International Journal of Research in Social Sciences
Vol. 9 Issue 4, April 2019,
ISSN: 2249-2496 Impact Factor: 7.081
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

A BIBLIOGRAPHICAL STUDY ON IMPORTANCE OF DATA PROFILING AND DATA MINING FOR EFFECTIVE BUSINESS TRANSACTIONS; A TECHNO-BUSINESS LEADERSHIP PERSPECTIVE

Prof Dr.C.Karthikeyan[*]

Asst Prof Krishna[**]

Ms Anna Benjamin[***]

Abstract: Data profiling is performed several times and with varying intensity throughout the data warehouse developing process. (Kimball et al). The light profiling assessment undertaken immediately after candidate source systems is identified and DW/BI business requirements are assessed till it is satisfactory. The purpose of this initial analysis is to clarify at an early stage, on the correctness of the data, and the detail are appropriate to the level required and that anomalies are comfortable enough to be handled subsequently, if it is otherwise, the project may be terminated. In addition, more in-depth profiling is done prior to the dimensional modeling process in order to assess what is required to convert data into a dimensional model. Detailed profiling extends into the ETL system design process in order to determine the appropriate data to extract and which filters to apply to the data set. Additionally, data may be conducted in the data warehouse development process after data has been loaded into staging, the data marts, etc. Conducting data at these stages helps ensure that data cleaning and transformations have been done correctly and in compliance of requirements.

Keywords: Data ; Leadership; Data; Profiling; Meta Data; Quality Control; Validity; Master Data; Computation ; Complexity

[*] Director and Professor, Management Studies, T.John College, Bangalore, Affiliated to Bangalore University, Bangalore, Karnataka, and Accredited by NAAC "A", and Approved by AICTE, New Delhi

[**] Asst Professor, Management Studies,T.John College, Bangalore, Affiliated to Bangalore University,Bangalore, Karnataka, and Accredited by NAAC "A", and Approved by AICTE, New Delhi

[***] Asst Prof, T.John Institute of Management Science, Bangalore, Affiliated to Bangalore University,Approved by AICTE, New Delhi.

Objectives of the Study:

(i) To evaluate the procedures of quality data profiling for effective business process.

(ii) To evaluate the modern data profiling techniques and its advancements.

(iii) To Examine the role of Professional Organisations in Data Quality Management and Governance.

(iv) To examine the advantages of Master Data for Business Process.

Scope: To get Conceptual Clarity on the growth and importance of Quality of Data, and its importance for Data Mining and Business Application

Review of Related Literature:

Girish Punj and David Stewart (1983) reviewed the applications of cluster analysis in marketing problems and recommended that, a two stage cluster analysis methodology and preliminary identification of clusters via Ward's minimum variance method found issues and problems related to the use and validation of cluster analysis, was useful to make decisions, which enabled the leaders to zero in on the decisions taken.

Agrawal et al. (1993) observed that in the recent past the exploratory analysis in particular of large sets of market basket data has become topic of pertinent research due to various publications on data mining and knowledge in databases and generated association rules from market basket data, which described relevant 19 interrelations like, and helped the sellers to identify that "If a consumer purchases fruit juice, then, in 40% of the cases they also purchase mineral water".

Piatetsky-Shapiro et al. (1996) found that a growing number of industrial applications and examined the existing data mining tools, described some representative applications like marketing, investment, manufacturing, fraud detection etc. and discussed the issues for deploying successful application and their adoption by business users. They also highlighted upon the fact of widespread realization of the potential value of data mining and a growing number of researchers and developers in this area.

Anand et al. (1996) focused on the organizations' need to investment in data mining solutions because of the phenomenal expansion of the data space and the resulting sharp increase in the size of typical data base. There is no alternative to heavy reliance on computer programs set to discover patterns for themselves with relatively little human intervention. The authors also proposed a general framework of data mining based on “Evidence Theory” that consisted of methods for representing data and knowledge, and methods for data manipulation and knowledge discovery.

Richard A. Spinello (1997) found Wal-Mart captures point-of-sale data from over 2,900 stores in 6 countries and transmits this data to its massive 7.5-terabyte data warehouse and uses it to identify customer-buying patterns, to manage local store inventory and identify new merchandizing opportunities using data mining techniques.

Peter Spiller and Gerald Lohse (1997) found a classification of on-line retail stores based upon convenience sample of 137 Internet retail stores. Cluster and factor analysis identified five distinct Web catalog interface categories which provide a better understanding of the strategies pursued in Internet-based marketing. Data mining and Knowledge discovery in databases (KDD) are concerned with extracting models and patterns of interest from large databases.

Fayyad and Stolorz (1997) in their paper, provide an overview of this growing multi-disciplinary research area, outline the 20 basic techniques, and provide a brief coverage of how they are used in some applications in science data analysis. Anand et al. (1998) used data mining to search for target customer.

Collier et al. (1998) specified that during the 1990s many companies reached to the conclusion that their data is a valuable asset. These companies moved quickly to build Data Ware houses and Data Marts. Furthermore, Companies such as Wal-Mart recognized the benefit of applying Data Mining to these rich stores of historical data. They further note that the data mining industry has, and will, provide real advantages to those who employ it.

Kleissner (1998) looked that there has been a growing gap between powerful data warehousing systems and the user’s ability to effectively analyze and act on the information they contain. Data mining tools and service provide the leap necessary to close this gap. Data mining offers automated discovery of previously unknown patterns as well as automated prediction of trends and behaviors; its technologies are complimentary to existing decision support tools and provide the business analyst and marketing professionals with a new way of analyzing the business. And more specifically in the retail industry, data mining tools can be used for market basket analysis (determining which products a customer typically purchases at the same time) and targeted marketing campaigns.

Weiss and Indurkhya (1998) stated that a business services company with national reach asked Data Miners to help them improve their resource allocation by understanding the characteristics of the places where they were most successful. After interviewing and brainstorming with stakeholders from several groups within the company they were able to develop logistic regression model of data mining to predict the likelihood that a given county would be classified as good based on a number of explanatory variables which resulted in more informed investment decisions.

Michel and John (1998) focused on fourteen data mining tools and then a standardize procedure and twenty evaluation criteria for assessing the tool qualities were developed 21 and applied. The traits were collected in five categories; capability, learn ability, Stability, Interoperability, Flexibility and Accuracy. The author then summarized the evaluation procedure and the scoring **of all component criteria.**

Holstein (1998) explored that at Wal-Mart, every scrap of information about who buys what is fed into a huge data warehouse. The information is sent to data warehouse, where company slice and dice it to glean who's buying what, and when and where they are buying it. This information is sent to manufacturers through the Internet, telling them what is selling and where it's selling best where they can quickly re supply or agree to cut prices if the product isn't moving by incorporating data mining technologies.

Brand and Gerritsen (1998) found large retail chains has turned itself into an information broker after embracing data mining and that leads supermarkets to purchases demographic data directly from its customers by offering them discounts in return for using a safe way savings club card. In order to obtain the card, shoppers voluntarily divulge personal information that is later used in predictive modeling.

Koga (1998) described that data mining techniques can discover latent rules and patterns. Koga suggests that techniques in data mining, such as association, classification, and clustering are useful for discovering the combination of products which have high probabilities of purchase at the same time, finding changing patterns of demand, identifying non-loyal customers, and targeting customers to increase response rate of Data mining.

Menon and Sharda (1999) a data mining technique like the decision tree helps to classify data into a finite number of classes by generating a hierarchy of 'IF-THEN' statements. Based on a series of 'IF-THEN' statements, they could predict the shopping and patronage behaviors of retail customers.

Jain et al. (1999) examined that clustering is the unsupervised classification of patterns into groups and reflected its broad appeal and usefulness as one of the steps in exploratory data analysis. Researchers present taxonomy of clustering techniques, and identify cross-cutting themes and recent advances.

Mitchell (1999) found that the data mining applications include predicting customer purchase behavior, customer retention, and quality of goods produced by a particular manufacturing line, if applied successfully promises more effectiveness.

Maimon and Last (2000) found that real value is added to data by multidimensional manipulation spreadsheets and query-based data mining tools are able to compete with full-fledged management information systems. These two features turn raw data into usable information.

Witten and Frank (2000) observed that the data mining is the task of identifying useful regularities and patterns in large bodies of data. The data mining task involves identifying patterns of consumer behavior from purchasing logs. In a typical case, the mining of supermarkets sales logs might lead to the discovery that consumers are more likely to buy a packet of washing powder.

According to Hackney (2000) data mining solutions are key weapons in an organization's business intelligence arsenal that reveal trends, relationships and predict future outcomes. Business intelligence consists of all the activities related to organizing and delivering Information and analysis to businesses and its decision makers.

Morrison et al. (2000) also mined customers' credit card data using the RFM (Recency, Frequency, Monetary) model to predict valuable hotel customers.

Russell and Petersen (2000) found that the market basket analysis focuses on the decision process by which a consumer selects items from a given set of product categories on the same shopping trip. It aims at the identification of interrelations between choices of different products purchased in a **specific retail store such as a supermarket.**

Berry and Linoff (2000) found that the information gained from data mining efforts can benefit an organization in increasing organizational profits by lowering costs and by increasing revenues. Data mining attempts to predict future business trends and customer behavior patterns from large data warehouses and, other 23 forms of data resources. Data mining is clearly of value to any organization where there are large amounts of data, and something worth learning from that data.

Ma et al. (2000) data mining is critical to the enterprise that wants to exploit operational and other available data to improve the quality of decision making and gain critical competitive advantages. This is where data mining can play a crucial role, by disclosing important **information in a cost effective and timely fashion.**

Yada et al. (2000) proposed a new approach to discover the loyal customers in the future from newcomers as early as possible, using data mining tool, "C5.0" on real purchase data.

