

A RELEVANT, BELIEVABLE APPROACH FOR DATA QUALITY ASSESSMENT

(Research-In-Progress)

G. Shankar

Boston University School of Management
gshankar@bu.edu

Stephanie Watts

Boston University School of Management
swatts@bu.edu

Abstract This research proposes the use of an IPMAP for providing metadata about the source and processing of primary data, in order to enhance its believability and fitness-for-use. We then derive a framework for assessing data quality-in-use from dual-process theories of human cognition. By applying a dual-process approach to data quality assessment, the model enables simultaneous evaluation of both objective and contextual data quality attributes. In addition to assessing the role of metadata for enhancing believability, we use our framework to investigate the data attributes of relevance, completeness, accuracy and timeliness. The model is the first to offer a theoretical explanation for the role of metadata in enhancing data quality.

Keywords: Data Believability, Data Relevance, Information Product Map, Information Validity Assessment, Dual-Process Theory

1. INTRODUCTION

Even the highest quality data must be used to be useful, hence the definition of data quality as “fitness for use” [15]. From this perspective, data that is of acceptable quality for one decision context (or use) may be considered to be of poor quality for another decision context – even by the same individual – when the data is applied to two different decision contexts. For example, course enrollment data obtained a few months before the start of a new semester indicating the number of students registered for a course may be accurate enough for ordering textbooks but not sufficiently accurate to decide the room (capacity) for class meetings – both decisions being made by the instructor of that course.

This notion, that some attributes of data quality are invariant while others vary depending on the context of use, makes the measurement of data quality problematic. Further, data quality is evaluated not as a whole, but along quality dimensions such as accuracy, timeliness, completeness, relevance, and believability, just to name a few. Research in data quality has described methods to *objectively* measure data quality attributes that lend themselves to this type of measurement, such as accuracy, timeliness, and completeness [1][14]. However, the quality of other data attributes often depends on how that data is used. In particular, data relevance lends itself to contextual evaluation, since data that is highly relevant for one task may be irrelevant for another task. Data believability, too, is quite difficult to evaluate objectively, since data that is believable to a novice may be unbelievable to an expert. To the extent that relevance and believability are difficult attributes to evaluate objectively, research in data quality has not

addressed these attributes in depth. This research seeks to address this oversight, in order to improve design of information systems for delivering high quality data along all data quality attributes – relevance and believability included.

To this end, we adopt the use of the Information Product map (IPMAP) – a form of metadata that augments the actual data used for making particular decisions. The IPMAP corresponding to an information product (IP) is an explicit, graphical representation of the manufacturing processes for that IP [13]. The IP-approach views information as a product that is manufactured from data elements obtained (input) from data sources and/or created by one/more associated manufacturing processes. These data elements are then assembled (another process) together to create the final IP. Any report (or output) that a decision-maker uses can be viewed as an IP. The IPMAP allows the decision-maker to visualize the source and flow of data elements, and the sequence by which these data elements have been combined, during construction of the information product being used for decision-making [13]. We believe that the provision of this metadata can significantly enhance the believability of the information product itself, just as higher levels of source credibility can enhance information adoption [16]

In order to evaluate the effect of the IPMAP on data believability, we utilize a theory-driven model of dual-process information validity assessment [5][16]. This theoretical framework enables us to integrate evaluation of the objective data quality attributes of accuracy, completeness, and timeliness with the contextual data quality attribute of believability. The model also reserves a central, mediating role for the key data quality attribute of relevance. Because it encompasses both objective and contextual approaches to validity assessment, it is ideally suited to the study of data quality *in use*.

Below we discuss the IPMAP concept in detail. Following this, we summarize the theory underlying dual-process approaches to information validity assessment, and use it to generate our hypotheses. We then present the methods we are using in this research-in-progress to apply our theoretical framework to the problem of assessing data quality *in use*. We then discuss the implications of this research for data quality assessment and the potential impact on information systems design.

2. THEORETICAL BACKGROUND

2.1 Data Believability and the IPMAP

Although clearly useful, conventional approaches to data quality management such as data cleansing [7], data tracking and statistical process control [12], data source calculus and algebra [10], data stewardship [6], and dimensional gap analysis [8] do not provide a systematic approach for managing data quality *in-use*. In this paper, an alternative approach based on the notion of an information product (IP) is developed for improving the believability of data quality in dynamic decision environments. The IP approach has gained considerable acceptance in organizations for several reasons. First, manufacturing an IP is akin to manufacturing a physical product. Raw materials, storage, assembly, processing, inspection, rework, and packaging (formatting) are all applicable concepts. Typical IPs (such as management reports, invoices, etc.) are “standard products” and hence can be “assembled” in a production line. Components and /or processes of an IP may be outsourced to an external agency (ASP), organization, or a different business-unit that uses a different set of computing resources. Second, IPs, like physical products, can be “grouped” based on similar characteristics and common data inputs, permitting the “group” to be managed as a whole. In other words, multiple IPs may share a subset of processes and data inputs, and may be created using a single “production line” with minor variations that distinguish each IP. Finally, proven methods for TQM (such as quality at the source and continuous improvement) that have been successfully applied in manufacturing can be adapted for total data quality management. To exploit these

