An Information Product Approach for Total Information Awareness

Richard Wang rwang@mit.edu
Thomas Allen tallen@mit.edu
Wesley Harris weslhar@mit.edu
Stuart Madnick smadnick@mit.edu

Abstract--To fight terrorism successfully, the quality of data must be considered to avoid garbage-in-garbage-out. Research has shown that data quality (DQ) goes beyond accuracy to include dimensions such as believability, timeliness, and accessibility. In collecting, processing, and analyzing a much broader array of data than we do currently, therefore, a comprehensive approach must be developed to ensure that DQ is incorporated in determining the most probable current or future scenario for preemption, national security warning and decision making. Additional data such as who was the data source, when was the data made available, how, where, and why also need to be included to judge the quality of the information assembled from these data.

We propose such an approach for Total Information Awareness with Quality (TIAQ), which includes concepts, models, and tools. Central to our approach is to manage information as a product with four principles. We have applied the information product approach to research sites where opportunities arise. For example, the Air Force Material Command uses requirements definition and forecasting processes to perform a number of functions. However, the Air Force experienced several complex problems due to DQ problems; as a result, fuel pumps were unavailable. Each engine needs a fuel pump; when a pump is not available, a military aircraft is grounded. We traced the fuel-pump throughout the process of remanufacture, and identified root causes such as delays by pump contractors and ordering problems. To a certain extent, detecting foreign terrorists and decipher their plots are analogous to tracing fuel pumps. Our research provides an interdisciplinary approach to facilitating Total Information Awareness.

KEY WORDS Total Information Awareness (TIA), Total Information Awareness with Quality (TIAQ), Data Quality (DQ), Information Product Map (IPMap), Quality Entity Relationship (QER).

1. INTRODUCTION

Much has been presented in the Congressional intelligence hearings regarding what had transpired prior to the September 11 tragic event. Many data [1] problems were revealed beyond the common problems of lack of human, hardware, and software assets to facilitate various counter-terrorism activities. Vital information failed to reach the right decision makers because they were not available or accessible, or accessible but not considered as relevant, credible, or accurate. Had the available data made accessible, represented appropriately to support the intelligence community in integrating the relevant, credible, and accurate information in a timely manner to make the connections, the tragic event might have been possibly averted.

Research Challenge

Most approaches to homeland security and Total Information Awareness (TIA) program focus on developing state-of-the-art, novel and practical computing environment to store a vast amount of data, and organizing well-trained personnel to collect, process, and analyze the vast amount of data. Conventional wisdom often dictates that data quality (DQ) is equal to accuracy; if the data is stored in the computer, it is of high quality. In contrast, our research has clearly shown that DQ goes beyond accuracy.

In our research [1-6], we have attributed timeliness, believability, relevance, and accuracy as part of the multi-dimensional DQ concept. Little research, to date, has been conducted to develop a cumulated body of DQ knowledge, to establish DQ as a discipline, and to transition research results to government and industry practice. Recently, DQ has become more visible as corporations learned from their costly experience. It is well accepted nowadays that for an Enterprise Resource Planning (ERP) or Data Warehouse project to be successful, firms must attend to DQ [7-12]. By the same token, in achieving TIA to fight terrorism successfully, DQ must be considered to avoid garbage-in-garbage-out. In collecting, processing, and analyzing a much broader array of data than we do currently, a comprehensive approach must be developed to ensure that DQ is incorporated in determining the most probable current or future scenario for preemption, national security warning and decision making. Additional metadata such as who was the data source, when was the data made available, and how were the data transmitted (via a secured mechanism or from the Internet say) need to be included for the intelligence

[1] We use “data” and “information” interchangeably throughout this paper.

community and decision makers to make their judgments. This, however, is a tremendously difficult and challenging problem considering the order of magnitude of data that must be dealt with and the complexity of how to facilitate the collection, processing, and analysis of the data in light of the evident needs to effectively manage this process. A comprehensive approach must be developed.

Research Approach

We propose an approach that begins to tackle systematically DQ problems such as those illustrated above. Specifically, we define DQ; we propose a Total Data Quality Management cycle [13]; we argue that the intelligence community must treat information as a product in developing technologies, components, and applications to produce prototype systems that will accelerate, integrate, broaden, and automate current approaches. Central to our approach is to manage information as a product with four principles [14]:

1. Understand data consumers' needs,
2. Manage information as the product of a well-defined information production process,
3. Manage the life cycle of the information product, and
4. Appoint information product managers.

Our research goal is to develop a pragmatic, theory-grounded methodology with the essential concepts, models, tools, and techniques to manage information as a product for TIA. This demands an interdisciplinary collaboration that builds upon cumulated knowledge from key fields such as Computer Science, Management of Technology & Policy, and Management Science. The solutions must address both technical and managerial issues in order to achieve TIA.

The remainder of this paper is organized as follows: Section 2 presents the concepts, tools and techniques needed for managing information as a product for TIA. We refer to the corresponding toolkit as ***TIAQ* - *Total Information Awareness with Quality***. Some of these tools and techniques have been developed, whereas others are still at their early stage of research and require substantial effort to be deployable. Section 3 illustrates how ***TIAQ*** may be developed via an Air Force Lean Sustainment Initiative project that traces an Air Force engine pump remanufacture process. Section 4 discusses technically some of the TIAQ tools can be created**.** Finally, concluding remarks are made in Section 5.

2. TIAQ RATIONALE AND REQUIREMENTS

As mentioned earlier, central to our research approach is the concept of managing information as a product. In our previous work, we have identified four principles in treating information as product based on many field studies in various organizational settings [14]. It became evident from our research results and industry practices that to achieve TIA, technical solutions alone will not suffice. Without the strategic and managerial components to direct the overall effort within and across organizational boundaries, data collected are often incompatible, inconsistent, and sometimes simply erroneous, leading to failures that costs dearly. A subtler but more complicated problem is the common practice of treating information as the by-product, with hardware and software upgrade as the main activities [14]. A key thrust in our research, therefore, is to develop the necessary tools and methods to begin to facilitate the institutionalization of the four principles of managing information as a product. Accordingly, below we discuss how the four principles can be supported.

To ***understand data consumers' needs***, the conventional systems analysis textbooks clearly suggest a user's requirements analysis followed entity-relationship (ER) modeling [15, 16], leading to database applications development. The problem with this approach in the asymmetric, counter-terrorism warfare, however, is the unknown factors and un-predictability of terrorists threats, and the necessity to dynamically re-configurate an information product manufacturing system to produce the on-demand information products for delivery timely and credibly – note that accuracy is not listed but rather timeliness and credibility. We did so because accuracy means that the information product delivered truly reflects the real-world state [1], which until one can verify, the decision makers must act based on the timely and credible information products available to them.

The preceding analysis leads to the following observations: In addition to leveraging the well-established system analysis and database management research and practice, we must develop facilities that capture, store, retrieve, and analyze information collected on-demand by adding data about the quality of data collected (including the currency of data, source of data, and other DQ dimensions), namely a mechanism to capture data about the quality of data, and the ensuing activities. This requires a capability beyond the conventional conceptual ER modeling capability. Preliminary research has begun to discuss the need for such a quality entity-relationship (QER) capability [16] and how to develop such a capability [17]. Much research is needed to convert these research ideas into deployable solutions.

As always, various types of data consumers will use the delivered information products, some at the very end of the information supply chain, i.e., the war fighters. Others include analysts in the intelligence community who make judgments and recommendations based on the information products they received. In so doing, they feed the information products into various decision tools, such as OLAP (on-line analytical processing) capabilities customarily used by business managers who have access to vast amount of data in data warehouses.

Other capabilities such as data mining, Bayesian Networks [18], and intelligent agent-oriented modeling are often discussed in the context of security of infrastructures, which involves the development of intelligent agents for monitoring the performance of networks of infrastructures and for communicating among themselves in order to avoid cascading effect of deliberate disruption of part of the system. These infrastructures include transportation, power, communication, water and energy.

To ***manage information as the product of a well-defined information production process***, it is essential to develop a mechanism for producing the Information Product Map (IPMap) [19]. Just like one needs a blueprint for an assembly line that produces a type of car (a physical product), it is logical to have an IPMap for the production of specified information products. Due to the unknown and unpredictable nature of information needed, however, capturing and managing an IPMap is more challenging than it appears to be [20]. Our research has found very few organizations that have IPMap for their information products, even for corporations whose main products are data. This is particularly evident when heterogeneous databases are involved. In the next section, we discuss one Air Force case in which an IPMap that overlays the physical supply chain for the fuel pump repair of a military aircraft is needed to identify the root causes of "dirty data", which leads to problems in requirements demand forecasting.

