IPMAP: Current State and Perspectives

Research-in-progress

G. Shankaranarayanan
Boston University School of Management
gshankar@bu.edu

Richard Y. Wang
Massachusetts Institute of Technology
rwang@mit.edu

Organizations are collecting large volumes of data and processing and analyzing data in many complex ways. In such environments, the quality of data can be affected by a number of different hazards. Visual representation of the data processing environments can significantly ease the task of managing data quality. In this paper, we examine the state of research of one such visual representation, the Information Product Map (IPMAP). Our examination here is set in the context of recent trends in data collection and processing. Our objective is to motivate research that can spur insightful ideas and specifications that ought to be part of the next version of the IPMAP and similar visual representations. We look at it from three perspectives: what is needed for visualizing data manufacture, for evaluating data quality in the context of the decision task, and for evaluating data quality in inter-organizational settings. We motivate the importance of each and examine how the IPMAP can assist with each perspective.

1. Introduction

The recent developments and trends in technology and information systems make it easy for organizations to acquire, collect, store, process, analyze, and deliver data. Organizations manipulate the data and analyze it in multiple complex ways to satisfy the need to gather business intelligence and to monitor internal processes. Through partnerships as well as through B2B web portals, organizations exchange data and use data from other organizations for mission-critical decisions. In these settings, there are numerous hazards that can potentially affect the quality of the data. Today, managing data quality is more critical than ever before.

Literature in data quality management, reflecting over three decades of research, has suggested many viable solutions for assessing, managing, and improving quality. The Total Data Quality Management (TDQM) approach for systematically managing data quality in organizations is a dominant paradigm (Wang 1998). This addresses not just data but also the processes that create that data. It is based on the perspective of managing data as a product and adopts several concepts from the manufacture of physical products. One of these is modeling and representing the manufacture of data products (Wang et al. 1998). In this paper, we survey the state of research on models for data manufacture, focusing on the Information Product Map (IPMAP) (Shankaranarayanan et al. 2003) and its predecessor, the Information Manufacturing System (IMS) (Ballou et al. 1998). Our objective in this paper is to examine the developments and trends in data collection and processing over the last decade and to identify the specifications that must be part of the "new and improved" models and tools for modeling data manufacture. We restrict

ourselves to three aspects: (a) visualizing data manufacture, (b) assessing data quality in context and interactively communicating data quality and process metadata, and (c) managing and understanding quality when exchanging data between organizations. For each aspect, we motivate the need for it and suggest the role the IPMAP can play to support it.

The key contributions of this paper are: (1) it highlights the benefits of data visualization and the role of the IPMAP for managing data quality in today's complex data environments. (2) The paper stresses the value of evaluating quality contextually and shows how the IPMAP can assist decision makers understand the quality of the data in the context where it is used.(3) It offers insights into using the IPMAP for managing the quality of data exchanged by organizations. (4) It highlights the extensions to IPMAP that make it more suitable for use in today's data management environments. We complete this section stressing the importance of understanding data manufacturing processes for managing data quality.

The reminder of this paper is organized as follows: in section 2, we present the need for visual representation for managing quality. We then briefly describe the IMS and the IPMAP, directing attention to the role of the constructs as well as the associated metadata. We further distinguish IPMAP from other modeling methods including data flow diagrams, work-flow models, and UML diagrams. In section 3, we motivate the need to contextually evaluate data quality and describe the role of IPMAP in assisting contextual evaluation. In section 4, we underscore the role of the IPMAP in evaluating quality in inter-organizational settings. We conclude, in section 5, by reiterating the opportunities and suggesting directions for further research.

Empirical studies have shown that the outcome of managerial decision making is highly influenced by the data's quality (Chengalur-Smith et al. 1999), (Fisher et al. 2003), Shankaranarayanan et al. 2006). The role of the decision-maker, the data consumer, is critical as this role assesses quality in context and adjusts the decision process and the use of data resources based on this assessment. It is important to support this role when designing tools for data quality management. Quality management paradigms emphasize the role of the manager in driving quality improvement. A key managerial role in many of those paradigms is the "process owner" – a person who assumes responsibility for the data and establishes the appropriate quality policies and leads the improvement efforts. Finally, the role of the data creator is no less important. Quality management in manufacturing has had great success when empowering the worker to manage quality at source. A key requirement for the success of Total Data Quality Management is having the data creator manage data quality at the data's source. Clearly, research in data quality management has to address techniques that can help all three roles cited above, to manage data quality effectively.

2. Visualizing the Creation of an Information Product

Given the volumes of data and the complexity in managing data within organizations, it is becoming increasingly evident that we need a formal modeling method that can alleviate the task of data quality managers. This can be accomplished by offering data quality managers the ability to represent, in an intuitive and easy manner, the complex "production" systems that are used to capture, store, create, and communicate data in organizations. A graphical representation of the different process elements (Ballou et al. 1998) can be used to create a visualized mapping of the

data process, similar to a Data Flow Diagram (DFD) – a visual format that is commonly supported by CASE tools. Other methodologies of documentation and visualization of data processes were discussed in the DQM literature (Redman 1996), (English 1999), (Loshin 2001) and mostly were influenced by other operational management paradigms – customer/supplier models, data flowcharts and the FIP (Function of Information Processing).

One such representation is the IPMAP. The IPMAP is an extension of the Information Manufacturing System or IMS proposed in (Ballou et al. 1998). Interestingly, in this research, the IMS appears to be a representation that supports the evaluation of data quality along data quality dimensions including accuracy, timeliness, and completeness. The basic constructs included in the IMS were the source block, the consumer/sink block, process block, inspection / quality block, and the storage block. The IMS and other similar models lacked a systematic method for representing the processes involved in manufacturing the IP. Furthermore, the existing modeling methods are often insufficient to capture the different manufacturing details.