properties of IPs and to manage data quality using the IP-approach, mechanisms for systematically representing the manufacturing stages, and for evaluating data quality at each stage are essential. For example, in order to apply total data quality concepts to the problem of poor-quality data, it is necessary to evaluate the impact of delays in one or more manufacturing stages, trace the quality-problem in an IP to the manufacturing stage(s) that may have caused it, and predict the IP(s) impacted by quality issues identified at the manufacturing step(s). The IP approach facilitates a comprehensive, intuitive, and visual representation of the manufacture of an IP. The IPMAP allows the decision-maker to visualize not only the widespread distribution of data and other resources but also the flow of data elements and the sequence by which these data elements are processed to create the required IPs. Combined with the capabilities for total data quality management, the IPMAP enables decision-makers to understand the sources, processes, systems, business units, and organizations involved in the creation of the IP.

The IPMAP provides users with detailed information on the manufacture of the information product. This is accomplished using a representation scheme that creates a systematic representation for capturing the details associated with the manufacture of an IP. The objectives of this representation are:

- To provide a set of constructs that facilitate representation of the steps involved in the manufacture of an IP. These constructs help model the various steps of the production process and assist the modeler in visualizing and representing these steps. The representation hence serves as a conceptual model for the manufacture of the IP.
- To allow the modeler to critically examine the steps in the manufacturing process. These steps include the arrival of the data elements (raw data), the locations for storing these data elements, the processes involved in creating, converting, and/or assembling the existing (or new) data elements, and the procedures for evaluating the IP and the work-in-progress for quality and correctness. Hence, the modeler/user can locate potential sources of information quality problems and more importantly, design procedures to rectify these problems thus ensuring a high quality IP.
- Further, the IPMAP allows the modeler to implement information quality-at-source (similar to quality-at-source in manufacturing). It permits the modeler to assign the responsibility for ensuring the quality of the IP - including the work-in-progress that comes into and leaves a “work area” - to individuals performing one or more steps in the manufacturing process within that “work area”.
- The IPMAP provides a formal representation that can be used to assess the quality of an IP based on selected information quality dimensions.
- It adopts a "top-down" approach, since the design requirements of the final product drive the design of the IPMAP. The IP designers and developers are required to precisely specify the raw or component data items that are needed to produce a particular IP. This specification provides the IP manager with the ability to look ahead and determine the feasibility of creating the IP. It is possible that some raw data items may not be available or cannot be produced in the current situation.
- Several extensions to the constructs (blocks) introduced by Ballou et al. [1] have been defined. These constructs facilitate the explicit representation of details in the manufacture of the IP, and include the *decision* block, the *organizational/business boundary* block, and the *information system boundary* block.
- A repository for capturing the metadata (described in more detail below) associated with the constructs in the IPMAP is also defined. The metadata adds to the ability of the IPMAP to comprehensively track and manage the information associated with the IP and serves to resolve issues concerning the quality of the IP.

2.2 The Composition of an IPMAP

Prior to explaining the constructs and modeling procedures, let us first examine the composition of an IP and distinguish it from a physical product, both in terms of its composition and its manufacturing process. Unlike a physical product where the *overall product and its quality* are of interest to the consumer, for an

information product it is the *data items that comprise the IP* and the *quality of each* that are of importance to the consumer. The IP must therefore be identified in terms of the data items used to manufacture it and must be specified by the final consumer of the IP. This breakdown drives the requirements of the raw information - *raw data (RD) items*- and the semi-processed information -*component data items (CD)*- needed to manufacture the IP. Raw data is defined as data (or information) that comes from outside the boundaries of the IPMAP. Component data is defined as data that might be generated within the IPMAP and used in creating the final IP.

Another distinct feature of IP manufacture compared to the manufacture of a physical product is that the raw data items and the component data items do not deplete when an IP is manufactured. While the raw data items may be stored for long periods of time, the component data items are stored temporarily until the final product is manufactured. The component items are regenerated each time an IP is needed. The same set of raw data and component data items may be used in the manufacture of several different products. The set of information products that uses the same set of raw and component data items may be considered a family (or group).