Current practices, in the absence of IPMap, typically use ER diagrams, Data Flow Diagrams (DFD), information systems diagrams (flow charts of which information system interfaces with which systems), etc. Although each of these methods has its merit, they are not sufficient in managing data quality. Making IPMap available is critical in many ways, for example data quality typically deteriorates when moving not within but across functional areas or organizational boundaries. Data quality inspection and improvement is another concept not captured explicitly in current methodologies. In short, much research and development is needed to advance this area.

To ***manage the life cycle of the information product***, it is critical to have the tools and techniques. We have developed and deployed two industrial-strength data quality assessment software tools for subjective and objective assessment of data quality [21]. By subjective we meant assessment by the stakeholders of an information product. For this, we have developed a software instrument called *Information Quality Assessment* (IQA) [4, 13, 22]. IQA has proven to be effective in assessing the stakeholders' evaluations of organizational data quality, which often are not congruent among the data collectors, custodians, and consumers (the 3 C's in our research). It also enables us to assess which data quality dimensions are more important than others for an information product. By objective we mean assessing data quality using the data integrity rules as proposed by Codd [23-26]. We have developed a data *Integrity Analyzer* (IA) software tool that can be embedded in the data quality management process (or TIAQ in the context of this paper) in practice. The Integrity Analyzer implements rigorously Codd's five data integrity rules coupled with Total Quality Management (TQM) principles such as control charts and the continuous improvement cycle [27-29].

IA and IQA are two well-researched software tools in meeting the difficult challenge of TIAQ. To develop a methodology for TIAQ, many other concepts, tools, and techniques must be developed, for example the underlying information infrastructure, software architecture, and data quality management tools. Some of our research offers direction in this endeavor [5], which we will elaborate later when we propose a software architecture and its components necessary for capturing quality about data in addition to data itself.

To ***appoint information product managers***, we first clarify that by appointing information product managers, we mean establish an entity that has the responsibility and authority to oversee the information product life cycle horizontally across functional areas of an organization or across organizational boundaries. This entity is charged with the responsibility such as proactively raising the awareness of the importance of managing information as a product, making business cases for information product management, and institutionalizing information product management in the organization [30].

It is important to realize that the management component is, if not more, important in achieving TIAQ. Without the proper personnel who are charged with the responsibility and authority, and equipped with the appropriate skill set and information infrastructure, data may be available but not accessible; data may be accessible, but not credible; and data may be credible but delivered timely, consistently, and completely and appropriately for the tasks at hand. The technical and management components must go hand-in-hand in the pursuit of TIA.

Cost-benefit analysis (cost justification), as it turns out, is a common theme not just for TIAQ but also for most of managerial activities. Business case analysis in good faith, therefore, is one of the early and continuous efforts that information product managers must be trained to do, in addition to other data quality and project management skills. Information product managers also need to understand that TIAQ is a journey that involves problem solving and continuous improvements. As such, they must be trained to initiate pilot projects, perform business case analysis based on the results from pilot projects, and scale the pilot projects up to implement throughout the organization (across various defense components for example). Conducting analysis of tangible and intangible benefits vs. direct, indirect, and opportunity costs is an inherent part of any project management, and no exception in achieving TIA. As the pilot projects are completed and best practices developed, training and education, policy and procedures, and organizational TIAQ institutionalization ensue.

We summarize in Table 1 some TIAQ requirements for achieving Total Information Awareness based on the above discussion. The ultimate capabilities and facilities necessary to sustain the operations for TIAQ will evolve as pilot projects are initiated, lessons learned, and scaled-up deployed. We have discussed various capabilities and facilities, at a high level, which are needed following the four principles of information product management as summarized in Table 1. In the next section, we illustrate concepts, tools and techniques, which may be applied to TIAQ. Specifically, we discuss why and how an IPMap would be useful in tracing the information supply chain, a U.S. Air Force project that we investigated, of a fuel pump, whose supply chain was problematic. We traced the information flow to determine if "dirty data" (data with poor quality) caused fuel pump shortage for the repair of engines, which in turn impact the mission capable rate (MCR) of military aircraft readiness.

Table 1: Information Product Management Requirements for the TIA Program

IP Management Principle	Capabilities and Facilities Required
Understand the war fighter's information needs	• Define Information Products • Perform QER Modeling • Support Information Consumer-oriented Tools, such as On-Line Analytic Processing (OLAP), Data mining, and Intelligent Agent Modeling with IQ
Manage information as the product of a well-defined information production process,	• Develop IPMap and the corresponding software tools to capture the IPMap metadata • Capture and manage data quality data about data • Develop Quality Database Management System (QDBMS) capable of performing SQL queries with IQ • Support Data Quality Judgment using Bayesian Networks or Utility theories
Manage the life cycle of the information product	• Conduct Information Quality Assessment (IQA) • Perform Integrity Analysis (IA) • Apply TQM tools and techniques such as SPC and Deming Cycle
Appoint IP managers	• Perform Cost Justification or Business Case Analysis • Initiate Pilot Projects • Conduct TIAQ Training and Education • Develop TIAQ Policy and Procedures • Institutionalize TIAQ throughout the organization

3. ILLUSTRATIVE CASE

Remanufacturing is a complex process, which involves repairing and refurbishing parts and products in the carcass. It has been a common practice for aircraft, railway locomotives, and heavy construction equipment. As landfill becomes a scarce resource, remanufacturing will undoubtedly be extended to other products and industries.

In this case study, we chose to trace the fuel booster pump, which had been identified to be the key problem for engine repair. The pump, as shown in Figure 1, plays a key role in the improvement of mission capability and aircraft flying hours.

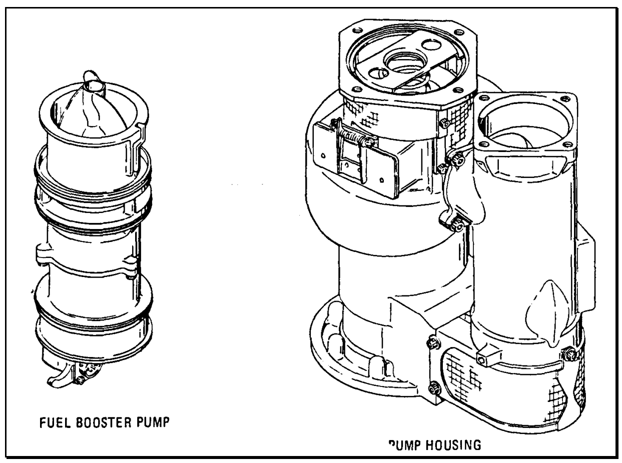


Figure 1: Fuel Booster Pump, Pump Housing.

A fighter aircraft is not ready unless its engine functions properly. An engine is not ready unless its pump works, which requires its stator installed. A stator is a system of stationary airfoils in the compressor of an aircraft fuel pump. An exploded view of the stator position in the fuel booster pump is shown in Figure 2.

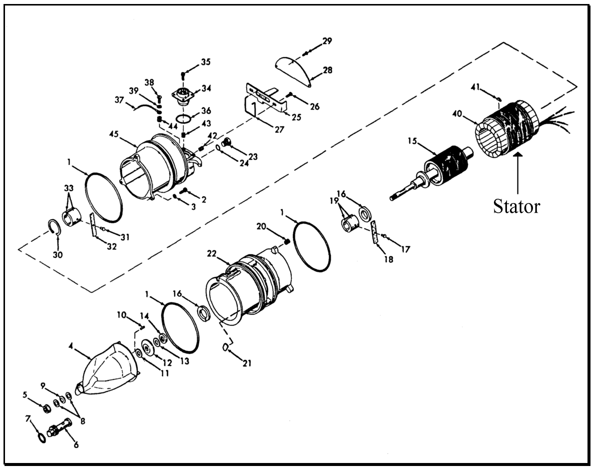


Figure 2: Stator position in the Fuel Booster Pump

Based on our initial field interviews and document review, we hypothesized that the root cause of problems with the fuel pump is the stator. Although inexpensive and easy to install, shortage of each stator means one less military aircraft that is mission capable. Therefore, we also traced the stator flow in detail. In so doing, we documented the work roles related to the pump and stator. We aim to show a clear picture of the entire remanufacture process of an aircraft fuel pump and the nature and various roles of information involved in the process. We will then be able to determine the required information products for this process. Tracing one specific part helps to understand the information needs in the remanufacture process. We refer to this as “one data element at a time” since an aircraft is composed of hundreds of thousands of parts and subparts, each identified by a National Stock Numbers (NSN), and tracked by numerous heterogeneous databases both in the Air Force and other DoD agencies such as DLA. Our particular purpose in mapping out the remanufacture process is to understand the characteristics of any discrepancies between the available information and information needs. This enables us to determine the required characteristics of information that should be embedded in the information for remanufacturing.