The IPMAP, besides adding specific constructs to extend the ability to represent manufacturing stages not offered in the IMS, incorporates the "manufacturing" view of the IP, thus allowing information managers to "visualize" how it is created. The constructs added in the IPMAP are: (1) the decision block - multiple information products may be created in a single "production line". Each product may have one or more specific production stages that are unique to that product, besides a large set of production stages that all products might share. The decision block is used to represent the stage after which one or more products may follow its separate and unique path. (2) System and Business boundary blocks – the processes and repositories involved in the manufacture of an IP, may each reside on a different system. Data might also flow across business units or even organizations. Data errors are typically introduced when data changes from one system to another or when transferred from one business unit to another. To facilitate the representation of a stage when data is transferred from one system to another, the system boundary block is used. Similarly, to represent the stage when data spans business units, the business boundary block is introduced. A detailed discussion of the constructs in an IPMAP along with a sample IPMAP is in (Shankaranarayanan et al. 2003). The procedure for creating an IPMAP along with examples of its successful implementation and use are described in (Lee et al. 2006).

The flow of data in the IPMAP is represented by the arrows between stages. Data in the IPMAP is classified into raw data and component data. Raw data refers to the data elements that are obtained from sources. Raw data that goes through any processing (including formatting, inspection, assembly, and transformation) is a component data. We further distinguish component data into simple component and an intermediate component. The former refers to data that has been processed by formatting or collating (assembly) and typically consists of raw data. Intermediate component refers to data that has been transformed or aggregated to create new data. An intermediate component can include raw data, simple component, and intermediate components. This distinction is important for evaluating quality of the data flows in the IPMAP. Quality of a raw data is *assigned* by the provider/manager or *intuitively estimated* by the decision maker. The methods for evaluating quality, such as those proposed by Ballou et al. (1998) are applicable to the component data only. Computing the quality of simple components can be achieved using sum-additive or weighted average techniques while computing the quality of

intermediate components, especially when the data undergoes transformation, requires more sophisticated methods.

Literature in data quality management has investigated managing the quality of an information product (IP) by using methods similar to those used in Total Quality Management (TQM) in manufacturing environments. To see how these methods and processes can be applied to managing the quality of an IP, it is helpful to accurately represent and view the manufacture of an IP as a sequence of processes. Today, organizations integrate data across business functions and units and the organizational systems support cross-functional business processes. Data collected by the sales department is used by manufacturing, data from sales and manufacturing are used for shipping, and data from all of these departments is consumed by accounting and finance. In such cross-functional settings, the IPMAP helps visualize not only the manufacturing stages, but also the flow of data across these stages.

Process documentation, specifically in a visual form, contributes to data quality improvement and provides an important tool to all information stakeholders – managers will find it important for capturing the entire process and understanding all the elements that are involved (Redman 1996), (Shankaranarayanan et al. 2003). The IPMAP helps the data quality manager (the custodian) see what raw materials are used (source blocks), what processing is performed and what new data is created (processing blocks and output data elements), what intermediate storages are involved (storage blocks), how data elements are assembled to create subcomponents and final IPs (assembly – variation of processing blocks), what quality checks are conducted (inspection blocks), whether a subcomponent is reworked (cyclic flows), how the final IP is formatted (variation of processing blocks) and who is using the IP (consumer block). 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 such as an application service provider (ASP), organization, or a different business-unit that uses a different set of computing resources. Today, organizations outsource or off-shore key business processes along with its associated data. Processing intensive tasks such as data cleansing, data enrichment, data aggregation are typically outsourced or off-shored. The cleansed, enriched, or aggregated data is returned to the organization for use within. The IPMAP can help visually represent all of the above. The organizational and system boundary blocks can be used to represent the outsourcing aspects (both data that goes to the external agency and the data that comes back). Decision blocks may be used to visualize the split between manufacturing stages shared by several IPs and stages unique to a specific IP. It allows data managers to capture the creation of several IPs in one IPMAP model. This permits them to manage the "group" of similar IPs as a unit. Lastly, the IPMAP permits data managers understand the implications of poor-quality data for total data quality management by tracing a 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 some manufacturing step(s).

Process documentation can improve the importance of quality from the perspective of information providers and collectors, by getting a more comprehensive picture of where the data goes and how it is used (Lee and Strong 2004). From a data provider's perspective, the IPMAP can help view the processing and transformations the provider's data goes through. More importantly, it can inform the provider of the different IPs that his/her data goes into.

Understanding the importance of such IPs, may serve as an incentive for the provider to manage data quality better. It also helps the information custodians to understand and manage the complexities and inter-dependencies with the process and can improve the sense of reliability and believability with end-users (Shankaranarayanan and Watts 2003).From a data consumer's perspective, the IPMAP helps visualize the sources, processing, and transformations that the data used by the decision-maker/consumer uses has gone through. It helps them gauge the quality in the context of the decision task based on believability, trust, integrity that can only be assessed if he/she is aware of the sources and processing associated with the IP consumed. This is discussed further in section 3. A human processes data using two types of processes: cognitive and perceptual. Literature, in human cognition and visualization in the information systems area, has shown that visual representations reduce cognitive load by diverting some of the processing load to the perceptual side (Jarvenpaa and Dickson 1998), (Vessy and Galetta 1991). Therefore the IPMAP plays a very important role as a visual tool that can significantly enhance not only data quality management but also organizational decision making and organizational performance (Lee et al. 2006). Whether this can be achieved only by an IPMAP or can other modeling methods do the same is a question that needs to be addressed. We now present a brief discussion on how IPMAP differs from existing modeling techniques and how the IPMAP supplements/complements these techniques.

2.1 Distinguishing IPMAP from Other Modeling Methods

The IPMAP, in data quality management, serves primarily as a management tool that helps analyze and understand data manufacturing processes. Hence, for comparing the IPMAP to other modeling methods, we resort to classifying them either as analysis models (used for improving understanding and analysis of data/process) or design models that can be converted into implementations (often times automatically). We first review some of these methods and then discuss the relative merits and demerits of the IPMAP. A summary of this comparison is presented in table 1 below.