An IP is created using data that is constantly being captured and stored. For example, the data on the inventory levels of products in a warehouse, order details for pending retail orders, pricing data, shipping and transportation charges, etc. (raw materials) are being captured in anticipation of the fact that the procurement manager of this organization may need the inventory information to make procurement decisions. It is possible that this report may never be requested (i.e. the same information obtained from elsewhere) and hence may never be created. The same raw data (material) could also be used for manufacturing a different IP such as an inventory volume (in units and \$\$) changes over a three-year period. Also, the raw material is never depleted even after multiple information products have been manufactured using this raw material. It is therefore imperative that a good representation accurately tracks the details of *what* triggered the capture of this data along with *how*, *who*, and *where* it was captured.

To address the “what” question we define events (or triggering events) that initiate data capture. Associated with an event are entities. For example, the arrival of a new shipment at a warehouse is an event. The dock-clerk who oversees the unloading and records the details is a data creator who creates the shipment data. The data creator may capture this on wireless tablet and beams it over to the server with the inventory management application. A module in the system may then validate this data and then update the inventory levels. This system is the data custodian. The users of this data (e.g. procurement manager) are the data consumers. The data creator, data collector, and the data consumer are all entities that answer the "who" question. In general, there are three kinds of roles under the data supplier category: data producer or creator, data collector, and data consumer. Often, however, a single person or system may assume several or all of these roles.

For an IP, the “where” question has two implications: the physical location (specific warehouse unloading dock, head office etc.) and the system used to capture the information about the product (paper form, computer system using a word processor, computer system with a specific application data entry interface, etc.). Both are important when tracking the IP as they serve to identify the "source" of the data and more importantly, the factors that (may) affect the quality of the data associated with each source. The proposed model is designed to capture this data (metadata) about the source explicitly.

Once the raw data items are received from their data sources they go through a sequence of steps that result in the IP. The steps include storage, quality evaluation, and processing, before ending up at the destination (or sink) of the IPMAP. Some of these steps result in the creation of component data items. To comprehensively capture and represent these steps and the information associated with each, the model supports constructs (blocks) and a repository for capturing the metadata associated with each block. The

five constructs supported by Ballou et al. [1] that are retained in the IPMAP are the source, sink, process, quality, and storage blocks. The IPMAP extends these constructs with three new constructs; the decision, the information system boundary, and the departmental/organizational boundary blocks. As the IPMAP attempts to capture more details and is more explicit in its representation, it provides a detailed description of the seven blocks along with the metadata associated with each. The constructs and the symbols used in to represent them are listed in table 1. Each construct is supplemented with metadata about the manufacturing stage that it represents. The metadata includes (1) a unique identifier (name or a number) for each stage, (2) the composition of the data unit when it exits the stage, (3) the role and business unit responsible for that stage, (4) individual(s) that may assume this role, (5) the processing steps to complete that manufacturing step, (6) the business rules/constraints associated with it, (7) a description of the technology used at this stage, (8) and the physical location where the step is performed. These help the decision-maker understand *what* is the output from this step, *how* was this achieved including business rules and constraints applicable, *where* (both physical location and the system used), and *who* is responsible for this stage in the manufacture.

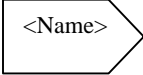
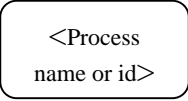
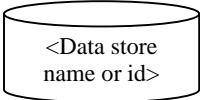
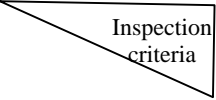
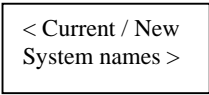
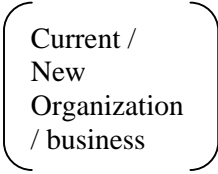
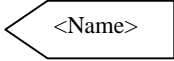
Construct	Description and Metadata associated
	Data Source Block: used to represent the source of each raw (input) data that must be available in order to produce the IP expected by the consumer.
	Processing Block: used to represent any manipulations, calculations, or combinations involving some or all of the raw input data items or component data items required to ultimately produce the IP. We allow for the specification of the processing requirements to be associated with the block.
	Data Storage Block: Storage blocks may be used to represent data items (raw and/or component) that wait for further processing or are captured as part of the information inventory in the organization.
	Inspection Block: While the quality of the data can be evaluated at all points on the IPMAP, the inspection block serves to represent specific pre-determined inspections (validity checks, checks for missing values etc., authorizations, approvals etc.). Even though it may be viewed as a process, we use this block to differentiate a transformation/transport from the inspection/validation process.
	Information System Boundary - used when a data unit (raw / component data) changes from one system (e.g., paper or computerized) to another (e.g., paper or computerized). This block is used to reflect the changes to the raw input (or component) data items as they move from one information system to another type of information system. These system changes could be intra or inter-business units. The information system boundary block is used to specify the two information systems involved.
	Business/Organizational Boundary: used to represent instances where the raw input (or component) data items are “handed over” by one business (or organizational) unit to an other unit. It is used to specify the movement of the IP (or raw / component data) across departmental or organizational boundaries. The role of this block is to highlight the data quality problems that might arise when crossing business unit boundaries and therefore assign accountability to the appropriate business unit.
	Data Sink (Consumer) Block: used by the consumer to specify the data elements that constitute the “finished” IP. Associated with this block are the name of the business / organizational / departmental unit in charge of the IP, the name of the entity that will actually use the information product, and the set of data items that make up the IP.