Dan Hopping (2000) emphasizes that the evolution of retailing reveals that technology has played a role as the primary enabler of change like knowledge management, data mining, customer relationship management, mathematical modeling, and data visualization that affect the future of retail. The decisions made by human affect the retailer's bottom line and the task of technology is to provide the information in a way that can help the decision makers to get the best way to get **ready for the future.**

Bharat Rao (2000) found that the Information technology solutions can reduce costs, increase flexibility and response and provide a more effective shopping experience for the customers. By utilizing and sharing technology, retailers can develop collaborative planning, forecasting, and replenishment programs.

Mattison (2000) found that the organizations to develop profiles using data mining systems, which is combined through data collected from various online and off-line customer touch points, and extract useful patterns or models from the data, and monitor behavior with the objective to identify and describe the main features of the distribution.

Berson et al. (2000) found that the application of data mining tools in CRM is an emerging trend with advancements in analyzing and understanding customer behaviors to evolve as competitive CRM strategy, to acquire and retain potential customers and maximize customer value and are one of the best supporting tools for making different CRM decisions.

Ling and Yen (2001) observed in the context of CRM, data mining can be seen as a business driven process aimed at the discovery and consistent use of profitable knowledge from organizational data, and a tool to guide decision making. Data mining increases the response rates of the marketing campaign by segmenting customers into groups with different characteristics and needs, and predicts how likely an existing customer is to take his/her business to a competitor.

Bose and Mahapatra (2001) found that data mining made it possible for organizations to accumulate massive amounts of data, and further, data mining offers an effective means of analyzing that data and converting it into valuable information or knowledge and can be applied to the field of dramatically growing business and the field of marketing including retailing sales analysis and so on.

Chopoorian (2001) found that data is the organizational challenge of the new millennium. They also state that competitive success will depend on the ability of companies to quickly and effectively convert their raw data into comprehensible information. It is impossible for managers to make the correct decisions if the needed information cannot be accessed or presented to them intelligibly.

Anon (2001) reveals the ability to access and retrieve data contributes to business success by increasing the knowledge of decision makers at all levels. The more efficiently managers can access value adding data, the better their chances of gaining insights into what drives their organizational activities which enables them to devise improved business strategy.

Melab (2001) found that with the recent development in computer science and particularly the coming of new technologies are creating data explosion. Drowning in data, and exploitation of the emerging data mining technology is promising and young field with a wide range of applications. It consists of discovering non-obvious knowledge, 'let us say gold', allowing to make better business decisions (jewels).

Shaw et al. (2001) describes the way to systematically apply data mining in marketing knowledge management framework to allow marketing personnel to know their customers better. **Song et al. (2001)** depicted a method to detect changes of customer behavior at different time from customer profiles and sales data.

Buck (2001) found that the retail analysis and financial institutions have been among the first to embrace data mining technologies, first with the analysis of data in the large corporate data warehouses, and more recently in the analysis of online web-based activities.

Charles et al. (2001) found that the data mining techniques could enable the leaders to have confidence in their decision as they had the data support that estimated potential increase in sales arising from changes to these critical attributes using predictive modeling and finally their analysis and modeling process was assisted by the description and visualization of shopper behavior.

Bounsaythip and Rinta-Runsala (2001) provided a review on some data mining methods that can be used for customer segmentation and profiling like K-nearest neighbor, neural networks, association rules and sequential pattern discovery and finally presented three well known examples of customer profiling from data mining literature.

Katraras et al. (2001) used cluster analysis data mining technique to segment grocery shoppers into six segments according to their preferences for 33 store and shopping experience characteristics. The segmentation was done on the basis of attitude and surprisingly as a result there were very few differences in the segments in terms of demographic factors.

Chris and Bhavani (2001) state that the data mining in a new and rapidly developing technology and given the wide variety of tasks data mining can perform, it is difficult to come up with a data mining standard. So this paper presents an overview of data mining and also includes the data mining model, data mining process and the data mining architecture standards that affect several stages of data mining project.

Michael et al. (2001) focus on the fact that much of the useful marketing insights into customer characteristics and their purchasing patterns are largely hidden and untapped. Thus a systematic methodology that uses data mining and knowledge management techniques is proposed to manage the marketing knowledge and is the basis for enhancing customer relationship management. The authors in this paper elaborate that how data mining can be integrated into a marketing knowledge management framework.

Chen and Lewis (2002) found that businesses today operate in an increasingly information or data driven economy consisting of more complex structures, data mining offers strategies for precision, more accurately addressing and dealing with problems, better consideration of bottom line issues, and more effective decision making.

Apte et al. (2002) explored that in recent years a number of trends have emerged to challenge the traditional approach of analyzing using statistical modeling and many industries have already adapted data mining techniques on customer data warehouses and thereby increase profits like insurance, Banking, mail order, communication, retail and medical industries.

Lavinson (2002) specifies that retailers have collected vast amounts of data for years, but they have not had the means to apply it effectively to their planning and buying because, until a few years ago, no computer or software application could process all of the data. 28 The data mining techniques available today offer retailers better results because they incorporate more than just historical sales data.

Kopanas et al. (2002) found that data mining techniques have been applied in many application areas like the automatic discovery of new knowledge from a vast amount of data. The author has described data mining as a continuous interaction between the implicit domain knowledge and the knowledge that is discovered through the use of data mining algorithm.

Nemati and Barko (2002) stated that information is quickly becoming one of the major differentiators between industry leading organizations and second rate organizations. Being able to extract the relevant knowledge from this information plays a vital role in enhancing enterprise decision making. Data mining is a key part of this methodology, and its successful implementation can lead to enhanced organizational decision making.

Koch (2002) found that computer systems have had a positive impact on the way that business is conducted today. In the past 30 years though, technology has been used mainly to automate manual tasks and to help organizations do business faster. Organizations must leverage technology in order to create value.

Coskun et al. (2002) data mining has a generic blanket definition that tends to include all the tools employed to help users analyze and understand their data. They also highlight that data mining differs from traditional statistical techniques by using the computer, rather than the analyst, to find patterns and relationships by identifying the underlying rules and features in the data.

Chris et al. (2002) give the overview of the concept of data mining and CRM. The authors offer a closer look at mainly two data mining techniques viz. Chi-Square Automatic Interaction Detection (CHAID) and Neural Networks. As a result of comparison of two techniques, the authors concluded that CHAID is much easier and quicker to construct and understand whereas neural networks provide more accurate models, especially for complex problems.

Joan Anderson (2002) emphasis on the struggle of the retailers in finding out various questions about their marketing campaign like who can I consider a loyal customer, what kind of marketing strategy is most likely to increase sales etc. So the purpose of his study is to critique data mining technology in comparison with the more familiar analytical tools for strategic decision making by small to medium size retailers. The context for this study includes current and future industry application for research performed in data mining applications within the retail sector.

Joyce Jackson (2002) elaborated the way to plan, evaluate and successfully refine a data mining project, particularly in terms of model building and evaluation. The author concludes with a major illustration of the data mining process methodology and the unsolved problems that offer opportunities for research.

Mild and Reutterer (2003) described that most of the earlier work on market basket analysis deals with either the traditional association coefficient based approaches, where by the interdependencies between the products are measured by means of cluster analysis 30 or multidimensional scaling, by means of correlation/ regression analysis or by means of econometric models such as multivariate logic.

Elovici and Braha (2003) found that independent data mining system can be combined that showed that the combined approach to data mining system can be used in the decision making process of the organization to increase payoff .

Labovitz (2003) found that data mining aims to add value to business organizations by applying highly rigorous statistical analysis to large masses of data in an effort to extract meaningful trends and relationships that are often hidden by the sheer volume and arrangement of the data. It further supports the transformation of data into information, knowledge and wisdom. This is a vital point to organizations who wish to make use of their data to establish a competitive advantage in their industries.

Mont and Plepys (2003) examined a great variety of methods for understanding and evaluating the consumer’s acceptance and satisfaction in different disciplines. They surveyed a range of tools for measuring and evaluating customer satisfaction using interviews, observations, and psychographic portrait of customers. They came up with some important benefits and drawbacks but both the research models and tools were found to be useful for product service systems applications.

Rosset and Neumann (2003) discussed in their paper, both theoretically and empirically, the optimal use of “customer value” in all phases of data analysis like model training, model evaluation and scoring stages. They have taken a concrete example, considering the problem of churn analysis in telephony where the task is to predict whether or not a customer will disconnect and switch to a competitor.

Yang and Padmanabhan (2003) discussed “pattern-based” clustering approaches to group customer transactions, and presented a new technique, YACA, that generates a highly effective clustering of transactions.

Sim Jaesung (2003) found that the concept of data mining is gaining acceptance in business as a means of seeking higher profits and lower costs. The author identifies critical success factors in data mining projects as implementing emerging information systems can be risky if the CSFs have not been researched and documented properly.

Carrier and Povel (2003) found that the data mining consists of the building of the model from data. Each data mining technique can perform one or more of the following types of data modeling like Association, Classification, Clustering, Forecasting, Regression, Sequence Discovery,Visualization. The authors focus on the point that the choice of data mining techniques should be based on the data characteristics and business requirements.

Youngsan and Yongmoo (2003) collected and analyzed the customer related data in order to resolve the problem of a stoppage to the customer visits to a Korean automobile repair service centre for unknown reasons. The authors first defined customer class based on RFM variables and then built a customer classification model using decision tree. They finally extracted the characteristics of each customer class for each kind of automobile and finally used the result of the analysis together in order to provide better services so that the total revenue of the centre can be increased.

Ryals Lynette (2003) discusses some of the common data analysis techniques used to identify the most and least profitable customers so that marketing managers can develop effective, appropriately targeted customer management strategies.

Kenneth et al. (2003) evaluated the issues surrounding the convergence of Data Mining and market research for deeper customer understanding. The authors begin with a review of the two disciplines i.e. SPSS Inc and The Kartner Group entitled "Two Rivers" and discuss where they fit within an overall Customer Intelligence environment.