Table 1: Comparison Summary - IPMAP with other modeling methods

Model / Software Tool	How does it differ from IPMAP?	Can it complement / substitute the IPAMP?
Process Flow Chart (top down chart or a detailed flow chart)	Shows the steps within a process. The arrows between stages capture the predecessor / successor association. The flow of data is not captured.	Can complement the IPMAP. Process stages within the IP and the business rules/logic associated with each processing stage can be made explicit using Process Flow Charts
Assembly Diagram (a popular use of the flow chart)	Shows the assembly stages that a physical product goes through as it assembled from raw materials to a finished product. The arrows represent the "product flow" through the different stages.	Can substitute the IPMAP for representing the "assembly" of the IP, i.e., constructs offered here can be used to depict the manufacture of the IP. It is designed to represent physical product manufacture and is therefore restricted in its ability to show the different types of processing and storage associated with creating IPs.
Conceptual Data Models (such as ERM)	Is data centric and offers a navigational view of data and data relationships. Represent facts about the real world and cannot represent the flow of data nor can it represent processing. These are not intuitive and require formal training to understand.	Can complement the IPMAP. Data storages in the IPMAP can be described in considerable detail using conceptual data models. Cannot substitute the IPMAP

Table 1: Comparison Summary - IPMAP with other modeling methods (concluded)

Model / Software Tool	How does it differ from IPMAP?	Can it complement / substitute the IPAMP?
UML – Use Case Models	Is a conceptual model of the functionality of a system and shows functional components along with actors/roles/other systems that interact with the system being examined.	From the perspective of an IPMAP, Use case models neither complement nor substitute it. Clearly, does not substitute the IPMAP.
UML – Class Diagrams	Shows the objects and the inter-relationships amongst these objects in a logical view of the system. While the data associated with the system is explicit, processes are hidden within object behavior and methods.	Like Conceptual Data models, the Class diagram can complement the IPMAP. It cannot substitute the IPMAP.
UML – Other Diagrams- State Transition and Interaction Diagrams	State transition diagrams show the state of objects, events that trigger changes in state, and the actions that result. Interaction diagrams show the interactions between objects and their temporal sequence.	Neither can substitute the IPMAP. Neither complements the IPMAP based on the purpose that the IPMAP is used for.
Data Flow Diagrams	Is a structured model that defines the scope of the system, processes within, data storages used by processes, and the flow of data. Is very similar to an IPMAP in this regard. But, the IPMAP is "product" centric while the DFD is a process centric model. DFD cannot distinguish between different types of processes needed to understand IP manufacture. DFD cannot represent the sequence of processes, a very important requirement to represent the manufacture of an IP.	DFD can complement the IPMAP by providing a superior understanding of the flow of data within processing stages in the IPMAP.
Work Flow Models and its predecessor, Work Flow Charts	Are similar to an IPMAP in many respects. Represents activities, data, and data flow in a business process and supports analyses and automation. Key benefit - can represent the checks and balances required to implement the flow of work within a business process. Can also assciate roles or inviduals with tasks and can specify control flows that define dependency relationships among tasks. Work flow models typically deal with a much deeper level of process granularity compared to IPMAPs.	Can substitute the IPMAP given certain restrictions due to the fact that they are not designed for this purpose. Offers a more process-centric view of the manufacture, while the IPMAP offers a product-centric view of the manufacture. Can also complement the IPMAP if used to represent a more granular descriptions of processes.
Microsoft Visio - a popular tool used to create models mentioned here	Offers a variety of process diagramming templates that can make the task of creating flow charts, data flow diagrams, some UML diagrams, and work flow digrams, easy. Does not offer the ability to capture and communicate metadata associated with model constructs, unlike specialized tools that support some of the other models (ERWin, Sybase Power Designer, Oracle CASE)	May be used to create preliminary representations of the IPMAP. Cannot support all the features of the IPMAP due to the lack of a backend metadata repository

Conceptual models (such as ER and all its extensions) were introduced to primarily help model data requirements and design databases. The purpose of using conceptual models was to accurately represent the data relationships that typically exist in database applications. These models hence define a more navigational view of the data relationships (Chen, 1976). Constructs (e.g. entity classes and relationship classes) in these models help represent facts about the real world that need to be captured in the database. Conceptual models are also used to understand the data requirements of an information system and to identify problems with existing databases. Thus conceptual models are both analysis and design models. The drawback of conceptual

models is that they are not intuitive and require some formal training before the model can be interpreted and understood. The IPMAP is an analysis model that represents processing that a data undergoes, besides the flow of data. Conceptual models represent neither of these two.

Data Flow Diagrams (DFD) is a structured modeling method for organizing and documenting the structure and flow of data through the processes. It helps specify the policies, logic, and procedures to be implemented within each process in an information system (Structured Analysis Wiki - Yourdan, E.) The constructs used in a DFD help define the scope of the system, the processes within the system, data storages, and the flow of data between them. A particular variation of the DFD, the "swimming lane diagram" allows processes to be represented within vertical "lanes" where each lane would represent a business-unit/department. Each process/data-storage shown within a given "lane" takes place/resides in the business-unit represented by that lane. The data flows that go across lanes may be interpreted as crossing business units. The "swimming-lane" diagrams serve both in analysis and in the physical design of an information system. The DFD is a process-centric model that is created from the perspective of a single system. An IPMAP is a product-centric model that is created from the perspective of one specific output created by one or more than one system.

The Unified Modeling Language (UML) is a modeling language for specifying, visualizing, constructing, and documenting the artifacts of a system-intensive process (OMG 2007). UML consists of several diagramming methods each with its own specific use. Use cases represent the functional components (or functionality) of the system while the actors represent users, roles, or other systems that interact with the system represented by the use cases. It is hence a conceptual model of the functionality of the system and its interactions and is used in analyzing system requirements. Use cases are closer to data flow diagrams than IPMAPs. Another UML diagram is the Class-diagram that models the objects (object-classes) in the system. It is used to show the existence of classes and inter-relationships between these classes in the logical view of the system (Blaha and Rumbaugh, 2004). This model can be used in both analysis (to capture the common roles and responsibilities of the objects that defines the system behavior) and in the design (to capture the structure of the classes that define the architecture of the system) of information systems. Yet another diagram in UML is the interaction diagram. It shows the interactions between the object classes and are useful in capturing the semantics of scenarios, usually early in the system life cycle (hence a conceptual model for analysis) (Blaha and Rumbaugh, 2004).