Table 1: IPMAP Constructs (adopted from [13])

The IPMAP allows us to track an information product as it goes through the various stages of an information manufacturing system. A single IP manufacturing system should be capable of manufacturing

all the variants of a product P. In some respects the idea is similar to that of a manufacturing cell created based on group technology. As the raw data requirements for the system that manufactures product P includes the raw data needs of all its variants, it is conceivable that the initial processing for P and its variants may be the same.

We illustrate with a simple example, how the IPMAP serves as a visual representation on the manufacture of an IP. Consider the case of an inventory status report that might be used by a bookstore that acquires books from publishers and sells these to consumers. The IPMAP for this IP is shown in figure 1. For simplicity, we assume that the bookstore has two warehouses, one of which is local. Current inventory levels of books (raw data elements RD¹ and RD²) are obtained from the two warehouses (data sources DS¹ and DS²). In the case of the local warehouse, this data (RD²) is processed (collated by process P³), transferred across to the head-office (spanning business units BB²), extracted by a process (P⁴), inspected (I) and stored in a database (S¹). In the case of the other warehouse, the raw data element (RD¹) is first collected (manually) and faxed over to the head office (spanning business unit boundary BB¹). Here is entered into a computer system (process P²) and hence changes system (paper to electronic) boundaries (shown by SB¹). The data is then inspected and stored in the same database. Similarly, the data on pending and in-process retail orders (RD³) and the data on prices (RD⁴) are obtained from the sales department and from within the procurement department respectively. These are processed (P⁵ and P⁶) and stored in a different data repository (S²). The inventory status report, the final IP, is generated by process P⁷ and sent to the consumer of the IP, the procurement manager. The IPMAP is presented using a GUI and the metadata associated with each construct can be viewed by a right-mouse-click on that construct in the GUI. In figure 1, the symbol “CD” represents a component data created by processing one or more raw or component data items.