Lijia Guo (2003) introduced data mining approach to modeling insurance risk and demonstrates two potential applications to property/ casualty actuarial practice. The author uses K-means clustering to better describe a group of drivers by segmentation. Then the decision tree algorithm is used to analyze the influence of the claim frequency risk factors which reveals that the credit score has the greatest impact on the claim frequency. Further logistic regression is applied to model claim frequency.

Israel Spiegler (2003) found the gap between the technology and knowledge with implicit and explicit methods for generating knowledge of technology like data mining which is the leading thrust to gain actionable knowledge from operational databases and is evident in direct marketing, customer relationship management, user profiling and e-commerce applications. Further two models are also reviewed, compared and discussed in the context of knowledge management, using data mining as an example.

Variar (2004) the importance of data mining to successful business intelligence is highlighted by the increased growth in organizational data volumes. The author states that information is growing faster than Moore’s law and without effective data mining, organizations cannot leverage their data to make better decisions.

Levy and Weitz (2004) found data mining assists organizations in creating value by providing functions such as forecasting, modeling and support for decision making. RFM (recency, frequency, monetary) analysis used by catalog retailers and direct marketers, is a scheme for segmenting customers according to how recent they have made a purchase, how frequent they make purchases, and how much they have bought.

Shimizu (2004) examined how FSP (Frequent Shopper Program) data should contribute to the retailers’ marketing strategy by using data mining techniques. The FSP data derived from a retailer for 3 months in 2002 and sampled 1296 transactions among the data ranked by amount of purchase to identify and classify customers based on real data with data mining techniques. It led to a process of identifying loyal customers, grouping loyal customers, assessing the profitability of each loyal customers segment, ranking the segments, and checking ranking of segments.

Abe (2004) evaluated the popularity of data mining techniques, which is categorized as computer-based discovery of rules. Instead, he developed an analytical model based on consumer behavior theories on traditional RF (recency and frequency) data analysis and estimated the probability of the unobserved defection of customers.

Tsai and Chiu (2004) found that the market segmentation is critical for a good marketing and customer relationship management program which is generally done using general variables such as customer demographic and lifestyle. So the author in this paper develops a novel market segmentation methodology based on product specific variables. Also purchase based similarity measure, clustering algorithm, and clustering quality function are described in this paper. After completing segmentation, a designated RFM model is used to analyze the relative profitability of each customer cluster.

Kim and Nick (2004) proposed a data mining approach for market managers using artificial neural networks guided by genetic algorithms. Their approach produces models that are easier to interpret by using smaller number of predictive features and allows the selection of an optimal target point where expected profits from direct mailing is maximized.

Syed Riaz Ahmed (2004) found the application of data mining in retail business, and its pros and cons on consumers. The author also discusses the various applications of data mining in retail business and focus that data mining is a powerful tool for increasing customer satisfaction, and provides best, safe and useful products at reasonable and economical prices.

Liu and Luo (2005) studied the implementation aspect of clustering data mining method to customer analysis of department store. They built the data warehouse based on OLTP database of Chongquing Liangbai Department Store and then two data mining models were applied for the analysis of customer characteristics and the relationship between customers and product categories and mining results were analyzed.

Chen et al. (2005) found Change mining can extract further value from customer, product and transaction databases. In this study, the behavioral variables, RFM, coupled with growth matrix of customer value, are applied to estimate the value that individual customers contribute to the business. Association rules are used to identify the association between customer profile and product items purchased. The improved measures of similarity and unexpectedness are developed for mining changes in customer behaviors at different time snapshots. Finally, an online query system provides marketing managers a tool for rapid information search, and valuable information based on prompt feedback. The developed system enables marketing managers to rapidly establish marketing strategies.

Kasindra and Robert (2005) illustrate, using a test dataset, how a data mining process can help to achieve an integrated understanding of consumers by marrying information of various types and from various sources, in a manner that ensures that the resulting segments are logically and strongly differentiated on all the types of information in the analysis.

Bart et al. (2005) emphasized on two types of random forests techniques that were applied to analyze the data. Random forests are used for binary classification and regression forests for the model with linear dependent variables. Their findings suggested that past customer behavior is more important to generate repeat purchasing and favorable profitability evolutions while the intermediary’s role has a great impact on the customer’s defection proneness.

Chad et al. (2005) specify that the loyalty of customers to a super market can be measured in a variety of ways. Regular visitors and spenders are most likely to be loyal to a supermarket. The authors describe the results of experiments attempted to identify customer loyalty based on transactional data obtained from a supermarket data collection program. Clustering was done on visits and total weekly expenditures using Kohonen 35 neural network and K-means methods. The authors also provide useful insights for the development of more sophisticated measures for studying customer loyalty.

Wencai and Yu (2005) study the implementation aspects of applying clustering data mining method to customer analysis of a department store. The author first builds a data warehouse of Chongquing Liangbai department store and then two data mining techniques are applied to the analysis of customer characteristics and the relationship between the customer and product categories. Then the mining results are analyzed.

Kuo et al. (2006) attempted to compare three clustering methods as cluster analysis is a common tool for market segmentation and compared with the proposed two-stage method that indicated that the proposed scheme is slightly better that conventional with respect to the rate of misclassification

Wong et al. (2006) applied the RFM model to decide upon the evaluation criteria of valuable customers. The study also uses data mining techniques like decision tree to classify valuable customers into different segments and market basket analysis to mine destinations for cress selling.

Wang and Wang (2006) analysed the importance of purchasing sequences in customer segmentation and proposed a new data mining model for online customer segmentation, and applied this method on an online nutrition product store.

Yen-Liang et al. (2006) attempted to explore the emerging area of merchandising problem using data mining. This paper is motivated in great part by the prominent beer and diapers example and uses data mining technique to discover the implicit, yet meaningful relationship between the relative spatial distance of displayed products and the items' unit sales in a retailer's store. This paper proposes a novel representation scheme and develops a robust algorithm based on association analysis.

Yiyang et al. (2006) proposes a fuzzy clustering based market segmentation approach keeping in to consideration the importance of customer purchase behavior for product family positioning. The authors have taken the application of the proposed methodology in a consumer electronics company producing vibration motors and discussed that using engineering characteristics as segmentation variables, fuzzy clustering based segmentation approach is likely to overcome the preference distortion resulting from other segmentation methods using general variables.

Bhasin (2006) found that the data mining tools in extracting important information from existing data to enable better decision making throughout the banking and retail industries have been successful in the areas like detecting frauds, predicting customer purchase behaviors, optimizing manufacturing process etc. Many retail industries are realizing that data mining can give them a competitive advantage. Data mining typically involves the use of predictive modeling, forecasting and descriptive modeling techniques. By using these techniques, an organization can proactively manage customer retention, identify cross sell and up-sell opportunities, profile and segment customers, set optimal pricing policies, and objectively measure and rank which suppliers are best suited for their needs.

Shin-Yuan et al. (2006) found that the best model of predictive churn from data warehouse to prevent the customers' turnover, further to enhance the competitive edge. Initially, the authors use K-means to model the customers into five clusters and then assess the performance of decision tree and back propagation neural network techniques and suggested that the data mining techniques can effectively assist telecom service providers to make more accurate churner prediction.

Shwu-Ing (2006) used factor and K-means cluster analysis to segment the sample into three clusters and to evaluate the discrimination among the cluster groups, ANOVA and discriminant analysis were used to assess the clustering effect. The analysis resulted in various relationships between attitudes, subjective norms, perceived behavioral control, behavioral intention and real behavior for different groups.

Wang and Hong (2006) described that the changes in customer behavior results in unpredictable customer profitability and cause inefficient and ineffective marketing planning. So the authors describe a customer profitability management system by using data mining techniques to achieve marketing goals. The mechanism has been applied to Telecom Company with promising results.

Waminee et al. (2006) used data mining techniques to analyze historical data of ebanking usages from a commercial bank in Thailand. The authors use K-means and RFM analysis to segment customers into groups according to their personal profiles and ebanking usages. Then Apriori algorithm is applied to detect the relationships within the features of e-banking services. The results can be used to generate new service packages customized to each segment of e-banking user.

Wang and Wang (2006) explained that the online purchasing behavior is characterized by purchasing sequences. Their study proposes a data mining method for customer segmentation and applies it to an online nutrition product store. The results indicate that the method is novel and effective for the online customer segmentation.

Sadic and Kayakutlu (2007) in their study illustrated the implementation of cognitive maps and decision trees in the development of customer segments to be used by sales and marketing departments. Their study contributed not only in the field of data mining but also in customer relations.

Ting Millette (2007) found that the need of a firm to determine 'who are the best prospects among the existing customers for a new product'. For this regression, decision tree and neural network models are built to use for scoring the prospective customers. A confusion matrix is then used to determine the cutoff point for the scores for customers to determine who qualifies as a good target. The models are compared by assessing model performance and validating the model to new data.

Doran Patrick (2007) found Business Intelligence (BI) systems have an important role to play in assisting retail management in the Irish grocery retail sector. The findings indicated that retail management are not getting the information and reports to make effective decisions and thus an information architecture was developed for different end user perspective and has numerous applications to retailers.

Buckinx et al. (2007) found that the customer database with a prediction of customer's behavioral use cluster analysis to locate the potential customers and association analysis to extract knowledge of loyal customers' purchasing behavior

Sung (2007) found that the most fundamental problem of customer segmentation analysis is deriving descriptive and predictive knowledge from customer segments, and attempts to make more accurate and reliable predictions about segment transitions.

Xindong (2008) presented the top ten data mining algorithms identified by IEEE International Conference on Data Mining (ICDM) covering Classification, clustering, statistical learning, association analysis, and link mining. The authors describe the top 10 algorithms including C4.5, K-means, SVM, Apriori, EM, PAgerank, Adaboost, KNN, Naïve Bayes and CART are among the most influential data mining algorithms in the research community. The authors provide the description, impact and review current and future research on the algorithms..