The IPMAP conceptually represents the processes involved in the manufacture of an information product. At a high level, it is a conceptual representation of how the different processes are "laid-out". It does not in anyway help design an information system. As it represents the "steps" or processes involved, it is closer to the process models (DFD and Use-cases) than to the data models (conceptual models, class diagrams etc.). The flows into a stage in an IPMAP represent the flows of data that form the "input" to that block, similar to a DFD data-flow. However, in DFD the end-point of each data-flow is one of the following: data storage, a process, or an external entity (something that is beyond the scope of the system represented by the diagram). In IPMAP, the blocks can not only represent processes, data sources/sinks (usually external entities), but also represent the transfer of data across departmental / organizational boundaries and the transfer of data from one information system to another. This concept is similar to

modeling the DFD in the form of “swimming-lane diagrams”. All processes are represented the same way in a DFD (and its variations). As the IPMAP attempts to create a conceptual representation of the different “steps” in the “product-line”, the representation must help differentiate the different types of processes. For instance it must differentiate between processes that transform input data to create a new data versus when the input data is checked for conformance to quality standards.

Another model similar to the IPMAP is the workflow model used in systems that support workflow management. A workflow management system is one that provides procedural automation of a business process by management of the sequence of work activities and by the invocation of appropriate human and/or IT resources associated with the different steps (The Workflow Reference Model, 1995). A workflow model creates a schematic representation of the activities, data, and data flows for a business process and helps both in analyzing the process as well as in automating it. Constructs in a typical workflow model include activities or tasks and data and its flow amongst the tasks. An activity (or task) may be elementary or complex, the latter consisting of more than one elementary task. In addition, workflow models offer the ability to associate a role or a specific individual with tasks to designate responsibility for tasks. Workflow models also have the ability to specify control flows that define precedence/dependency relationships between the tasks. Instead of or in addition to specifying roles associated with tasks, workflow models also permit the specification of software application modules/programs that perform these tasks (when a task is automated or performed using software).

Workflow models are similar to the IPMAP in several respects. They model the tasks that makeup a process (typically a business process but can be applied to other processes as well) and help define controls between tasks. Instead of tasks, the IPMAP captures the manufacturing processes that create an information product and define the sequence or “layout” of these processes. Workflow models allow the representation of data and its flow between the activities in the process being modeled just as the IPMAP represents the data elements that flow between manufacturing processes that define the creation of the information product. The workflow model helps specify the role/individual responsible for each of the activities captured in the model and the same information is specified as metadata within each relevant block in the IPMAP. In fact the workflow models also permit the capture of semantic constraints associated with tasks, task-termination dependencies and others as metadata associated with tasks. Workflow models differ from the IPMAP representations in two important respects. The latter is an information-product centric approach in which the model (IPMAP) captures the processes and data quality related activities involved in the creation of an information product while the workflow model is a process-centric approach. Secondly, workflow models deal with a much lower level of process granularity in that they breakdown a specific business process into activities while the IPMAP deals with all the processes involved in the manufacture of an information product. In the IPMAP, the activities that make-up a specific manufacturing process are specified as metadata associated with that process block and these activities could be represented using a workflow model. The workflow model hence supplements that IPMAP if the information manager wants to analyze one or more steps in the manufacture of the information product.

Work flow charts, the predecessor of today's work flow models are a variation of the most commonly used technique, flow charts. Besides work flow charts, there are two types of flow charts that are relevant here. First is the process flow chart, a model that represents the logical steps within a process (De Marco 1979). The arrows between steps represent the logical flow and capture the predecessor / successor association. These do not show the flow of data at all. In that respect, they are different from the IPMAP. The second type of flow chart is the assembly flow chart that is used to represent the "assembly" of physical products from raw materials through finished goods. This feature makes an assembly chart very similar to an IPMAP. However, since they were intended for use with physical products, they are restricted in their ability to represent the different stages associated with the creation of an IP. Like other flow charts, an assembly chart cannot represent the flow of data.

There are several software tools to assist users build models discussed here. A popular tool that is easily available and extensively used to develop models is Microsoft Visio (Microsoft Visio – Wikipedia, 2007). Visio supports a very large variety of modeling methods and provides templates to make it easy for modelers to develop their models. However, Visio, lacks the backend repository support that many of the more advanced (and expensive) modeling tools (e.g., Oracle CASE (www.oracle.com), ERWin (www.ca.com), Sybase Power Designer (www.sybase.com)) have. Hence its ability to capture metadata associated with the model is restricted, making it a static visual representation with no interactive capabilities. However, it is arguably the most popular tool and is extensively used by organizations to create models similar to an IPMAP. Its power lies in its user friendly interfaces and its easy accessibility.

3. Evaluating Data Quality in Context

Data quality is evaluated along quality dimensions such as timeliness, completeness, accuracy, and relevance, each requiring a different technique and using different metadata for evaluation. Further, it has also been shown that data quality may be evaluated impartially, based on its structure (called structure-based by (Ballou and Pazer, 2003)) or evaluated contextually based on the content and evaluated within a specific usage-context (called as content-based by (Ballou and Pazer, 2003)). Ballou et al. (1998) have shown how timeliness and accuracy can be computed impartially. Today, for evaluating data quality, it is not sufficient to offer decision makers just the impartial measurements and the methods used to derive them, but, it is also necessary to allow decision makers to evaluate quality in the context of the task that the data is used for.

Literature has acknowledged that some attributes of data quality are invariant while others vary depending on the context of use (e.g., (Redman 1996), (Wang and Strong, 1996), (Jarke et al., 1999), (Ballou and Pazer 2003)). This makes the measurement of data quality problematic, since it means that aspects of the user's interaction with the data while performing a task must be accounted for. Only the user can do so and hence the need to allow users to gauge the quality in the context of the task. This is not to be misconstrued as saying that contextual assessments are more critical. We posit that both are important and the users must be given the impartial assessments and allowed to evaluate assessments contextually. For example, an accuracy assessment that was categorized as impartial may have contextual interpretation and the same may be true for completeness or timeliness assessments. Some of the attributes under other categories (such as interpretability or access security) may have both impartial and contextual

aspects. It is agreed that some quality attributes are more impartial in nature, while other are more contextual, but many allow both types of assessment. A recent study (Shankaranarayanan, Even, and Watts 2005) suggests a model of sequential impact - impartial data quality assessments impact the assessment of contextual quality. This, in turn, allows more efficient decision making and leads to better decision making outcome. This chain of impacts, including mediation effects of the intermediate links are supported by an empirical study.