Sources of “Dirty” Data

Based on our initial observations from field interviews, two related areas need further investigation in order to identify the sources of poor quality data. One is the area of obtaining quality data for effectively predicting the need for parts; and the other is the area of effectively providing and recording work activities performed on the parts and in other remanufacture processes.

Unlike initial manufacture where all of the parts needed to assemble a product are known well in advance, remanufacture has far less predictability. Instead, remanufacture involves many unscheduled, variable, and evolving activities. Much of the uncertainty in the process stems from the fact that there are two possible supply chains. One is similar to that found in the initial manufacture, in which new parts are fabricated and delivered by suppliers internal or external to the organization. Unlike initial manufacture, however, this is not the only source of parts. A second “supply line” delivers the parts that are contained in the “carcass,” or product that is to be repaired. Particularly, the unpredictable quality of the parts contained in the carcass that is the major source of uncertainty. Not knowing whether the parts delivered from the carcass are workable or not makes the need for additional parts through the normal supply chain unpredictable. It follows that one way to reduce this uncertainty is to find better ways to predict the state of the parts contained in the carcass. The predictive capability of these models is, of course, highly dependent upon the quality of the stored information. It is the interaction of the two supply chains that makes remanufacture the complex setting. This is where understanding the information process becomes critical.

Our interviews with those involved in the overhaul process led us to hypothesize that much is needed in order to improve Requirements Determination Forecast (RDF). The principal complaint voiced in interviews had to do with DQ: the RDF models were unable to accurately predict parts needs because the data on which they are based is faulty and questionable – referred to as the “dirty data”. Where and how did these errors enter into the information process? Some had suspicions about how past demands and future predictions were calculated, to state a few. These suspicions had never been verified for their validity.

Pump repair is performed in two geographically dispersed places: depots and fields. Depots conduct regular scheduled overhaul, whereas fields handle surprises, the immediate problems at hand. Throughout the repair process, information about the work process and the physical products or parts are isolated from each other by operational procedures. The direct impact is that it becomes difficult to connect the two kinds of information. One needs the ability to retrieve and understand the physical parts as well as the work information. For example, when a pump needs to be repaired, the repair history is not easily accessible. As such, the repair history, the supply information, and the conformance-testing information is stored and used separately. Most of the relevant information is stored and categorized meticulously, but without consideration of cross-area retrieval and access. The connection between physical parts and process information part is missing. Currently, one has to contact multiple agents and places over phone and email to track down the information needed to make this connection. An integrated pump and stator flows is depicted in Figure 3 [31].

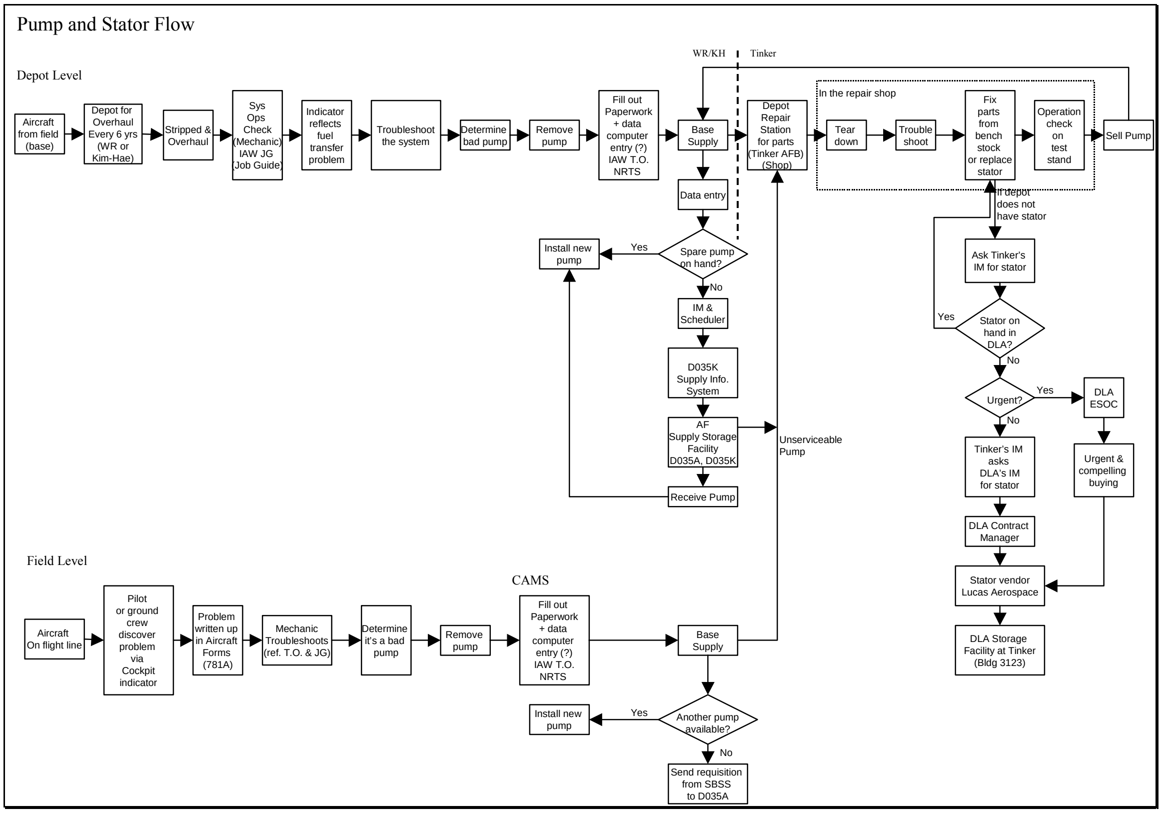


Figure 3: Integrated Pump and Stator Flows

We observed yet another area for improvement. A supply vendor initially produced the engineering specification of the stator, for the pump. We encountered some opinions that the engineering specifications were not consistent with the stators manufactured and delivered to the Air Force. After going through revisions, the updated specification document and drawings were not stored most effectively by the Air Force and the vendor. In the process, different people could develop different understandings of what the official engineering specifications for a stator should be. Meanwhile, the aircraft has to fly. Some "work-arounds" might have been performed to meet the demand using parts of questionable quality. In short, the lack of management of data products (in this case, the blue prints of the stator) led to a possible lack of quality in physical products. We hypothesize that changes in engineering specifications over time have been poorly communicated between the Air Force and vendors in terms of specific design problems and resolutions.

Discussion

Viewing information as a product implies two essential information management requirements. For historical and future use requirements, information must be stored and protected against undesired change. For current use, information must be kept as current as possible. Information stored in databases is typically safeguarded to preserve these two aspects of quality among others.

This case illustrates that in a complex system such as an aircraft repair that involves hundreds of thousands of parts, a data warehouse with thousands of data elements, or the war against terrorism, it is very difficult to ensure that the quality of the data used by data consumers unless a systematic discipline is developed to deal with the quality of the information. Indeed, in many a case that we have studied, it is common to hear comments such as, "I don't like the data, but that's the best I can get." It does not have to be like that. By fundamentally treating information as a product instead of a by-product from various source systems, data consumers will be better served. IPMap is a blueprint of how the raw data came into the information manufacturing system, how it flows through each of the job shops (an information system), and eventually how it is packaged to become an information product for delivery to be applied in some decision making or security analysis activities. Little research has been conducted to provide the facilities necessary to capture the pertinent information as shown in Figure 3 with computer-aided software engineering (CASE) tools, which will provide a friendly user interface for information product managers to develop an information product map (IPMap) given the specification of an information product (which may be dynamically re-configurated). This illustrative pump case also demonstrates the needs for the capabilities and facilities as summarized in Table 1.

4. TOWARDS A RESEARCH PROTOTYPE

As always, in developing capabilities for deployable industrial-strength applications, it would be wise to leverage Commercial of the shelf (COTS) technologies to the maximum extent as possible. In Table 1, IA and IQA are two example COTS tools that can be applied to the TIAQ context. There are other applications that can be leveraged towards the development of a demonstrable research prototype for TIAQ. For example, conceivably, MSBNx, A Component-Centric Toolkit for Modeling and Inference with Bayesian Networks [18], may well be useful to help an analyst in automatically filter a vast amount of possibly important data into the most relevant, credible, and timely information for his or her task at hand.