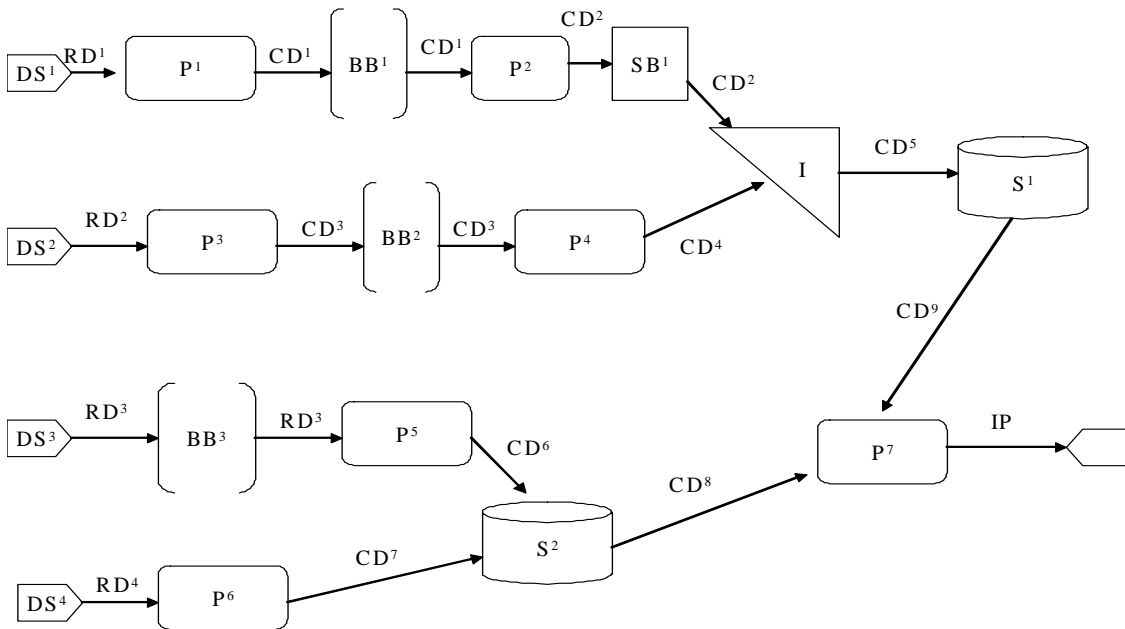


Figure 1: IPMAP for the Inventory Status Report Example

We are convinced of the merits of the IP concept and the IPMAP for augmenting data quality *in-use*. We now describe a theoretically-derived framework for evaluating the contribution that can be made by an IPMAP. To the extent that the IPMAP is meta-data, its role is one of augmentation of primary data quality. Since the IPMAP provides information on the sources and paths of primary data, it can greatly contribute to the believability of primary data. Thus the theoretical framework presented below integrates

the assessment of believability – to be operationalized using the IPMAP – with the more objective data quality attributes of completeness, accuracy and timeliness. In this model also, the contextual data quality attribute of relevance plays a primary role in aggregating assessments of the other attributes.

2.3 Dual-Process Approaches to Information Validity Assessment

In order to understand how people assess contextual data quality attributes such as believability and relevance in conjunction with more objective data attributes, we look to cognitive theories of how individuals process received information. Dual-process theories of information processing originated from individual-level, laboratory-based research in cognition and social-psychology [11][5][3]. Over the past 20 years they have been applied to many domains as a way of understanding how people process received information, including IS research [4][9][2][16]. Dual-process approaches to information processing encompass a family of theories, all of which examine both the information content of received information, and factors in the surrounding context. Under the dual-process perspective, people assess received information – in this case data quality – in two ways. First, decision-makers assess content using systematic processing, analyzing the data for its inherent merits. Second, people use many additional cues or heuristics to assess data quality (for example, the believability of the data source), in a process labeled heuristic processing. Systematic processing requires more cognitive workload than the relatively easier route of heuristic processing. Assessment and consequent utilization of particular data takes place via either or both of these cognitive processes depending on how *motivated or able* the decision maker is to undertake the cognitive effort of systematic processing. Under certain conditions, a tradeoff between the two types of cognitive processing can take place, such that a greater reliance on heuristics corresponds to lower levels of systematic processing. At other times, the two cognitive processes can bias one another. This theoretical orientation has been used to understand individual’s information adoption in a variety of computer-mediated domains [16]), but has not been applied to the assessment of data quality.

Since decision-makers adopt information via both systematic and heuristic processes, neither process is inherently superior to the other. It is the best-fit of cognitive processes to decision tasks that suggest strategies for optimizing data quality for decision making. Based on this theoretical framework, we have developed the following model (shown in figure 2) for understanding both objective and perceptual aspects of data quality assessment.

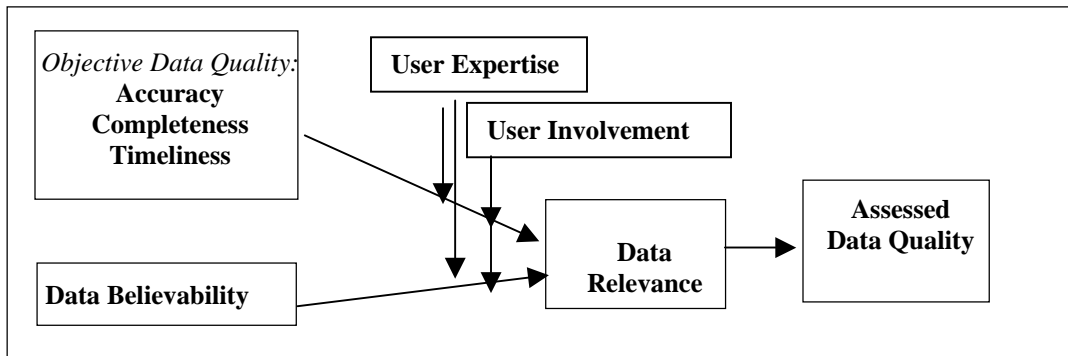


Figure 2: Theoretical Model of Data Quality Assessment

In this model, users process received data both systematically and heuristically. While either or both types of processing can influence data quality assessments, the way that the process plays out depends on moderator variables that affect availability of cognitive resources. When a lot of cognitive resources are deployed on the problem, systematic processing will tend to dominate, and when fewer cognitive resources are available for data quality assessments, heuristic processing will prevail. Higher levels of