Lee (2003) emphasizes the need for contextual understanding from a different perspective. Her study shows that data quality professionals in organizations undertake a thorough investigation of the context of an IP to solve data quality issues with it. This process connects and more importantly, contextually associates, the different processing stages (acquisition, collection, processing, storage, and delivery) in the manufacture of an IP. Further more, this research highlights the importance of these contextual associations for solving data quality problems and for data quality management.

The IPMAP supports contextual evaluation by allowing the decision-maker to gauge the quality of the data in the context of the decision-task. It does so by communicating the impartial data quality measurements associated with the data at each stage of the manufacture. These measurements are referred to as quality metadata, data quality information, or data tags. Further, the details of the data sources and manufacturing processes, process metadata, is also communicated to the user to permit the user to understand the lineage and to get a sense of the believability and reliability of the data sources and manufacturing processes. The process metadata also communicates the assumptions and logic associated with the transformations that the data goes through at the corresponding processing stages. Lastly, the IPMAP allows users to use established methods for computing data quality. Examples of such methods include those proposed in (Ballou et al. 1998), (Pipino et al. 2002), and (Shankaranarayanan and Cai 2006).

The IPMAP uses a comprehensive metadata repository to capture the metadata and communicate it with the users. The conceptual schema of the metadata repository that serves as the back-end of the IPMAP is shown in figure 2 using an Entity-Relationship model. Each construct in the IPMAP 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 requirements for that manufacturing step, (6) the business rules/constraints associated with it, (7) a description of the technology used, (8) the physical location where the step is performed, (9) and the type (data source, processing, storage, inspection, system boundary, business boundary, or data sink) of manufacturing stage represented by the construct. 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. The final metadata element helps determine the type of computation necessary for evaluating some specific quality dimension (e.g., completeness, accuracy).

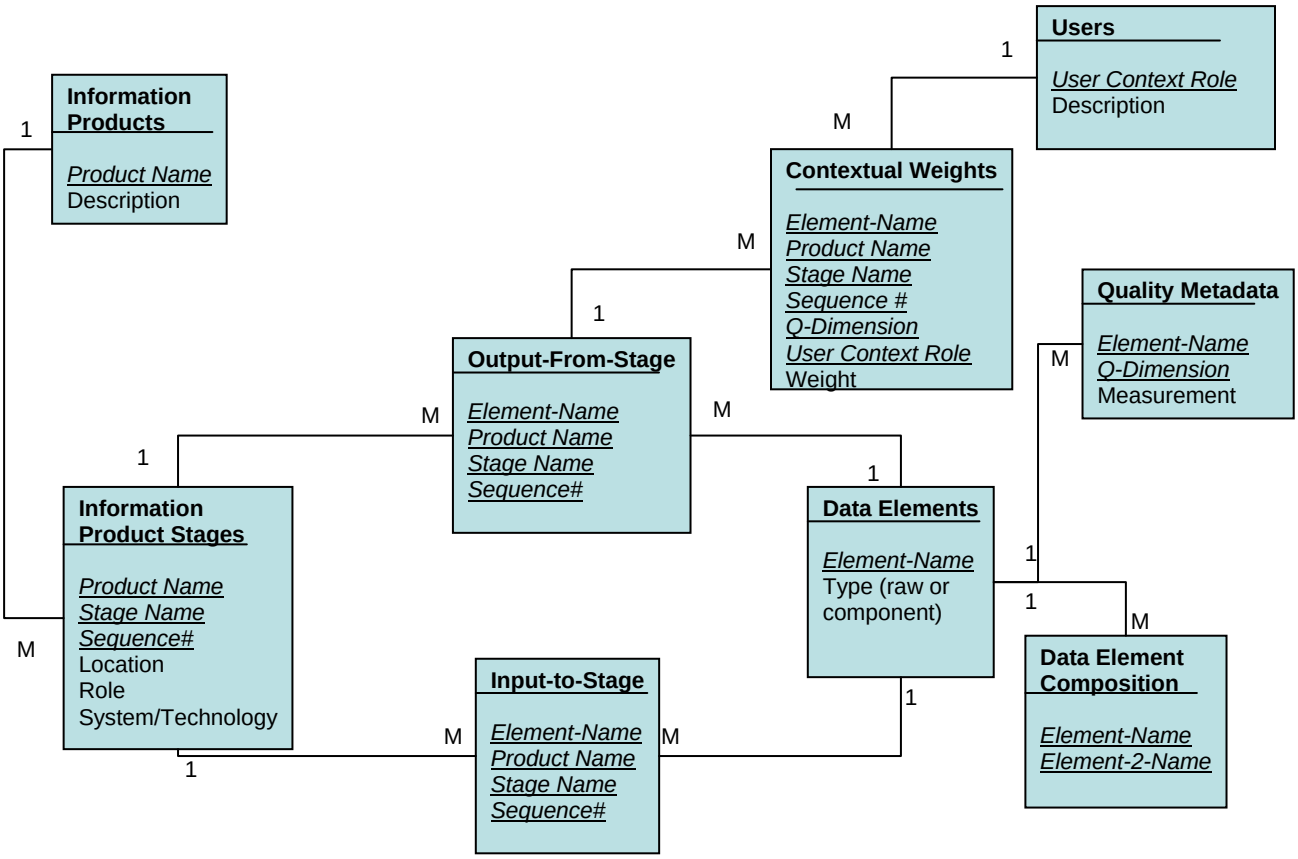

Figure 1: Logical Model of the Metadata Repository (illustrative sample subset is shown)

An illustrative sample of the metadata repository in the IPMAP is shown using a logical model in figure 1. Information products are associated with multiple product manufacturing stages. A data element is associated with one or more of these stages. It may be an input to the stage or may be the output of a stage. The metadata also captures the composition of each data element, in terms of other data elements that are part of it. Associated with each data element ***j*** is its impartial measurement $\mathbf{M}_I^X(j)$ along a specific quality dimension (e.g., $A_I^X(j)$ for accuracy or $C_I^X(j)$ for completeness) determined using methods proposed in literature. *(X* can be D, SC, or IC depending on whether the data element is a raw data element (D), simple component (SC), or an intermediate component(IC)*)***.** This value is used to determine the context-dependent measures. The context dependent measurements can be evaluated using techniques proposed in literature (e.g., (Even and Shankaranarayanan, 2007)) and captured (subscript of *C instead of I).* The measured values (context-dependent and independent) of each component are associated with the specific stage in the IPMAP of which the component is the output. The weights that can be assigned and changed to define the context by the decision-maker associated with each data element or component is also maintained at this stage of a specific IP and associated with a specific decision-maker.

