A SYSTEMS APPROACH TO MONITORING FINANCIAL DATA QUALITY ASSESSMENT AND IMPROVEMENT

(Completed Research Paper)
IQ Concepts, Tools, Metrics, Measures, Models, and Methodologies
IQ Practices: Case Studies and Experience Reports

Donna Fletcher
Bentley College, Waltham, MA
dfletcher@bentley.edu

Mary Ann Robbert
Bentley College. Waltham, MA
mrobbert@bentley.edu

Kurda Mohamad
kurda_mail@yahoo.co.uk

Phillip Middleton
philmid@ntlworld.com

Abstract: Producing financial data of high accuracy requires knowledge about the collection process, the storage process and the utilization process. Data collectors, data custodians and data users require a proactive approach, sharing knowledge to improve data quality. This paper defines a process approach to monitoring, assessing and improving financial data quality, exemplified by potential exposure calculations.

Key Words: Data Quality, Information Quality, Risk Management

I. INTRODUCTION

The business of all financial institutions involves information systems and the data that flow through them. Data links policy to operations; i.e., it translates objectives into measurable performance indicators [1]. One such measure used by financial institutions in credit risk management is potential exposure (PE). PE is an estimate of the exposure that a customer with a portfolio of trading transactions (e.g., foreign exchange and derivatives) may owe over the life of the portfolio.

The objective of this research is to use a systems approach to monitoring, assessing and improving financial data quality used in calculating potential exposure for measuring credit risk. While we are not offering a methodology that can be easily extracted for application by other practitioners, in this paper, we present examples of the data processes and data quality analysis used by a large financial institution[1]. We detail the error map that describes the PE calculation process and identifies when and where errors enter the system.

[1] For reasons of confidentiality, we present examples of the data processes and data quality analysis that cannot identify the financial institution.

Two main approaches to avoiding a significant rate of errors in data are (1) validating data as they are input to or stored in databases and (2) depending on users to detect and correct errors [2]. Our research indicates that while the latter approach is utilized by the data quality teams, timely, complete and accurate data for assessing PE is dependent on validating data as they are input and stored in databases.

Moreover, a combination of objective and subjective quality assessment greatly improves the detection of errors within a firm and enables institutionalizing its data quality improvement program [1, 5]. We suggest several stages within the data assessment process where such a combination is beneficial. Finally, we find that knowledge sharing by data collectors, custodians and consumers, while essential to production of high quality data [3, 4], is not fully realized in the firm studied in this research.

The remaining sections of this paper present (1) a review of prior research related to data quality processes, dimensions and analysis, (2) the PE models and data processes supporting the models, (3) the data quality assessment results and analysis and (4) conclusions and suggestions for further research.

II. RELATED WORK

According to Lee and Strong [3], data are produced by a process that starts with the collection of raw data and ends with the utilization of information[2] products by information consumers working on various tasks, as depicted in Figure 1.

The quality of the data produced is determined by the activities performed as part of the data production process. According to the conventional wisdom about processes, the ability to meet the goals and objectives of a process depends on workers who are knowledgeable about the entire process, beyond their individual work activities.

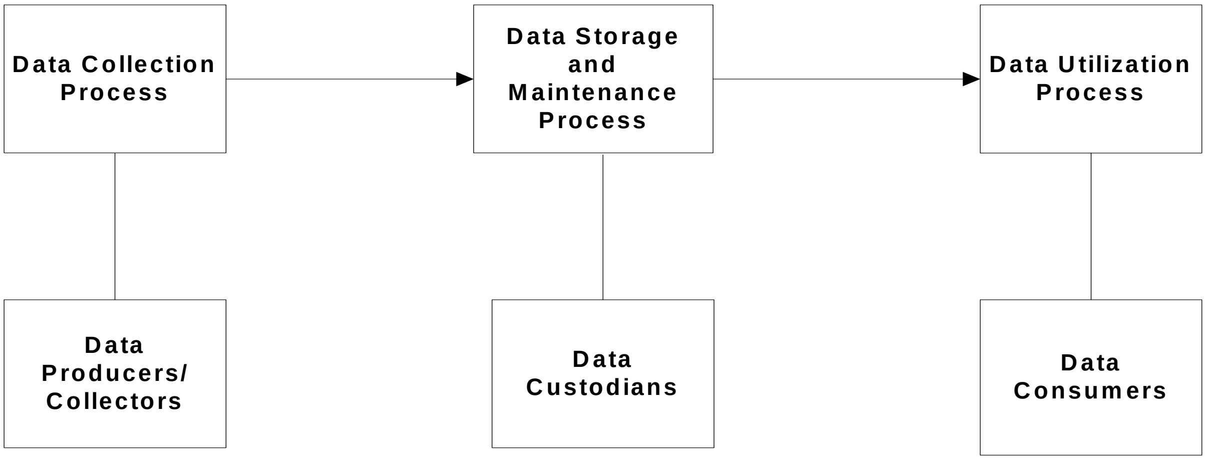


Figure 1 Data Production Process

[2] Following Strong, Lee and Wang (1997) we use data and information synonymously.