topic-related user expertise and involvement increase available cognitive resources, so the presence of these factors will tend to induce systematic processing and hence increase the role of objective data quality attributes in determining both relevance and quality. For novices and uninvolved users, we would expect heuristic processing to dominate assessments of relevance and quality. The *objective data quality* using quality attributes (or dimensions) is *computed* objectively (i.e. independent of the decision-task/decision-maker) and is not measured. The *assessed data quality* is measured by how the users evaluate the quality based on the decision-task and is impacted by the factors described earlier. Note that this model is an adaptation of Watts-Sussman and Siegal's [16] framework for assessing the information quality of received emails. In their research, information usefulness was found to play a central, mediating role in information assessment. Based on these author's findings, we position data relevance here as mediating between the other data attributes and data quality assessment. Thus factors describing the data itself inform how relevant it is for the task at hand, and irrelevant data is not assessed as high quality regardless of how accurate, complete and timely it is. In this way our theoretical framework of data quality assessment reflects *quality in-use*.

In assessment of information quality, source credibility is an important and widely studied heuristic [5]. However, source credibility generally reflects the influence of a single author on text-based information. And since delivered data often reflects the influence of multiple source databases, processing algorithms, aggregation and distillation, etc., source credibility of a "single-author" is not an important heuristic in the data quality context. However, the concept of data *believability* is analogous to the source credibility heuristic in the context of data quality assessment, since it reflects concern for how the data was constructed and the path it has traveled to reach the decision-maker. For this reason, we operationalize data believability here using the IPMAP, since it is theoretically consistent with the source credibility heuristic, reflecting how the data was constructed and the path it took to reaching the decision-maker.

Based on the theoretical model above, perceptions of high data relevance will be associated with high data quality assessment, and data relevance will mediate the relationship between data quality assessment and the other objective and perceptual data attributes, as follows:

Hypothesis 1a: High levels of perceived data relevance will be associated with high data quality assessment.

Hypothesis 1b: Perceived data relevance will mediate the relationship between objective data quality attributes and data quality assessment.

Hypothesis 1c: Perceived data relevance will mediate the relationship between perceived data believability and data quality assessment.

When the user is an expert in the topic of the received data, he or she is more *able* to undertake systematic processing, given available cognitive resources, relative to the cognitive abilities of a novice in the topic. Thus for experts, systematic processing will dominate and objective data attributes will be more influential on assessment of data relevance than will be perceptual attributes. Since examination of the IPMAP will tend to take greater cognitive resources than simply reading reported levels of data completeness, accuracy and timeliness, in this operationalization the contextual attribute of believability will require systematic processing, while the objective attributes will serve as heuristics, thus:

Hypothesis 2a: When the user is an expert in the data topic, data believability will be more strongly associated with perceived data completeness, accuracy and timeliness than perceived data believability will be.

Hypothesis 2b: When the user is a novice in the data topic, perceived data accuracy, completeness and timeliness will be more strongly associated with perceived data relevance than data believability will be.

Hypothesis 2 reflects the fact that this theoretical framework allows us to take characteristics of the data user into account. We are also able to use it to take into account factors relating to the nature of the task. Tasks that *motivate* high levels of involvement on the part of the user tend to induce systematic processing rather than heuristic processing, as the following hypothesis reflects:

Hypothesis 3a: When the task is a highly involved one, data believability will be more strongly associated with perceived data relevance than perceived data accuracy, completeness and timeliness will be.

Hypothesis 3b: When the task is not a highly involving one, perceived data accuracy, completeness and timeliness will be more strongly associated with perceived data relevance than data believability will be.

3. METHOD

This research in progress will measure the impact of data quality both objectively and perceptually. In an experimental setting, users (decision-makers) are asked to make a recommendation on a stock given the performance report of the same stock over a five-year period. The issue here is whether the users consider (or assess) the overall quality of the performance report (the information product) to be sufficient to make a recommendation. The recommendation (to buy or not to buy) in itself is not the focal point. The objective of this research is to identify the impact (if any) of the following important factors on the overall assessed quality of the information product: an objective assessment along the quality dimensions (accuracy, timeliness, and completeness) of the quality of an information product (the performance report), the users' perception of the believability and relevance of this information product, individual expertise in the domain of the task, and the extent of involvement in the task.