Yu and Wang (2008) found that the use of K-means algorithm to classify the customers and products according to their demographic data to predict their return patterns have performed the association mining process across customer, product, and marketing strategy dimensions.

Henry and Beverley (2008) presented that the Knowledge Discovery in databases in a field of research that studies the development and use of various data analysis tools and techniques and Data mining is one of such tool. However, nearly two third of the IT 40 managers say that data mining products are too difficult to use in a business context. So the authors discuss how advances in data mining translate into the business context.

Yasemin and Reutterer (2008) proposed and illustrated a two stage procedure combining the features form exploratory and model based approaches of market basket analysis. Retail marketing managers can make use of this information and thus can be assisted in designing targeted direct marketing actions within their loyalty programs.

Hsieh and Chu (2009) propose an integrated data mining and behavioral scoring model to manage existing credit card customers in a bank. A self-organizing neural network was used to identify groups of customers based on repayment behavior and RFM behavioral scoring predicators. The authors divided the bank customers into three major profitable groups which were then profiled by customer's feature attributes using an Apriori association rule indicator that further facilitates marketing strategy development.

Shu-Hsien et al. (2008) developed a relational database and proposes Apriori algorithm and K-means as methodologies for association rule and cluster analysis for data mining, which is then implemented to mine customer knowledge from household customers. Knowledge extraction by data mining results is illustrated as knowledge patterns/rules and clusters in order to propose suggestions and solutions to the case firm for product line and brand extensions and knowledge management.

Popovic and Dalbela (2009) found that the data mining methods based on fuzzy logic could be successfully applied in retail banking analysis and they have used application of fuzzy clustering in churn prediction for retail banking.

Yaganag et al. (2009) found the concept of cluster analysis data mining technique to resolve quick and exact detection of fault components and fault sections, and finally accomplish fault analysis.

Nagi and Trappey et al. (2009) found that the historical sales data and demographic attributes clustered for the first level of segmentation, to analyze each sub segment based on menu choice preferences to provide customized coupons and price discounts for each customer based on their previous behavior.

Fang and Lee (2009) evaluated 565 valid responses of the questionnaire distributed and then using two step cluster analysis, four distinct food-related consumer life style segments were identified and the segments differ in their attitude and behavior towards food consumption. Further the profiles of the segments are achieved by observing the socio-demographic characteristics of typical segment members.

Sheu et al. (2009) contributed to integrate data mining and experiential marketing to segment online game customers. The authors first located and modified the important influential factors and then by applying the techniques of decision tree data mining, they have explored the potential relationship between these factors and customer loyalty.

Irko et al. (2009) found that the customer segmentation is typically done by applying some form of cluster analysis to obtain a set of segments to which future customers are assigned to. The frequent item set discovery, combined with tracking the temporal development of support and the 42 application of an change-based interestingness notion used for detecting and monitoring customer segments provides detailed knowledge about how customer behavior evolves over time.

Musa (2010) described the way to utilize Microsoft's Excel Data Mining add-ins as a front-end to Microsoft's Cloud Computing and SQL Server 2008 Business Intelligence Platform as the back-end. The contents presented have broader applications in the areas such as accounting, finance, general business and marketing. The paper provides an understanding of data mining algorithms and tools to perform (1) elementary data analysis (2) configure and use data mining computing engines to build, test, compare, and evaluate various data mining models, and (3) use the mining models to analyze data and predict outcomes for the purpose of decision support.

Dhanpal et al. (2010) presented an original methodological approach of customer satisfaction evaluation by combining multicriteria preference disaggregation analysis and rule induction data mining.

Introduction: Data profiling is the process of examining data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that data. The purpose of these statistics is to find out whether existing data can easily be used for other purposes, and improve the ability to search data by tagging it with keywords, descriptions, or assigning it to a category. Data profiling refers to the analysis of information for use in a data warehouse in order to clarify the structure, content, relationships, and derivation rules of the data. The Profiling helps to not only understand anomalies and assess data quality, but also to discover, register, and assess enterprise metadata. The result of the analysis is used to determine the suitability of the candidate source systems, usually giving the basis for an early go/no-go decision, and also to identify problems for later solution design.

Objective; (i); To evaluate the procedures of quality data profiling for effective business process: The Profiling is Conducted with the methods of descriptive statistics such as minimum, maximum, mean, mode, percentile, standard deviation, frequency, variation, aggregates such as count and sum, and additional metadata information obtained during data profiling such as data type, length, discrete values, uniqueness, occurrence of null values, typical string patterns, and abstract type recognition. Normally, purpose-built tools are used for data profiling to ease the process. The computation complexity increases when going from single column, to single table, to cross-table structural profiling. Therefore, performance is an evaluation criterion for profiling tools. **Benefits of data profiling for business use**; The benefits of data profiling are to improve data quality, shorten the implementation cycle of major projects, and improve users' understanding of data. Discovering business knowledge embedded in data itself is one of the significant benefits derived from data profiling. Data profiling is one of the most effective technologies for improving data accuracy in corporate databases.

Data Profiling Tools; Some tools are free software, or open source, however, many, but not all free data profiling tools are open source projects. In general, their functionality is more limited than that of commercial products, and they may not offer free telephone or online support. Furthermore, documentation is not always thorough. However, some small companies still use these free tools instead of expensive commercial software, considering the benefits that free tools provide.

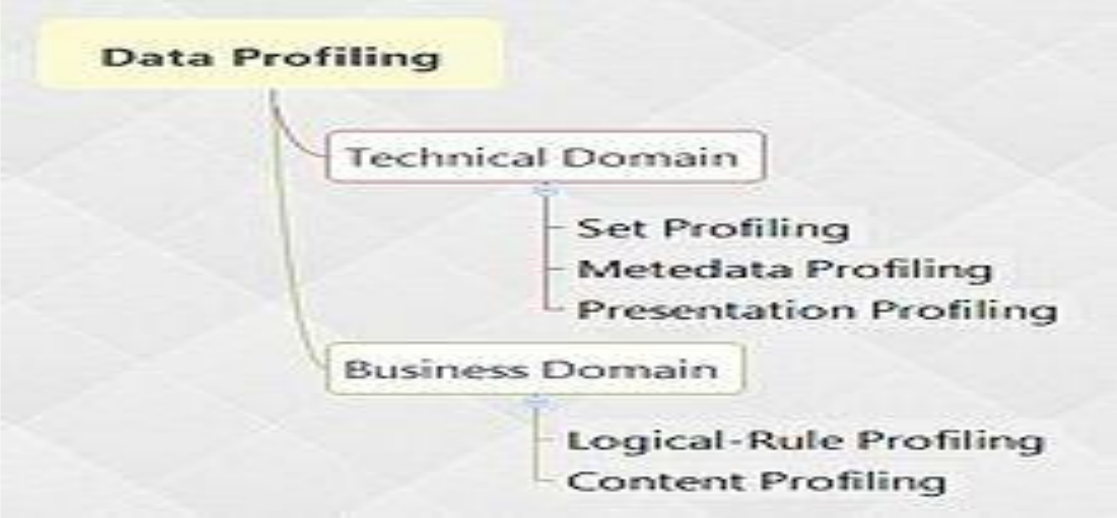

Figure: 1: Data Profiling Process: Source ; www.wikipedia.org

Objective; (ii): To evaluate the modern data profiling techniques and its advancements:

Modern data profiling process starts with the data quality assessment: The Corporate companies with an emphasis on marketing often focus on their quality efforts on name and address information, but data quality is recognized as an important property of all types of data. Principles of data quality can be applied to supply chain data, transactional data, and nearly every other category of data found. For example, making supply chain data conform to a certain standard has value to an organization by: 1) avoiding overstocking of similar but slightly different stock; 2) avoiding false stock-out; 3) improving the understanding of vendor purchases to negotiate volume discounts; and 4) avoiding logistics costs in stocking and shipping parts across a large organization. For companies with significant research efforts, data quality can include developing protocols for research methods, reducing measurement error, bounds checking of data, cross tabulation, modeling and outlier detection, verifying data integrity, etc. **Examples of characteristics are**: completeness, validity, accuracy, consistency, availability and timeliness. Requirements are defined as the need or expectation that is stated, generally implied or obligatory.

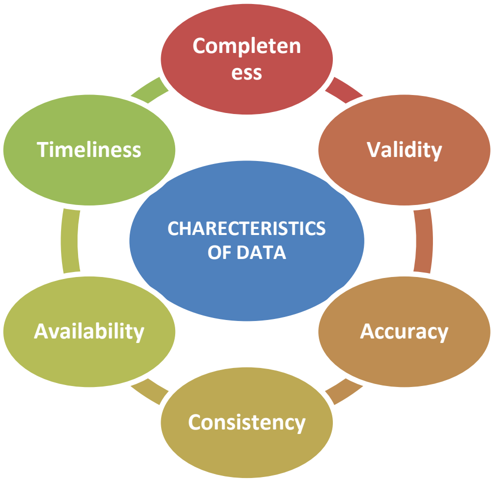

Fig:2: Characteristics of Data Quality: Concept Design and Source: Prof Dr.C.Karthikeyan

Problems with data quality not only arise from incorrect data but with inconsistent data as well. The data shadow is another problem with quality, which a company can take to ensure data consistency. Enterprises, scientists, and researchers needs to handle data curation to improve the quality of their common data. The market thrives with data quality assurance, and number of vendors makes tools for analyzing and repairing poor quality data on a contract basis and consultants can advise on fixing processes or systems to avoid data quality problems in the first place. There are several well-known authors and self-styled experts, with Larry English perhaps the most popular guru. In addition, IQ International - the International Association for

Information and Data Quality was established in 2004 to provide a focal point for professionals and researchers in this field. ISO 8000 is an international standard for data quality.