These researchers find that while knowledge about data collection is significant, it is the least visible or important of the three domains in terms of managerial attention. In designing any information system, much of the importance is placed on what users need in their utilization process and on designing the computer system. Typically, little attention is given to the data collection process, which is generally the responsibility of clerks or is a by-product of organizational transactions.

Strong, Lee and Wang (1997) identify three roles within an information production system:

1. Information producers generate and provide information
2. Information custodians provide and manage computing resources for storing, maintaining and securing information
3. Information consumers access and utilize information for their tasks. Utilization may involve additional information aggregation and integration.

Each stage of the data production process is scrutinized in the data quality analysis and assessment.

These researchers also categorize information quality into four major aspects with associated quality dimensions as follows:

Category	***Dimension***
Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation
Accessibility IQ	Accessibility, Security
Contextual IQ	Relevancy, Value-Added, Timeliness, Amount of Information
Representational IQ	Interpretability, Ease of Understanding, Concise Representation, Consistent Representation

We focus on intrinsic and contextual data quality dimensions in this research study. Systematic errors in information production can lead to lost information, resulting in errors of correctness, completeness and relevancy. Systematic errors during production are especially important because they affect the entire system. Timeliness and value-added are also affected, as errors occurring early on in the data production process that are not assessed at this stage lead to a delay in value-added information in deriving the credit risk measurement.

Organizational knowledge about the data production process also impacts data quality assessment. Lee and Strong [4] find that organizations with seemingly knowledgeable IS groups and well-established organizational rules; procedures and routines are not exempt from producing poor-quality data and being affected by them. Whereas an IS group is typically very knowledgeable about storage and maintenance of data in its systems, it may know little about how and why data consumers use data. Knowledge about user processes may help IS groups understand the reasons why they store and maintain organizational data and thus contribute to the production of higher-quality data for data consumers.

An interesting finding from their study is that there is no overlap between the performance dimensions associated with custodian’s knowledge and those correlated with consumers’ knowledge. One conjecture from these findings is that a key function of data collectors is to understand the needs of data consumers, the relevancy dimension, and to collect accurate and complete data for storage by custodians. Thus, the data collectors may serve the role of data quality brokers or intermediaries between custodians and consumers. Data collectors seem to hold key data quality knowledge, but in todays IS requirement process, they rarely play a significant role.

Since the focus of our study is on financial data quality, it is important to define what is meant by financial data. According to De Amicis and Batini [1], financial data can be classified into four main categories: (a) registry data used to describe financial instruments; (b) daily data that refers to prices and exchange rates; (c) historical data mainly related to times series, and (d) theoretical data that corresponds to output of financial models. Arguably, the data used in calculating PE involves all four categories, since the exposure is dependent on the type of financial instrument and its market value and because internal credit models use and analyze historical and theoretical data.

These researchers also support Lee and Strong's [3, 4] finding that data collection and input are very significant to data quality. De Amicis and Batini note that knowledge on data loading and updating process has an important impact on data quality dimensions. For example, when a data loading process is not optimized, and then timeliness and uniqueness as data quality dimensions are affected by errors.

Improving organizational data quality also requires subjective assessment of data. Pipino, Lee and Wang [5] demonstrate that in order to use the subjective and objective metrics to improve organizational data quality requires three steps:

1. Performing subjective and objective data quality assessments
2. Comparing the results of the assessments, identifying discrepancies, and determining the root causes of discrepancies
3. Determining and taking necessary actions for improvement