On the other hand, there are some software architecture, system components, and fundamental database issues that may require much research effort. In this section, we illustate two inter-related research activities that need to be further developed to be deployable in practice.

IPMap

We sumarize research that we have performed in [6, 19] in this sub-section. The modeling constructs in the IPMap consist of eight types of construct blocks, as summarized in Table 2. Each construct block is identified by a unique and non-null name. Each construct block is described by a set of attributes. The composition of this set varies depending on the type of construct block it is associated with.

Table 2: IPMap Building Blocks

SYMBOL	REPRESENTS
<Name>	**Data Source / Data Vendor / Point-of-Origin**
< Process Identifier>	**Process**
<Storage Name>	**Data / Information Storage**
<Criteria>	**Decision**
Quality Criteria	**Quality / Evaluation / Check**
< Current / New System Names >	**Information System Boundary -** used when a data unit (raw data, component data) changes from one system (paper or computerized) to another (paper or computerized**)**
<Current / New Org/dept names>	**Organizational Boundary -** used when a data unit (raw, component) moves across departments or across organizations
<Name>	**Data Sink / Consumer Block / Point-of-Destination.**

In the preceding section we described the constructs that are useful in creating a conceptual modeling method to represent the manufacture of an IP. We use an example to illustrate how these constructs are used in defining the IPMap. Let us examine some important reports (IPs) typically produced in a hospital. For the purpose of this illustration we will consider a small subset of the operations (and processes) of a major hospital including only the in-patient admissions, treatment, and discharge sections. There are five products associated with these operations. All five use information that is gathered from two important sources: the patient and the team of hospital employees (doctors, nurses, lab technicians, radiologists, therapists, and administrative staff) involved (directly or indirectly) in the admission, treatment, or discharge of the patient. Each uses a subset of the large set of information. The first product (IP_1) is the *admissions report* submitted to the management of the hospital on a daily, weekly, and monthly basis. It provides a description of the number of patients admitted, expected duration of stay, along with patient information and serves as a monitoring instrument that reflects how busy the units are. The second product (IP_2) is the *patient treatment report* generated on a daily basis and appended to the patient's chart. Care providers (nurses/doctors) use it to monitor the patient's response(s) to treatments and procedures administered. These two are information products used internally within the hospital. The final three products are sent to external agencies. The *birth/death report* (IP_3) is submitted to the registry of vital statistics, and the *health report* (IP_4) is a bi-annual report required by the department of public health on the types of patients treated and released, ailments, treatments, and the reason for discharge. The final product (IP_5) is the *patient bill* submitted to the HMOs for payment. This is an itemized list of services, equipment charges (if any), medications, tests, and procedures provided to the patient.

The IPMap representing the manufacture of the patient admission report (IP_1) is shown in Figure 4. An inpatient may be admitted at any one of three locations: the admissions office, emergency room, or in the department of maternal and fetal medicine.

The patient (or an accompanying adult) provides the patient information (raw data RD_1 from data source DS_1) by completing a form. The admissions clerk enters this data into the Patient Medical Office System using a form-based interface (process P_1). In this process the data changes from a paper-based system to an electronic system shown by the system boundary block SB_1. The software module associated with the interface checks the form for completeness, and verifies the guarantor/HMO and this check is shown as QB_1. The raw data elements examined along with the authorization is sent for storage and is shown by the component data CD_1.

Upon admission, the ward nurse responsible for admitting the patient assigns a bed number that specifies the type of ward (cardiac ICU, general, etc.) and also examines the general condition and disposition of the patient. The nurse (treated as two data sources DS_3 and DS_4 as the two tasks may be done by more than one nurse) records this information (RD_3 and RD_4) on the chart and subsequently enters it into the computer system (process blocks P_3 and P_4). As the underlying system changes a system boundary block (SB_3) is shown to represent this change. The patient's past medical records (source block DS_2) is obtained and the information (RD_2) is used to update the patient's medical record in the system (process P_2). The records are verified to ensure that they come from the right source authorized by the patient and, if necessary, the information is verified with the doctor/medical office that created the record. Quality block QB_2 represents this check. The resulting component data (CD_2) is then sent for storage. All of this information is captured in the data storage of the medical office system shown by the storage block STO_1. The information product IP_1, generated by process P_5, uses a collection of component data items cumulatively identified as CD_5. It is sent to the hospital management as shown by the consumer block CB_1.

Once the admission is complete, a record of the treatments / procedures recommended and performed on the patient is created as shown by the IPMap in Figure 5. The specialists and attending physicians (data sources DS_7 and DS_8) recommend the course of treatment and procedures/tests to be performed. This information is then recorded (RD_7) on the charts. Prior to its capture, it is verified by the attending physicians and modified (if needed) in consultation with the specialist. The quality block QB_4 represents this check. The resulting authorized treatments/procedure information (CD_5) is captured in the computer system by process P_8. The attending physicians also report on the progress of the patient and sign off on the recommended treatments/procedures completed as shown by RD_8 which is captured in the system by process P_9. The change of system from paper-based to electronic is represented by SB_5. The reports from the labs and radiology (data source DS_5) are collected and the information (RD_5) is entered into the computer. The change in system is represented by SB_4. Process P_6 captures this and a module in this process verifies the source of the report as well. The component data CD_6 generated by P_6 is matched with the patient's record shown by QB_4 and sent for storage.