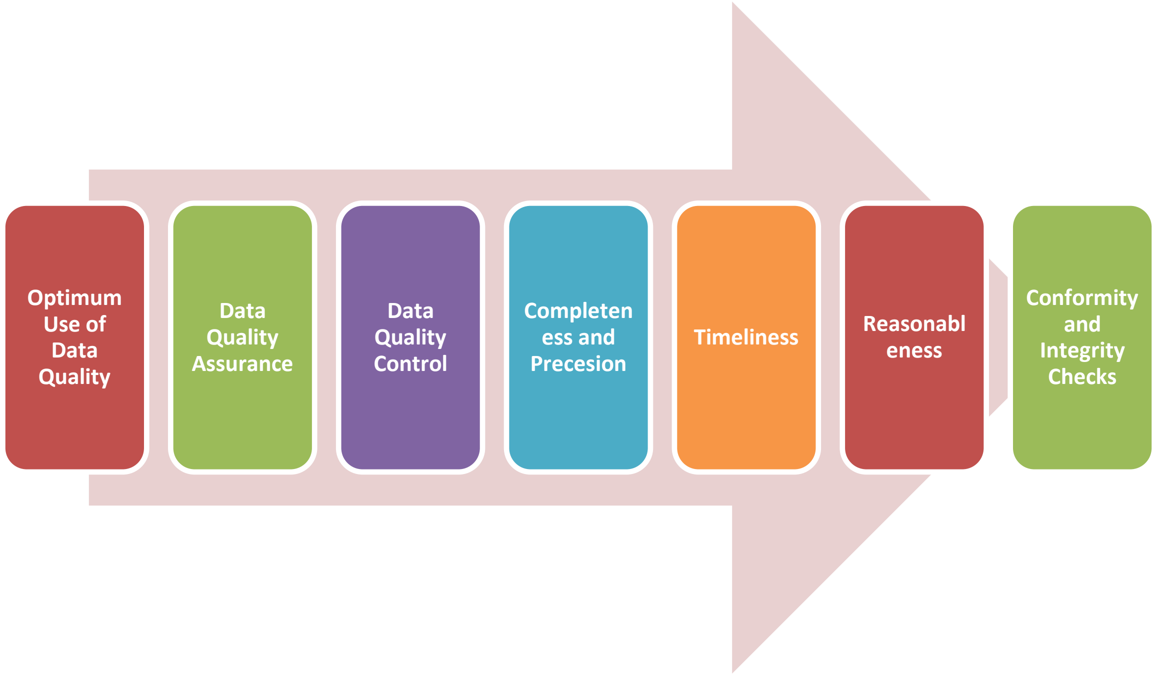
Optimum Use of Data Quality
Data Quality Assurance
Data Quality Control
Completeness and Precesion
Timeliness
Reasonableness
Conformity and Integrity Checks

Fig:3: The cumulative problems of Data Quality: Concept Design and Source: Prof Dr.C.Karthikeyan.

Optimum use of data quality; Data Quality (DQ) is a niche area required for the integrity of the data management by covering gaps of data issues. This is one of the key functions that aid data governance by monitoring data to find exceptions undiscovered by current data management operations. Data Quality checks may be defined at attribute level to have full control on its remediation steps. DQ checks and business rules may easily overlap if an organization is not attentive of its DQ scope. Business teams should understand the DQ scope thoroughly in order to avoid overlap. Data quality checks are redundant if **business logic** covers the same functionality and fulfills the same purpose as DQ.

Data quality assurance; Data quality assurance is the process of data profiling to discover inconsistencies and other anomalies in the data, as well as performing data cleansing activities (e.g. removing outliers, missing data interpolation) to improve the data quality. These activities can be undertaken as part of data warehousing or as part of the database administration with leading application software.

Data quality control is the process of controlling the usage of data with known quality measurements for an application or a process. This process is usually done after a Data Quality Assurance (QA) process, which consists of discovery of data inconsistency and correction. **The Data QC process** uses the information from the QA process to decide to use the data for analysis or in an application or business process. Establishing data QC process provides the protection of usage of data control and establishes safe information usage.

Completeness and **precision**: All data having attributes referring to *Reference Data* in the organization may be validated against the set of well-defined valid values of Reference Data to discover new or discrepant values through the **validity** DQ check. Results may be used to update *Reference Data* administered under *Master Data Management (MDM)*. All data sourced from a *third party* to organization's internal teams may undergo **accuracy** (DQ) check against the third party data. These DQ check results are valuable when administered on data that made multiple hops after the point of entry of that data but before that data becomes authorized or stored for enterprise intelligence. All data columns that refer to *Master Data* may be validated for its **consistency** check. A DQ check administered on the data at the point of entry discovers new data for the MDM process, but a DQ check administered after the point of entry discovers the failure (not exceptions) of consistency. As data transforms, multiple timestamps and the positions of that timestamps are captured and may be compared against each other and its leeway to validate its value, decay, operational significance against a defined SLA (service level agreement).

Timeliness DQ check can be utilized to decrease data value decay rate and optimize the policies of data movement timeline. In an organization complex logic is usually segregated into simpler logic across multiple processes.

Reasonableness DQ checks on such complex logic yielding to a logical result within a specific range of values or static interrelationships (aggregated business rules) may be validated to discover complicated but crucial business processes and outliers of the data, its drift from BAU (business as usual) expectations, and may provide possible exceptions eventually resulting into data issues. This check may be a simple generic aggregation rule engulfed by large chunk of data or it can be a complicated logic on a group of attributes of a transaction pertaining to the core business of the organization. This DQ check requires high degree of business knowledge and acumen. Discovery of reasonableness issues may aid for policy and strategy changes by either business or data governance or both.

Conformity checks and **integrity checks** need not covered in all business needs, it's strictly under the database architecture's discretion. There are many places in the data movement where DQ checks may not be required. For instance, DQ check for completeness and precision on not–null columns is redundant for the data sourced from database. Similarly, data should be validated for its accuracy with respect to time when the data is stitched across disparate sources. However, that is a business rule and should not be in the DQ scope. Regretfully, from a software development perspective, Data Quality is often seen as a non functional requirement. And as such, key data quality checks/processes are not factored into the final software solution. Within Healthcare, wearable technologies or Body Area Networks, generate large volumes of data. The level of detail required to ensure data quality is extremely high and is often under estimated. This is also true for the vast majority of mHealth apps, EHRs and other health related software solutions. However, some open source tools exist that examine data quality. The primary reason for this, stems from the extra cost involved is added a higher degree of rigor within the software architecture.

Objective: (iii); Examine the role of Professional Organisations in Data Quality Management and Governance.

Professional associations; **IQ International—the International Association for Information and Data Quality;** IQ International is a not-for-profit, vendor neutral, professional association formed in 2004, dedicated to building the information and data quality profession.

Data governance is a control that ensures that the data entry by an operations team member or by automated processes meets precise standards, such as a business rule, a data definition and data integrity constraints in the data model. The data governor uses data quality monitoring against production data to communicate errors in data back to operational team members, or to the technical support team, for corrective action. Data governance is used by organizations to exercise control over processes and methods used by their data stewards and data custodians in order to improve data quality. Data governance is a set of processes that ensures that important data assets are formally managed throughout the enterprise. Data governance ensures that data can be trusted and that people can be made accountable for any adverse event that happens because of low data quality. It is about putting people in charge of fixing and preventing issues with data so that the enterprise can become more efficient. Data governance also describes an evolutionary process for a company, altering the company's way of thinking and setting up the processes to handle information so that it may be utilized by the entire organization. It's about using technology when necessary in many forms to help aid the process. When companies desire, or are required, to gain control of their data, they empower their people, set up processes and get help from technology to do it. According to one vendor, data governance is a quality control discipline for assessing, managing, using, improving, monitoring, maintaining, and protecting organizational information. It is a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods. Data governance encompasses the people, processes, and information technology required to create a consistent and proper handling of an organization's data across the business enterprise. Goals may be defined at all levels of the enterprise and doing so may aid in acceptance of processes by those who will use them.

Data governance drivers; While data governance initiatives can be driven by a desire to improve data quality, they are more often driven by C-Level leaders responding to external regulations. Examples of these regulations include Sarbanes-Oxley, Basel I, Basel II, HIPAA, GDPR and a number of data privacy regulations. To achieve compliance with these regulations, business processes and controls require formal management processes to govern the data subject to these regulations. Successful programs identify drivers meaningful to both supervisory and executive leadership. Common themes among the external regulations center on the need to manage risk. The risks can be financial misstatement, inadvertent release of sensitive data, or poor data quality for key decisions. Methods to manage these risks vary from industry to industry. Examples of commonly referenced best practices and guidelines include COBIT, ISO/IEC 38500, and others. The proliferation of regulations and standards creates challenges for data governance professionals, particularly when multiple regulations overlap the data being managed. Organizations often launch data governance initiatives to address these challenges.

Data governance initiatives (Dimensions); Data governance initiatives improve data quality by assigning a team responsible for data's accuracy, accessibility, consistency, and completeness, among other metrics. This team usually consists of executive leadership, project management, line-of-business managers, and data stewards. The team usually employs some form of methodology for tracking and improving enterprise data, such as Six Sigma, and tools for data mapping, profiling, cleansing, and monitoring data. Data governance initiatives may be aimed at achieving a number of objectives including offering better visibility to internal and external customers (such as supply chain management), compliance with regulatory law, improving operations after rapid company growth or corporate mergers, or to aid the efficiency of enterprise knowledge workers by reducing confusion and error and increasing their scope of knowledge. Many data governance initiatives are also inspired by past attempts to fix information quality at the departmental level, leading to incongruent and redundant data quality processes. Most large companies have many applications and databases that can't easily share information. Therefore, knowledge workers within large organizations often don't have access to the information they need to best do their jobs. When they do have access to the data, the data quality may be poor. By setting up a data governance practice or Corporate Data Authority, these problems can be mitigated. The structure of a data governance initiative will vary not only with the size of the organization, but with the desired objectives or the 'focus areas' of the effort.