Users are provided with objective data about the accuracy, completeness and timeliness of the information product that is used for the decision task. Methods for objectively evaluating accuracy and timeliness have been prescribed in [1]. Objective evaluation for completeness has been described in [14]. We will adopt these methods to define objective evaluations of accuracy, completeness, and timeliness for the information product used in the empirical evaluation. By weighting accuracy, timeliness, and completeness equally, we can derive the overall quality assessment (based on these three quality dimensions) of the information product using the method prescribed in the literature. Using the values derived, we will classify the overall quality of the performance report as being high or low.

The IPVIEW is a graphical modeling tool that is used in this experimental setup. The IPVIEW supports two functionalities. First it serves as a drawing tool to create/edit an IPMAP and to capture the metadata elements associated with this IPMAP. Second, it serves as a visualization tool to communicate the metadata about the manufacture of an IP. The IPVIEW consists of a canvas or drawing area on which users can create a new IPMAP or view/modify an existing IPMAP. The constructs for creating the IPMAP are available as icons on a tool-bar. To define a construct when creating an IPMAP, users can drag-and-drop the corresponding icon from the tool-bar onto the drawing area. The flows between the constructs can be defined in a similar fashion. Upon creating each new construct, users are prompted using a pop-up text-entry interface to capture the metadata corresponding to the block that this construct represents. This metadata is stored in a back-end SQL Server database. The visual components (tool-bar, canvas along with the drawing capabilities) of IPVIEW are implemented using Java Swing API and the JGraph library. Complete implementation details are omitted here for brevity and relevance.

Users can visually examine the metadata associated with each report (IP) using the IPVIEW. The GUI permits users to view the entire IPMAP corresponding to an IP that is of interest. All of the metadata

(described earlier in the section titled “Composition of an IPMAP”) at *each stage/block* in this IPMAP can be examined by a right-click on a specific IPMAP-block displayed on the drawing canvas. Further, values associated with data quality dimensions (accuracy, timeliness, and completeness are currently supported) at each stage of the IPMAP can also be viewed. The computed value corresponding to any of these quality dimensions is displayed at each stage of manufacture and for the final IP. The users have the ability to evaluate the overall quality of the IP (a combination of the dimensions available). Further more, if a quality problem (a low value along one/more dimensions, or a suspiciously high value) is identified at some stage (say A) in the IPMAP, the IPVIEW also offers the ability to trace back and pin-point the stages that may have caused this and offers the ability to look ahead and target the stages and IP(s) that are likely to be affected by this problem stage. This visualization tool will be used to communicate the metadata to the users in the experimental setup to test the impact of data believability upon the perceived overall data quality of the report (IP). Using the metadata and the quality evaluations, users can decide between alternate data sources, alternate manufacturing steps, and examine other “what if” scenarios. Based on these examinations users can decide on appropriate data sources, manufacturing stages, and even between alternate (assuming that these satisfy the basic needs) IPs.

4. RESEARCH IMPLICATIONS

Assessment of data quality *in use* is problematic because data that is relevant and believable to one person may be irrelevant and unbelievable to the next, even when the data is objectively complete, accurate, and timely. This research aims to empirically tease apart the relationships between objective and contextual data quality attributes for the purpose of enhancing data quality *in use*. This work is particularly challenging when we consider that data quality attributes that are easily measured objectively – accuracy for example – may still be judged to inaccurate by an uninformed user, and hence deemed not fit-for-use. This example highlights why it is time that data quality research acknowledges the importance of both objective *and* contextual aspects of data quality. For designing high quality data delivery systems, fitness-for-use begs research, and the dual-process framework here offers a theoretically-based means of understanding it.

We are not suggesting that we should abandon our quest for high levels of data quality along traditional, objective data quality attributes. On the contrary, this research highlights the importance of integrating contextual and objective data quality assessment processes. The contribution of the theoretical framework presented here is that it enables just such integration, simultaneously examining both contextual and objective data quality attributes, reflecting how decision-makers actually think about received information. The dual process theories take their name from the fact that they embody both contextualized, heuristic-based thinking, and analytical cognitive processes. Further, they offer clear mechanisms for understanding when and why one type of processing is more likely to predominate in particular contexts. By using this referent theory, we can account for variations across decision makers and decision contexts.

Several different data quality dimensions have been proposed in the data quality literature. Research has attempted to evaluate data quality along many of these dimensions, however these evaluation methods have utilized only objective, de-contextualized measures. Such objective evaluations are necessary and useful for establishing a consistent and unbiased view of data quality. An important issue to understand is how useful such objective evaluations of data quality are in the context of a decision-task. If useful, we would expect them to have a significant impact on the decision-maker’s assessment (perception) of overall data quality. If not, then it may be that other factors, contextual and/or objective, may impact the usefulness of objective evaluations and thereby affect overall perceptions of data quality. Intuition and experience as decision-makers leads us to believe that contextual factors do have a role in the assessment

of data quality, and hence are important to investigate. This research takes a first step towards understanding the contextual dimensions of data quality (data believability and data relevance) and how they interact with objective data quality to impact perceived data quality assessments.