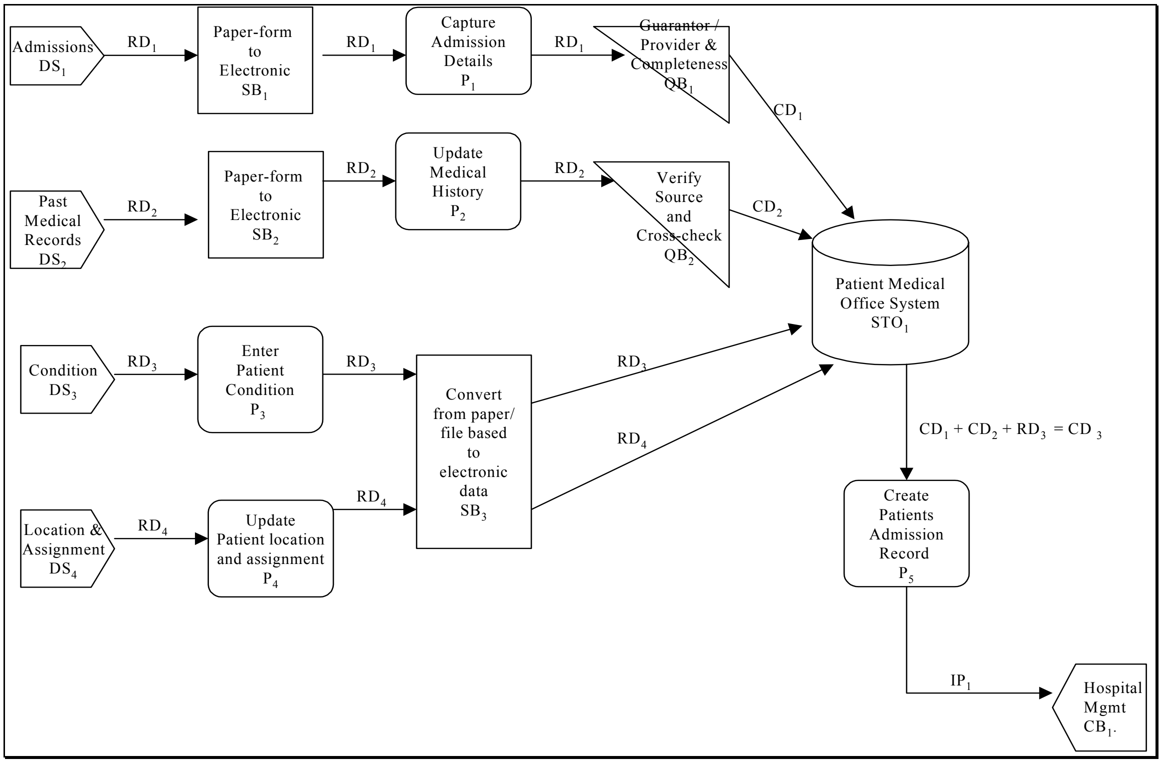


Figure 4: IPMap for the Patients Admission Record

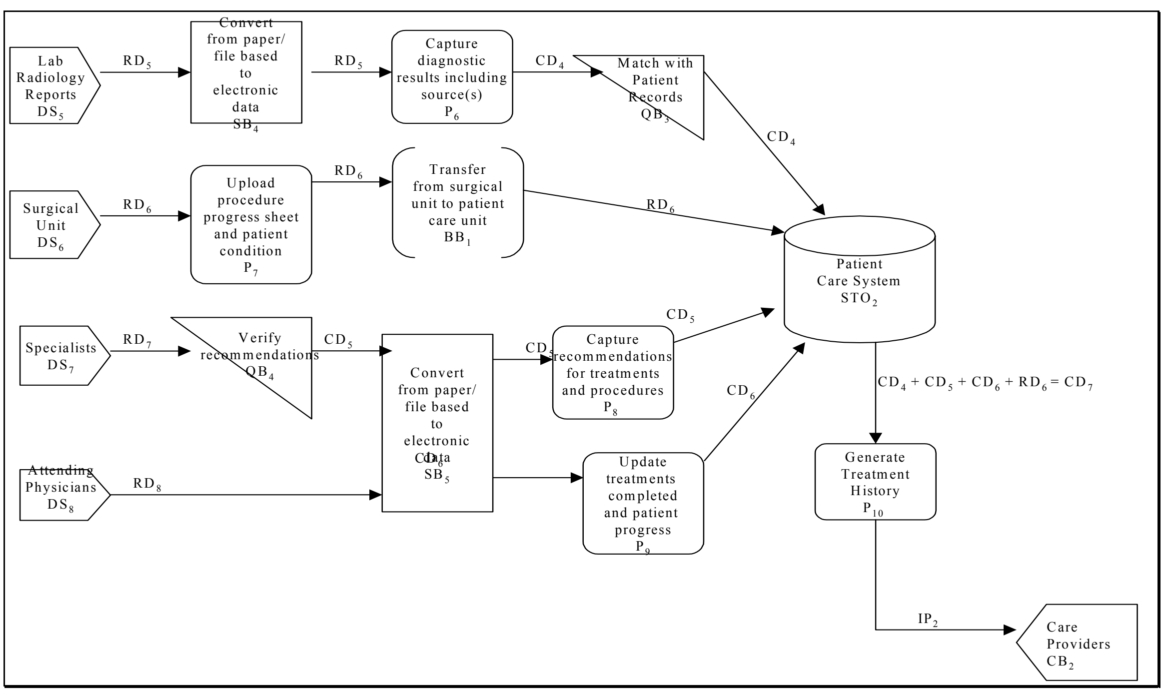


Figure 5: IPMap for Patient's Treatment History

The comments and reports from the surgical unit (different from the patient care facility) are electronically uploaded by process P_7. The business boundary block BB_1 represents the transfer of information across business units. The storage location for all the above information is the Patient Care System database shown by storage block STO_2. The treatment report (IP_2) is created by process P_{10} and sent to care givers (customer block CB_2).

The manufacture of the information products IP_3, IP_4, and IP_5 is represented by the IP-MAP in Figure 6. The information in the admissions office system and the patient care system is uploaded into the administrative system by processes P_{11} and P_{12}. The records from each are matched to ensure that the right admission is combined with the right treatment (shown by quality block QB_6) and the resulting component data CD_{10} is stored in the administrative system database represented by STO_3. As all three systems are different, we need to show the change in the underlying system during this transfer. We also need to capture the fact that the information changes business boundaries as well. We use the combined system and business boundary blocks BSB_1 and BSB_2 to represent the transfer. Process P_{13} generates the report on vital statistics (IP_3) which is sent to the consumer (CB_3) and the Registry of Vital Statistics. Processes P_{14} and P_{15} generate the hospital health report (IP_4) and the patient bill (IP_5) respectively. The state department of health (CB_4) and the HMO (CB_5) are the consumers of the two information products in that order. For each of the three products, the set of data items used to generate each is different and is shown by the component data items CD_{11}, CD_{12}, and CD_{13}.

To complete the representation, we need to capture the information about each of the blocks and the data elements included in each flow in the model(s) above. This is akin to the data dictionary for a data flow diagram and we refer to this as the metadata associated with the model. The metadata is captured in a repository; the complete metadata for the above model is too large for this paper and therefore only a sample is shown in Table 3. We have illustrated some of the key concepts for the development of a user-freindly IPMap tool. Next we discuss how to facilitate data quality judgment automatically via a data quality reasoner.

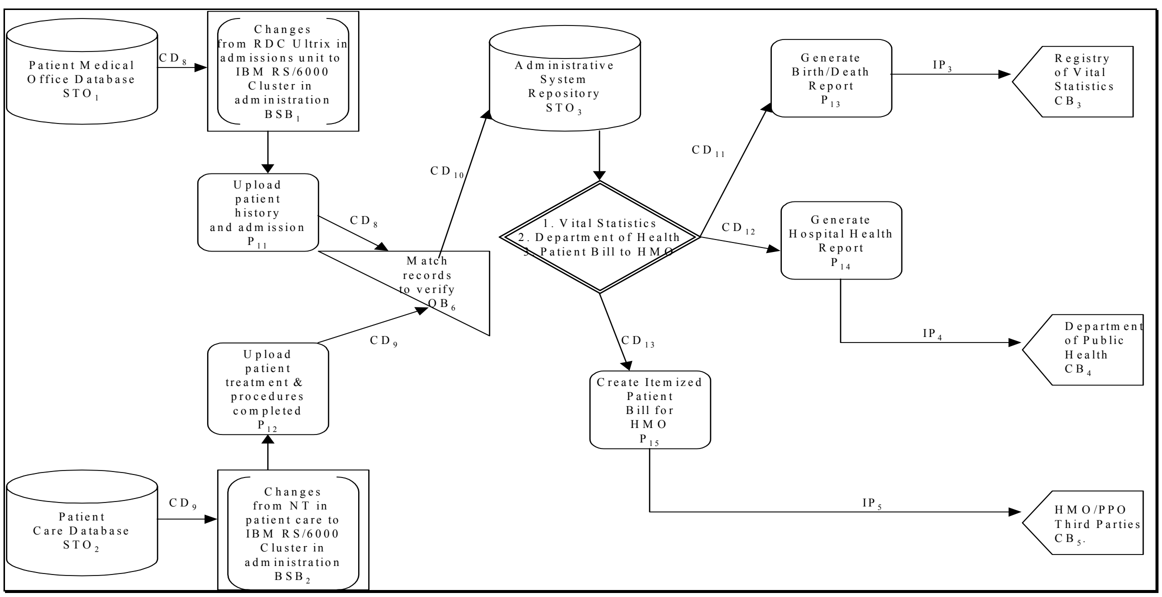


Figure 6: IPMap for Vital Statistics Report, Hospital Health Report, and Bill

Table 3: Sample metadata for IPMap in Figure 4

Name/Type	*Department/Role*	*Location*	*Business Process*	*Composed Of*	*Base System*
Admissions/DS_1	Admissions Office/ Patient	Admissions, OB/GYN, Emergency	Standard Form (#1101P)		Paper-based - Patient File
Past Medical Records / DS_2	Admissions Office / Admissions clerk	Admissions Bldg., Records Room	Contact source and request with patient authorization.		Paper-based - patient file

Data Quality Reasoner

We summarize in this sub-section research that we have performed in [32-35]. In a simple homogeneous database management system environment, data consumers are generally knowledgeable about the characteristics of the data they use. As such, DQ is handled through personal familiarity. As the integration of information systems has enabled data consumers to gain access to not only familiar but also unfamiliar data sources, however, such an informal personal-familiarity-based approach becomes increasingly ineffective, often resulting in serious economic and social damage. This problem clearly exists in the TIA context where an analyst must deal with data from various sources. To reduce such damage, there has been growing interest and activity in the area of data quality maintenance and judgment.

Research has been conducted to provide data consumers with "meta-data," i.e., data about data that an facilitate the judgment of data quality such as data source, creation time, and collection method [36-38]. We refer to such characteristics of the "data manufacturing process" as *data quality indicators*, or simply *quality indicators*. Table 4 shows several examples of quality indicators.

Table 4: Data Quality Indicators

Data Quality Indicator	data #1	data #2	data #3
Source	DB#1	DB#2	DB#3
Creation-time	6/11/92	6/9/92	6/2/92
Update-frequency	daily	weekly	monthly
Collection-method	bar code	entry clerk	radio freq.

Data-quality judgment, however, has generally been left to data consumers. Unfortunately, the vast amount of data such as those in the TIA context would make it difficult to analyze such data and draw useful conclusions about overall data quality for subsequent analysis, recommendation and decision activities. In an attempt to assist data consumers in judging if the "quality" of data meets a data consumer's requirements, we propose a framework for deriving an overall data quality level from various local relationships between the factors, which affect data quality [33, 35]. We focus on the problem of assessing levels of data quality, *i.e.*, the degree to which data characteristics match those required by a data consumer. Even if each individual data supplier were to guarantee the integrity and consistency of data, data from different suppliers may still be of different quality levels due, for example, to different data maintenance policies. As a result, data consumers are increasingly forced to make decisions based on a set of data instances of varying levels of data quality [8]. A systematic way of assessing data quality levels is thus needed by data consumers. Toward this end, this paper investigates a mechanism which can generate an insightful answer to the question of "What would be a level of data quality for a given data instance?"