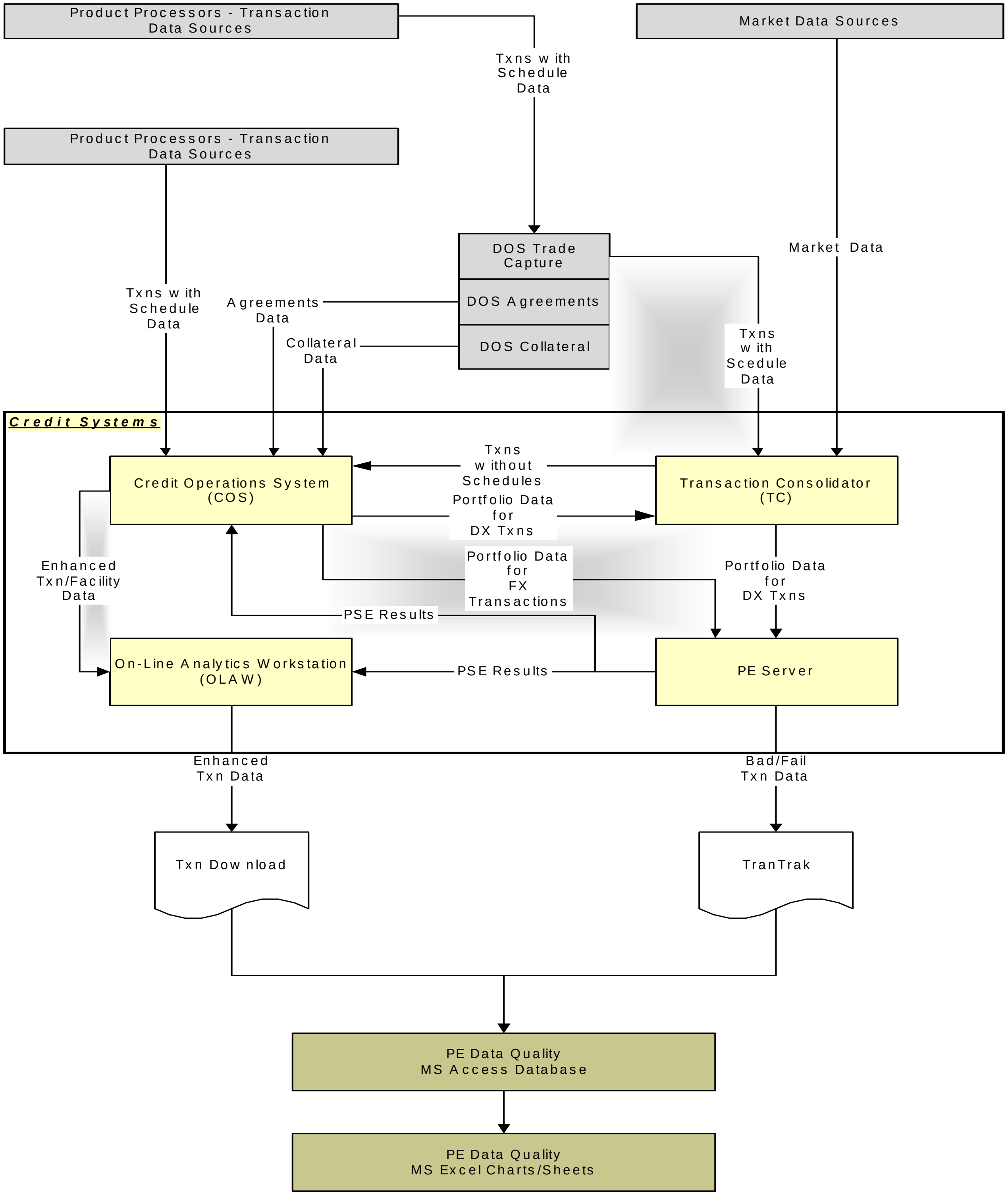
Finally, Klein and Davis [2, 3] note that there is strong evidence that data stored in organizational databases have a significant rate of errors. As computerized databases continue to proliferate and as organizations become increasingly dependent upon these databases to support business processes and decision making, the number of errors in stored data and the organizational impact of these errors are likely to increase. To solve this problem, focus is made either on validating data as they are input to or stored in databases or depending on users to detect and correct errors.

III. THE PE MODELS AND DATA QUALITY PROCESS

The financial institution studied in this research study measures credit risk (or PE) using both a transactions based method and a portfolio based method. The latter is preferred, as it allows for netting agreements among counterparties and contracts and represents a time varying risk assessment over the life of the portfolio, with a peak risk exposure occurring at a particular point in time. For reasons listed below, an unacceptably high number of transactions cannot be leveraged by the Credit systems for PE calculations using the preferred portfolio based Monte Carlo simulation methodology. Consequently, the institution must add the transaction based PE for these incomplete contracts to the portfolio based result to arrive at the overall credit risk exposure measurement:

- ✓ Unsupported products (products for which pricing models and market data simulation do not exist in the PE Server).
- ✓ Source system flawed handoff logic to the Credit systems for transaction and market data.
- ✓ No comprehensive transaction and market data model(s) to describe the data required by credit necessary to calculate exposure accurately.
- ✓ Credit infrastructure design issues.
- ✓ Missing Service Level Agreements (SLAs) for market data
- ✓ Correct data validation and error trapping doesn't occur at the earliest point possible within the Credit systems.

Figure 2 depicts the data process involved in calculating PE under the two methods and the supporting information systems.