A data unit typically has time-tags (also data quality metadata) specifying when it was obtained (Wang and Storey, 1995). Time estimates for the processing duration at each stage can be obtained using the time-tag and knowing the time when the output was created at this stage. These estimates may be revised over time. The tags also help estimate the elapsed time between

data capture (e.g. using PDAs or RF receivers) and the time when data becomes accessible (a PDA's synchronized with a networked computer or a receiver pushing data into the network).

The IPMAP offers a comprehensive view of the data used in a decision-task by informing the consumer about the sources or providers of the data, storages, transformations and processing, logic and assumptions associated with these transformations and processing, and the custodians associated with each of these stages. It further provides access to the methods for evaluating quality. The consumer or decision-maker now has the ability to compute and gauge the quality of the data in the context of the task in which the data is to be used. The decision maker would do so by assigning weights to the data, reflecting the perceived importance of that data for the task it is used.

4. Need for Managing DQ in Inter-organizational settings

The sharing and exchange of data between organizations is an important part of business relationships. Business operations require routine exchange of data. For instance, to improve operational efficiency and achieve competitive advantage in supply chain relationships, organizations acquire upstream and downstream inventory information, production, logistics and storage capacities from their partners besides transaction data. In the past, it has been difficult and expensive to seamlessly exchange and integrate large volumes of data across organizational boundaries. The advances in information technology (IT) greatly facilitate inter-organizational data exchange. IT reduces the data collection, transfer and processing costs and makes data assets more attractive and valuable to create, own, and manage. Daniel and White (2005) suggest that the inter-organizational data linkages will become ubiquitous in the future.

Data networks for inter-organizational data exchange are characterized by multiple, independent data sources from which this data is extracted, and multiple, independent data repositories in which the data is captured / stored. Data management for decision-making in such environments involves gathering relevant data from outside the organization and integrating it with local data. Organizations appear to implicitly assume that the quality of the data obtained from other organizations is acceptable. Research indicates that poor data quality (DQ) is a serious problem within many organizations (e.g., Eckerson 2002). This study also states that a major source of data quality problems is *external* data. If an organization has no accepted standard for data quality, how does it know what the quality of the data from external sources is and whether it conforms to its data quality standard(s)? In a network supporting data exchange among organizations, it is important to assure organizations of the quality of data they get from other organizations. A prerequisite is that organizations must first manage DQ internally. Further, organizations use data received from another organization as inputs (either directly or after processing) to their business operations and decision-making. The issue of data quality is therefore not local to or isolated within one specific organization.

To manage and evaluate DQ in data networks, we first need a standard based on which organizations can develop, implement, and understand data quality. This would enable an organization to understand the implications of data quality measurements associated with the data it receives from other organizations. However, this does not assume that two or more organizations that share/exchange data have agreed upon a single data quality standard. Second,

we need a method for evaluating the quality of data exchanged by organizations by which the receiving organization can understand the quality of the data obtained, in the context of its own data quality standards. Further, an organization needs to not only evaluate the quality of the data it generates but also understand how the *quality of that data* is impacted by the quality of data received from other organizations. To achieve this, besides exchanging data, organizations must be able to exchange data quality information (or data quality metadata) associated with the exchanged data. They must be able to integrate their own data quality information with that received from other organizations. Finally, the inter-organizational data network is characterized by multiple different stakeholders. Research has shown that data quality evaluation is context-sensitive (e.g., (Fox et al. 1994), (Wang and Strong 1996)). Each stakeholder must therefore be permitted to evaluate the quality of the data, independently and within the specific context in which he/she uses that data.

We can adopt the information product approach to conceptualize and define the foundation of the data quality management model. The IP approach considers data and processing, to systematically evaluate and manage data quality in information systems (Wang 1998), (Wang et al. 1998). From this perspective, the data exchanged across organizational boundaries can be viewed as IPs created in one organization and consumed by another. A special case of IP exchange in inter-organizational networks is the purchase of information goods in dedicated marketplaces. The data quality management model proposed here is also applicable for facilitating transactions in marketplaces dedicated to information goods.

To exploit the characteristics of 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. The IPMAP, corresponding to some IP, is a graphical model that represents the manufacture of that IP. As mentioned earlier, it is supplemented by a comprehensive metadata repository that can help decision-makers gauge the quality of data. The IPMAP also includes quality metadata associated with the data to which the IPMAP corresponds. These measurements are captured at each stage of the IPMAP including the final stage.

For each IP, the process metadata and the quality metadata specified in its corresponding IPMAP, as well as the structure of the IPMAP, can be defined using XML. When data, an IP, is exchanged, organizations can exchange the metadata associated with that data as well as the structure of the IPMAP associated with that data (IP). This would enable organizations to integrate not just the data obtained from outside its boundaries, but also integrate the associated quality metadata and process metadata to efficiently manage data quality in inter-organizational data networks.

Adopting the IPMAP offers three major benefits for managing data quality in inter-organizational data networks. First, it offers an intuitive way of integrating one or more IPs received from external organizations with the manufacture of an IP created within – very similar to physical components/subassemblies purchased from suppliers and incorporated into products that a company manufactures. Second, using the IPMAP, techniques have been described for evaluating data quality in *a single information system* (Ballou et al., 1998), (Shankaranarayanan and Cai, 2006). These techniques can be adapted for evaluating data quality in inter-organizational settings that typically spans multiple systems. Third, as a visualization tool, the

IPMAP can help the decision-makers visualize the exchange of data across organizational boundaries and its integration with local data.

Metadata exchange has been discussed extensively in data management literature. We use Context Interchange (COIN) Project (Goh et al., 1997) and the Semantic Web (Berners-Lee, 2001) as exemplars of research in metadata exchange. Supporting quality management in inter-organizational settings requires the exchange of quality metadata associated with the exchanged data. It is hence appropriate to distinguish the metadata exchange proposed here with these two well known initiatives describing the exchange of metadata.