In considering the data quality assessment problem, our analysis has identified several related theoretical and practical issues:

1) What are data quality requirements?
2) How can relationships between these requirements be represented?
3) What information about overall data quality can one derive from such relationships, and how?

Studies on data quality requirements, such as accuracy and timeliness, can be found in [4, 6, 22, 39]. Such data quality requirements are referred to as *data quality parameters*, or simply *quality parameters* in this section. Table 5 shows some examples of data parameters.

Table 5: Data Quality Parameters

DQy Parameter	data #1	data #2	data #3
Credibility	High	Medium	Medium
Timeliness	High	Low	Low
Accuracy	High	Medium	Medium

The essential distinction between quality indicators and quality parameters is that quality indicators are intended primarily to represent objective information about the data manufacturing process, such data creation time, while quality parameters can be user- or application-specific, and are derived from either underlying quality indicators or other quality parameters.

In this discussion, we assume that there is a "quality parameter hierarchy," where a single quality parameter is derived from *n* underlying quality parameters. Each underlying quality parameter, in turn, could be derived from other underlying quality parameters or quality indicators. For an intuitive understanding, let us consider the hierarchy shown in Figure 7, which shows that a user may conceptualize quality parameter Credibility as depending on underlying quality parameters such as Timeliness and Source-reputation. The quality parameter Source-reputation, in turn, can be derived from quality indicators such as the frequency with which a source supplies obsolete data. For the purpose of this presentation, we assume that such derivations are given and complete, and thus relevant quality parameter values are always available.

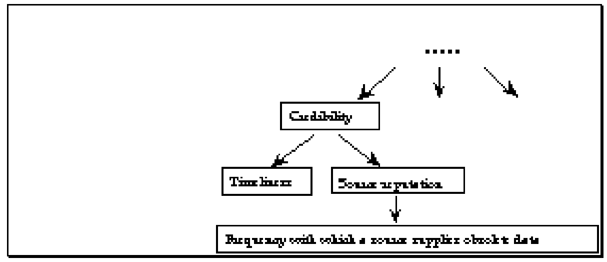


Figure 7: Example quality parameter hierarchy

We introduce a *data quality calculus,* which is a simple data quality judgment model based on the notion of a "census of needs." The intention is to provide flexibility in dealing with the subjective, decision-analytic nature of data quality judgments. The data quality calculus provides a framework for representing and reasoning with basic relationships among quality parameters, and deriving overall data quality levels.

In general, several quality parameters may be involved in determining overall data quality. This raises the issue of how to specify the degree to which each quality parameter contributes to overall data quality. One approach is to specify the degree, in certain absolute terms, for each quality parameter. It may not, however, be practical to completely specify such values. Instead, people often conceptualize "qualified" local relationships, such as "Timeliness is more important than the credibility of a source for this data, except when timeless is low." So that, if timeliness is high and Source-credibility is medium, the data may still be of high quality. The data quality calculus provides a formal specification of such local "dominance relationships" between quality parameters.

Since each local relationship between quality parameters specifies the local relative significance of quality parameters, one way to use local dominance relationships would be to rank and enumerate quality parameters in the order of implied significance. Finding a total ordering of quality parameters that is consistent with local relative significance, however, can be computationally intensive. In addition, a complete enumeration of quality parameters may contain too much raw information to convey any insights about overall data quality. Instead of trying to find a total ordering of quality parameters, the data quality calculus attempts to infer the overall data quality implied by the local relationships between quality parameters.

The *data quality reasoner* (DQR) is a data quality judgment model [35] that derives an overall data quality value for a particular data element, based on the following information:

1. A set QP of underlying quality parameters qi that affect data quality: QP = $\{q_1, q_2, ..., q_n\}$.
2. A set DR of local dominance relationships between quality parameters in QP.

For any quality parameter q_i, let V_i denote the set of values that q_i can take. In addition, the following notation is used to describe value assignments for quality parameters. For any quality parameter q_i, the value assignment $q_i := v$ (for example, Timeliness := High) represents the instantiation of the value of q_i as v, for some v in V_i. Value assignments for quality parameters, such as $q_i := v$, are called "quality-parameter value assignments". A quality parameter that has a particular value assigned to it is referred to as an *instantiated* quality parameter.

For some quality parameters $q_1, q_2, ..., q_n$, and for some integer $n \leq 1$, $q_1.q_2.....q_n$ represents a conjunction of quality parameters. Similarly, $q_1:=v_1.q_2:=v_2.....q_n:=v_n$, for some v_i in V_i, for all i = 1,2, ... , n, represents a conjunction of quality-parameter value assignments. The notation '⊕' is used to state that data quality is affected by quality parameters. It is represented as $\oplus(q_1.q_2.....q_n)$ to mean that data quality is affected by quality parameters $q_1, q_2, ...,$ and q_n. The statement $\oplus(q_1.q_2.....q_n)$ is called a *quality-merge statement*, and is read as "the quality merge of $q_1, q_2, ...,$ *and* q_n." The simpler notation, $\oplus(q_1, q_2, ..., q_n)$ can also be used. A quality-merge statement is said to be instantiated, if all quality parameters in a quality-merge statement are instantiated to certain values. For example, statement $\oplus(q_1:=v_1.q_2:=v_2.....q_n:=v_n)$ is an instantiated quality-merge statement of $\oplus(q_1, q_2, ..., q_n)$, for some v_i in V_i and for all i = 1, 2, ..., n.

A first-order data quality reasoner that guarantees the well-defined reduction of quality-merge statements has been proposed in [32-35]. The reasoner requires that the dominance relation be transitive. This implies that for any conjunctions of quality-parameter value assignments, *E1, E2,* and *E3,* if $E_1 >_d E_2$ and $E_2 >_d E_3$, then $E_1 >_d E_3$. Transitivity of the dominance relation implies the need for an algorithm to verify that, when presented with an instance of the quality-estimating problem ($\oplus(q_1, q_2, ..., q_n)$, DR), dominance relationships in DR do not conflict with each other. Well-known graph algorithms can be used for performing this check. Quality-merge statements in Figure 8 can be classified into groups, with respect to levels of the reducibility.

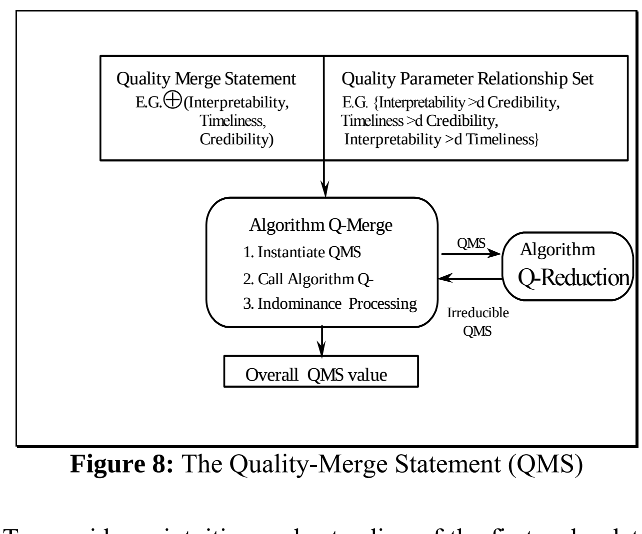


Figure 8: The Quality-Merge Statement (QMS)

To provide an intuitive understanding of the first-order data quality reasoner, we show how the first-order data quality calculus can be applied to compute overall data quality with respect to the believability of data. Let A:v (for example, Profit:$3.9 million) denote a data instance of an attribute A.

Suppose that management of a hypothetical firm XYZ is going over the profit figures shown in Table 6, in order to allocate resources for the coming year, and is wondering if these figures are believable enough to make decisions based on them.

Table 6: Hypothetical Profit Records

Business-Unit	Profit	...
BU1	$1.2 million	...
BU2	$1.4 million	...
BU3	$3.9 million	...
BU4	$1.5 million	...
BU5	$2.0 million	...
BU6	$1.8 million	...
BU7	$1.5 million	...
BU8	$1.9 million	...
BU9	$2.1 million	...