Product Processors - Transaction Data Sources
Market Data Sources
Txns w ith Schedule Data
Product Processors - Transaction Data Sources
DOS Trade Capture
DOS Agreements
DOS Collateral
Market Data
Txns w ith Schedule Data
Agreements Data
Collateral Data
Txns w ith Scedule Data
Credit Systems
Credit Operations System (COS)
Txns w ithout Schedules
Transaction Consolidator (TC)
Portfolio Data for DX Txns
Enhanced Txn/Facility Data
Portfolio Data for FX Transactions
Portfolio Data for DX Txns
PSE Results
On-Line Analytics Workstation (OLAW)
PSE Results
PE Server
Enhanced Txn Data
Bad/Fail Txn Data
Txn Dow nload
TranTrak
PE Data Quality MS Access Database
PE Data Quality MS Excel Charts/Sheets

Figure 2 Data Process and Information Systems for PE Calculations

Transaction Input

Transactions are booked by manual entry into the Product Processors (PP). When transaction data is sourced from a PP system, the receiving Credit system will be the Credit Operations System (COS).

- Transaction data is input manually into PP system. It is not known to the PE data quality analysts what specific data validation exists at this stage. Furthermore data validation that does occur at this stage typically satisfies the requirement of the PP system but might not satisfy the requirements of the Credit systems. Even if data validation does occur, it will only satisfy the requirement of the PP system and not the Credit systems. The PP system uses this data to price transactions, but errors occur when the data is mapped into the Credit system requirements because the required documentation, which the PP system technology group uses, is not available or incomplete. The product processes have a lack of transaction data models and different credit processes generate incorrect data rather than actual value input.
- The PP system will create a process that will extract data out of the PP system into file(s) to be sent to the COS. The files are created in accordance with the file format specification for the COS. The process, which maps data out of the PP system into file(s), is where errors often occur due to the lack of transaction data models.
- Files arrive at the COS by agreed Cut-off time.
- As part of a loading process the COS loads the data into its database. Some validation takes place as part of the loading process. Most of the validation is general validation, i.e. they are not product specific. Examples of validation checks are ensuring that each field conforms to the correct data type; checking that mandatory fields are provided; and validating currencies against a list of valid currencies.

The Derivatives Operations System (DOS) is the global back office system for derivative transactions. As a result, all derivative transactions have to exist on the DOS either through shadow booking (i.e. re-entering the data manually into the DOS) or through building a technical interface between the PP system and the DOS. The DOS is the sole source of transaction data for derivative transactions. When transaction data is sourced from the DOS, the receiving Credit system is the Transaction Consolidator (TC).

- Transaction data is either input manually into the DOS (shadow booking) or it's loaded into the DOS through a technology interface. There are also issues with the data that exists on the DOS. The DOS does NOT price transactions (the price is received from the PP systems) and hence it does not care about the quality of the data that it does not use. It only cares about the data that it uses for back office functionality.
- Through a process called the Credit Feed, the data is extracted out of the DOS and sent to the TC as Tibco messages. The transaction prices received by the DOS from the PP systems are also fed into the TC as Tibco messages.
- As part of a loading process using Tibco, the data is loaded into the TC's database. It is assumed that the COS will perform the validation checks.
- The TC extracts and sends the transaction data to the COS.

As part of a loading process the COS loads the data into its database. Most of the validation is general validation. Examples of validation checks are ensuring that each field conforms to the correct data type; check that mandatory fields are provided; validating currencies against a list of valid currencies etc.

Transaction Data Processing

The following is a list of the relevant functions performed by the COS:

- Match transactions to facilities/portfolios.
- Match transactions to Netting and Margin agreements.
- Calculate notional based PE at transaction level.
- Calculate Settlement risk at transaction level.
- Calculate Lending Risk at transaction level.
- Aggregate all risk types under facilities.
- Calculate availability at facility level.
- Extract transactions data for PE Server in accordance with PE Server file format requirements. Transactions are grouped under facilities/portfolios to which they match since the PE Server calculates PE at facility/portfolio level (not transaction level).
- The COS sends FX PE extract files directly to the PE Server. Derivative extract files are sent through the TC. The TC does not send the schedules' data[3], which are required for PE Server processing only, to the COS. Therefore, there is a requirement for the TC to add the schedules' data to the PE extract files once received from the COS. Once the schedules' data are added, the TC sends DX PE extract files to the PE Server.
- The PE Server sends the results of its processing to the COS so that the data can be made available to users.
- The COS generates a large number of reports.