The Internet, combined with mobile technologies and wireless devices, enables decision-makers to access data that spans departmental and organizational boundaries. In such decision environments, decision-makers often have little or no control over the sources of data and the manner in which data is managed and maintained. Decision-makers are well aware of this. Data believability has become a more important factor than ever before for today's decision-makers, as evidenced by the widespread adoption of document reputation systems. This research examines data believability and its impacts as a quality dimension, providing a theoretical foundation for better understanding the implications of data believability.

5. CONCLUSIONS

Decision-making is significantly affected by the quality of the data used in the decision-task. Today's decision environments are characterized by widely distributed data sources that span organizational boundaries. It is necessary to both inform the decision-makers of the quality of the data that they use and also to permit them to gauge data quality on their own as it relates to their decision task. Contextual factors such as data believability and relevance are very important for gauging data quality in such environments. This research proposes the concept of an IPMAP to communicate metadata about the source and processing of primary data in order to enhance its believability and fitness-for-use. It then derives a framework for assessing data quality-in-use from dual-process theories of human cognition. By applying a dual-process approach to data quality assessment, the model enables simultaneous evaluation of both objective and contextual data quality attributes. In addition to believability, other key data attributes measured in this research are relevance, completeness, accuracy and timeliness. The research model is the first to offer a theoretical explanation for the role of metadata in enhancing data quality.

REFERENCES

1. Ballou, D., Wang, R. Y., Pazer, H., and Tayi, G. K. (1998) *Modeling Information Manufacturing Systems to Determine Information Product Quality*, Management Science, 44 (4), 462-484
2. Briggs, P., Burford, B., De Angeli, A., & Lynch, P. (2002). *Trust in online advice*. Social Science Computer Review, 20(3), 321-332.
3. Chaiken, S., Liberman, A. and Eagly, A.H. "Heuristic and systematic information processing within and beyond the persuasion context," In *Unintended thought.*, J. S. Uleman and J. A. Bargh (Ed.), Guilford Press, New York, NY, USA, 1989, pp. 212-252
4. *Dijkstra, J. J. (1999). User agreement with incorrect expert system advice, Behavior & Information Technology, 18(6), 399-411*
5. Eagly, A.H. and Chaiken, S. *The psychology of attitudes*, Hartcourt Brace College Publishers, New York, 1993.
6. English, L. P. *Improving Data Warehouse and Business Information Quality – Methods for Reducing Costs and Increasing Profits*, Joh Wiley and Sons, New York, NY, 1999.

7. Hernandez, M. A. and Stolfo, S. J. (1998) *Real World Data is Dirty: Data Cleansing and the Merge/Purge Problem*, Journal of Data Mining and Knowledge Discovery, 2(1), 9-37
8. Kahn, B. K., Strong, D. M. and Wang, R. Y. (2002), *Data quality Benchmarks: Product and Service Performance*, Communications of ACM, 45 (4), 184-193
9. Mak, B., Schmitt, B. H., & Lyytinen, K. (1997). User participation in knowledge update of expert systems, *Information & Management*, 32(2), 55-63
10. Parsanian, A., Sarkar, S., and Jacobs, V. S. (1999) *Assessing Data Quality For Information Products*, Proceedings of the International Conference on Information Systems, Charlotte, NC, 1999
11. Petty, R.E. and Cacioppo, J.T. *Communication and persuasion: Central and peripheral routes to attitude change*, Springer-Verlag, New York, 1986.
12. Redman, T.C. (Ed.), *Data Quality for the Information Age*, Artech House, Boston, MA, 1996
13. Shankaranarayanan, G., Wang, R. and Ziad, M. (2000), *IPMAP: Representing the Manufacture of an Information Product*, The Proceedings of MIT Data quality Conference (IQ 2000), Boston, MA, 2000
14. Shankaranarayanan, G., Ziad, M. and Wang, R (2003), *Managing Data Quality in Dynamic Decision Environment: An Information Product Approach*, Forthcoming in the *Journal of Database Management*, 2003
15. Tayi, G. K. and Ballou, D. P. (1998) *Examining Data Quality*, Communications of the ACM, 41(2), 54-57
16. Watts Sussman, S.A. and Siegal, W. "Informational Influence in Organizations: An Integrated Approach to Knowledge Adoption," *Information Systems Research* (14:1), 2003.