One factor that affects believability of data is time. The effect of time on believability, in general varies from data instance to data instance. For some data instances, believability is static over time. For example, a person's gender is hardly affected by the age of this data. For other data instances, however, believability can change dynamically over time. Consider, for example, a person's age. If a person's age is 40 years in a database, this age will no longer be valid a year later, but is highly likely to be valid a day after the data is entered. A temporal effect, denoted by Temporal-effect, can be a quality parameter that affects data believability.

Another factor, which can affect the believability of data, is the credibility of the source that provided the data. For example, an address provided by the Internal Revenue Service is more believable than one supplied by a mail-order firm. Let Credibility denote the credibility of a data originator. In addition to a temporal effect and the credibility of a data source, the degree to which the semantics of an attribute are consistent can affect believability of data. Let Semantic-consistency denote the quality parameter that indicates how consistently the semantics of an attribute are captured.

Suppose that Temporal-effect can take on one of Tolerable, Moderate, Intolerable, and that Credibility and Semantic-consistency can take on one of the following values: High, Moderate, and Low. While mechanisms are needed for determining, data instance by data instance, values of these quality parameters, this paper assumes that the values for these quality parameters are available for use. In addition, even though other factors such as fidelity or consistency of a data source can also affect believability of data, for expository simplicity, let us suppose that the profit believability is affected only by the credibility of the data source, a temporal effect, and the Semantic-consistency of attribute Profit: In other words,

QP = {Temporal-effect, Credibility, Semantic-consistency}.

Given QP, the first-order data quality reasoner can infer that overall believability is the result of quality merge of Temporal-effect, Credibility, and Semantic-consistency, *i.e.,* $\oplus$(Temporal-effect, Credibility, Semantic-consistency). For the purposes of this example, we assume that Semantic-consistency dominates Credibility, Temporal-effect dominates Credibility, and Semantic-consistency dominates Temporal-effect. In other words, DR consists of the following dominance relationships between Credibility, Temporal-effect, and Semantic-consistency:

Semantic-consistency := $v_1 >_d$ Credibility:= v_2, for all $v1$ and v_2 in {High, Moderate, Low},

Temporal-effect:= $v_1 >_d$ Credibility:= v_2, for all v_1 in {Tolerable, Moderate, Intolerable} and for all v_2 in {High, Moderate, Low},

Semantic-consistency:= $v_1 >_d$ Temporal-Effect:= v_2, for all v_1 in {High, Moderate, Low} and for all v_2 in {Tolerable, Moderate, Intolerable}

We exemplify how the reasoner can compute believability values of data instances based on information available in QP and DR. Let Symbol B denote the believability of data. Suppose that management of XYZ is concerned about the believability of data instance Profit:$3.9 million. A query to the data quality reasoner can be made in a form similar to the following: Data consumer = XYZ; Quality dimension = B; Data of concern = Profit:$3.9 million; Subject to source = BU3. Given this query, the reasoner computes the believability value for the data instance Profit:$3.9 million. Suppose that for data instance Profit:$3.9 million, the temporal effect is intolerable, the credibility of BU3 is low, and the semantics of Profit in BU3 are captured with high consistency, in other words, Temporal-effect:=Intolerable; Credibility:=Low; Semantic-consistency:=High. Figure 9 attempts to graphically illustrate how the query made by XYZ can be processed. Given the query for the believability of data instance Profit:$3.9 million, the data quality reasoner infers, given QP, that the believability of this data instance is determined by the temporal effect, credibility of BU3, and the semantics of Profit captured by BU3. The first-order data quality reasoner returns High as the believability value of Profit:$3.9 million (which is rather counterintuitive).

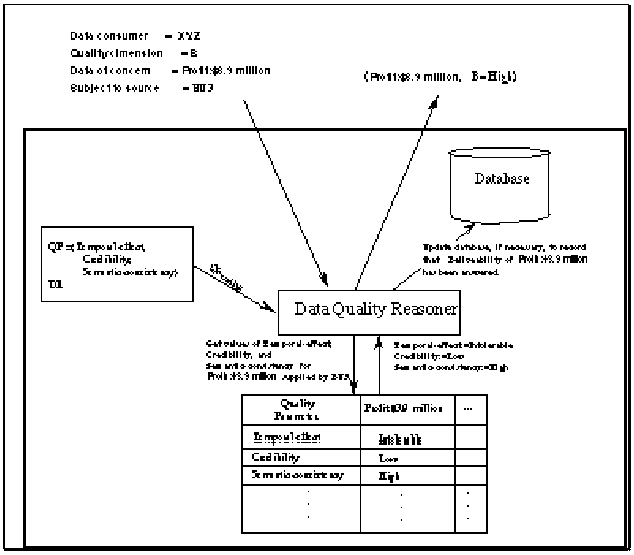


Figure 9: Data flow to generate an answer to a query.
(A solid arrowed line represents flow of data)

The following describes how the reasoner concludes that the believability value of Profit:$3.9 million is High. In computing the believability of data instance Profit:$3.9 million, the reasoner reduces the instantiated quality-merge statement as follows:

1. $\Omega \leftarrow$ {Temporal-effect:=Intolerable, Credibility:=Low, Semantic-consistency:=High}
2. $\Omega \leftarrow \Omega -$ {Credibility:=Low}, since dominance relations of Semantic-consistency over Credibility and of Temporal-effect over Credibility asserted in DR. Now, Ω ={Temporal-effect:=Low, Semantic-consistency:=High}.
3. $\Omega \leftarrow \Omega -$ {Temporal-effect:=Low}, since the dominance relation of Semantic-consistency over Temporal-effect is found in DR. Ω then consists of only one element Semantic-consistency:=High.
4. Return the value of Semantic-consistency as the reduction-based value of the instantiated quality-merge statement$\oplus$(Temporal-effect:=Intolerable, Credibility:=Low, Semantic-consistency:=High).

In other words, for data instance Profit:$3.9 million, the instantiated quality-merge statement $\oplus$(Temporal-effect:=Intolerable, Credibility:=Low, Semantic-consistency:=High) is reducible to $\oplus$(Semantic-consistency:=High), even when the value of Temporal-effect is Intolerable and that of Credibility is Low. This reduction implies that XYZ is only concerned about how consistently the semantics of Profit are captured by data suppliers. As a result, for data instance Profit:$3.9 million, the data quality reasoner returns to XYZ the believability value of High. A more complex set of relationships between quality parameters would likely yield a different believability value. It is worth noting that no changes are needed to the data quality reasoner itself if a different set of relationships are presented. All that changes is the input set DR. It is precisely this fact that allows easy "tailoring" of our methods to the needs of specific data consumers.

CONCLUDING REMARKS

We have illustrated two research components that would be useful in achieving TIAQ. Due to time and space constraint, we have not presented many other components such as the underlying Quality Database Management System (QDBMS) component [38]. In any case, much research is needed to advance the research into deployable TIAQ toolkit and methodology.

More importantly, based on four principles of managing information as a product, we have proposed a set of capabilities and facilities required to facilitate Total Information Awareness with Quality (TIAQ). These capabilities will enable us to deal with the inherent trade offs must be made between dimensions of DQ and vast amount of data. We envision that TIAQ software tools need to be developed with a web-based, front-end client intelligent agent and data quality judgment facilities that would made use of the information products tagged with data quality information such as how timely and credible the information is. Concurrently, we envision a back-end server based on an innovative quality database management system (QDBMS) capable of storing not only data per se but the data-quality data, as well as the query language and corresponding data quality management utilities for managing the life cycle of information products. The conventional ER conceptual modeling work must also be extended to incorporate data quality attributes, which will capture data quality values in the underlying QDBMS.

REFERENCES

[1] Y. Wand and R. Y. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM*, vol. 39, pp. 86-95, 1996.

[2] Y. R. Wang and S. E. Madnick, "A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective," presented at the 16th International Conference on Very Large Data bases (VLDB), Brisbane, Australia, 1990.

[3] R. Y. Wang, V. C. Storey, and C. P. Firth, "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, pp. 623-640, 1995.

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

[5] R. Wang, M. Ziad, and Y. Lee, *Data Quality*. Norwell, Massachusetts: Kluwer Academic Publishers, 2001.

[6] D. P. Ballou, R. Y. Wang, H. Pazer, and G. K. Tayi, "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science*, vol. 44, pp. 462-484, 1998.

[7] R. Y. Wang, "Proceedings of the 1996 Conference on Information Quality." Cambridge, MA: Total Data Quality Management Program, 1996, pp. 270.

[8] D. M. Strong and B. K.Kahn, "Proceedings of the 1997 Conference on Information Quality." Cambridge, MA: Total Data Quality Management Program, 1997, pp. 372.