PE Server Processing

- The PE Server processes the transaction files received (FX from the COS and DX from the TC) along with the market data to calculate PE using Monte Carlo methodology.
- The PE Server produces result files, which are then sent to the COS so that they can be made available to users.
- The PE server validates each transaction before it decides whether to include it in the simulation process or not. The validation check used is to price every transaction and compare it against the price calculated by the source system (i.e. product processor). The result of the comparison is then tested against the tolerance specified for each product to ensure the accuracy of the price calculated by the PE Server.
- Transactions with "Bad" and "Fail" result status are logged and reported in a report called "TranTrak" (TranTrak = Transaction Tracking).
- The PE Data Quality team receives the TranTrak report on daily basis. The data in this report is used to help identify DQ issues and are reported on through PE EDQ graphs/reports.

[3] Schedules data take place on a pre-determined date and are only effective for a period of time, over which the scheduled event affects the price of the transaction. This is known as the 'lifetime' of the Scheduled Event. The lifetime of each Scheduled Event is demarcated by 'key-date' and 'next key-date', where next key-date is the key-date for the next Scheduled Event or the Maturity Date of the transaction.

For example:

1) Rate Reset Schedules describe the floating rate resets for the floating leg of an interest rate swap.

2) Settlement Schedules describe the fixed and floating cash flows for the fixed and floating leg of an interest rate swap.

Result Status	**Notes**
Bad	Indicates that the PE Server was not able to price the transaction, which is normally as a result of missing/incomplete transaction and/or market data. The reason for such issues can normally be determined by the PE server and are logged in a report (i.e. TranTrak report), which is used by the PE DQ team to identify issues. This results status indicates a problem and transactions with 'Bad' statuses are not included in the simulation process.
Fail	Indicates that the transaction failed the tolerance test when comparing the price calculated by the PE Server against the source system price. The PE server is not able to determine the reason for failed transactions. Therefore, although 'failed' transactions are logged in a report (i.e. TranTrak report) the reason for their failure will not be known. Failed transactions require detailed and lengthy analysis in order to identify the root cause for failure. This results status indicates a problem and transactions with 'Failed' statuses are not included in the simulation process.
Pass	Indicates that the results of both the pricing and the tolerance test were successful. This result status indicates that we have no problem and transactions with 'Pass' statuses are included in the simulation process.
Pass/ Proxied	Indicates that some guessing work was involved in generating some of the input data. For certain products the PE Server tries to compensate for missing data by generating the data itself. If as a result the transaction can be priced and it passes the tolerance test, the "Pass" status is qualified as a Proxy "Pass". This results status indicates a problem but since the transaction is included in the simulation run, this result status is not treated as a high priority issue.

On-Line Analytics Workstation (OLAW) Processing

The main functionality of the OLAW is to display the PE Server results to the users. The PE Server results are enhanced by supporting data received from the COS. The PE Data Quality team downloads transactions data from the OLAW to produce PE DQ graphs and reports.

Market Data

Market data is used by the PE Server for pricing transactions and PE calculations using Monte Carlo simulation.

Market Data Sourcing

The Credit systems source market data from a large number of systems and the TC receives this market data. Normally, and ideally, the Credit systems would use the market data that the PP system uses to price transactions and hence the majority of market data is sourced from PP systems (i.e. marking system). The following describes the process:

- Market data source systems prepare files in accordance with the TC's file specification.
- Files arrive at the TC by agreed Cut-off time.

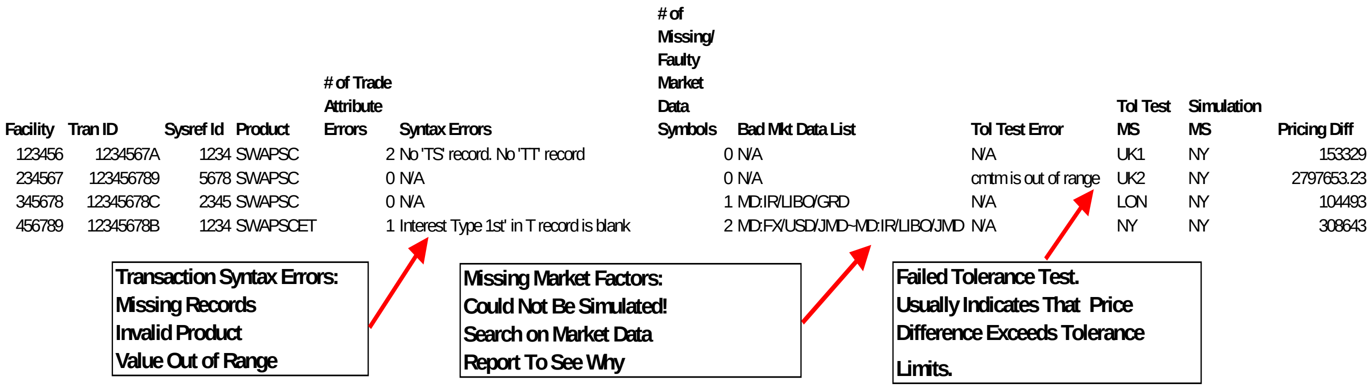
- Through the loading process, files are entered into the TC's database. It is not known whether validation checks exist at this stage, although such validation checks would be general in nature.
- The TC performs some added functionality on the market data received, for example reformatting the data, creating supersets etc.
- The TC sends the market data to the PE Server by agreed cut-off time.