COIN is an initiative that supports the intelligent integration of data from multiple heterogeneous sources (Goh et al. 1997). There are several directions within this initiative including integration of information with source attribution (e.g., (Goh et al. 1997), (Lee et al. 1998), (Lee et al. 1999)), integrating information by mediating the context to provide an uniform view of the integrated data (e.g., (Goh et al. 1999), (Moulton et al. 1998)), integrating information and incorporating it into business process using pattern matching (e.g., (Bressan and Goh, 1997), (Bressan and Bonnet, 1997), (Bressan et al. 1997)), and the development of a SQL-based query interface for information integration with context mediation (Bressan and Goh, 1998). Of these directions, the one most relevant to this discussion is source attribution (e.g., (Lee et al. 1998), (Lee et al. 1999)). The authors state that there are instances where it is important to associate each data element with the source(s) that contributed to its creation or from which the data element was extracted. Attribution is defined as the association between a data element and its source and the attribution framework includes a conceptual attribution model and an algebra to define a formal query language. The research leaves it to the user to gauge the results (data) based on the source associated. The IPMAP is similar to this research in that it too, associates the sources with the data elements in the final information product. However, the IPMAP goes a step further to include the processing as well as the processing logic involved with the creation of the data element in the final product from the data element(s) obtained from its source(s). Moreover, the IPMAP communicates quality metadata associated with the different data elements and how these were used to define the overall quality of the final information product, the data that is exchanged between two organizations. Similar to the attribution research within COIN, the IPMAP provides the metadata to the user but does not integrate this data, allowing the user to apply his/her own heuristics to understand the overall quality. The COIN initiative also offers methods that synchronize the contexts when information is integrated from multiple sources. Currently, the IPMAP does not offer methods to mediate the contexts. Context mediation methods can used to offer a uniform and integrated view of the quality metadata. However, since quality needs to be interpreted in the context of the task, it might be difficult to specify the mediation constraints a priori.

The Semantic Web (recently termed as the Web of Data) is based on the Resource Description Framework (RDF). Its purpose is to provide a common framework that allows applications, organizations, and communities to share and reuse the data. It attempts to offer a common format for integrating and combining data pulled from multiple different sources as well as a language that links this data to real-world objects of interest (http://www.w3.org/2001/sw/; Berners-Lee in Business Week 2007). The IPMAP has much the same idea as the Semantic Web. The purpose is to permit users to understand the integrated quality of the data that is combined from multiple

sources. It is unlike the semantic web in that it does not rely on the meaning/semantics of the data, but, proposes a way to communicate the semantics related to the quality of data – so as to permit users interpret the quality in a manner that fits contextual assessment of data quality. The Semantic Web is more useful for unstructured data that is typical of the Web. The metadata associated with the IPMAP is more formally structured, making it easier to integrate and interpret. In this regard, the use of the IPMAP for integrating quality metadata is less complex. The Semantic Web is founded on the RDF – a specification framework for the exchange of metadata – and on its extension, the Web Ontology Language (OWL). Although the RDF was meant to describe the resources associated with the Web, this foundation of the Semantic Web can be used to define the quality and process metadata exchange specifications to be implemented with the IPMAP.

The organization receiving data must be aware of the standard that the sending organization has used to define its data quality. Two general ideas come to mind for supporting this. A universal standard (like the ISO 9000 in manufacturing) may be established to serve as the benchmark. Alternately, an organization can communicate its standard for data quality along with its data to the receiving organization. The first requires the definition of a standard that must be accepted by all participating organizations. A body must define and continuously refine this standard and encourage organizations to embrace it. The second can be supported by communicating the standard using metadata exchange and supporting the interpretation of the sender's standard in the context of the receiver. COIN and similar initiatives can effectively support the latter case.

5. Conclusion

Modeling the manufacture of information products can play a very important role in managing data quality. In this paper we present the state of research on the Information Product map (IPMAP) and its predecessor, the Information Manufacturing System (IMS). We examine the developments and trends in data collection and processing over the last decade and describe three key needs that must be addressed by models for data manufacture. These needs are *visualizing data manufacture* and its importance for managing data quality, *contextual assessment* of data quality and its role in data quality management, and the importance of evaluating data quality in inter-organizational settings. Each of these needs offer opportunities to extend the IPMAP. Organizations are using generic models for modeling data quality. These representations are not based on any standard and each organization and functional units within adopt different and unique standards based on their own needs. We believe that there is a strong need to define some standard for models that represent data manufacture. In this paper, we have revisited the standards that the IPMAP proposes and identified the areas for extending the IPMAP and defining the corresponding standards for it. Our intent is to help define specifications that must be part of the "new and improved" models and tools for modeling data manufacture.

Besides the extensions discussed here, there are others that can be useful. We mention some of these here. First, can we integrate the IPMAPs corresponding to two or more IPs that are combined to create a new product? The IPMAP, in its current form can support this integration. The outputs of the two (or more) IPs can be directed into a process (or a set of manufacturing stages) that creates the new IP. Second, can we construct a timeline for how long it takes to manufacture an IP? Capturing processing time at each stage of the IPMAP and communicating it

to the user will permit the user to build a timeline and estimate completion time. A method for estimating completion time using time-tags has been proposed in (Shankaranarayanan et al., 2003). Third, can we build a CASE tool for building the IPMAP and for animating it (using simulation)? We are in the process of building a comprehensive tool for the IPMAP and examining the possibility of including animation into this tool.

References

Blaha, M. R. and Rumbaugh, J. R (2004) Object Oriented Modeling and Design with UML (2nd Edition), *Prentice Hall, New York, NY*

Berners-Lee, T. and Miller, E. (2002), The Semantic Web Lifts Off, *ERCIM News,* No. 51, October, 2002 *http://www.ercim.org/publication/Ercim_News/enw51/berners-lee.html*

Bressan, S., Fynn, K., Goh, C., Jakobisiak, M., Hussein, K., Kon, H., Lee T., Madnick, S., Pena, T., Qu, J., Shum, A. and Siegel, M. (1997), The COntext INterchange Mediator Prototype, *Proceedings of the ACM SIGMOD International Conference on Management of Data*

Bressan, S. and Bonnet, Ph. (1997) Extraction and Integration of Data from Semi-structured Documents into Business Applications, *Proceedings of the Conference on the Industrial Applications of Prolog*

Bressan, S. and Goh, C. (1998), Answering Queries in Context, *Proceedings of the International Conference on Flexible Query Answering*

Ballou, D. and Pazer, H. (2003), Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts, *IEEE Transactions on Knowledge and Data Engineering,* 15 (1) , 240-243

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

Chen, P.P., "The Entity-Relationship Model: Toward a Unified View of Data," *ACM Trans. on Database Systems,* Vol.1, No.1, March 1976, pp. 1-36.