[9] Y. W. Lee and G. K. Tayi, "Proceedings of the 1999 Conference on Information Quality," Cambridge, MA, 1999.

[10] I. Chengalur-Smith and L. L. Pipino, "Proceedings of the 1998 Conference on Information Quality," Cambridge, MA, 1998.

[11] B. Klein and D. Rossin, "Proceedings of the 1999 Conference on Information Quality," Cambridge, MA, 2000.

[12] W. W. Eckerson, "Data Quality and The Bottom Line: Achieving Business Success Through a Commitment to High Quality Data," The Data Warehousing Institute, Seattle 2002 2002.

[13] R. Y. Wang, "A Product Perspective on Total Data Quality Management," *Communications of the ACM*, vol. 41, pp. 58-65, 1998.

[14] R. Y. Wang, Y. L. Lee, L. Pipino, and D. M. Strong, "Manage Your Information as a Product," *Sloan Management Review*, vol. 39, pp. 95-105, 1998.

[15] P. P. Chen, "The Entity-Relationship Model - Toward a Unified View of Data," *ACM Transactions on Database Systems*, vol. 1, pp. 166-193, 1976.

[16] P. S. Chen, "The Entity-Relationship Approach," in *Information Technology in Action: Trends and Perspectives*, R. Y. Wang, Ed. Englewood Cliffs: Prentice Hall, 1993, pp. 13-36.

[17] V. C. Storey and R. Y. Wang, "Modeling Quality Requirements in Conceptual Database Design," presented at the 1998 Conference on Information Quality, Massachusetts Institute of Technology, 1998.

[18] C. M. Kadie, D. Hovel, and E. Horvitz, "MSBNx: A Component-Centric Toolkit for Modeling and Inference with Bayesian Networks," Microsoft Research, Richmond, WA, Technical Report MSR-TR-2001-67, 28 July 2001.

[19] G. Shankaranarayan, R. Y. Wang, and M. Ziad, "Modeling the Manufacture of an Information Product with IP-MAP," presented at The 5th International Conference on Information Quality, Massachusetts Institute of Technology, 2000.

[20] Y. Lee, T. Allen, and R. Wang, "Information Products for Remanufacturing: Tracing the Repair of an Aircraft Fuel-Pump," presented at Sixth International Conference on Information Quality, Cambridge, MA, 2001.

[21] L. Pipino, Y. Lee, and R. Wang, "Data Quality Assessment," *Communications of ACM, forthcoming*, vol. April, 2002.

[22] K. Huang, Y. Lee, and R. Wang, *Quality Information and Knowledge*. Upper Saddle River: N.J.: Prentice Hall, 1999.

[23] E. F. Codd, "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM*, vol. 13, pp. 377-387, 1970.

[24] E. F. Codd, " Relational database: A Practical Foundation for Productivity, the 1981 ACM Turing Award Lecture," *Communications of the ACM*, vol. 25, pp. 109-117, 1982.

[25] E. F. Codd, *The Relational Model for Database Management: Version 2*. Reading, MA: Addison-Wesley, 1990.

[26] E. F. Codd, "An evaluation scheme for database management systems that are claimed to be relational," presented at the Second International Conference on Data Engineering, Los Angeles, CA, 1986.

[27] J. Funk, Y. Lee, and R. Wang, "Institutionalizing Information Quality Practice," presented at Conference on Information Quality, Cambridge, MA, 1998.

[28] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: A Methodology for Information Quality Assessment," *Forthcoming, Information & Management*, 2002.

[29] E. W. Deming, *Out of the Crisis*. Cambridge, MA: Center for Advanced Engineering Study, MIT, 1986.

[30] R. Wang, J. Funk, Y. Lee, and L. Pipino, *Journey to Data Quality*. Cambridge, Massachusetts: MIT Press, 2001.

[31] R. Wang, T. Allen, and W. Harris, "One Part at a Time in Lean Sustainment: Tracing the F-15 Fuel Pump," CTPID, Massachusetts Institute of Technology, Cambridge, MA, Technical Report TDQM-2001-07, 25 October 2001.

[32] Y. Jang and Y. R. Wang, "Data Quality Calculus: A data-consumer-based approach to delivering quality data,"

Composite Information Systems Laboratory, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 02139 CISL-91-08, November 1991.

[33] Y. Jang, H. B. Kon, and R. Y. Wang, "A Data Consumer-Based Approach to Data Quality Judgment," presented at the Second Annual Workshop on Information Technologies and Systems (WITS), Dallas, Texas, 1992.

[34] Y. Jang, H. Kon, and R. Wang, "A Knowledge-Based Approach to Assisting In Data Quality Judgment." Cambridge, MA: Massachusetts Institute of Technology, 1992.

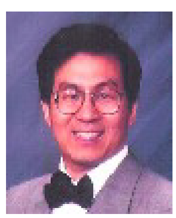
[35] Y. Jang, A. T. Ishii, and R. Y. Wang, "A Qualitative Approach to Automatic Data Quality Judgment," *Journal of Organizational Computing*, vol. 5, pp. 101-121, 1995.

[36] M. P. Reddy and R. Y. Wang, "Estimating Data Accuracy in a Federated Database Environment," presented at 6th International Conference, CISMOD (Also in Lecture Notes in Computer Science), Bombay, India, 1995.

[37] R. Y. Wang, H. B. Kon, and S. E. Madnick, "Data Quality Requirements Analysis and Modeling," presented at the 9th International Conference on Data Engineering, Vienna, 1993.

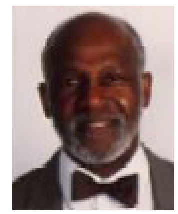
[38] R. Y. Wang, M. P. Reddy, and H. B. Kon, "Toward quality data: An attribute-based approach," *Decision Support Systems (DSS)*, vol. 13, pp. 349-372, 1995.

[39] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data Quality in Context," *Communications of the ACM*, vol. 40, pp. 103-110, 1997.



Richard Y. Wang *is a Principle Investigator at the Center for Technology, Policy, and Industrial Development (CTPID) and Co-Director for the Total Data Quality Management (TDQM) Program at the Massachusetts Institute of Technology, where he has been a professor for a decade. He is also a visiting professor at the CITM, University of California at Berkeley. Prof. Wang also served as a professor at Boston University and the Unviersity of Arisona, Tucson. He has published extensively in top journals to develop concepts, principles, tools, methods, and techniques related to data quality. He co-authored Information Technology in Action: Trends and Perspectives (Prentice Hall, 1993), Data Quality Systems (CMI, 1995), Quality Information and Knowledge (Prentice Hall, 1999), Data Quality (Kluwer Academic Publisher, 2001), Journey to Data Quality (MIT Press, forthcoming), and Principles of Data Quality (textbook in preparation). Prof. Wang can be reached at 617-739-7234.*

Thomas J. Allen*, a former Sloan Senior Associate Dean, is a specialist in organizational psychology and management and nine-year veteran of Boeing, succeeds Professor William Pounds as the MIT Lead of the Lean Aircraft Initiative's (LAI) Integration Team. Allen joins LAI on the heels of its creation of the prototype Lean Enterprise Model, which contains a combination of the Initiative's research results. Allen was appointed to the Sloan faculty in 1966, where he is currently the Howard W. Johnson Professor of Management in the Management of Technological Innovation Group. In 1993, MIT appointed Allen to a ten-year term as a MacVicar Fellow. He is the author of Managing the Flow of Technology.*



Wesley L. Harris *is Charles Stark Draper Professor of Aeronautics and Astronautics, and Director of the Lean Sustainment Initiative at MIT. He recently returned to MIT from a sabbatical at Arizona State University where he was the first engineer and the first African-American to hold the prestigious Barry Goldwater Chair of American Universities. Professor Harris and his students, through their research, have made seminal contributions to the understanding of helicopter rotor noise, unsteady aerodynamics, maintenance of capital assets, and sickle cell disease.*

Stuart E. Madnick *has been a faculty member at M.I.T. since 1972. He has served as the head of MIT's Information Technologies Group for more than ten years. During that time the group has been consistently rated #1 in the nation among business school information technology programs (U.S. News & World Reports, BusinessWeek, and ComputerWorld). He is also an affiliate member of MIT's Laboratory for Computer Science, a member of the research advisory committee of the International Financial Services Research Center, and a member of the executive committee of the Center for Information Systems Research. He has been active in industry, making significant contributions as a key designer and developer of projects such as IBM's VM/370 operating system and Lockheed's DIALOG information retrieval system. He has served as a consultant to many major corporations, such as IBM, AT&T, and Citicorp.*