Market Data Processing

- As part of its processing, the PE Server performs validation checks on the market data by pricing every transaction and comparing against the price calculated by the source system. If the PE Server cannot price a transaction, then the result status will be set to "Bad" for the transaction. If the PE Server can price the transaction but the price is not within the tolerance test specified for the product (when comparing to the source price) the result status is set to 'Fail' for the transaction in a similar manner to the PE server processing.

IV. Data Quality Assessment: Results and Analysis

Figure 3 below provides an example of the errors that are found by the data quality teams at the PE Server processing stage and the market data processing stage.



Facility	Tran ID	Sysref Id	Product	# of Trade Attribute Errors	Syntax Errors	# of Missing/ Faulty Market Data Symbols	Bad Mkt Data List	Tol Test Error	Tol Test MS	Simulation MS	Pricing Diff
123456	1234567A	1234	SWAPSC	2	No 'TS' record. No 'TT' record	0	N/A	N/A	UK1	NY	153329
234567	123456789	5678	SWAPSC	0	N/A	0	N/A	cmtm is out of range	UK2	NY	2797653.23
345678	12345678C	2345	SWAPSC	0	N/A	1	MD:IR/LIBO/GRD	N/A	LON	NY	104493
456789	12345678B	1234	SWAPSCET	1	Interest Type 1st' in T record is blank	2	MD:FX/USD/JMD~MD:IR/LIBO/JMD	N/A	NY	NY	308643

Transaction Syntax Errors:
Missing Records
Invalid Product
Value Out of Range

Missing Market Factors:
Could Not Be Simulated!
Search on Market Data
Report To See Why

Failed Tolerance Test.
Usually Indicates That Price
Difference Exceeds Tolerance
Limits.

Figure 3 PE Transactions Data Errors

As depicted in Figure 3, the errors uncovered by the data quality teams involve completeness (missing data), relevancy (data is inapplicable as presented) and accuracy (out of tolerance range). Mapping these errors to Figure 2, it appears that the source of the errors occur at the transaction input stage in product processors. This is not timely, however, as the errors are not uncovered until processed in the PE Server.

The data quality team assesses quality and produces error reports while processing transaction and market price data in the PE Server. De Amicis and Batini [1] note that selection and inspection of data quality dimensions is related to process analysis, with the final goal of discovering the main causes of erroneous data, such as unstructured and uncontrolled data loading and data loading and data updating processes. The final result of data quality analysis is the identification of errors. The data quality team extends this process analysis to the stages that occur after the errors have been uncovered. This is step 1 in Figure 5, which provides the steps taken by the data quality team after it has uncovered and assessed these errors.