Chengalur-Smith, I., Ballou, D. and Pazer, H. (1999), The Impact of Data Quality Information on Decision-Making: An Exploratory Analysis, *IEEE Transactions on Knowledge and Data Engineering,* 11(6), 853-864

Daniel, E. and White, A. (2005) "The future of inter-organizational system linkages: findings of an international Delphi study" *European Journal of Information Systems,* 14, 188-203.

De Marco, T., (1979) Structured Analysis and System Specification, Yourdon Press Computing Series, A Prentice Hall PTR Title, Chicago, IL

Eckerson, W.W. (2002), Data Quality and the Bottom Line: Achieving Business Success Through a Commitment to High Quality Data, *The Data Warehousing Institute Report Series*, No.101, Chatsworth, USA.

English, L. (1999) *Improving Data Warehouse and Business Information Quality*: *Methods for Reducing Costs and Increasing Profits,* Wiley, New York, NY.

Fisher, C. W., Chengalur-Smith, I., and Ballou, D. (2003) The Impact of Experience and Time on the Use of Data Quality Information in Decision Making. *Information Systems Research,* 14 (2), 170-188

Fox, C., Levitin A., and Redman, T. C. (1994) The Notion of Data and its Quality Dimensions, *Information Processing & Management,* 30 (1), 9-19

Goh, C., Bressan, S., Lee, T., Madnick, S. and Siegel, M. (1997) A Procedure for the Context Mediation of Queries to Disparate Sources, *Proceedings of the International Logic Programming Symposium*

Jarke, M., Jeusfeld, M., Quix, C. and Vassiliadis, P. (1999) Architecture and Quality in Data Warehouses: An Extended Repository Approach, *Information Systems*, 24 (3), 229-253

Jarvenpaa, S. L. and Dickson, G. W. (1998), Graphics and Managerial Decision Making: Research Based Guidelines, *Communications of the ACM*, 31(6), 764-774

Lee, T. and Bressan, S. (1997) Multimodal Integration of Disparate Information Sources with Attribution, *Proceedings of the Entity Relationship Workshop on Information Retrieval and Conceptual Modeling*

Lee, T., Bressan, S., and Madnick, S. (1998), Source Attribution for Querying Against Semi-structured Documents**,** *Proceedings of the Workshop on Web Information and Data Management, ACM Conference on Information and Knowledge Management*

Lee, T., Chams, M., Nado, R., Madnick, S., and Siegel, M. (1999), Information Integration with Attribution Support for Corporate Profiles, *Proceedings of the ACM Conference on Information and Knowledge Management*

Lee, Y. W. (2004) Crafting Rules: Context-Reflective Data Quality Problem Solving, *Journal of Management Information Systems*, 20(3), (Winter), 93-120.

Lee, Y. W., Pipino, L. L., Funk, J. D. and Wang, R. Y. (2006), *Journey to Data Quality*, MIT Press, Cambridge, MA

Lee, Y. W. and Strong, D. M. (2004), Knowing Why about Data Processes and Data Quality, *Journal of Management Information Systems*, 20(3), (Winter), 13-49

Loshin, D. (2001) *Enterprise Knowledge Management: The Data Quality Approach*, Morgan Kaufmann, San Francisco, CA.

Microsoft Visio (2007) – Wikipedia, the Free Encyclopedia at http://en.wikipedia.org/wiki/Microsoft_Visio - accessed September 2007.

Moulton, A., Madnick, S. E., and Siegel, M. D. (1998), Context Mediation on Wall Street, *Proceedings of the Cooperative Information Systems*

OMG (2007), Introduction to OMG's Unified Modeling Language, *Object Management Group* (accessed September 2007 - *http://www.omg.org/technology/documents/formal/uml.htm)*

Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002) Data Quality Assessment, *Communications of the ACM*, 45 (4), 212-218

Redman, T.C. (1996) Data Quality for the Information Age. *Boston, MA: Artech House*, 1996.

Shankaranarayanan G. and Watts, S. (2003), A Relevant, Believable Theory of Data Quality Assessment, *Proceedings of the 8th International Conference on Information Quality*, Cambridge, MA

Shankaranarayanan, G., Ziad, M. and Wang, R. Y. (2003) Managing Date Quality in Dynamic Decision Environment: An Information Product Approach *Journal of Database Management*, 14 (4) 14-32

Shankaranarayanan, G., Even, A. and Watts, S. (2006) The Role of Process Metadata and Data Quality Perceptions in Decision Making: An Empirical Framework and Investigation, *Journal of Information Technology Management*, XVII (1), 50-67

Shankaranarayanan, G. and Cai, Y. (2006) Supporting Data Quality Management in Decision Making, *Decision Support Systems*, 2(1), 302-317

Vessey, I. and Galletta, D. (1991), Cognitive Fit: An Empirical Study of Information Acquisition" *Information Systems Research*, 2(1), 63-85

Yourdan, E., (2007) Data Flow Diagrams: Chapter 9, Structured Analysis Wiki, http://yourdon.com/strucanalysis/wiki/index.php?title=Chapter_9 (accessed September 2007)

Wang, R. Y. and Strong, D. M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems* 12(4), 5-34.

Wang, R. Y., Lee, Y. W., Pipino, L. L. and Strong, D. M. (1998), Manage your Information as a Product, *Sloan Management Review* 39(4), 95-105.

Wang, R. Y. (1998) A Product Perspective on Total Data Quality Management, *Communications of the ACM*, 41(2), 56-65

Workflow Reference Model version 1.1 (1995), Workflow Management Coalition, *http://www.wfmc.org/standards/referencemodel.htm*