses are expected. Weight gains are errors which need to be controlled. Weight "gains" usually occur when a weigh scale is down and the calculation model being used to predict a charge or finish weight is in error. Model errors are monitored through weight discrepancies and reported via a Shewhart chart, as shown in Figure 5.

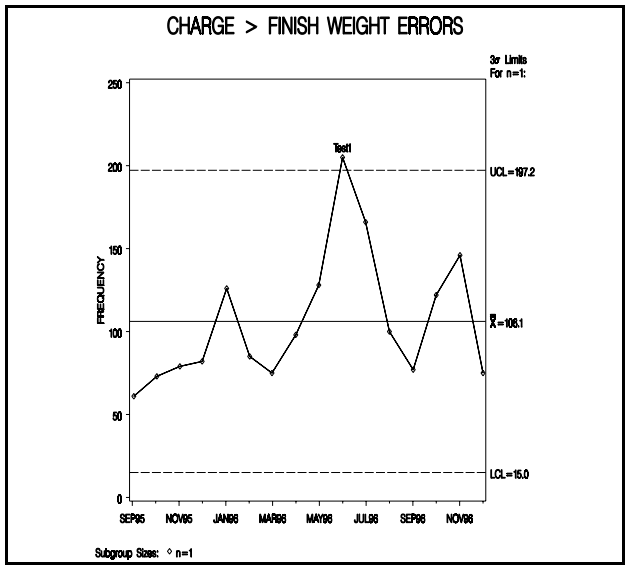

Figure 5 : Control Chart of Errors

5.0 Summary

Developing an Information Quality program begins with the realization that there is a huge cost in not having one. While it can be a challenge to quantify, most companies can point to instances where poor information led to poor decisions or missed opportunities. Once the need for an information quality program has been realized, then building the program begins with a foundation of data principles and data architecture. The addition of a data stewardship program, and the required information quality processes provides the elements required.

The SAS System has a variety of analysis tools that can be used within the information quality processes to identify, analyze, and present quality problems, and to track quality improvements.

Dofasco has successfully combined the use of the SAS System with it's information quality program to both make gains in data quality improvement, and to develop and expand it's information quality program.

W. Droogendyk, Dofasco Inc.
Box 2460, Hamilton, Ontario, Canada, L8N 3J5
(905)544-3761 ext. 3359
bill_droogendyk@dofasco.ca

L. Harschnitz, Dofasco Inc.
Box 2460, Hamilton, Ontario, Canada, L8N 3J5
(905)544-3761 ext. 6415
lesley_harschnitz@dofasco.ca