Implementation; Implementation of a Data Governance initiative may vary in scope as well as origin. Sometimes, an executive mandate will arise to initiate an enterprise wide effort, sometimes the mandate will be to create a pilot project or projects, limited in scope and objectives, aimed at either resolving existing issues or demonstrating value. Sometimes an initiative will originate lower down in the organization’s hierarchy, and will be deployed in a limited scope to demonstrate value to potential sponsors higher up in the organization. The initial scope of an implementation can vary greatly as well, from review of a one-off IT system, to a cross-organization initiative.

Data governance tools; Leaders of successful data governance programs declared in December 2006 at the Data Governance Conferenc e in Orlando, Fl, that data governance is between 80 and 95 percent communication." That stated, it is a given that many of the objectives of a Data Governance program must be accomplished with appropriate tools. Many vendors are now positioning their products as Data Governance tools; due to the different focus areas of various data governance initiatives, any given tool may or may not be appropriate, in addition, many tools that are not marketed as governance tools address governance needs.

Data governance organizations; **DAMA International**; DAMA (the Data Management Association) is a not-for-profit, vendor-independent, international association of technical and business professionals dedicated to advancing the concepts and practices of information resource management (IRM) and data resource management (DRM).

Data Governance Professionals Organization (DGPO); The Data Governance Professionals Organization (DGPO) is a non-profit, vendor neutral, association of business, IT and data professionals dedicated to advancing the discipline of data governance. The objective of the DGPO is to provide a forum that fosters discussion and networking for members and to encourage, develop and advance the skills of members working in the data governance discipline.

The Data Governance Society; The Data Governance Society, Inc. is dedicated to fostering a new paradigm for the effective use and protection of information in which Data is governed and leveraged as a unique corporate asset.

The Data Governance Council; The Data Governance Council is an organization formed by IBM consisting of companies, institutions and technology solution providers with the stated objective to build consistency and quality control in governance, which will help companies better protect critical data." **IQ International -- the International Association for Information and Data Quality;** IQ International is a not-for-profit, vendor neutral, professional association formed in 2004, dedicated to building the information and data quality profession. **Master data**

management; In business, **master data management** (**MDM**) comprises the processes, governance, policies, standards and tools that consistently define and manage the critical data of anorganization to provide a single point of reference.

Objective :(iv): To examine the advantages of Master Data for Business Process

The data that is mastered may include: reference data – the business objects for transactions, and the dimensions for analysis; analytical data – supports decision making;. In computing, a master data management tool can be used to support master data management by removing duplicates, standardizing data (mass maintaining), and incorporating rules to eliminate incorrect data from entering the system in order to create an authoritative source of master data. Master data are the products, accounts and parties for which the business transactions are completed. The root cause problem stems from business unit and product line segmentation, in which the same customer will be serviced by different product lines, with redundant data being entered about the customer (a.k.a. party in the role of customer) and account in order to process the transaction. The redundancy of party and account data is compounded in the front to back office life cycle, where the authoritative single source for the party, account and product data is needed but is often once again redundantly entered or augmented. Master data management has the objective of providing processes for collecting, aggregating, matching, consolidating, quality-assuring, persisting and distributing such data throughout an organization to ensure consistency and control in the ongoing maintenance and application use of this information. The term recalls the concept of a *master file* from an earlier computing era.

Master data management (MDM) is a comprehensive method of enabling an enterprise to link all of its critical data to one file, called a master file that provides a common point of reference. When properly done, master data management streamlines data sharing among personnel and departments. In addition, master data management can facilitate computing in multiple system architectures, platforms and applications.

Master Data Management (MDM) can be viewed as a "discipline for specialized quality improvement" defined by the policies and procedures put in place by a data governance organization. At a basic level, master data management seeks to ensure that an organization does not use multiple (potentially inconsistent) versions of the same master data in different parts of its operations, which can occur in large organizations. A typical example of poor master data management is the scenario of a bank at which a customer has taken out a mortgage and the bank begins to send mortgage solicitations to that customer, ignoring the fact that the person already has a mortgage account relationship with the bank. This happens because the customer information used by the marketing section within the bank lacks integration with the customer information used by the customer services section of the bank. Thus the two groups remain unaware that an existing customer is also considered a sales lead. The process of record linkage is used to associate different records that correspond to the same entity, in this case the same person.

As with other Extract, Transform, Load-based data movement, these processes are expensive and inefficient to develop and to maintain which greatly reduces the return on investment for the master data management product. One of the most common reasons some large corporations experience massive issues with master data management is growth through mergers or acquisitions. Any organizations which merge will typically create an entity with duplicate master data (since each likely had at least one master database of its own prior to the merger). Ideally, database resolves this problem through reduplication of the master data as part of the merger. In practice, however, reconciling several master data systems can present difficulties because of the dependencies that existing applications have on the master databases. As a result, more often than not the two systems do not fully merge, but remain separate, with a special reconciliation process defined that ensures consistency between the data stored in the two systems. Over time, however, as further mergers and acquisitions occur, the problem multiplies, more and more master databases appear, and data-reconciliation processes become extremely complex, and consequently unmanageable and unreliable. Because of this trend, one can find organizations with 10, 15, or even as many as 100 separate, poorly integrated master databases, which can cause serious operational problems in the areas of customer satisfaction, operational efficiency, decision support, and regulatory compliance.

Transmission of master data; There are several ways in which master data may be collated and distributed to other systems. This includes: Data consolidation – The process of capturing master data from multiple sources and integrating into a single hub (operational data store) for replication to other destination systems. Data propagation – The process of copying master data from one system to another, typically through point-to-point interfaces in legacy systems.

Analysis paralysis; Analysis paralysis or paralysis by analysis is the state of over-analyzing (or over-thinking) a situation so that a decision or action is never taken, in effect paralyzing the outcome. A decision can be treated as over-complicated, with too many detailed options, so that a choice is never made, rather than try something and change if a major problem arises. A person might be seeking the optimal or "perfect" solution upfront, and fear making any decision which could lead to erroneous results, while on the way to a better solution. The phrase describes a situation in which the opportunity cost of decision analysis exceeds the benefits that could be gained by enacting some decision, or an informal or non-deterministic situation where the sheer quantity of analysis overwhelms the decision-making process itself, thus preventing a decision. The phrase applies to any situation where analysis may be applied to help make a decision and may be a dysfunctional element of organizational behavior. This is often phrased as *paralysis by analysis*, in contrast toextinct by instinct (making a fatal decision based on hasty judgment or a gut-reaction).

Personal analysis; Casual analysis paralysis can occur during the process of trying to make personal decisions if the decision-maker overanalyzes the circumstance with which they are faced. When this happens, the sheer volume of analysis overwhelms the decision-maker, weighing him or her down so much that they feel overwhelmed with the task, unable to make a rational conclusion. In some cases, the decision-maker can analyze every possible outcome of an action and write it all out, but then delete it because of how they analyze the outcome to be and how they may be viewed.

Conversational analysis; Although analysis paralysis can actually occur at any time, regarding any issue in typical conversation, it is particularly likely to occur during elevated, intellectual discussions. During such intellectual discussion, analysis paralysis involves the over-analysis of a specific issue to the point where that issue can no longer be recognized, and the subject of the conversation is lost. Usually, this happens because complex issues (which are often the basis of elevated, intellectual conversation) are intricately connected with various other issues, and the pursuit of these various issues makes logical sense to the participants.

Information architecture (**IA**) is the structural design of shared information environments; the art and science of organizing and labelling websites, intranets, online communities and software to support usability and findability; and an emerging community of practice focused on bringing principles of design and architecture to the digital landscape. Typically, it involves a model or conceptof information that is used and applied to activities which require explicit details of complex information systems. These activities include library systems and database development. Information architecture is considered to have been founded by Richard Saul Wurman.[2] Today there is a growing network of active IA specialists who constitute the Information Architecture Institute.

Information and technology (**IT**) **governance** is a subset discipline of corporate governance, focused on information and technology (IT) and its performance and risk management. The interest in IT governance is due to the ongoing need within organizations to focus value creation efforts on an organization's strategic objectives and to better manage the performance of those responsible for creating this value in the best interest of all stakeholders.

Findings and Conclusion:

From the study, it can be concluded that the Data Profiling Mechanism have grown phenomenally and professionally and also have grown internationally. The findings apart from the data quality and profiling the latest developments are. Professional certification; Certified in the Governance of Enterprise Information Technology (CGEIT) is a certification created in 2007 by the Information Systems Audit and Control Association (ISACA). It is designed for experienced professionals, who can demonstrate 5 or more years experience, serving in a managing or advisory role focused on the governance and control of IT at an enterprise level. It also requires passing a 4-hour test, designed to evaluate an applicant's understanding of enterprise IT management. The first examination was held in December 2008. COBIT; COBIT (Control Objectives for Information and Related Technologies) is a good-practice framework created by international professional association ISACA for information technology (IT) management and IT governance. COBIT provides an implementable "set of controls over information technology and organizes them around a logical framework of IT-related processes and enablers." ISACA first released COBIT in 1996, originally as a set of control objectives to help the financial audit community better maneuver in IT-related environments. The COBIT framework; COBIT was initially "Control Objectives for Information and Related Technologies," though before the release of the framework people talked of "CobiT" as "Control Objectives for IT" or "Control Objectives for Information and Related Technology." The framework defines a set of generic processes for the management of IT, with each process defined together with process inputs and outputs, key process-activities, process objectives, performance measures and an elementary maturity model. COBIT also provides a set of recommended best practices for governance and control process of information systems and technology with the essence of aligning IT with business. COBIT 5 consolidates COBIT 4.1, Val IT and Risk IT into a single framework acting as an enterprise framework aligned and interoperable with other frameworks and standards. Framework and components; The business orientation of COBIT consists of linking business goals to IT goals, providing metrics and maturity models to measure their achievement, and identifying the associated responsibilities of business and IT process owners.The process focus of COBIT is illustrated by a process model that subdivides IT into four domains (Plan and Organize; Acquire and Implement; Deliver and Support; and Monitor and Evaluate) and 34 processes in line with the responsibility areas of plan, build, run, and monitor. It is positioned at a high level and has been aligned and harmonized with other, more detailed IT standards and good practices such as COSO, ITIL, BiSL, ISO 27000, CMMI, TOGAF and PMBOK. COBIT acts as an integrator of these different guidance materials, summarizing key objectives under one umbrella framework that link the good practice models with governance and business requirements.[1] COBIT 5 further consolidated and integrated the COBIT 4.1, Val IT 2.0 and Risk IT frameworks and drew from ISACA's *IT Assurance Framework* (ITAF) and the *Business Model for Information Security* (BMIS).The framework and its components can, when utilized well, also contribute to ensuring regulatory compliance. It can encourage less wasteful information management, improve retention schedules, increase business agility, and lower costs while better complying with data retention and management regulations. The standard; ISO/IEC 38500 is applicable to organizations of all sizes, including public and private companies, government entities, and not-for-profit organizations. This standard provides guiding principles for directors of organizations on the effective, efficient, and acceptable use of Information Technology (IT) within their organizations. It is organized into three prime sections: Scope, Framework and Guidance. The framework comprises definitions, principles and a model. It sets out six principles for good corporate governance of IT: Responsibility, Strategy Acquisition, Performance, Conformance, Human behavior. It also provides guidance to those advising, informing, or assisting directors. ISO/TC 215; The ISO/TC 215 is the International Organization for Standardization's (ISO) Technical Committee (TC) on health informatics. TC 215 works on the standardization of Health Information and Communications Technology (ICT), to allow for compatibility and interoperability between independent systems. Unique Identifiers Rule (National Provider Identifier); HIPAA covered entities such as providers completing electronic transactions, healthcare clearing houses, and large health plans, must use only the National Provider Identifier (NPI) to identify covered healthcare providers in standard transactions by May 23, 2007. Small health plans must use only the NPI by May 23, 2008. Effective from May 2006 (May 2007 for small health plans), all covered entities using electronic communications (e.g., physicians, hospitals, health insurance companies, and so forth) must use a single new NPI. The NPI replaces all other identifiers used by health plans, Medicare, Medicaid, and other government programs. However, the NPI does not replace a provider's DEA number, state license number, or tax identification number. The NPI is 10 digits (may be alphanumeric), with the last digit being a checksum. The NPI cannot contain any embedded intelligence; in other words, the NPI is simply a number that does not itself have any additional meaning. The NPI is unique and national, never re-used, and except for institutions, a provider usually can have only one. An institution may obtain multiple NPIs for different "sub-parts" such as a free-standing cancer center or rehab facility.

References ;

1. Theodore Johnson (2009), "Data Profiling", in Encyclopedia of Database Systems, Springer, Heidelberg]

2. Woodall, Philip; Oberhofer, Martin; Borek, Alexander (2014). "A classification of data quality assessment and improvement methods".International Journal of Information Quality. **3** (4). doi:10.1504/ijiq.2014.068656.

3. Ralph Kimball et al. (2008), “The Data Warehouse Lifecycle Toolkit”, Second Edition, Wiley Publishing, Inc., ISBN 9780470149775], (p. 297) (p. 376)

4. David Loshin (2009), “Master Data Management”, Morgan Kaufmann Publishers, ISBN 9780123742254], (pp. 94–96)

5. David Loshin (2003), “Business Intelligence: The Savvy Manager’s Guide, Getting Onboard with Emerging IT”, Morgan Kaufmann Publishers, ISBN 9781558609167], (pp. 110–111)]

6. Erhard Rahm and Hong Hai Do (2000), “Data Cleaning: Problems and Current Approaches” in “Bulletin of the Technical Committee on Data Engineering”, IEEE Computer Society, Vol. 23, No. 4, December 2000]

7. Ranjit Singh, Dr Kawaljeet Singh et al. (2010), “A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing”, IJCSI International Journal of Computer Science Issue, Vol. 7, Issue 3, No. 2, May 2010]

8. Ralph Kimball (2004), “Kimball Design Tip #59: Surprising Value of Data Profiling”, Kimball Group, Number 59, September 14, 2004, (www.rkimball.com/html/designtipsPDF/ KimballDT59 SurprisingValue.pdf)]

9. Jack E. Olson (2003), “Data Quality: The Accuracy dimension”, Morgan Kaufmann Publishers], (pp. 140–142)

10. Redman, Thomas C. (30 December 2013). Data Driven: Profiting from Your Most Important Business Asset. Harvard Business Press. ISBN 978-1-4221-6364-1.

11. "What is data scrubbing (data cleansing)? - Definition from WhatIs.com".

12. "Data Quality: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors". Retrieved 5 February 2013.

13. "IAIDQ--glossary".

14. Government of British Columbia

15. REFERENCE-QUALITY WATER SAMPLE DATA: NOTES ON ACQUISITION, RECORD KEEPING, AND EVALUATION

16. "ISTA Con - Innovations in Software Technologies and Automation.".

17. Anonymous (23 December 2014). "Data Quality".

18. "Liability and Leverage - A Case for Data Quality".

19. "Address Management for Mail-Order and Retail".

20. E. Curry, A. Freitas, and S. O'Riáin, "The Role of Community-Driven Data Curation for Enterprises," in Linking Enterprise Data, D. Wood, Ed. Boston, MA: Springer US, 2010, pp. 25-47.

21. "ISO/TS 8000-1:2011 Data quality -- Part 1: Overview". International Organization for Standardization. Retrieved 8 December 2016.

22. "Can you trust the quality of your data?".

23. "What is Data Cleansing? - Experian Data Quality". 13 February 2015.

24. "Lecture 23 Data Quality Concepts Tutorial – Data Warehousing". Watch Free Video Training Online. Retrieved 8 December 2016.

25. O'donoghue, John, and John Herbert. "Data management within mHealth environments: Patient sensors, mobile devices, and databases." Journal of Data and Information Quality (JDIQ) 4.1 (2012): 5.

26. Huser, Vojtech; DeFalco, Frank J; Schuemie, Martijn; Ryan, Patrick B; Shang, Ning; Velez, Mark; Park, Rae Woong; Boyce, Richard D; Duke, Jon; Khare, Ritu; Utidjian, Levon; Bailey, Charles (30 November 2016). "Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets". eGEMs (Generating Evidence & Methods to improve patient outcomes). **4** (1). doi:10.13063/2327-9214.1239.

27. "IQ International - the International Association for Information and Data Quality". IQ International website. Retrieved 2016-08-05.

28. Baškarada, S; Koronios, A (2014). "A Critical Success Factors Framework for Information Quality Management". Information Systems Management. **31** (4): 1–20.doi:10.1080/10580530.2014.958023.

29. Baamann, Katharina, "Data Quality Aspects of Revenue Assurance", Article

30. Eckerson, W. (2002) "Data Warehousing Special Report: Data quality and the bottom line", Article

31. Ivanov, K. (1972) "Quality-control of information: On the concept of accuracy of information in data banks and in management information systems". The University of Stockholm and The Royal Institute of Technology. Doctoral dissertation.

32. Hansen, M. (1991) Zero Defect Data, MIT. Masters thesis [1]

33. Kahn, B., Strong, D., Wang, R. (2002) "Information Quality Benchmarks: Product and Service Performance," Communications of the ACM, April 2002. pp. 184–192. Article

34. Price, R. and Shanks, G. (2004) A Semiotic Information Quality Framework, Proc. IFIP International Conference on Decision Support Systems (DSS2004): Decision Support in an Uncertain and Complex World, Prato. Article

35. Redman, T. C. (2008) Data Driven: Profiting From Our Most Important Business Asset

36. Wand, Y. and Wang, R. (1996) "Anchoring Data Quality Dimensions in Ontological Foundations," Communications of the ACM, November 1996. pp. 86–95. Article

37. Wang, R., Kon, H. & Madnick, S. (1993), Data Quality Requirements Analysis and Modelling, Ninth International Conference of Data Engineering, Vienna, Austria. Article

38. Fournel Michel, Accroitre la qualité et la valeur des données de vos clients, éditions Publibook, 2007. ISBN 978-2-7483-3847-8.

39. Daniel F., Casati F., Palpanas T., Chayka O., Cappiello C. (2008) "Enabling Better Decisions through Quality-aware Reports", International Conference on Information Quality (ICIQ), MIT. Article

40. Jack E. Olson (2003), "Data Quality: The Accuracy dimension", Morgan Kaufmann Publishers

41. Woodall P., Oberhofer M., and Borek A. (2014), "A Classification of Data Quality Assessment and Improvement Methods". International Journal of Information Quality **3** (4), 298–321. doi:10.1504/ijiq.2014.068656.

42. Woodall, P., Borek, A., and Parlikad, A. (2013), "Data Quality Assessment: The Hybrid Approach." Information & Management **50** (7), 369–382.

43. Jeff Boss, How To Overcome The 'Analysis Paralysis' Of Decision-Making, Forbes, March 20, 2015

44. Parvini, Neema. "'And Reason Panders Will': Another Look at Hamlet's Analysis Paralysis". Shakespeare and Cognition: Thinking Fast and Slow through Character. Palgrave Macmillan UK. pp. 52–62. ISBN 9781349713080.

45. Roberts, Lon (January–February 2010). "Analysis Paralysis: A Case of Terminological Inexactitude" (PDF). Defense AT&L: 18–22. Retrieved 13 May 2016.

46. Ansoff, H. Igor (1965). Corporate Strategy: an Analytic Approach to Business Policy for Growth and Expansion. New York: McGraw-Hill.

47. Kennedy, Carol (2006). Guide to the management gurus : the best guide to business thinkers (5th ed.). London: Random House Business. ISBN 1905211023.