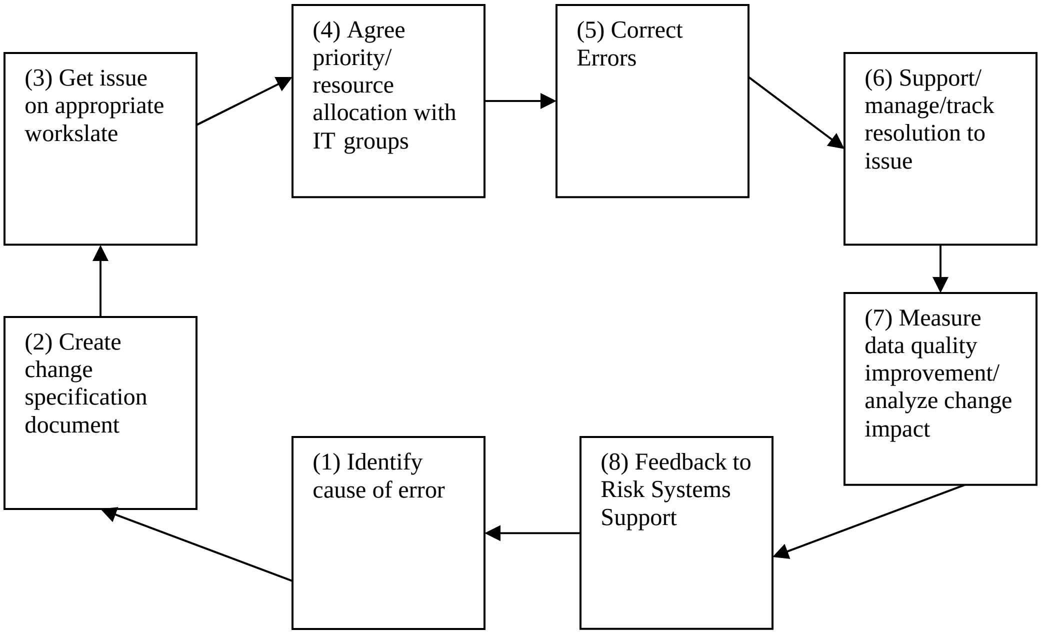
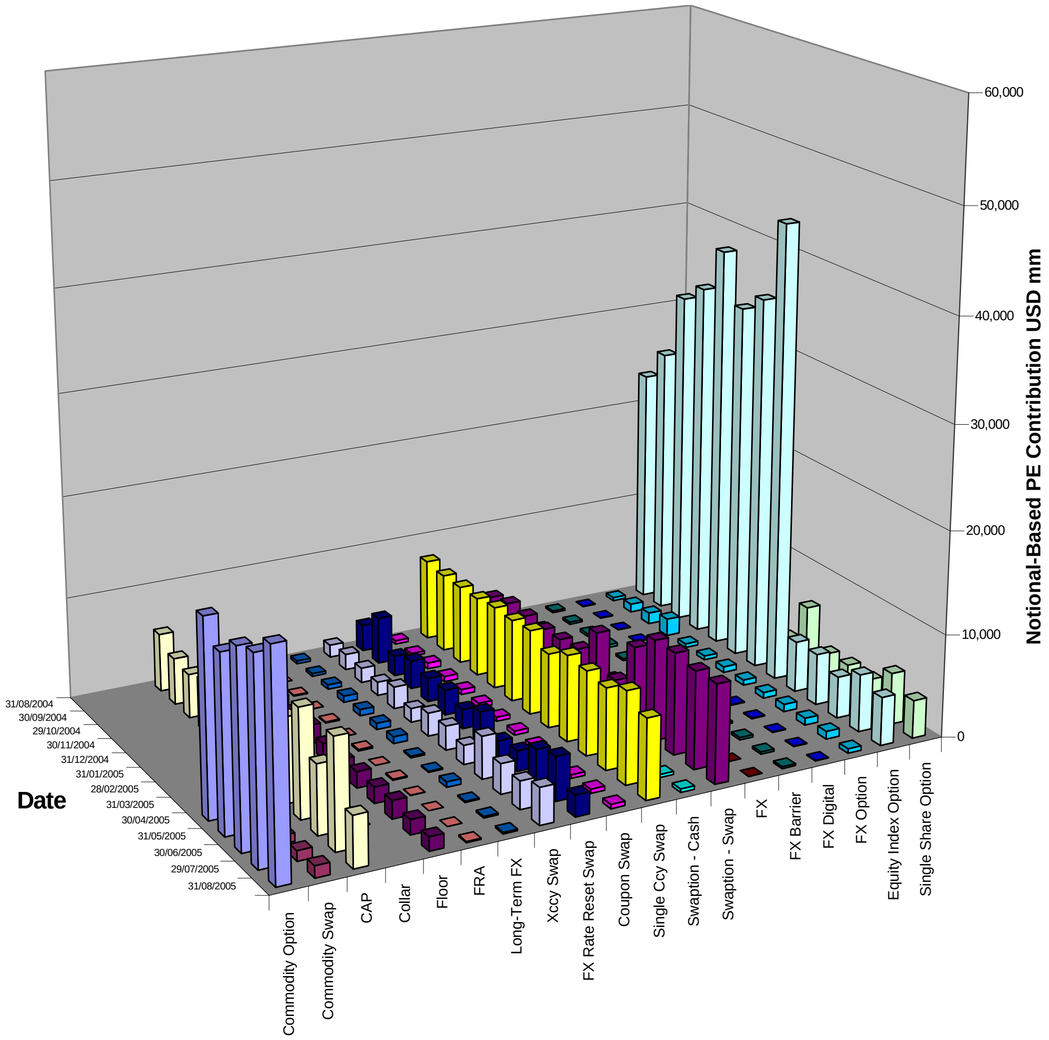

(3) Get issue on appropriate workslate
(4) Agree priority/ resource allocation with IT groups
(5) Correct Errors
(6) Support/ manage/track resolution to issue
(7) Measure data quality improvement/ analyze change impact
(8) Feedback to Risk Systems Support
(1) Identify cause of error
(2) Create change specification document

Figure 4 Data Quality Assessment Process

As depicted in Figure 4, once the data quality assessment team has identified the error, they create a change specification document that gets forwarded to the IT department. The IT department serves as custodians of data. Priorities are set by the User groups but, typically given the number of requests in the change queue and competing priorities of the requests, backlogs occur. Also the PPs providing data do not always view the change requests with the same priority as the Credit users. This lower prioritization can be further compounded by a lack of knowledge of the data utilization. As noted by Lee and Strong [4], whereas an IS group is typically very knowledgeable about storage and maintenance of data in its systems, it may know little about how and why data consumers use data. Knowledge about user processes may help IS groups understand the reasons why they store and maintain organizational data and thus contribute to the production of higher-quality data for data consumers.

V. DATA QUALITY PROCESS IMPROVEMENT

Once the IT group provides the model support for the failed transactions per the change specification documentation, the transaction data collection is improved as indicated by fewer transactions classified as bad/fail. Figure 5 shows the improvement in The Equity Index Option product pricing model that was corrected in March. The drop in the Equity Index Option bar shows an obvious improvement in the USD PE contribution of failed transactions.



PE Data Quality
Non-Pass Notional-Based PE by Product (USD mm, Supported Products)

Figure 5 PE Data Quality

VI. CONCLUSIONS AND FURTHER RESEARCH

In this paper we have used a process approach to monitoring financial data quality assessment and improvement. Notably, the data quality team is to be commended for their assessment beyond generating error reports to generation of the ultimate resolution. We have described in detail the error map designed to reveal the current process. Our interest is to improve this process we described with an eye towards proactive assessment early on in the data production process. Such a proactive approach requires that data validation occur at the collection stage, noting where missing data on pricing models is first encountered (within the product processors). When a transaction is processed, the firm has the opportunity to validate the data within its rich contextual information, permitting analysis of the reason for the missing information. Presently, transaction records that are missing too much information can't be used and are not processed. A proactive approach would result in error reports that include these unprocessed transactions generated for the data collectors, so that they are aware of the implications of the errors and for the custodians, or IT department, so that it prioritizes the automation and support of the missing pricing model.

Further, subjective assessment should occur in parallel with the objective data quality assessment. It would be beneficial for the data quality team to perform subjective assessment of the missing data so that issues can be resolved prior to detection of the errors further along the production process, i.e., the PE server phase. One suggestion is to have staff in the PE DQ group review and approve transaction documentation that is utilized in both the product processors and the credit systems. Identifying and resolving issues before they result in failed transactions will ultimately lead to complete utilization of the portfolio measure of PE exposure.

Further, our research indicates that data collectors and data custodians and data users need to share knowledge to improve data quality in the three dimensions. Producing financial data of high accuracy requires knowledge about the collection process, the storage process and the utilization process. While the data collection process is typically the least visible to managers, it is clearly very important to the accuracy of the data quality produced. It is possible, for example that the data quality team can act as a type of information broker between the collection (product processors) and the custodians (IT department). The ability to have a robust system in place to validate the accuracy and consistency of rating systems, processes and all relevant risk components depends upon workers who are knowledgeable about the entire process. We plan to continue this investigation into the role of knowledge sharing and data quality in the next phase of our research, focusing on differing perspectives and goals of the data collectors, custodians and users that contribute to the problem of attaining high-quality data.

VII. REFERENCES

[1] De Amicis, F., C. Batini. “A Methodology for Data Quality Assessment on Financial Data.” *Studies in Communication Sciences,* 4. (2). 2004. pp115-136.

[2] Klein, B., D. Goodhue and G. Davis. “Conditions for the Detection of Data Errors in Organizational Settings: Preliminary Results from a Field Study”. *Proceedings of 1996 International Conference of Information Quality. Cambridge, MA: MIT* 1996.

[3] Klein, B.D., D.L. Goodhue, and G.B. Davis, "Can Humans Detect Errors in Data? Impact of Base Rates, Incentives, and Goals," MIS Quarterly (June 1997) pp. 169-194.

[4] Lee, Y. and D. Strong. “Process Knowledge and Data Quality Outcomes”. *Proceedings of the Eighth International Conference on Information Quality. Cambridge, MA: MIT.* 2003.

[5] Lee, Y. and D. Strong. “Knowing-Why About Data Processes and Data Quality”. *Journal of Management Information Systems,* 20 (3). Winter 2003-4. pp 13-39.

[6] Pipino,L., Y. Lee and R. Wang. (April 2002). “Data Quality Assessment”. *Communications of the ACM.* 45 (4). 2002. pp211-218.

[7] Strong, D., Y. Lee and R. Wang. (1997). “10 Potholes in the Road to Information Quality”. Computer, 30 (8). 1997.

[8] Wang, R. and D. Strong. “Beyond Accuracy: What Data Quality Means to Data Consumers”. *Journal of Management Information Systems.* 12 (4). 1996. pp 5-34.