48. "Igor Ansoff". The Economist.

49. Silver, Henry K.; Hecker, James A. (March 1970). "The Pediatric Nurse Practitioner and the Health Associate: New Types of Health Professionals". Journal of Medical Education. **45**: 171–176. Retrieved 10 May 2016.

50. "analysis paralysis: definition of analysis paralysis in Oxford dictionary (American English) (US)". Oxford Dictionaries. Retrieved 10 May 2016.

51. "Managing Analysis Paralysis". Business Analyst Learnings. Retrieved 15 May 2016.

52. "Analysis Paralysis". Sourcemaking. Retrieved 14 May 2016.

53. "Board Game Resource - How to deal with Analysis Paralysis?". Board Game Resource. 26 October 2015. Retrieved 15 May 2016.

54. "Designing Games to Prevent Analysis Paralysis - Part 1 | The Best Games Are Yet To Be Made". www.leagueofgamemakers.com. Retrieved 15 May 2016.

55. "GDC Vault - Overcoming Analysis Paralysis: Experimenting with Bears vs. Art". www.gdcvault.com.

56. Kane, Becky (8 July 2015). "The Science of Analysis Paralysis: How Overthinking Kills Your Productivity & What You Can Do About It". Todoist Blog. Retrieved 14 May 2016.

57. Langley, Ann (April 15, 1995). "Between "Paralysis by Analysis" and "Extinction by Instinct"". MIT Sloan Management Review. **36** (3).

58. Smallwood, R.F. (2014). "Chapter 10: Information Governance and Information Technology Functions". Information Governance: Concepts, Strategies, and Best Practices. John Wiley & Sons, Inc. pp. 189–206. ISBN 9781118421017. Retrieved 23 June 2016.

59. Toomey, M. (20 November 2008). "A Significant Achievement" (PDF). The Informatics Letter. Infonomics Pty Ltd. Retrieved 23 June 2016.

60. Juiz, C.; Toomey, M. (2015). "To Govern IT, or Not to Govern IT?". Communications of the ACM. **58** (2): 58–64. doi:10.1145/2656385.

61. McKay, A. (2007). "Australia leads the world on ICT governance" (PDF). Up. **8** (Summer 2007): 3. Retrieved 23 June 2016.

62. Feltus, C. (21 July 2010). "ISO/IEC 29382 - The new standard for ICT governance". SlideShare. LinkedIn Corporation. pp. 8–10. Retrieved 23 June 2016.

63. "ISO/IEC DIS 29382: 2007 Edition, February 1, 2007". IHS Standards Store. IHS, Inc. Archived from the original on 23 June 2016. Retrieved 23 June 2016.

64. Jones, B. (29 January 2007). "Explanation of the ISO "Fast-Track" process". Microsoft Developer Network Blog. Microsoft. Retrieved 23 June 2016.

65. "JTC1/SC7 List of Documents: N3851 - N3900". ISO/IEC. 18 January 2008. Archived from the original on 23 June 2016. Retrieved 23 June 2016.

66. "IT Governance and The International Standard, ISO/IEC 38500". IT Governance. IT Governance Ltd. Retrieved 23 June 2016.

67. "ISO 38500 IT Governance Standard". 38500.org. 2008. Retrieved 23 June 2016.

68. Garcia-Menendez, M. (1 June 2009). "ISO/IEC 38500:2008. Un año difundiendo el concepto de 'Buen Gobierno Corporativo de las TIC'". Gobernanza de TI. Retrieved 23 June 2016.

69. "ISO/IEC 38500:2008". ISO. Retrieved 23 June 2016.

70. "2015 Edition of ISO/IEC 38500 Published" (PDF). Standards Australia. 23 March 2015. Retrieved 23 June 2016.

71. Haes, S.D.; Grembergen, W.V. (2015). "Chapter 5: COBIT as a Framework for Enterprise Governance of IT". Enterprise Governance of Information Technology: Achieving Alignment and Value, Featuring COBIT 5 (2nd ed.). Springer. pp. 103–128. ISBN 9783319145471. Retrieved 24 June 2016.

72. Stroud, R.E. (2012). "Introduction to COBIT 5" (PDF). ISACA. Retrieved 24 June 2016.

73. da Cruz, M. (2006). "10: AS 8015-2005 - Australian Standard for Corporate Governance of ICT". In van Bon, J.; Verheijen, T. Frameworks for IT Management. Van Haren Publishing. pp. 95–102. ISBN 9789077212905. Retrieved 23 June 2016.

74. "ISO/IEC DIS 29382: 2007 Edition, February 1, 2007". IHS Standards Store. IHS, Inc. Archived from the original on 23 June 2016. Retrieved 23 June 2016.

75. "ISACA Releases COBIT 5: Updated Framework for the Governance and Management of IT" (PDF). Provitivi, Inc. 18 May 2012. Retrieved 23 Jan 2017.

76. "COBIT 5 for Information Security". ISACA. Retrieved 24 June 2016.

77. "COBIT 5 for Assurance". ISACA. Retrieved 24 June 2016.

78. Katsikas, S.; Gritzalis, D., eds. (1996). Information Systems Security: Facing the Information Society of the 21st Century. IFIP Advances in Information and Communication Technology. Springer. p. 358. ISBN 9780412781209. The McCumber model has great similarities with the CobiT - Control Objectives for IT - framework (CobiT 1995).

79. "Welcome to the ISACA/F". ISACA. 18 October 1996. Archived from the original on 7 November 1996. Retrieved 24 June 2016.

80. Luellig, L.; Frazier, J. (2013). "A COBIT Approach to Regulatory Compliance and Defensible Disposal". ISACA Journal. **5**. Retrieved 24 June 2016.

81. Weill, P. & Ross, J. W., 2004, IT Governance: How Top Performers Manage IT Decision Rights for Superior Results", Harvard Business School Press, Boston.

82. Blitstein, Ron, 2012. "IT Governance: Bureaucratic Logjam or Business Enabler", Cutter Consortium.

83. Brown, Allen E. and Grant, Gerald G. (2005) "Framing the Frameworks: A Review of IT Governance Research," Communications of the Association for Information Systems: Vol. 15, Article 38.

84. S. De Haes, and W. Van Grembergen, "Exploring the relationship between IT governance practices and business/IT alignment through extreme case analysis in Belgian mid-to-large size financial enterprises", Journal of Enterprise Information Management, Vol. 22, No. 5, 2009, pp. 615–637.

85. Georgel F., IT Gouvernance : Maitrise d'un systeme d'information, Dunod, 2004(Ed1) 2006(Ed2), 2009(Ed3), ISBN 2-10-052574-3. "Gouvernance, audit et securite des TI", CCH, 2008(Ed1) ISBN 978-2-89366-577-1

86. Lutchen, M. (2004). Managing IT as a business : a survival guide for CEOs. Hoboken, N.J., J. Wiley., ISBN 0-471-47104-6

87. Renz, Patrick S. (2007). "Project Governance." Heidelberg, Physica-Verl. (Contributions to Economics) ISBN 978-3-7908-1926-7

88. Van Grembergen, W., Strategies for Information technology Governance, IDEA Group Publishing, 2004, ISBN 1-59140-284-0

89. Van Grembergen, W., and S. De Haes, Enterprise Governance of IT: Achieving Strategic Alignment and Value, Springer, 2009.

90. Wim Van Grembergen, and S. De Haes, "A Research Journey into Enterprise Governance of IT, Business/IT Alignment and Value Creation", International Journal of IT/Business Alignment and Governance, Vol. No. 1, 2010, pp. 1–13.

91. Weill, P. and Ross, J.W. (2004). IT Governance: How Top Performers Manage IT Decision Rights for Superior Results, Boston, MA, Harvard Business School Publishing,ISBN 1-59139-253-5

92. Wilkin, C.L. and Chenhall, R.H. (2010). A Review of IT Governance: A Taxonomy to Inform AIS, Journal of Information Systems, 24 (2), 107–146.

93. Wood, David J., 2011. "Assessing IT Governance Maturity: The Case of San Marcos, Texas". Applied Research Projects, Texas State University-San Marcos. (This paper applies a modified COBIT framework to a medium sized city.) "Richard Saul Wurman awarded for Lifetime Achievement". Smithsonian Cooper-Hewitt, National Design Museum. Retrieved 19 April 2014.

94. "Join the IA Network". Information Architecture Institute.

95. Morville & Rosenfeld 2007.

96. Morville & Rosenfeld (2007). p. 4. "The art and science of shaping information products and experienced to support usability and findability."

97. Resmini, A. & Rosati, L. (2012). A Brief History of Information Architecture. Journal of Information Architecture. Vol. 3, No. 2. [Available at http://journalofia.org/volume3/issue2/03-resmini/]. Originally published in Resmini, A. & Rosati L. (2011). Pervasive Information Architecture. Morgan Kauffman. (Edited by the authors).

98. Toms, Elaine (17 May 2012). "Information interaction: Providing a framework for information architecture". Journal of the American Society for Information Science and Technology. **53**(10.1002/asi.10094).

99. "Information Architecture". Mozilla Developer Network.

100. Dillon, A (2002). "Information Architecture in JASIST: Just where did we come from?". Journal of the American Society for Information Science and Technology. **53** (10): 821–23.doi:10.1002/asi.10090.

101. Wurman, "Introduction", in: Information Architects (1997). p. 16.

102. "What is Master Data" SearchDataManagement, TechTarget, 22 November 2010, http://searchdatamanagement.techtarget.com/definition/master-data-management

103. "Introduction to Master Data Management", Mark Rittman, Director, Rittman Mead Consulting, 9 May ""Defining Master Data", David Loshin, BeyeNetwork, May 2006

104. "Master data management". IBM.

105. DAMA-DMBOK Guide,2010 DAMA International

106. "Creating the Golden Record: Better Data Through Chemistry", DAMA, slide 26, Donald J. Soulsby, 22 October 2009